INFORMACIONI SISTEMI ZA PODRŠKU MENADŽMENTU

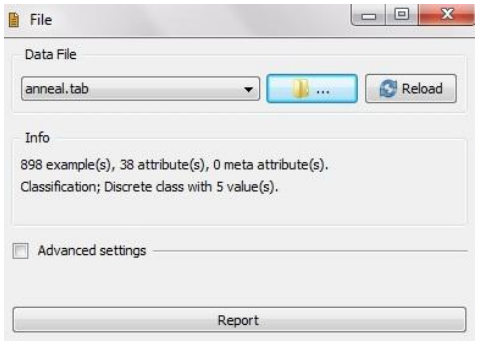
OBLAST:	**Classification**
ČVOROVI (WIDGET):	**SVM, Majority, Test learners, Predictions**
SKUPOVI PODATAKA:	**Anneal**
AUTOR:	**Slobodan Đorđević 254/06**

2011, Beograd

Učitavanje skupa podataka se vrši pomoću čvora File 

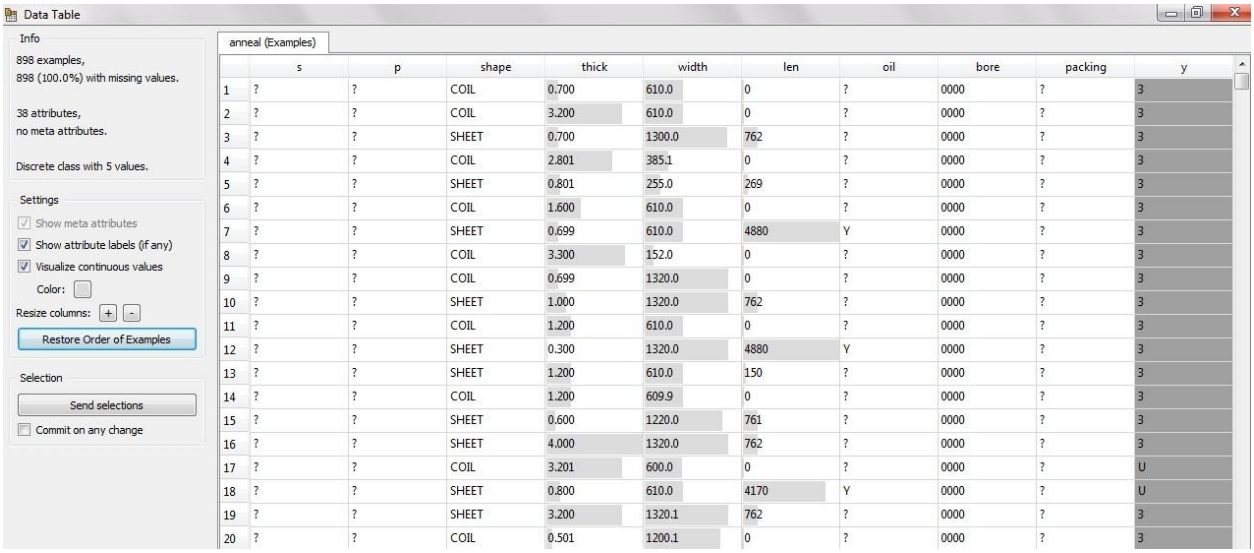
.

U ovom primeru će biti korišćen skup podataka Anneal.



Čvor Data Table 

nam omogućuje da pregledamo podatke.



	s	p	shape	thick	width	len	oil	bore	packing	y
1	?	?	COIL	0.700	610.0	0	?	0000	?	3
2	?	?	COIL	3.200	610.0	0	?	0000	?	3
3	?	?	SHEET	0.700	1300.0	762	?	0000	?	3
4	?	?	COIL	2.801	385.1	0	?	0000	?	3
5	?	?	SHEET	0.801	255.0	269	?	0000	?	3
6	?	?	COIL	1.600	610.0	0	?	0000	?	3
7	?	?	SHEET	0.699	610.0	4880	Y	0000	?	3
8	?	?	COIL	3.300	152.0	0	?	0000	?	3
9	?	?	COIL	0.699	1320.0	0	?	0000	?	3
10	?	?	SHEET	1.000	1320.0	762	?	0000	?	3
11	?	?	COIL	1.200	610.0	0	?	0000	?	3
12	?	?	SHEET	0.300	1320.0	4880	Y	0000	?	3
13	?	?	SHEET	1.200	610.0	150	?	0000	?	3
14	?	?	COIL	1.200	609.9	0	?	0000	?	3
15	?	?	SHEET	0.600	1220.0	761	?	0000	?	3
16	?	?	SHEET	4.000	1320.0	762	?	0000	?	3
17	?	?	COIL	3.201	600.0	0	?	0000	?	U
18	?	?	SHEET	0.800	610.0	4170	Y	0000	?	U
19	?	?	SHEET	3.200	1320.1	762	?	0000	?	3
20	?	?	COIL	0.501	1200.1	0	?	0000	?	3

