Računarstvo usluga i analiza podataka

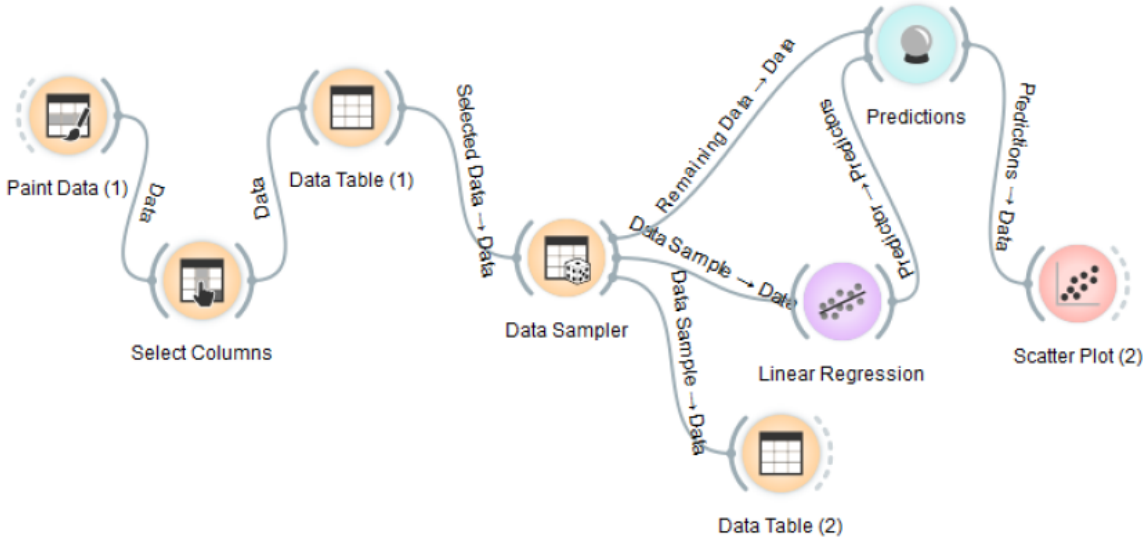
Izvještaj pete laboratorijske vježbe

Marko Ćosić

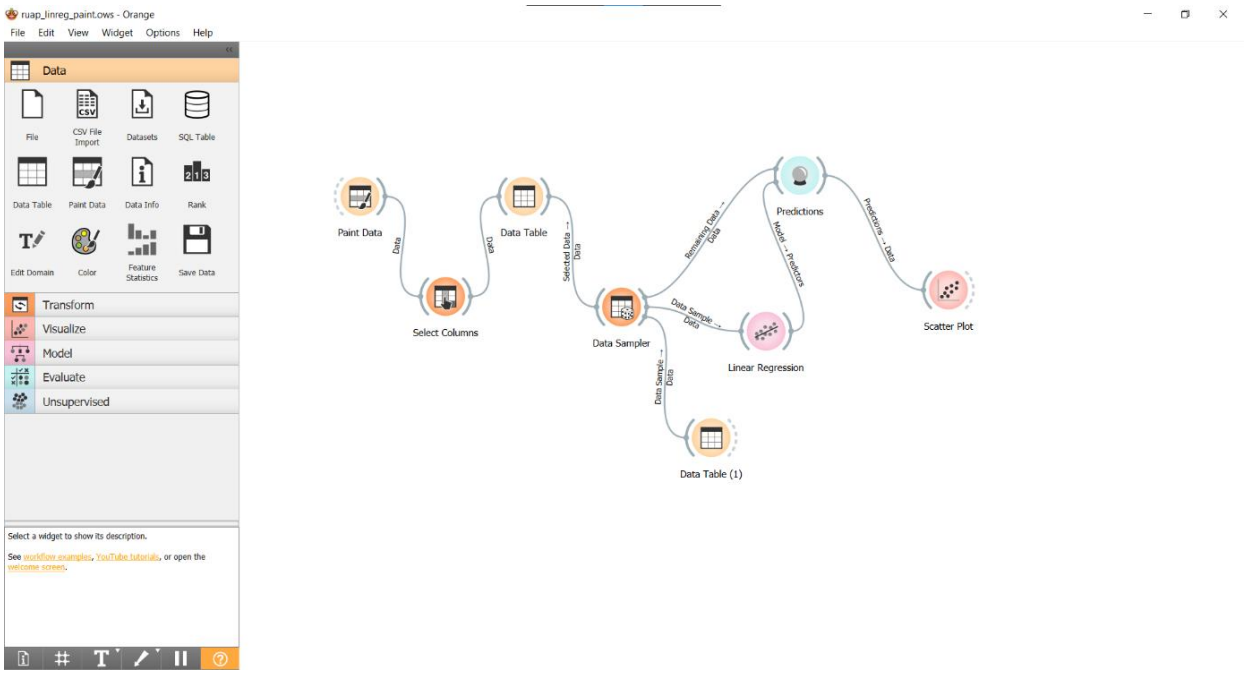
Zadatak 1.

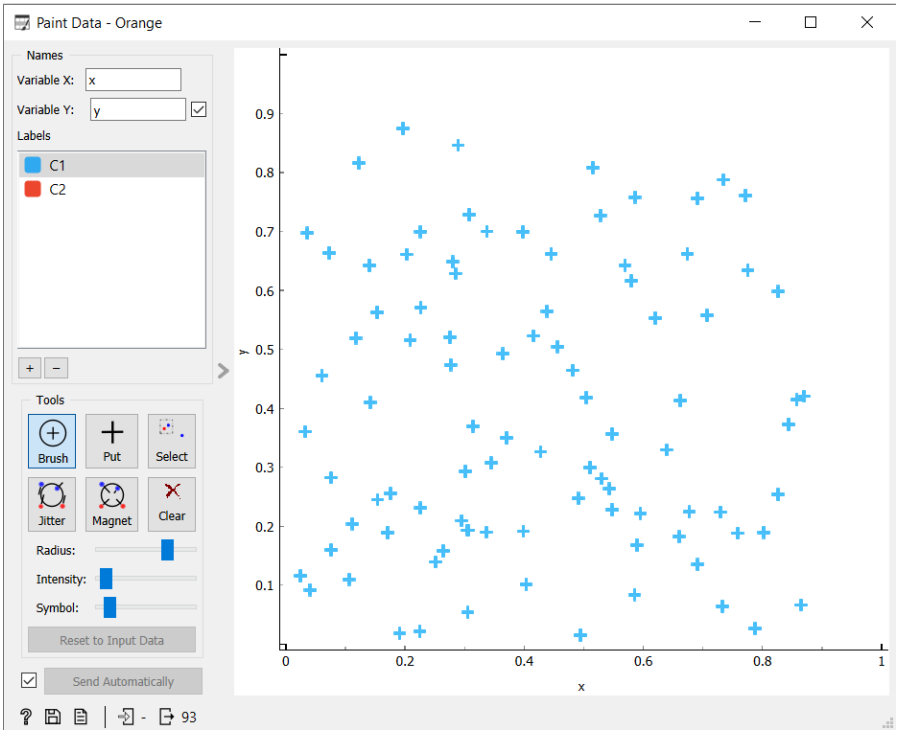
Izraditi linearni regresijski model za predviđanje umjetno generiranih podataka u 2D.

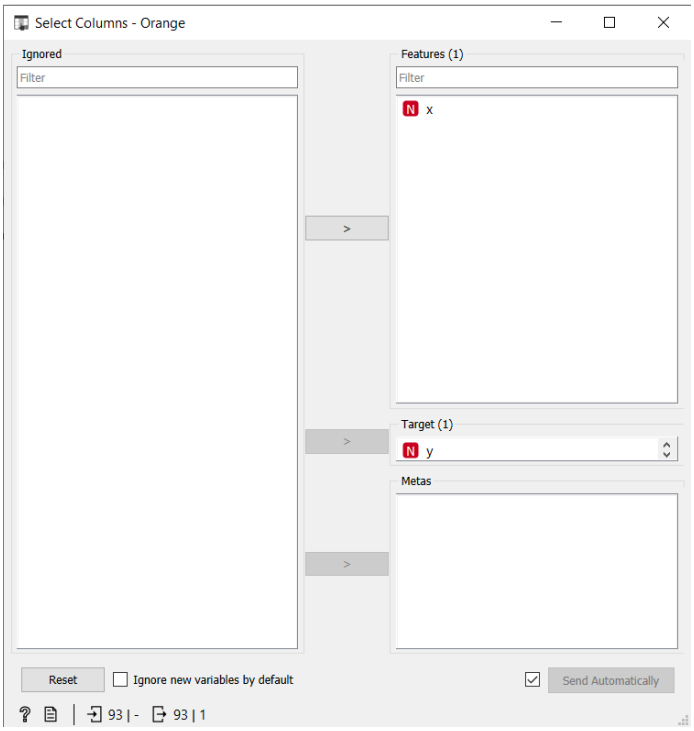
1. Otvoriti Orange alat
2. Odabrati novi dokument i dati mu ime *ruap_linreg_paint*
3. Izraditi eksperiment prema slici
4. "Nacrtati" podatke za jednu klasu
5. Odabrati podatke x-osi kao značajke (na temelju čega se predviđa), a podatke y-osi kao ciljane (vrijednost koju se predviđa)
6. Podijeliti skup na 0.75:0.25 za učenje i testiranje
7. Izvršiti eksperiment
8. Komentirati dobivene rezultate

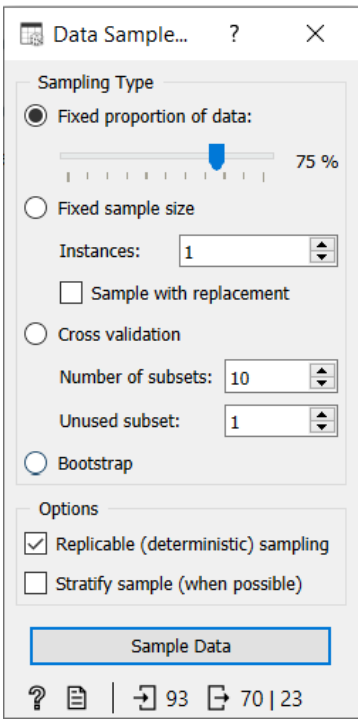


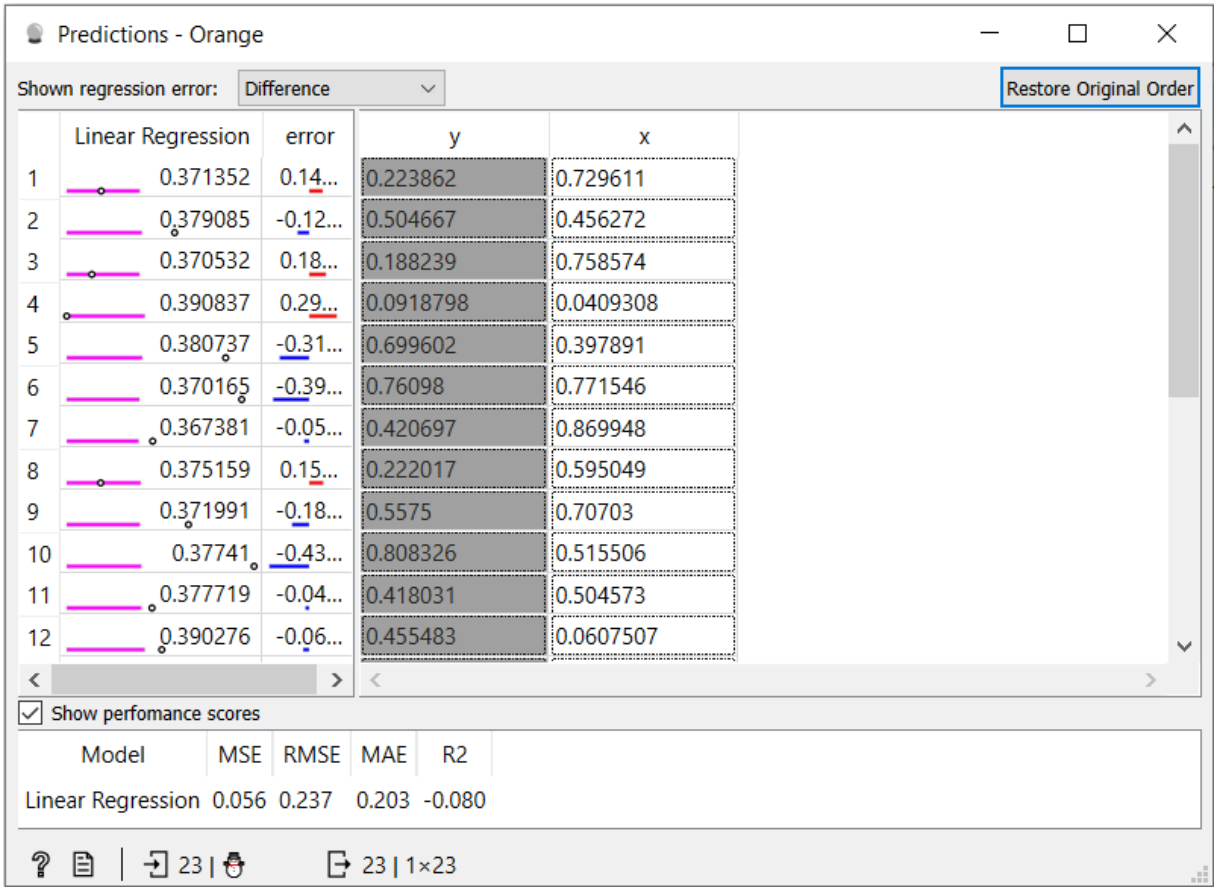
Rješenje:




Paint Data - Orange
Names
Variable X:
x
Variable Y:
y
Labels
C1
C2
Tools
Brush
Put
Select
Jitter
Magnet
Clear
Radius:
Intensity:
Symbol:
Reset to Input Data
Send Automatically
93
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
y
0
0.2
0.4
0.6
0.8
1
x


Select Columns - Orange
Ignored
Filter
Features (1)
Filter
x
Target (1)
y
Metas
Reset
Ignore new variables by default
Send Automatically
93 | -
93 | 1


Data Sample...
Sampling Type
Fixed proportion of data:
75 %
Fixed sample size
Instances:
1
Sample with replacement
Cross validation
Number of subsets:
10
Unused subset:
1
Bootstrap
Options
Replicable (deterministic) sampling
Stratify sample (when possible)
Sample Data
93
70 | 23



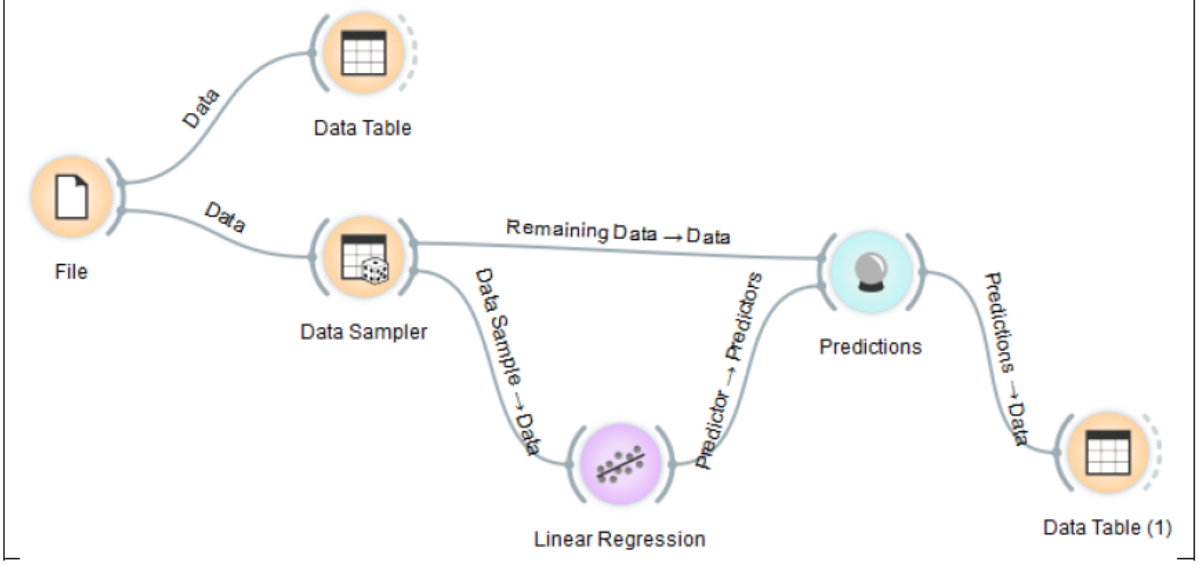
Komentar:

U ovome zadatku nacrtan je dataset, odnosno podatci za jednu klasu. Podatci koji se nalaze na x-osi odabrani su kao značajke (na temelju čega se predviđa), a podatci na y-osi kao ciljane značajke (vrijednosti koje se predviđaju). Može se vidjeti da je nakon podjele skupa 75% : 25% te nakon testiranja linearni regresijski model dao slijedeće rezultate: MSE 0.056, RMSE 0.237 i MAE 0.203.

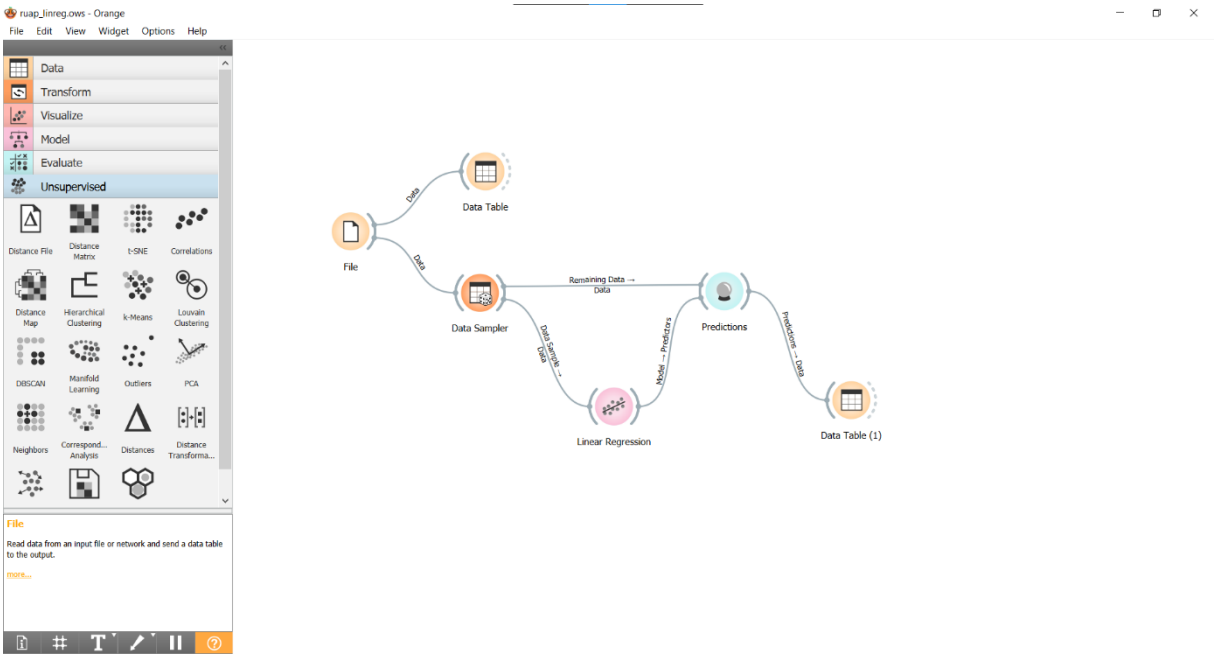
Zadatak 2.

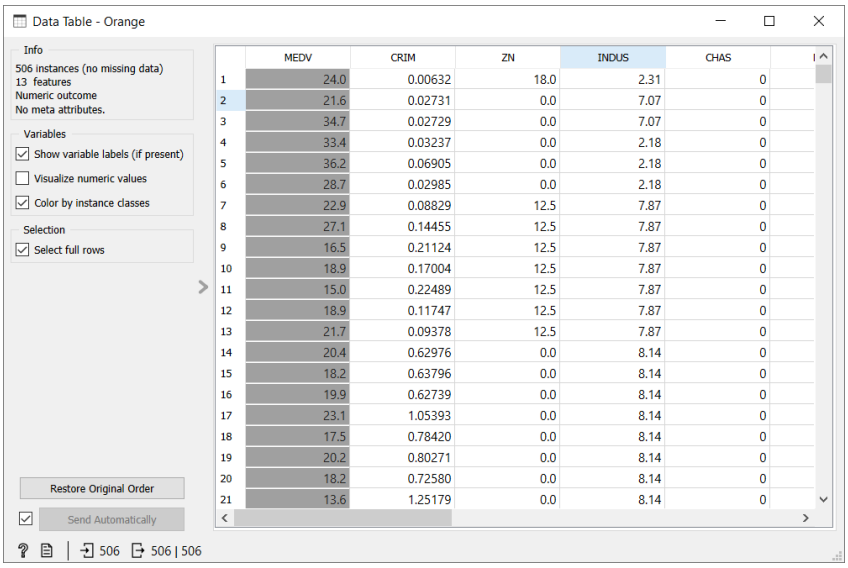
Izraditi linearni regresijski model za predviđanje cijena nekretnina (https://archive.ics.uci.edu/ml/datasets/Housing).

1. Otvoriti Orange alat
2. Odabrati novi dokument i dati mu ime *ruap_linreg*
3. Izraditi eksperiment prema slici
4. Iscrtati u Excelu stvarne vrijednosti i one predviđene modelom
5. Komentirati dobivene rezultate

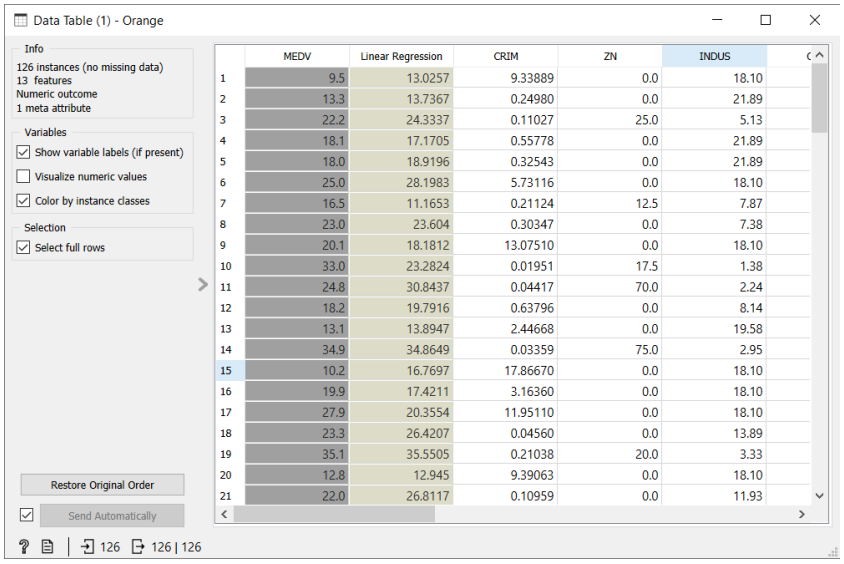


Rješenje:

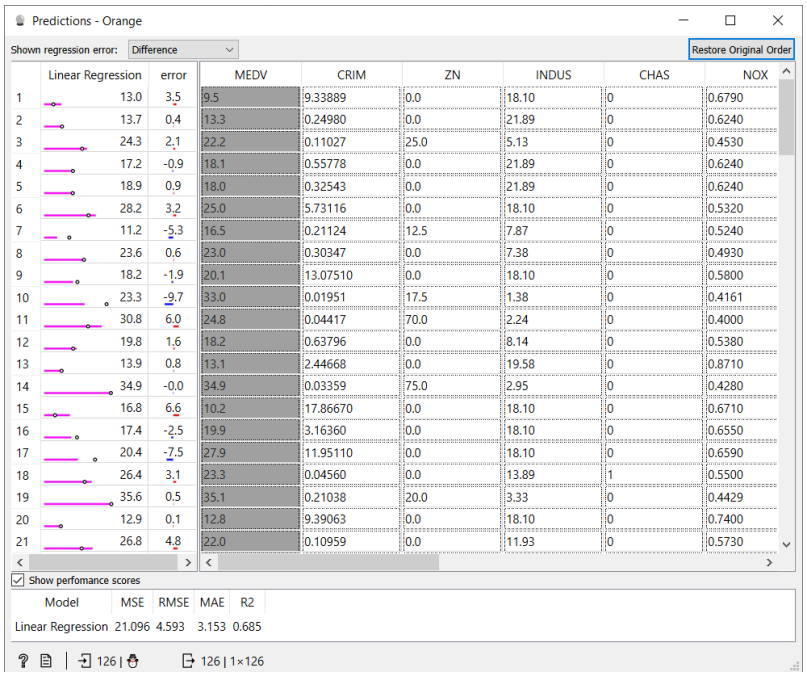




	MEDV	CRIM	ZN	INDUS	CHAS
1	24.0	0.00632	18.0	2.31	0
2	21.6	0.02731	0.0	7.07	0
3	34.7	0.02729	0.0	7.07	0
4	33.4	0.03237	0.0	2.18	0
5	36.2	0.06905	0.0	2.18	0
6	28.7	0.02985	0.0	2.18	0
7	22.9	0.08829	12.5	7.87	0
8	27.1	0.14455	12.5	7.87	0
9	16.5	0.21124	12.5	7.87	0
10	18.9	0.17004	12.5	7.87	0
11	15.0	0.22489	12.5	7.87	0
12	18.9	0.11747	12.5	7.87	0
13	21.7	0.09378	12.5	7.87	0
14	20.4	0.62976	0.0	8.14	0
15	18.2	0.63796	0.0	8.14	0
16	19.9	0.62739	0.0	8.14	0
17	23.1	1.05393	0.0	8.14	0
18	17.5	0.78420	0.0	8.14	0
19	20.2	0.80271	0.0	8.14	0
20	18.2	0.72580	0.0	8.14	0
21	13.6	1.25179	0.0	8.14	0



	MEDV	Linear Regression	CRIM	ZN	INDUS
1	9.5	13.0257	9.33889	0.0	18.10
2	13.3	13.7367	0.24980	0.0	21.89
3	22.2	24.3337	0.11027	25.0	5.13
4	18.1	17.1705	0.55778	0.0	21.89
5	18.0	18.9196	0.32543	0.0	21.89
6	25.0	28.1983	5.73116	0.0	18.10
7	16.5	11.1653	0.21124	12.5	7.87
8	23.0	23.604	0.30347	0.0	7.38
9	20.1	18.1812	13.07510	0.0	18.10
10	33.0	23.2824	0.01951	17.5	1.38
11	24.8	30.8437	0.04417	70.0	2.24
12	18.2	19.7916	0.63796	0.0	8.14
13	13.1	13.8947	2.44668	0.0	19.58
14	34.9	34.8649	0.03359	75.0	2.95
15	10.2	16.7697	17.86670	0.0	18.10
16	19.9	17.4211	3.16360	0.0	18.10
17	27.9	20.3554	11.95110	0.0	18.10
18	23.3	26.4207	0.04560	0.0	13.89
19	35.1	35.5505	0.21038	20.0	3.33
20	12.8	12.945	9.39063	0.0	18.10
21	22.0	26.8117	0.10959	0.0	11.93