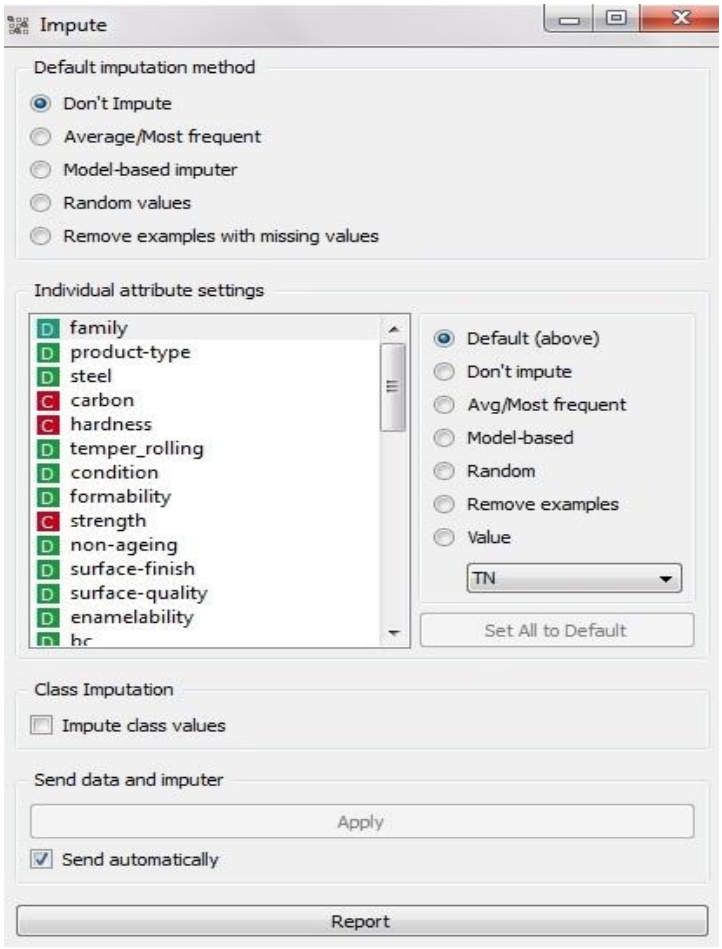
Ovaj Data Set ima 898 primera, odnosno slučaja i tiče se prekaljivanja. Neki od atributa su debljina, širina, dužina (oni su numerički i realni), tvrdoća (celi brojevi) i oblik (kategorički). Annealing Data Set ima mnogo nedostajućih vrednosti i nijedan slučaj nije kompletno opisan.

Čvorovi Majority i SVM koji će u ovom primeru biti korišćeni za učenje rade sa nedostajućim vrednostima.

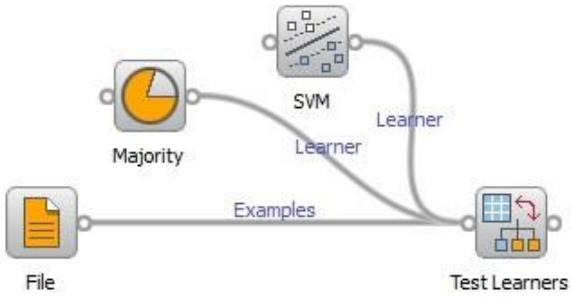
Problem nedostajućih vrednosti se može rešiti korišćenjem čvora Impute. 