	Linear Regression	error	MEDV	CRIM	ZN	INDUS	CHAS	NOX
1	13.0	3.5	9.5	9.33889	0.0	18.10	0	0.6790
2	13.7	0.4	13.3	0.24980	0.0	21.89	0	0.6240
3	24.3	2.1	22.2	0.11027	25.0	5.13	0	0.4530
4	17.2	-0.9	18.1	0.55778	0.0	21.89	0	0.6240
5	18.9	0.9	18.0	0.32543	0.0	21.89	0	0.6240
6	28.2	3.2	25.0	5.73116	0.0	18.10	0	0.5320
7	11.2	-5.3	16.5	0.21124	12.5	7.87	0	0.5240
8	23.6	0.6	23.0	0.30347	0.0	7.38	0	0.4930
9	18.2	-1.9	20.1	13.07510	0.0	18.10	0	0.5800
10	23.3	-9.7	33.0	0.01951	17.5	1.38	0	0.4161
11	30.8	6.0	24.8	0.04417	70.0	2.24	0	0.4000
12	19.8	1.6	18.2	0.63796	0.0	8.14	0	0.5380
13	13.9	0.8	13.1	2.44668	0.0	19.58	0	0.8710
14	34.9	-0.0	34.9	0.03359	75.0	2.95	0	0.4280
15	16.8	6.6	10.2	17.86670	0.0	18.10	0	0.6710
16	17.4	-2.5	19.9	3.16360	0.0	18.10	0	0.6550
17	20.4	-7.5	27.9	11.95110	0.0	18.10	0	0.6590
18	26.4	3.1	23.3	0.04560	0.0	13.89	1	0.5500
19	35.6	0.5	35.1	0.21038	20.0	3.33	0	0.4429
20	12.9	0.1	12.8	9.39063	0.0	18.10	0	0.7400
21	26.8	4.8	22.0	0.10959	0.0	11.93	0	0.5730

,

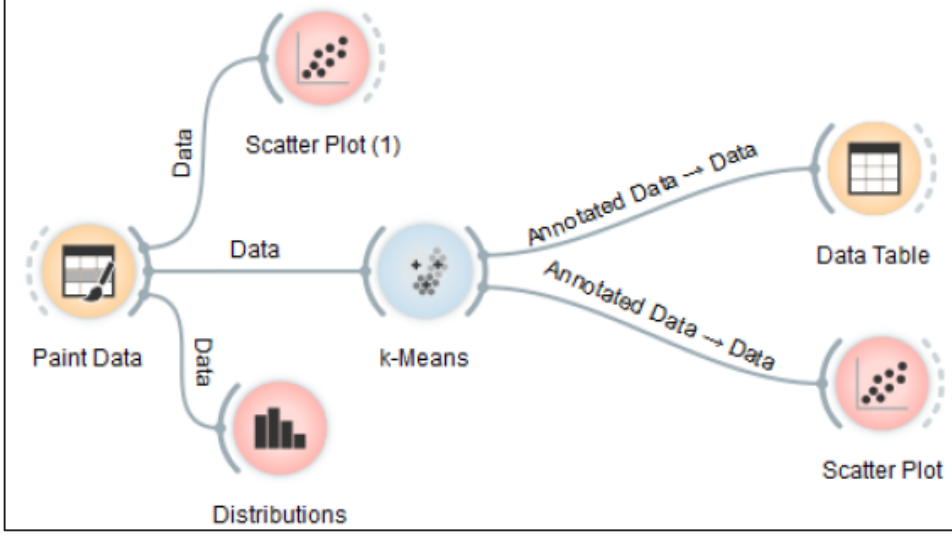
Komentar:

U odnosnu na prethodni zadatak, korišten je Housing dataset koji se učitava prije testiranja. Nakon testiranja može se vidjeti da stvarne vrijednosti iscrtane u Excelu i one predviđene modelom imaju manje razlike (usporedba Data Table i Data Table(1)). Nakon testiranja dobiveni su slijedeći rezultati: MSE 21.096, RMSE 4.593 i MAE 3.153. U usporedni s prošlim zadatkom, u ovome imamo više značajki na osnovu kojih radimo testiranje, stoga su i rezultati drugačiji.

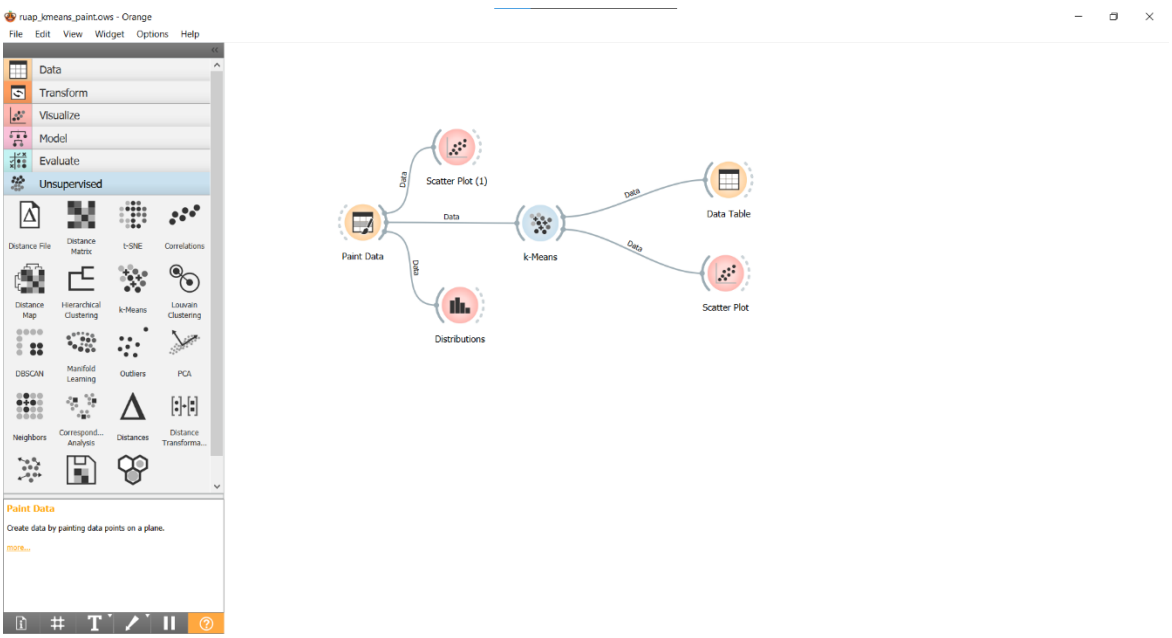
Zadatak 3.

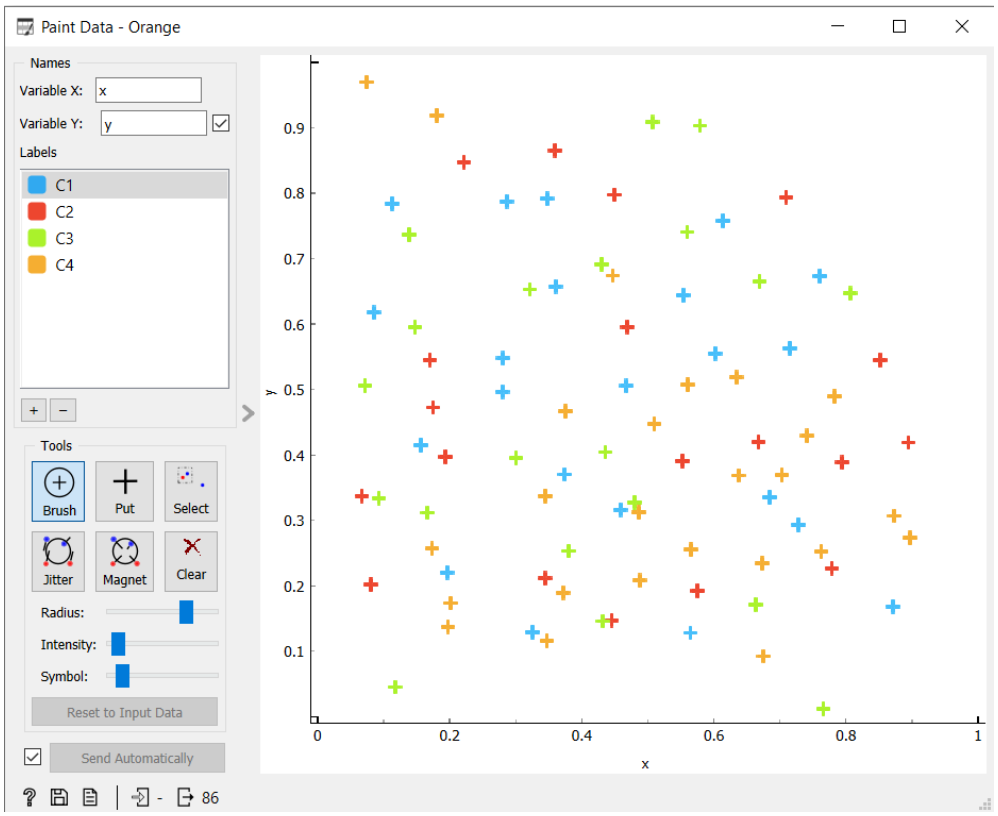
Stvoriti particiju umjetno generiranog skupa podataka pomoću k-means algoritma.

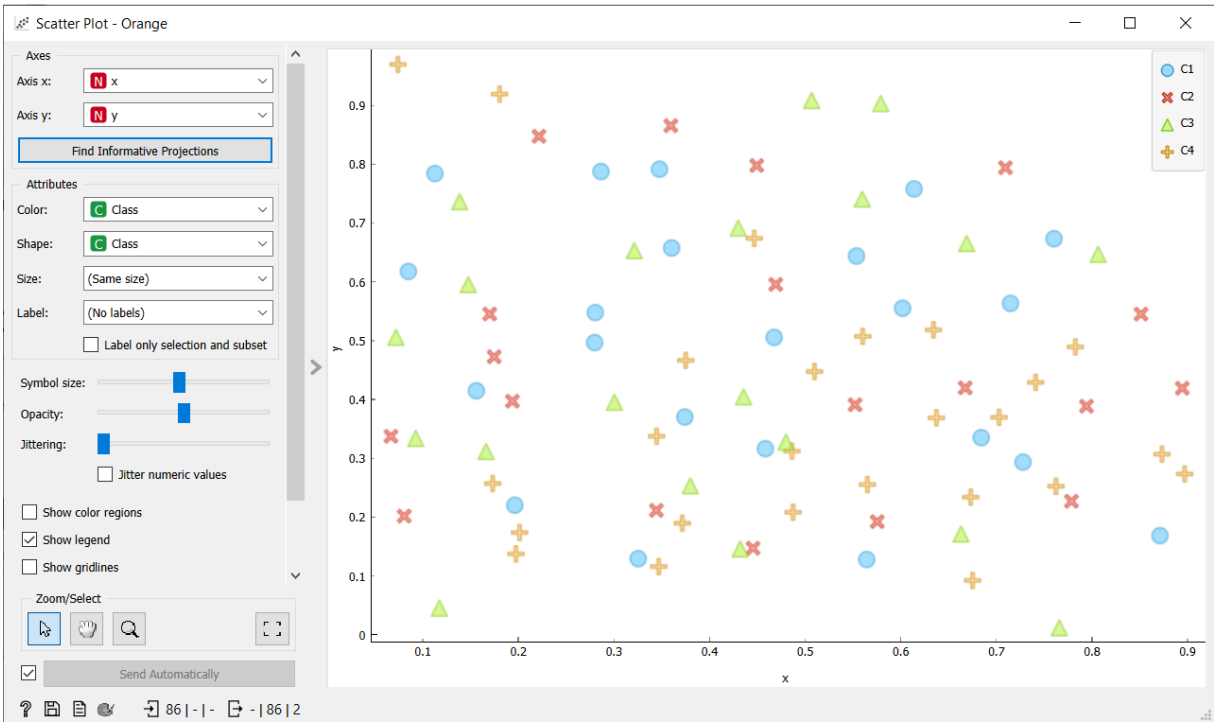
1. Otvoriti Orange alat
2. Odabrati novi dokument i dati mu ime *ruap_kmeans_paint*
3. Izraditi eksperiment prema slici
4. "Nacrtati" podatke za 4 klase
5. Izračunati 4-particiju generiranog skupa pomoću k-means algoritma
6. Komentirati dobivene rezultate

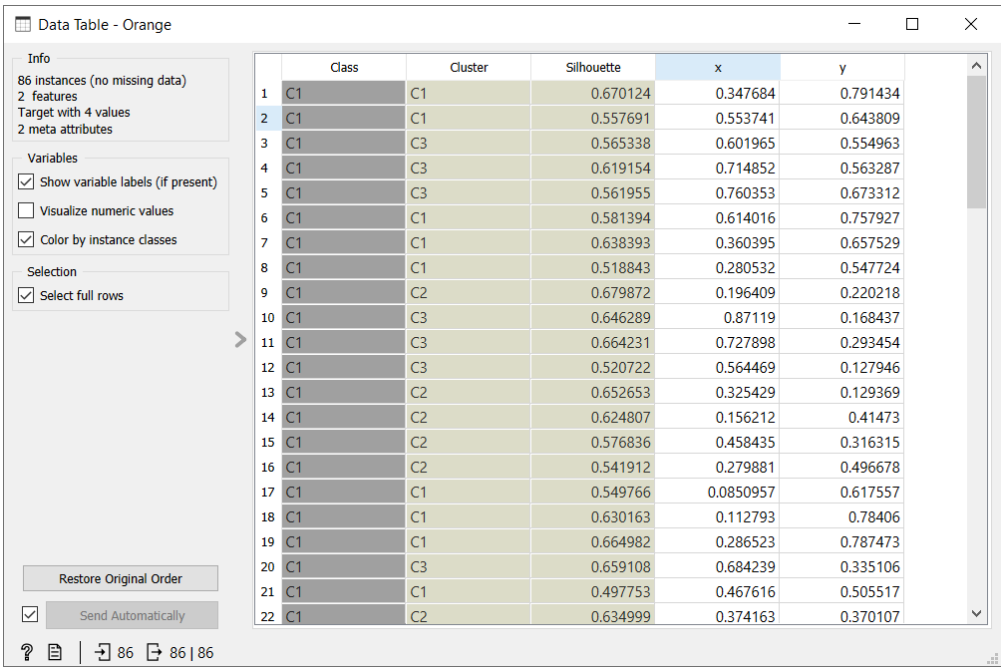


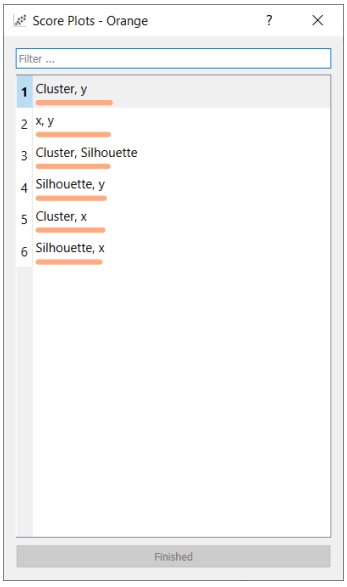
Rješenje:

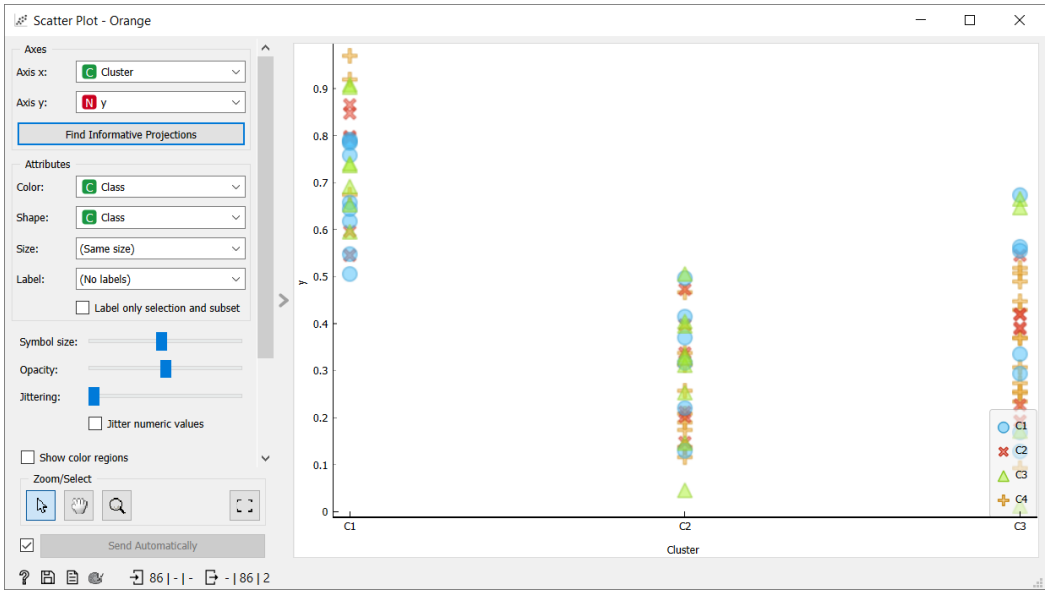



Paint Data - Orange
Names
Variable X:
x
Variable Y:
y
Labels
C1
C2
C3
C4
Tools
Brush
Put
Select
Jitter
Magnet
Clear
Radius:
Intensity:
Symbol:
Reset to Input Data
Send Automatically
86
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
y
0
0.2
0.4
0.6
0.8
1
x


Scatter Plot - Orange
Axes
Axis x:
x
Axis y:
y
Find Informative Projections
Attributes
Color:
Class
Shape:
Class
Size:
(Same size)
Label:
(No labels)
Label only selection and subset
Symbol size:
Opacity:
Jittering:
Jitter numeric values
Show color regions
Show legend
Show gridlines
Zoom/Select
Send Automatically
86 | - | -
- | 86 | 2
C1
C2
C3
C4
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
y
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
x







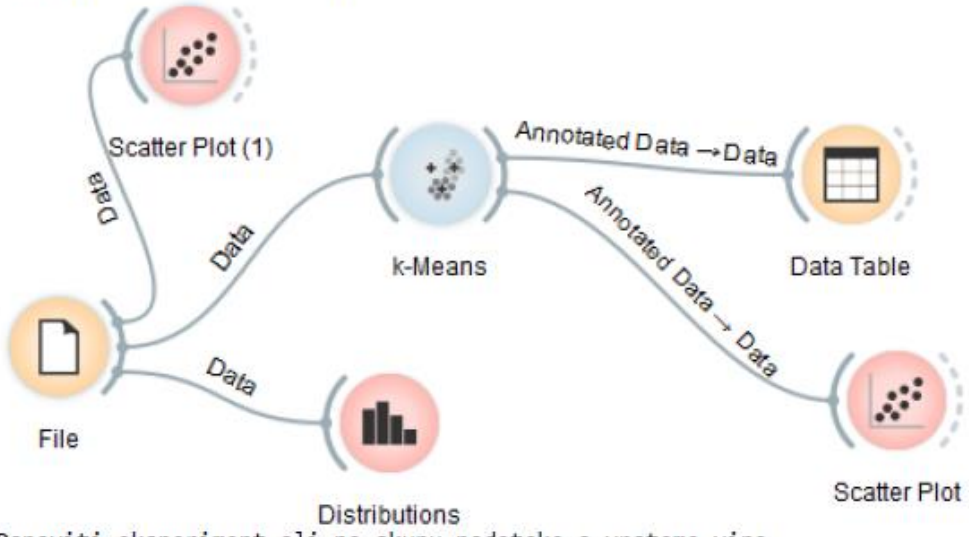
Komentar:

Kmeans je algoritam alternirajuće optimizacije koji na temelju zadanih centara određuje particiju, a na temelju particije centre. Glavni nedostatak Kmeans algoritma, ali i mnogih drugih je što broj grupa mora biti unaprijed poznat. Budući da taj podatak često nije poznat i potrebno je odrediti najprikladniji broj grupa, Kmeans algoritam može se više puta izvesti s različitim brojem grupa. Na prethodnim slikama prikazani su rezultati nakon testiranja, ali i rezultati nakon odabira opcije „Find Informative Projections“.

Zadatak 4.

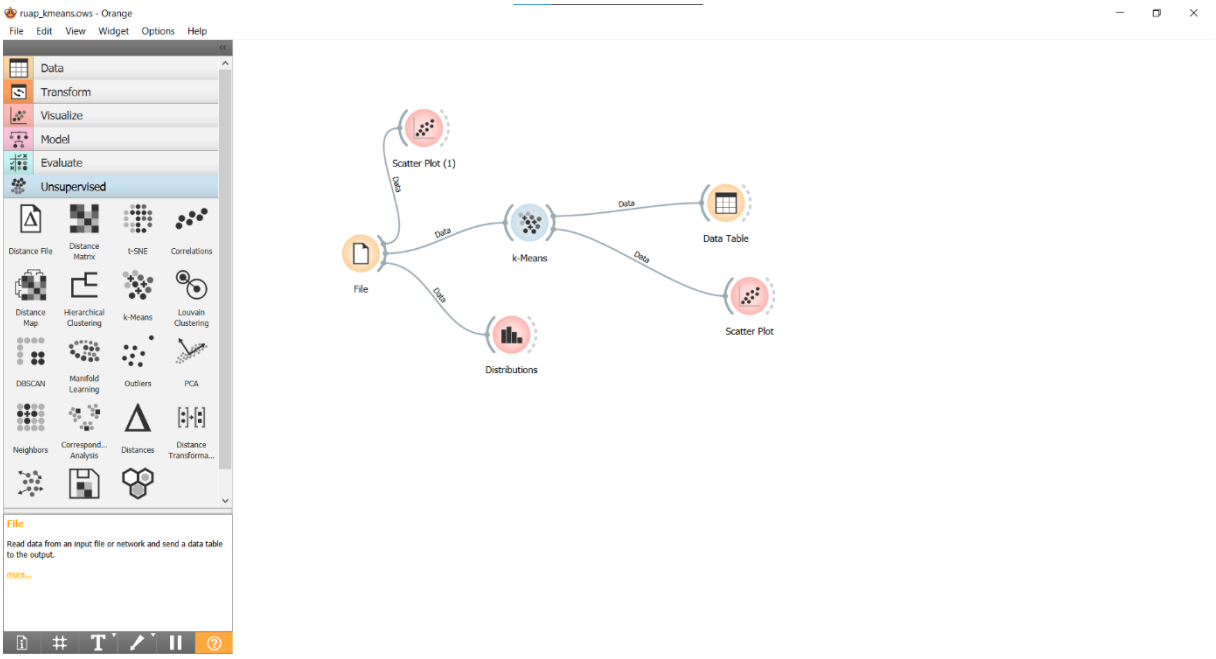
Stvoriti particiju skupa podataka o cvijetu perunike (https://archive.ics.uci.edu/ml/datasets/Iris) pomoću k-means algoritma.

1. Otvoriti Orange alat
2. Odabrati novi dokument i dati mu ime *ruap_kmeans*
3. Izraditi eksperiment prema slici
4. Komentirati dobivene rezultate

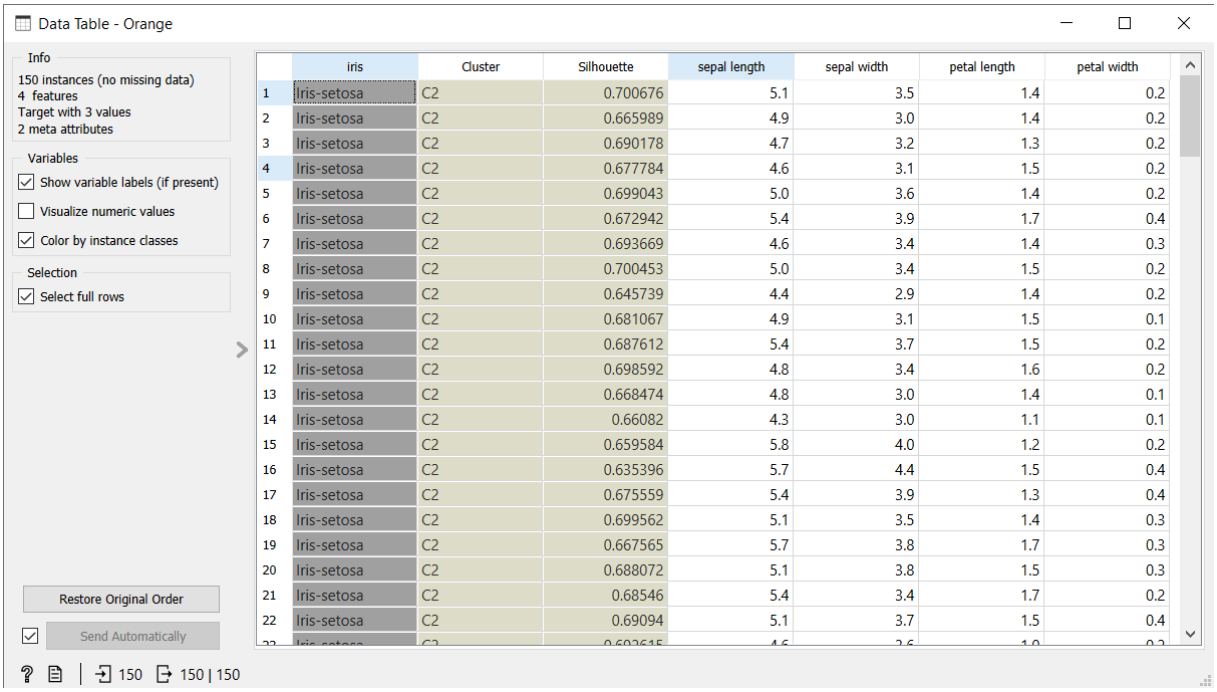


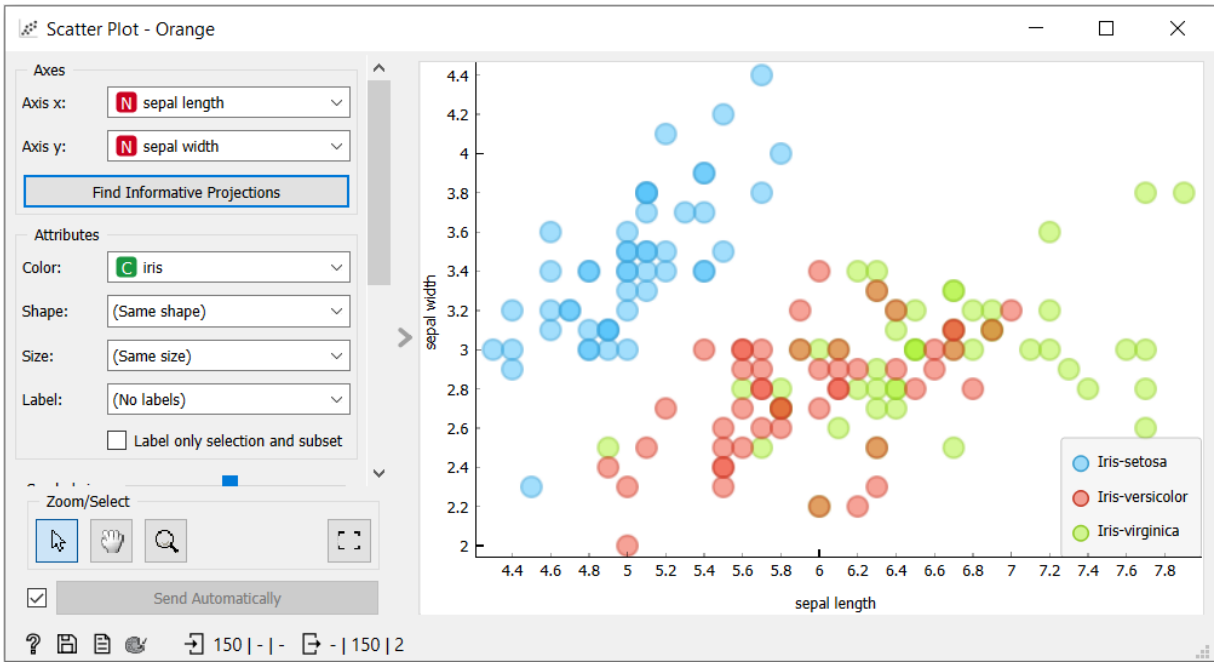
5. Ponoviti eksperiment ali na skupu podataka o vrstama vina (https://archive.ics.uci.edu/ml/datasets/Wine)
6. Pokusati odrediti odgovarajući broj grupa traženjem (optimizacijom) uporabom različith kriterija vrednovanja