Polja koja nemaju definisanu vrednost moguće je popuniti na sledeće načine:

- prosečnom ili najčešćom vrednošću
- pomoću modela koji uzima u obzir i vrednosti drugih atributa
- slučajnom vrednošću, odnosno nekom od onih koje su korišćene u ocenjivanju atributa
- izbacivanjem slučaja koji ima nedostajuće vrednosti
- tačno određenom vrednošću.



Za učenje će biti korišćeni čvorovi Majority i SVM, a za evaluaciju Test Learners.
Šema izgleda ovako.

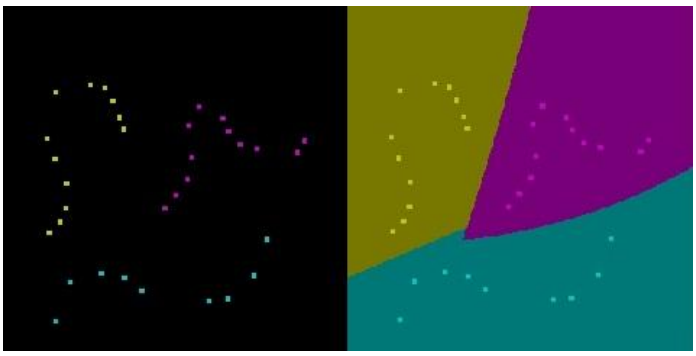


Čvor Majority uzima najzastupljenije rešenje skupa podataka i prilikom svakog upita daje uvek taj isti odgovor. Kao takav, ovaj metod učenja nema veliku vrednost i koristi se za poređenje sa drugim metodama. Jedino što mu se može menjati jeste ime.

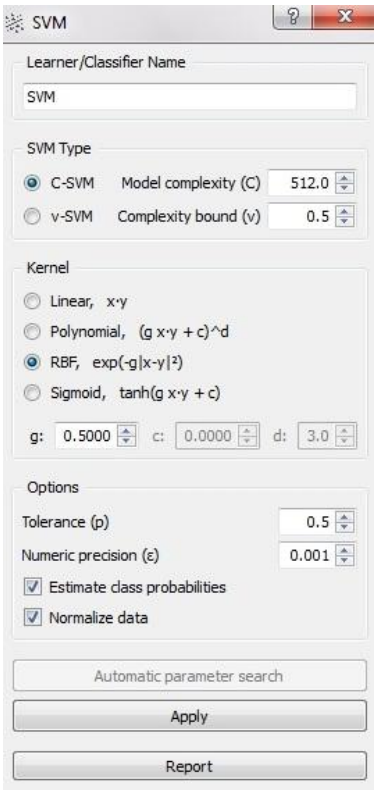
Čvor SVM 

je daleko složeniji. Na atributskom prostoru konstruiše odvojene hiperravni koje maksimiziraju marginu tj. razliku između slučajeva različitih klasa.

Grafički prikaz toga bi izgledao ovako: unosimo slučajeve različitih klasa, u ovom primeru boja, i SVM pomoću Kernel-a pravi hiperravni i vrši klasifikaciju.



Podešavanje SVM-a: Prvo definišemo kompleksnost modela, kompleksnost granice, toleranciju i numeričku preciznost. Ovi parametri definišu funkciju optimizacije.



Kompleksnost granice određuje broj podržavajućih vektora u konstruisanju margina i presudno utiče na kompleksnost modela.

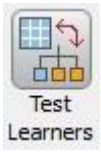
Tolerancija određuje u koju će klasu upasti primer koji se nalazi u blizini granice.

Numerička preciznost određuje decimalu zaokruživanja i ključna je u stvaranju dobrog algoritma. Ako je loše postavljena može dovesti do greške u proračunima. Treba voditi računa da kompjuter može prikazati broj samo do neke određene preciznosti.