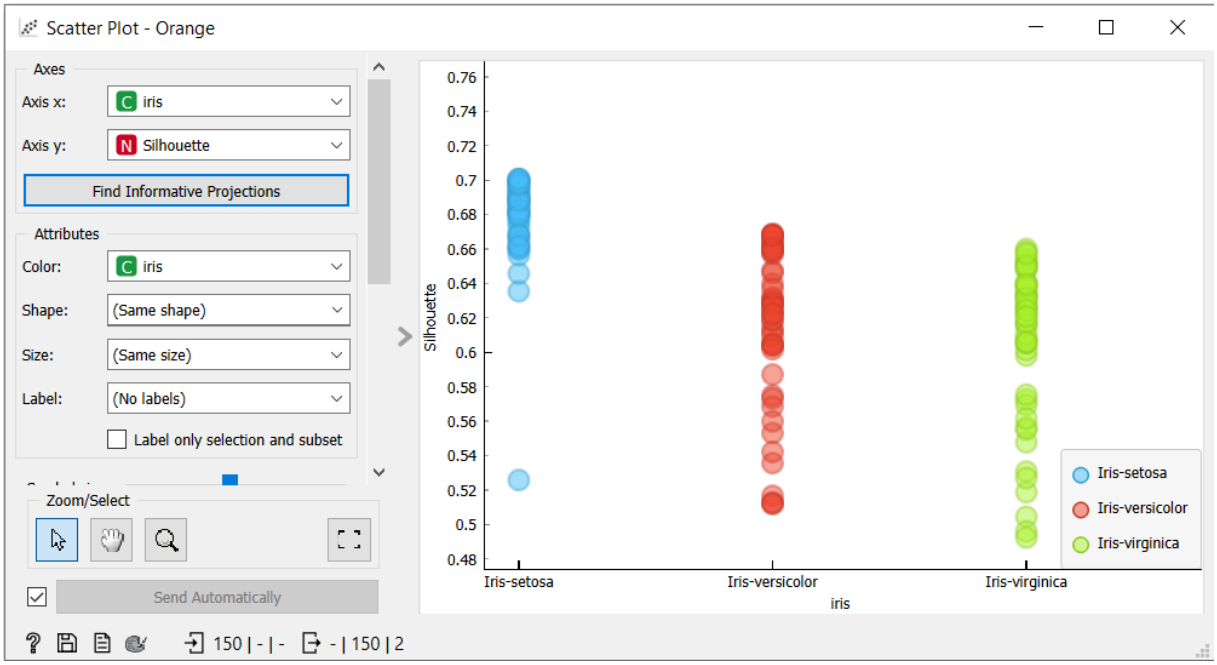
Rješenje:



Rezultati za Iris dataset





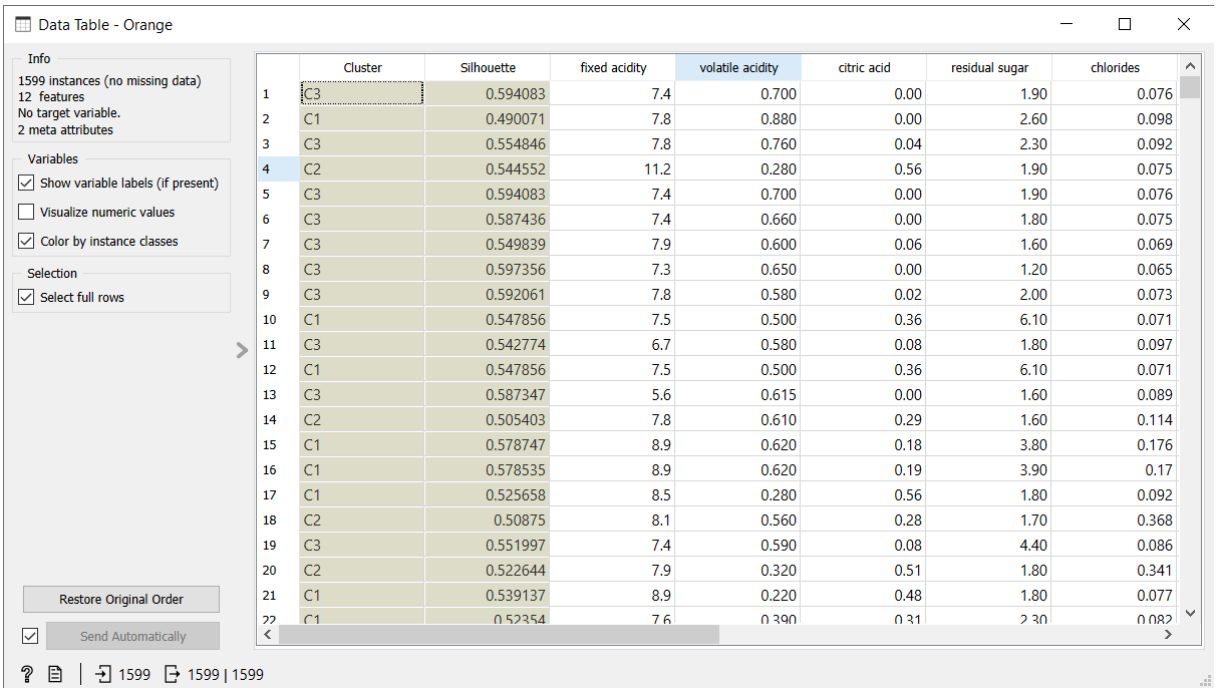


Komentar:

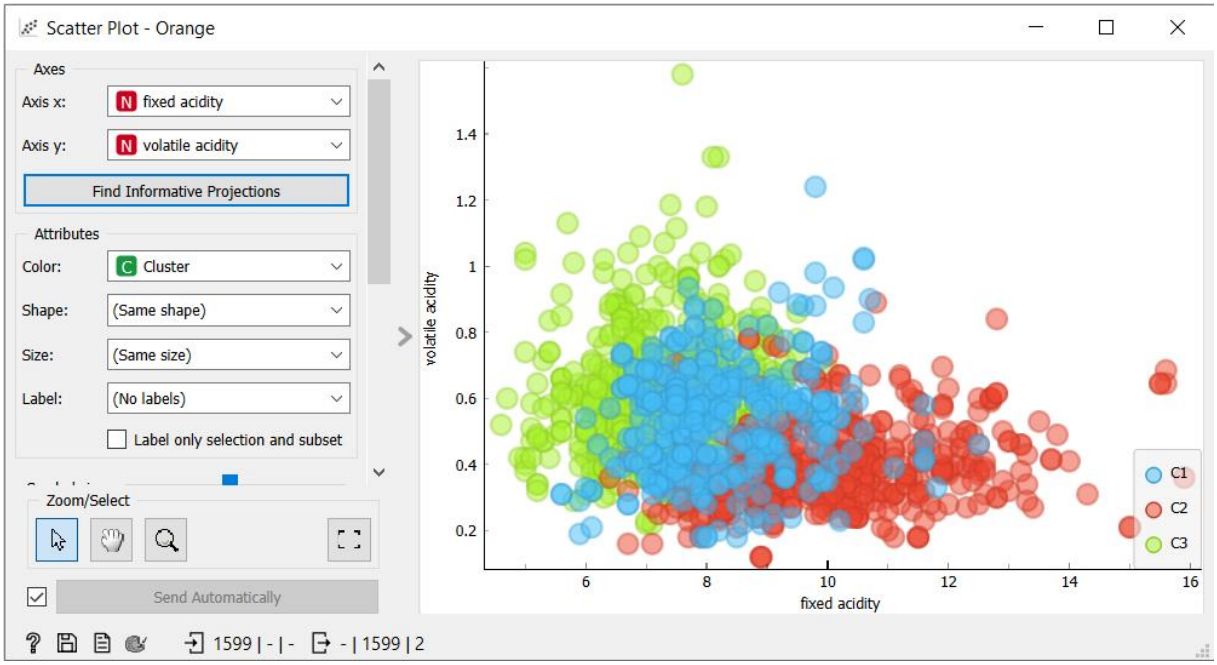
Silhouette analiza koristi se kako bismo proučili razdvajanje udaljenosti između clustera. Silhouette plot prikazuje kako je blizu svaka vrijednost u clusteru vrijednostima u susjednim clusterima te pruža način za vizualnu procjenu parametara poput broja clustera. Pogledamo li zadnju sliku, može se vidjeti da je vrijednost Silhouette otprilike 0.7.

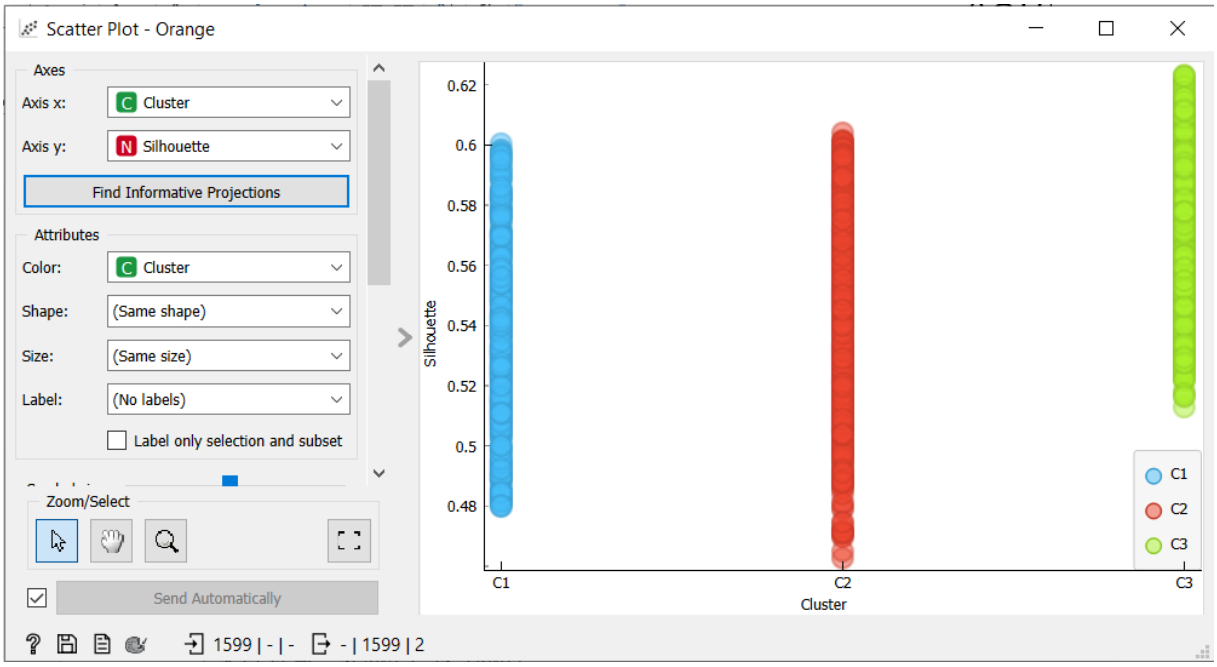
Rezultati za Wine dataset

Dataset preuzet sa stranice Kaggle datasets -> https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009?resource=download



	Cluster	Silhouette	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides
1	C3	0.594083	7.4	0.700	0.00	1.90	0.076
2	C1	0.490071	7.8	0.880	0.00	2.60	0.098
3	C3	0.554846	7.8	0.760	0.04	2.30	0.092
4	C2	0.544552	11.2	0.280	0.56	1.90	0.075
5	C3	0.594083	7.4	0.700	0.00	1.90	0.076
6	C3	0.587436	7.4	0.660	0.00	1.80	0.075
7	C3	0.549839	7.9	0.600	0.06	1.60	0.069
8	C3	0.597356	7.3	0.650	0.00	1.20	0.065
9	C3	0.592061	7.8	0.580	0.02	2.00	0.073
10	C1	0.547856	7.5	0.500	0.36	6.10	0.071
11	C3	0.542774	6.7	0.580	0.08	1.80	0.097
12	C1	0.547856	7.5	0.500	0.36	6.10	0.071
13	C3	0.587347	5.6	0.615	0.00	1.60	0.089
14	C2	0.505403	7.8	0.610	0.29	1.60	0.114
15	C1	0.578747	8.9	0.620	0.18	3.80	0.176
16	C1	0.578535	8.9	0.620	0.19	3.90	0.17
17	C1	0.525658	8.5	0.280	0.56	1.80	0.092
18	C2	0.50875	8.1	0.560	0.28	1.70	0.368
19	C3	0.551997	7.4	0.590	0.08	4.40	0.086
20	C2	0.522644	7.9	0.320	0.51	1.80	0.341
21	C1	0.539137	8.9	0.220	0.48	1.80	0.077
22	C1	0.52354	7.6	0.390	0.31	2.30	0.082





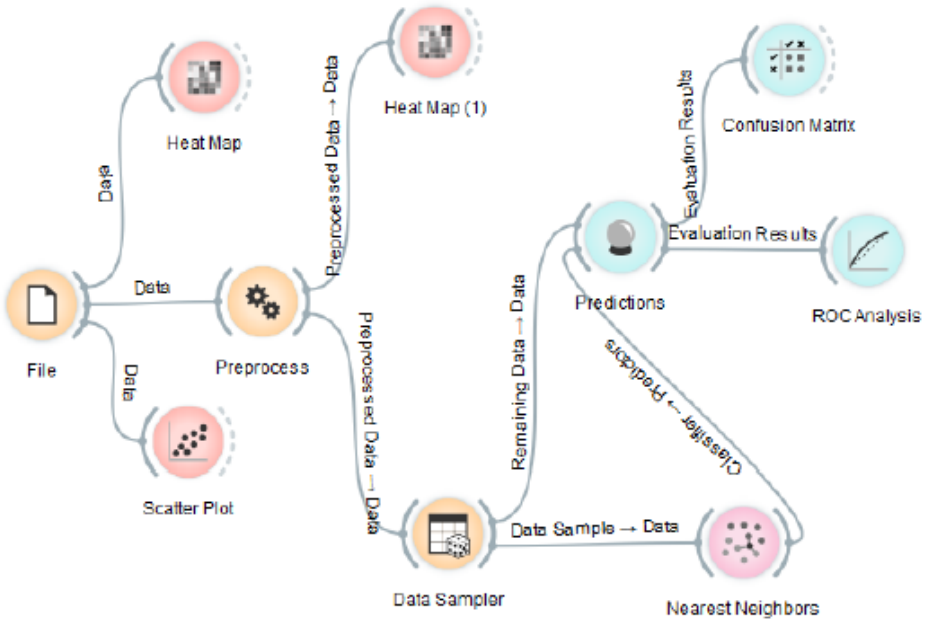
Komentar:

Usporedimo li rezultate s prethodnim datasetom, može se vidjeti da je u ovome primjeru Silhouette vrijednost bila otprilike 0.62, što je manje nego u prethodnom primjeru. Budući da se Silhouette vrijednost nalazi u intervalu [-1, 1], a vrijednosti koje se nalaze bliže vrijednosti 1 ukazuju da je uzorak daleko od susjednih clustera, može se zaključiti da je ovaj primjer imao lošiji rezultat u odnosu na prethodni.

Zadatak 5.

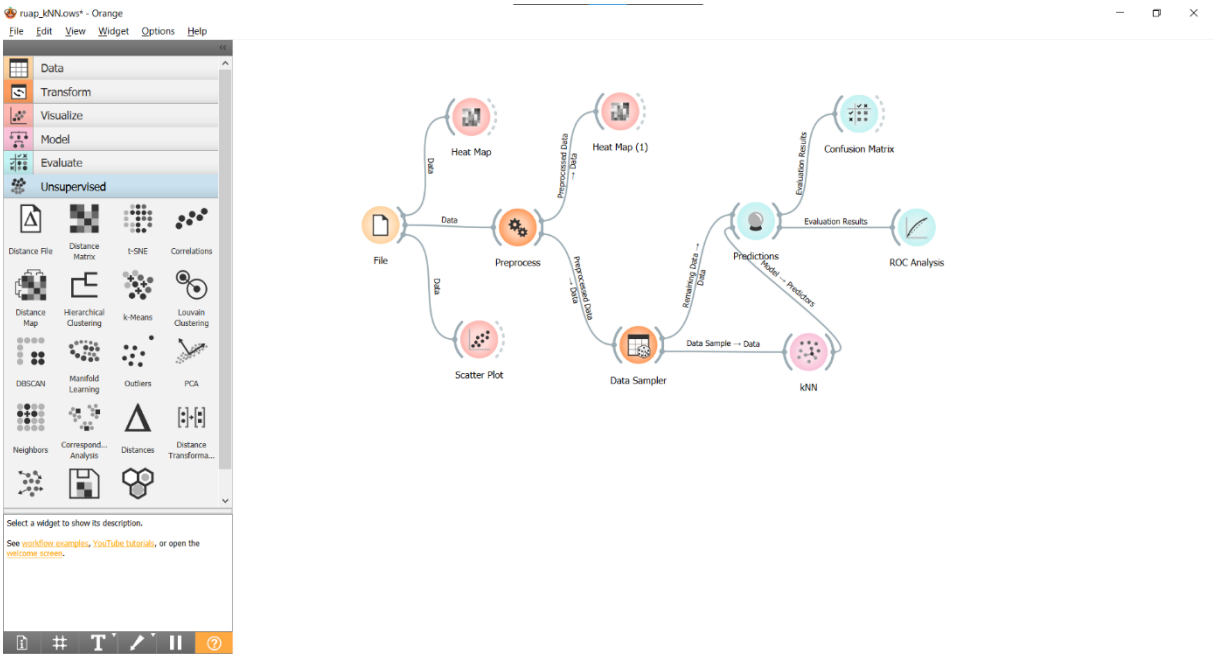
Na primjeru podatkovnog skupa provedite klasifikaciju korištenjem algoritma najbližih susjeda.

1. Otvoriti Orange alat
2. Odabrati novi dokument i dati mu ime ruap_kNN
3. Izraditi eksperiment prema slici

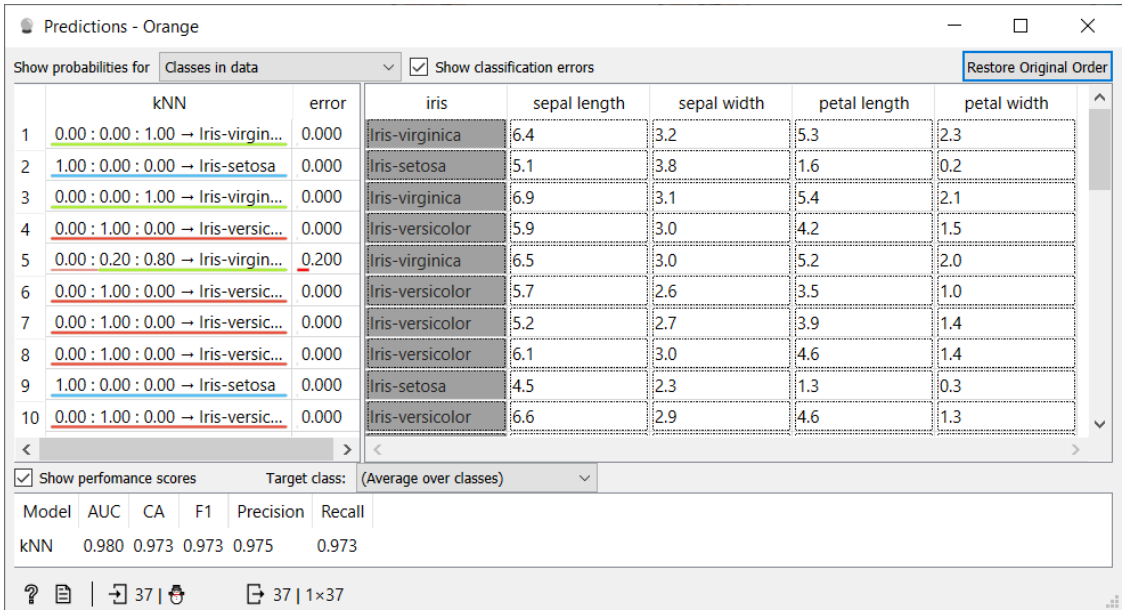


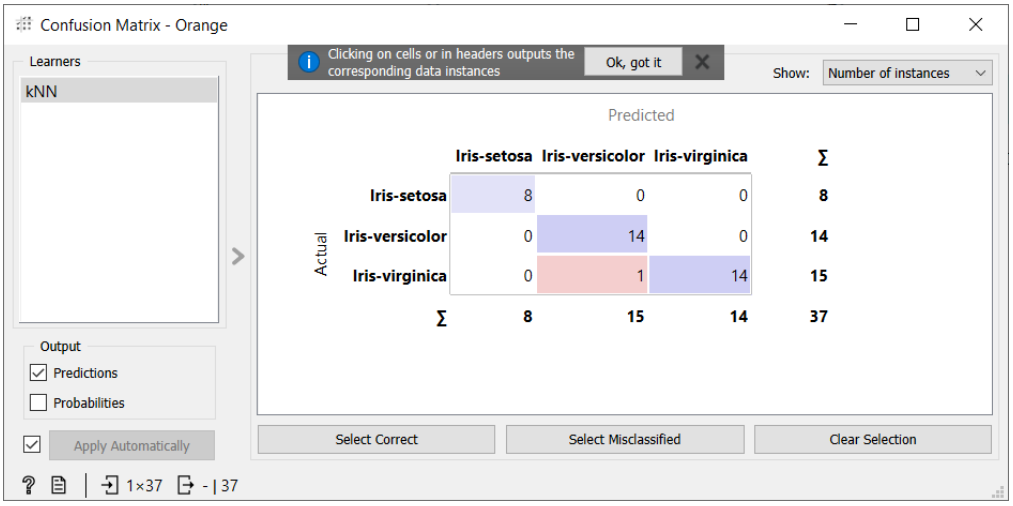
4. Za ulazne podatke odabrati Iris podatkovni skup dostupan kroz Orange alat
5. Komentirati efekt pred-obrade (omogućiti normalizaciju) podataka (Heat-mape)
6. Pokušati klasifikaciju s i bez normalizacije podataka
7. Pokušati klasifikaciju s različitim vrijednostima parametra k, različitom mjerom udaljenosti, različitim podatkovnim skupom
8. Komentirati rezultate

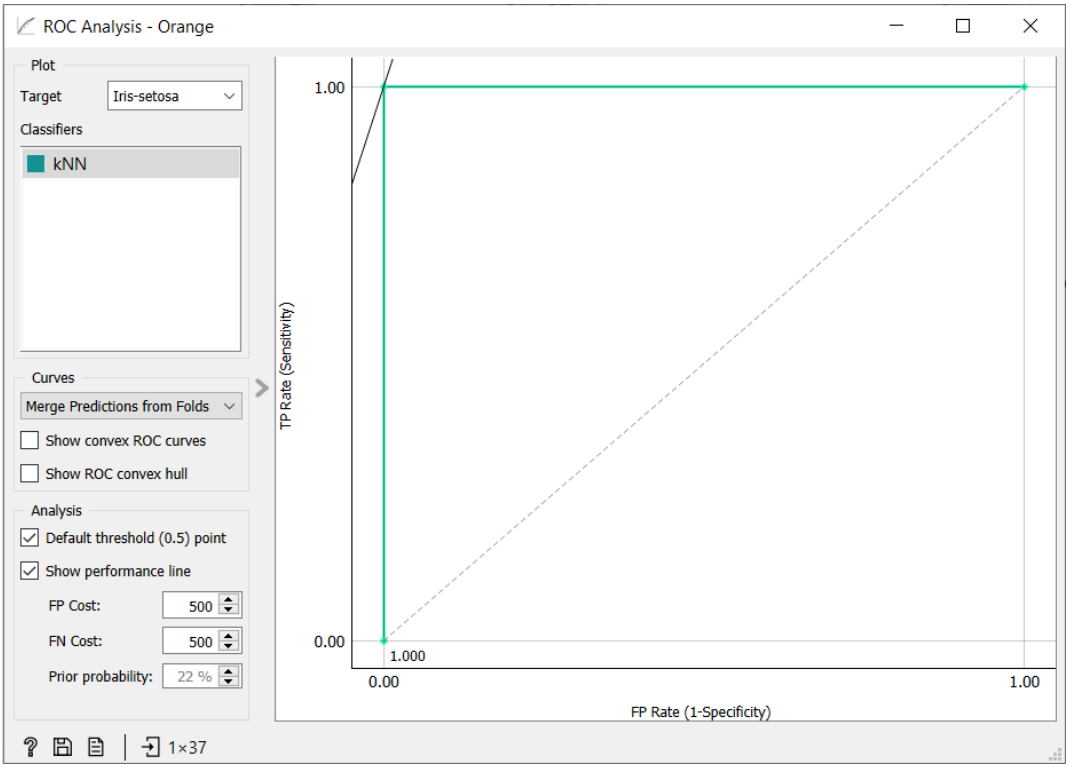
Rješenje:



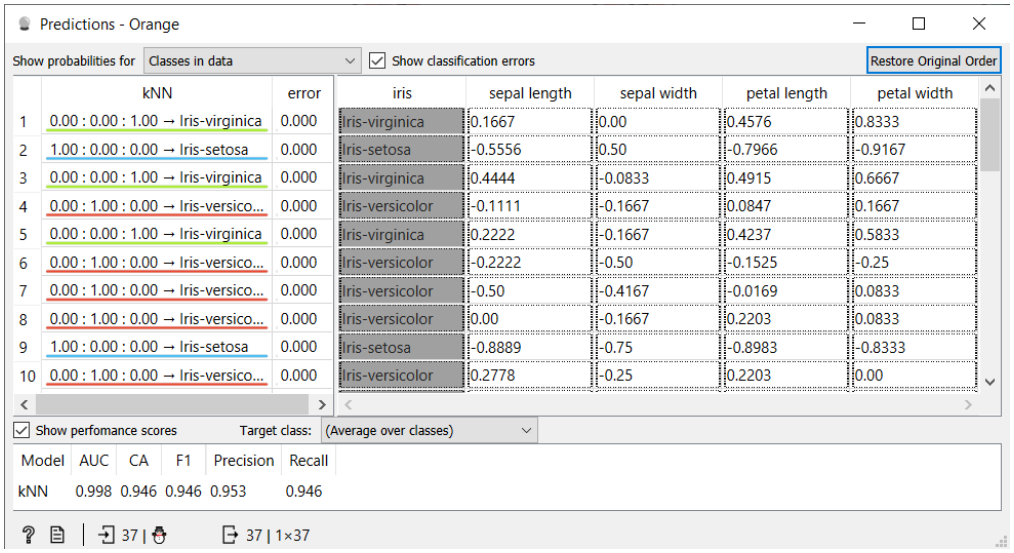
Rezultati bez normalizacije

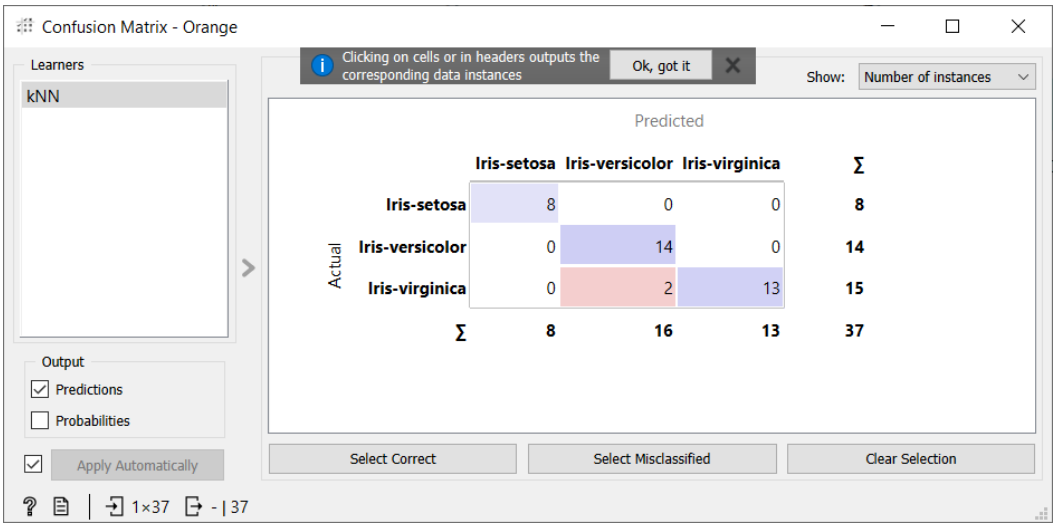


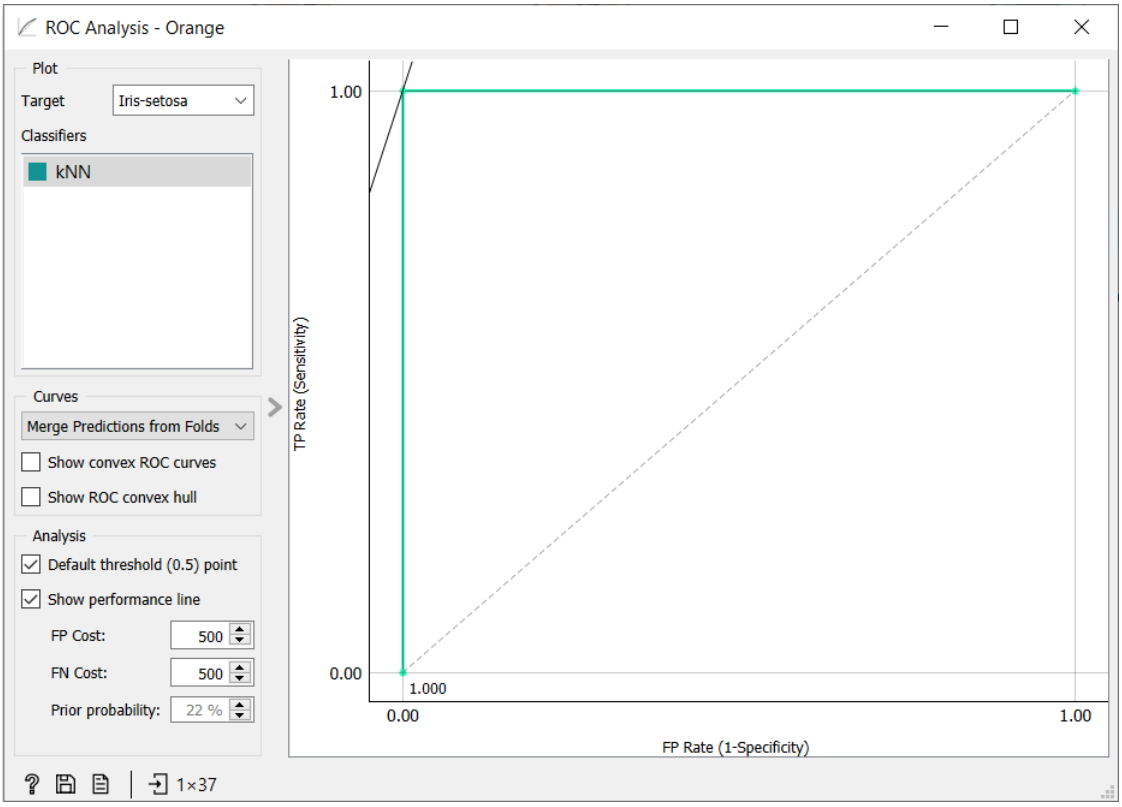




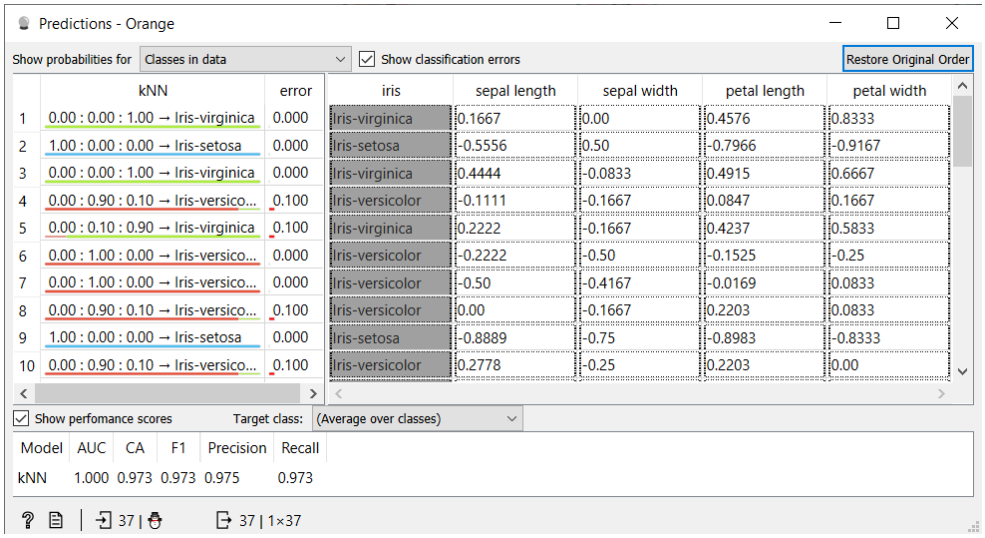
Rezultati nakon omogućene normalizacije:

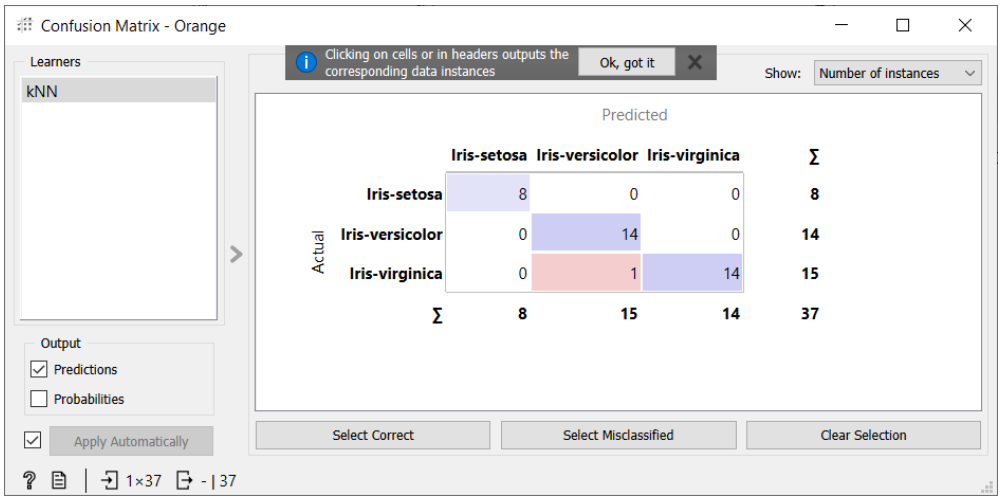


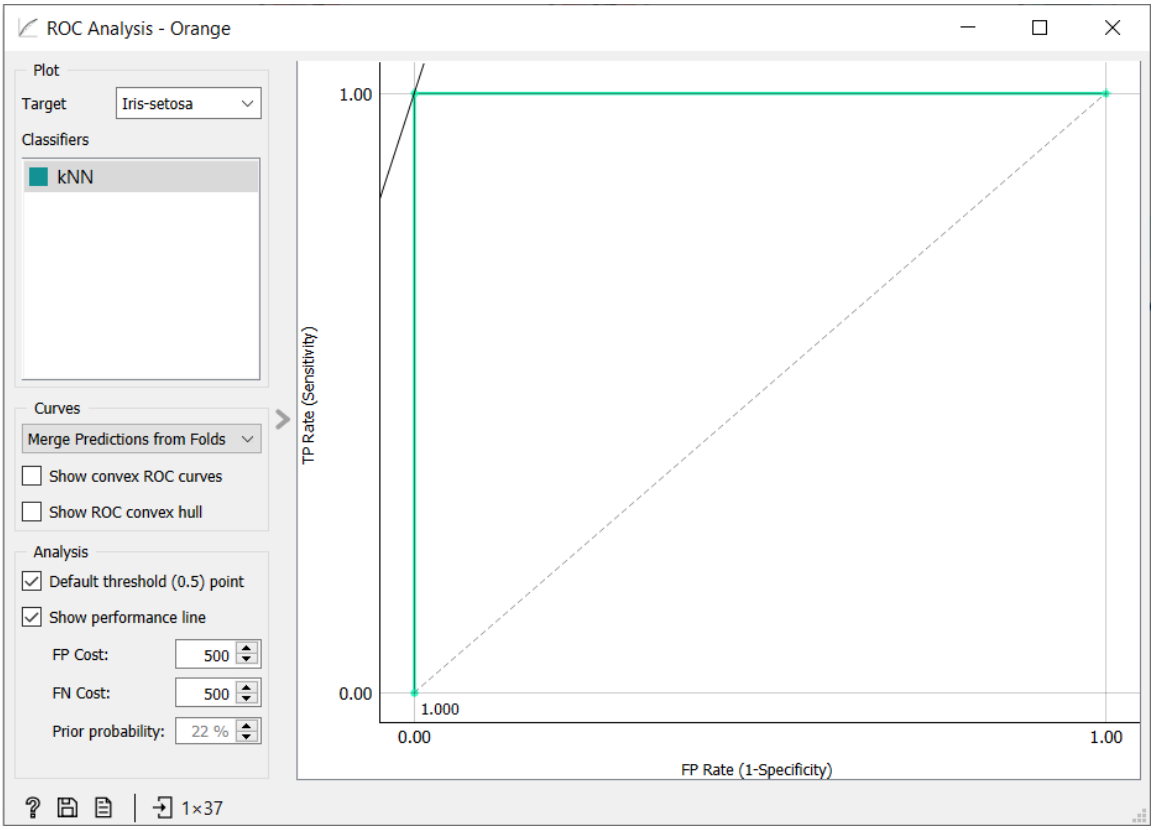




Rezultati nakon omogućene normalizacije i postavljanja parametra k = 10:







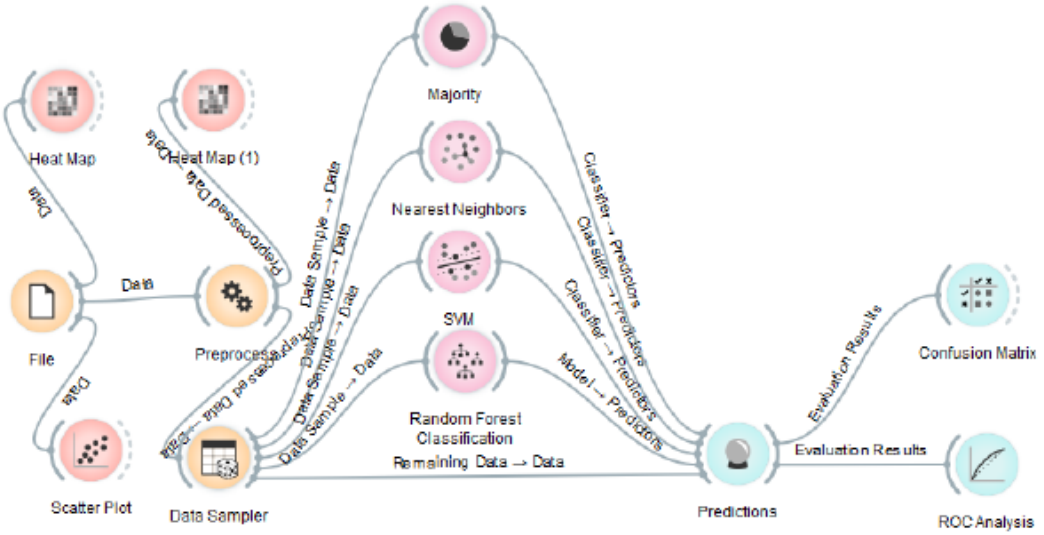
Komentar:

Usporedimo li dobivene rezultate, može se vidjeti da su rezultati nakon omogućene normalizacije nešto lošiji u usporedbi s druga dva rezultata. Uzmemo li slučaj bez normalizacije i slučaj s uključenom normalizacijom i povećanim parametrom k, vidimo da se dobiju bolji, odnosno precizniji rezultati. Analiza rezultata ostvarena je usporedbom „Precision“ vrijednosti, odnosno preciznosti, te analizom matrice gdje je cilj imati što manje vrijednosti unutar false-negative i false-positive polja.

Zadatak 6.