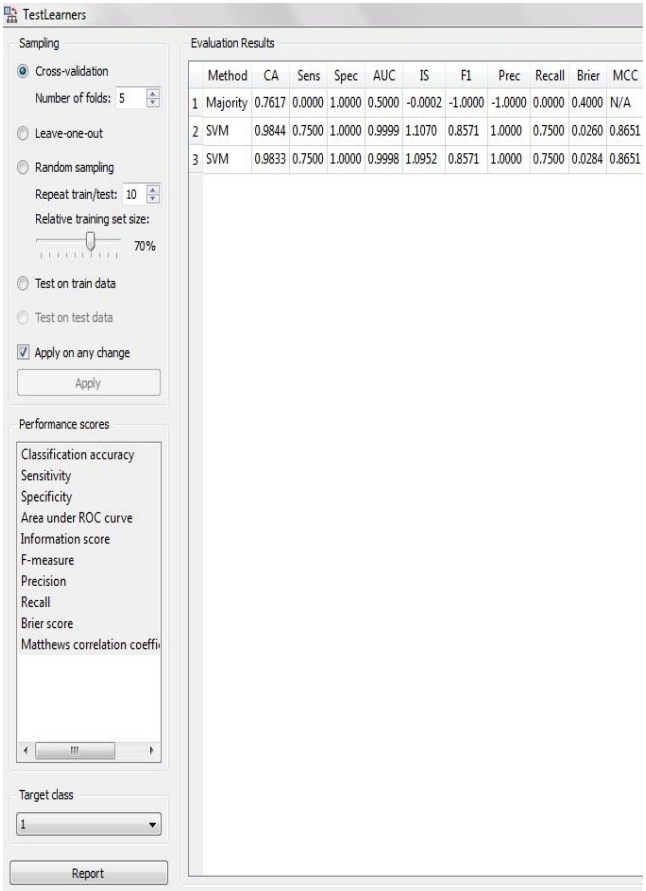
Pomoću SVM-a je moguće proceniti verovatnoću izbora klase i izvršiti normalizaciju podataka.

Što se tiče Kernel-a, to je funkcija koja pretvara atributski prostor u novi oblik kako bi se uklopio u hiperravni maksimiziranih margina. To algoritmu omogućava da kreira nelinearne klasifikatore. Prvi Kernel je linearni i on

ne treba ovaj trik. Pored svakog Kernel-a se nalazi i funkcija koja mu odgovara. G je gama konstanta i jednaka je 1/broj atributa. Po default-u je 0 jer nije urađen trening. C je takođe konstanta, dok D predstavlja ugao Kernel-a i po default-u je 3.

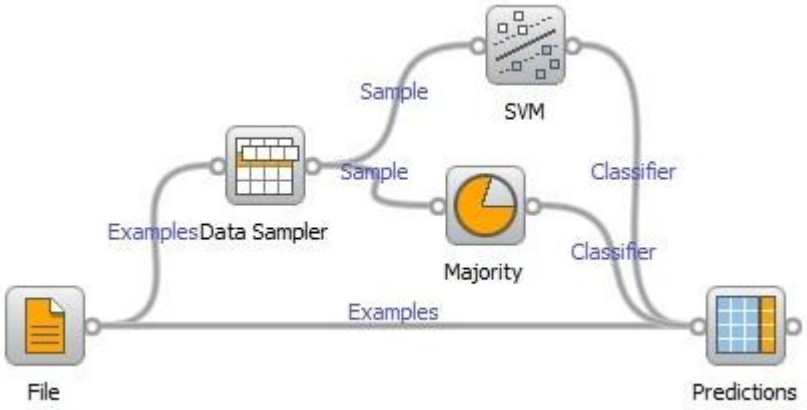


Test Learners omogućuje poređenje metoda učenja.
Podržava različite načine uzorkovanja. Kros-validacija deli podatke u više grupa. Algoritam se testira pomoću jedne grupe i kako nova grupa dođe na red tako se prethodna klasifikuje. Leave One Out je sličan samo što drži jednu grupu, a onda je klasifikuje na osnovu učenja iz svih ostalih. Moguće je odabrati veličinu uzorka, kao i sprovesti test na podacima koji su korišćeni za trening.



- **CA** je procenat tačno klasifikovanih primera
- **Sensitivity** pokazuje broj otkrivenih pozitivnih primera među svim pozitivnim primerima
- **Specificity** pokazuje broj otkrivenih negativnih primera među svim negativnim primerima
- **AUC** je oblast pod ROC krivom
- **IS** je prosečna količina informacija po klasifikovanom primeru
- **F1** je mera koja predstavlja ponderisanu harmonijsku sredinu Precision i Recall
- **Precision** je broj pozitivnih primera među svim primerima klasifikovanim kao pozitivni
- **Recall** je ista mera kao i Sensitivity, ali ovde ima više medicinsko značenje
- **Brier** meri preciznost verovatnoće procene, odnosno odstupanje između predviđene verovatnoće događaja i stvarnog događaja
- **MCC** pokazuje tačnost predviđanja -1 suprotno 0 prosečno 1 perfektano.

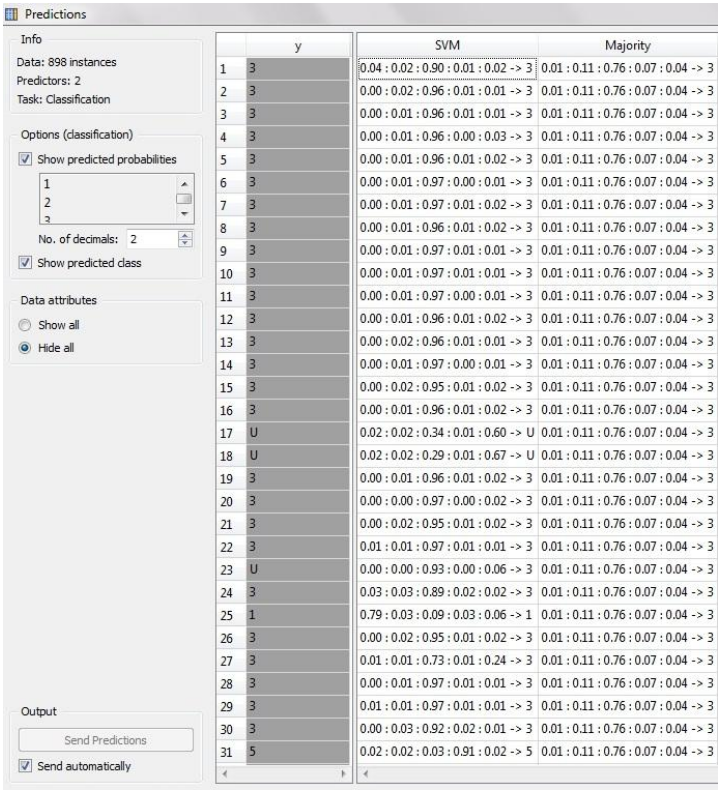
Radi lakšeg shvatanja ovih metoda učenja u nastavku će biti prikazan drugi način evaluacije.
Šema izgleda ovako.



Čvor Data Sampler ovde ima sledeću ulogu: uzima zadatu količinu podataka, odnosno uzorak i na njemu čvorovi učenja Majority i SVM treniraju tj. vežbaju, a onda se na ostatku skupa podataka sprovodi testiranje.

Za pogled na rezultate se u ovom slučaju koristi čvor Predictions 

koji prikazuje predviđanje čvorova učenja i verovatnoće koje su postavili za svaku klasu.



	y	SVM	Majority
1	3	0.04 : 0.02 : 0.90 : 0.01 : 0.02 -> 3	0.01 : 0.11 : 0.76 : 0.07 : 0.04 -> 3
2	3	0.00 : 0.02 : 0.96 : 0.01 : 0.01 -> 3	0.01 : 0.11 : 0.76 : 0.07 : 0.04 -> 3
3	3	0.00 : 0.01 : 0.96 : 0.01 : 0.01 -> 3	0.01 : 0.11 : 0.76 : 0.07 : 0.04 -> 3
4	3	0.00 : 0.01 : 0.96 : 0.00 : 0.03 -> 3	0.01 : 0.11 : 0.76 : 0.07 : 0.04 -> 3
5	3	0.00 : 0.01 : 0.96 : 0.01 : 0.02 -> 3	0.01 : 0.11 : 0.76 : 0.07 : 0.04 -> 3
6	3	0.00 : 0.01 : 0.97 : 0.00 : 0.01 -> 3	0.01 : 0.11 : 0.76 : 0.07 : 0.04 -> 3
7	3	0.00 : 0.01 : 0.97 : 0.01 : 0.02 -> 3	0.01 : 0.11 : 0.76 : 0.07 : 0.04 -> 3
8	3	0.00 : 0.01 : 0.96 : 0.01 : 0.02 -> 3	0.01 : 0.11 : 0.76 : 0.07 : 0.04 -> 3
9	3	0.00 : 0.01 : 0.97 : 0.01 : 0.01 -> 3	0.01 : 0.11 : 0.76 : 0.07 : 0.04 -> 3
10	3	0.00 : 0.01 : 0.97 : 0.01 : 0.01 -> 3	0.01 : 0.11 : 0.76 : 0.07 : 0.04 -> 3
11	3	0.00 : 0.01 : 0.97 : 0.00 : 0.01 -> 3	0.01 : 0.11 : 0.76 : 0.07 : 0.04 -> 3
12	3	0.00 : 0.01 : 0.96 : 0.01 : 0.02 -> 3	0.01 : 0.11 : 0.76 : 0.07 : 0.04 -> 3
13	3	0.00 : 0.02 : 0.96 : 0.01 : 0.01 -> 3	0.01 : 0.11 : 0.76 : 0.07 : 0.04 -> 3
14	3	0.00 : 0.01 : 0.97 : 0.00 : 0.01 -> 3	0.01 : 0.11 : 0.76 : 0.07 : 0.04 -> 3
15	3	0.00 : 0.02 : 0.95 : 0.01 : 0.02 -> 3	0.01 : 0.11 : 0.76 : 0.07 : 0.04 -> 3
16	3	0.00 : 0.01 : 0.96 : 0.01 : 0.02 -> 3	0.01 : 0.11 : 0.76 : 0.07 : 0.04 -> 3
17	U	0.02 : 0.02 : 0.34 : 0.01 : 0.60 -> U	0.01 : 0.11 : 0.76 : 0.07 : 0.04 -> 3
18	U	0.02 : 0.02 : 0.29 : 0.01 : 0.67 -> U	0.01 : 0.11 : 0.76 : 0.07 : 0.04 -> 3
19	3	0.00 : 0.01 : 0.96 : 0.01 : 0.02 -> 3	0.01 : 0.11 : 0.76 : 0.07 : 0.04 -> 3
20	3	0.00 : 0.00 : 0.97 : 0.00 : 0.02 -> 3	0.01 : 0.11 : 0.76 : 0.07 : 0.04 -> 3
21	3	0.00 : 0.02 : 0.95 : 0.01 : 0.02 -> 3	0.01 : 0.11 : 0.76 : 0.07 : 0.04 -> 3
22	3	0.01 : 0.01 : 0.97 : 0.01 : 0.01 -> 3	0.01 : 0.11 : 0.76 : 0.07 : 0.04 -> 3
23	U	0.00 : 0.00 : 0.93 : 0.00 : 0.06 -> 3	0.01 : 0.11 : 0.76 : 0.07 : 0.04 -> 3
24	3	0.03 : 0.03 : 0.89 : 0.02 : 0.02 -> 3	0.01 : 0.11 : 0.76 : 0.07 : 0.04 -> 3
25	1	0.79 : 0.03 : 0.09 : 0.03 : 0.06 -> 1	0.01 : 0.11 : 0.76 : 0.07 : 0.04 -> 3
26	3	0.00 : 0.02 : 0.95 : 0.01 : 0.02 -> 3	0.01 : 0.11 : 0.76 : 0.07 : 0.04 -> 3
27	3	0.01 : 0.01 : 0.73 : 0.01 : 0.24 -> 3	0.01 : 0.11 : 0.76 : 0.07 : 0.04 -> 3
28	3	0.00 : 0.01 : 0.97 : 0.01 : 0.01 -> 3	0.01 : 0.11 : 0.76 : 0.07 : 0.04 -> 3
29	3	0.01 : 0.01 : 0.97 : 0.01 : 0.01 -> 3	0.01 : 0.11 : 0.76 : 0.07 : 0.04 -> 3
30	3	0.00 : 0.03 : 0.92 : 0.02 : 0.01 -> 3	0.01 : 0.11 : 0.76 : 0.07 : 0.04 -> 3
31	5	0.02 : 0.02 : 0.03 : 0.91 : 0.02 -> 5	0.01 : 0.11 : 0.76 : 0.07 : 0.04 -> 3

Kao što je ranije objašnjeno čvor Majority uzima najzastupljenije rešenje skupa podataka i kao što se na slici može videti daje uvek taj isti odgovor. Za razliku od njega SVM traži sličnosti među slučajevima i sa povećanjem uzorka teži perfekciji. SVM je neuporedivo bolji vid učenja,

dok Majority ima primenu u poređenju sa drugim metodama i kod traženja najzastupljenije klase u velikim skupovima podataka.