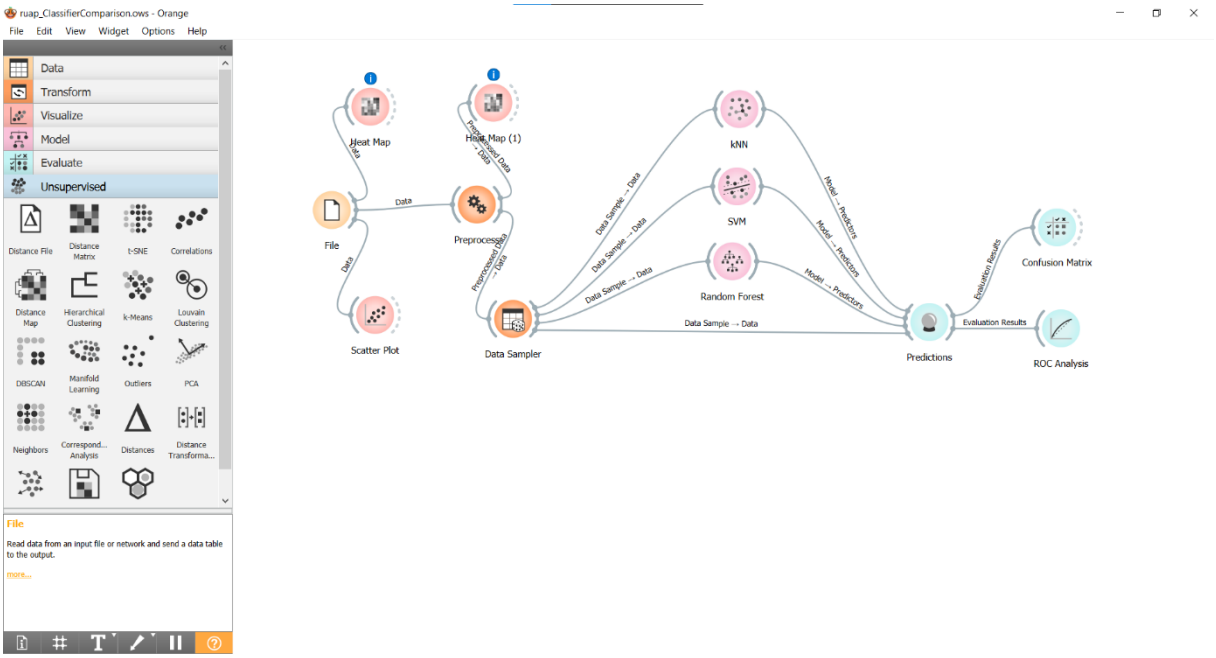
Usporedba nekoliko klasifikatora.

1. Otvoriti Orange alat
2. Odabrati novi dokument i dati mu ime ruap_ClassifierComparison
3. Izraditi eksperiment prema slici

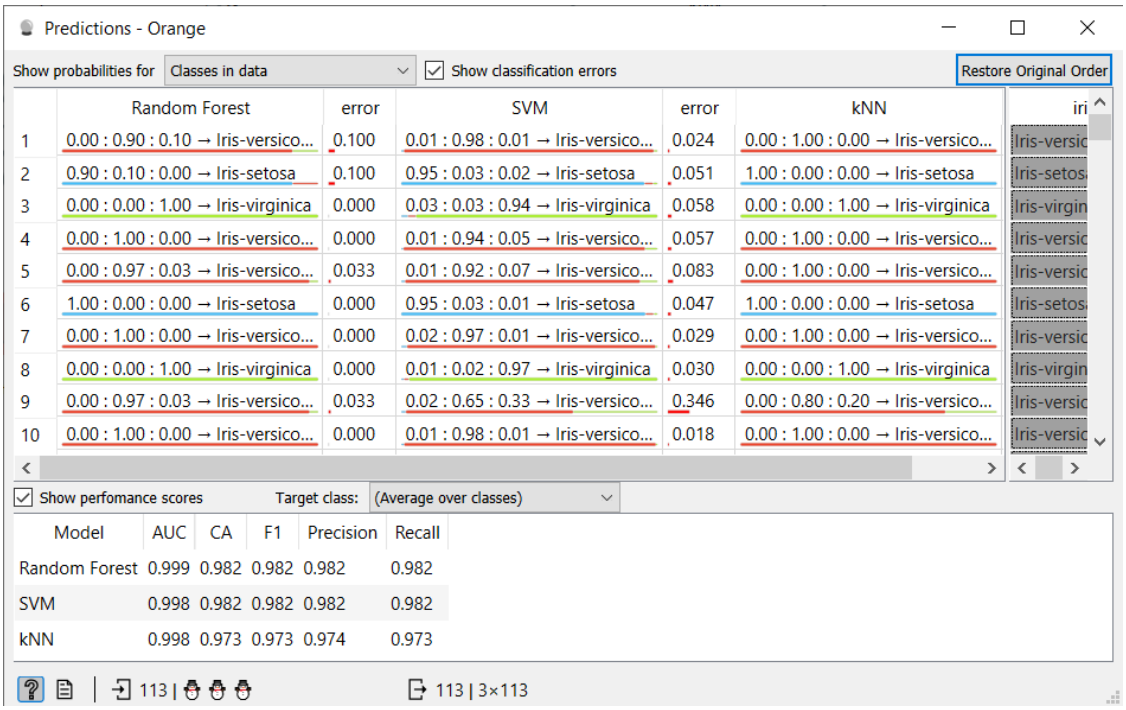


4. Provesti analizu korištenjem nekoliko klasifikatora na tri podatkovna skupa preuzeta s UCI repozitorija. Navesti u tablici informacije o podatkovnim skupovima koji su korišteni.
5. Komentirati rezultate. Koji klasifikator smatrate najboljim, a koji najgorim i zašto?

Rješenje:



Rezultati za Iris dataset



Predictions - Orange

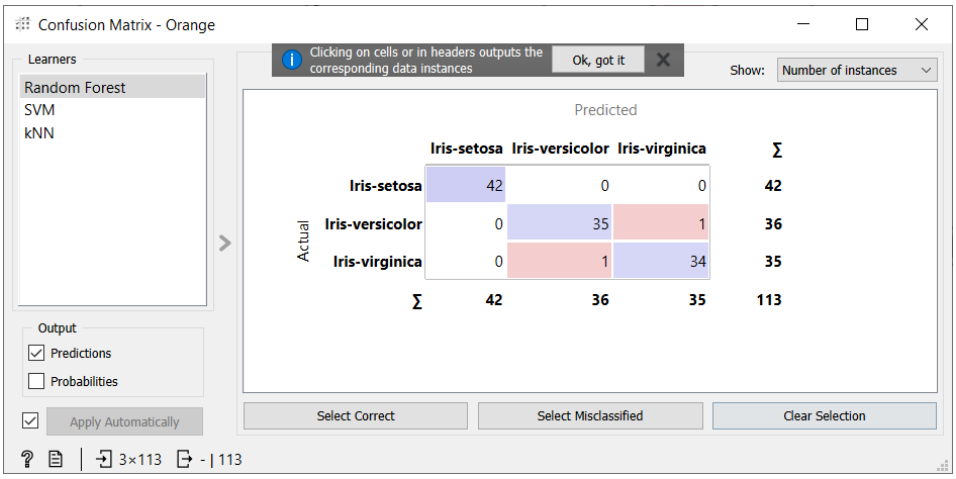
Show probabilities for: Classes in data | Show classification errors | Restore Original Order

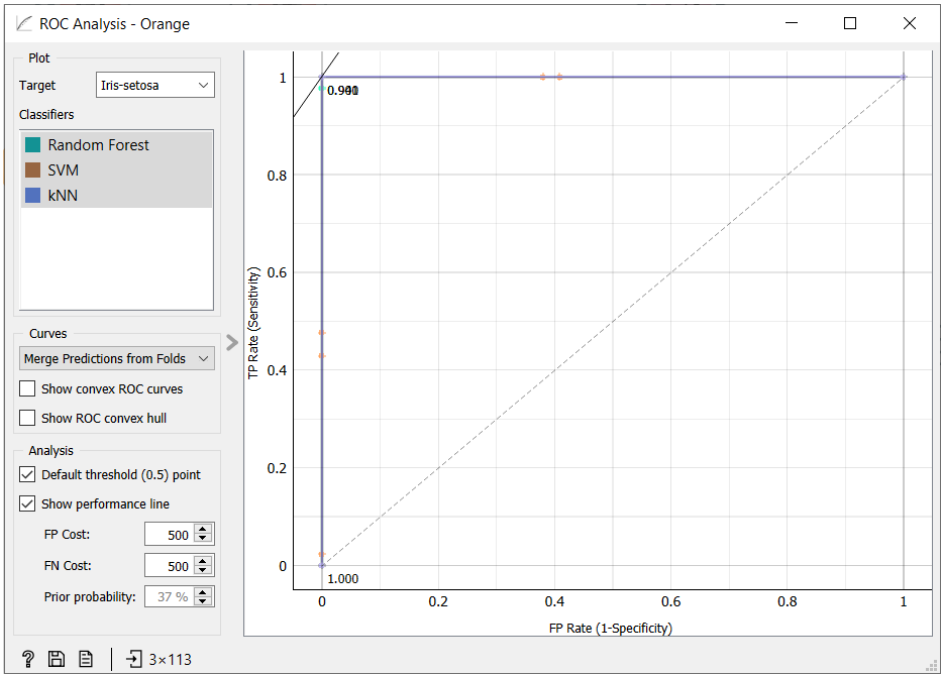
	Random Forest	error	SVM	error	kNN	iri
1	0.00 : 0.90 : 0.10 → Iris-versico...	0.100	0.01 : 0.98 : 0.01 → Iris-versico...	0.024	0.00 : 1.00 : 0.00 → Iris-versico...	Iris-versic
2	0.90 : 0.10 : 0.00 → Iris-setosa	0.100	0.95 : 0.03 : 0.02 → Iris-setosa	0.051	1.00 : 0.00 : 0.00 → Iris-setosa	Iris-setos
3	0.00 : 0.00 : 1.00 → Iris-virginica	0.000	0.03 : 0.03 : 0.94 → Iris-virginica	0.058	0.00 : 0.00 : 1.00 → Iris-virginica	Iris-virgin
4	0.00 : 1.00 : 0.00 → Iris-versico...	0.000	0.01 : 0.94 : 0.05 → Iris-versico...	0.057	0.00 : 1.00 : 0.00 → Iris-versico...	Iris-versic
5	0.00 : 0.97 : 0.03 → Iris-versico...	0.033	0.01 : 0.92 : 0.07 → Iris-versico...	0.083	0.00 : 1.00 : 0.00 → Iris-versico...	Iris-versic
6	1.00 : 0.00 : 0.00 → Iris-setosa	0.000	0.95 : 0.03 : 0.01 → Iris-setosa	0.047	1.00 : 0.00 : 0.00 → Iris-setosa	Iris-setos
7	0.00 : 1.00 : 0.00 → Iris-versico...	0.000	0.02 : 0.97 : 0.01 → Iris-versico...	0.029	0.00 : 1.00 : 0.00 → Iris-versico...	Iris-versic
8	0.00 : 0.00 : 1.00 → Iris-virginica	0.000	0.01 : 0.02 : 0.97 → Iris-virginica	0.030	0.00 : 0.00 : 1.00 → Iris-virginica	Iris-virgin
9	0.00 : 0.97 : 0.03 → Iris-versico...	0.033	0.02 : 0.65 : 0.33 → Iris-versico...	0.346	0.00 : 0.80 : 0.20 → Iris-versico...	Iris-versic
10	0.00 : 1.00 : 0.00 → Iris-versico...	0.000	0.01 : 0.98 : 0.01 → Iris-versico...	0.018	0.00 : 1.00 : 0.00 → Iris-versico...	Iris-versic

Show perfomance scores | Target class: (Average over classes)

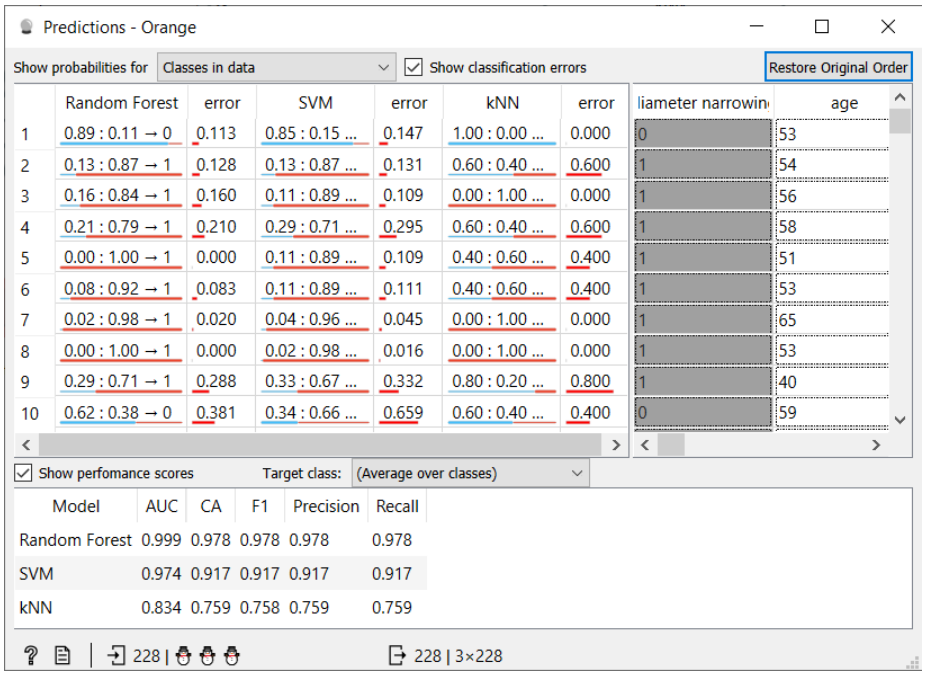
Model	AUC	CA	F1	Precision	Recall
Random Forest	0.999	0.982	0.982	0.982	0.982
SVM	0.998	0.982	0.982	0.982	0.982
kNN	0.998	0.973	0.973	0.974	0.973

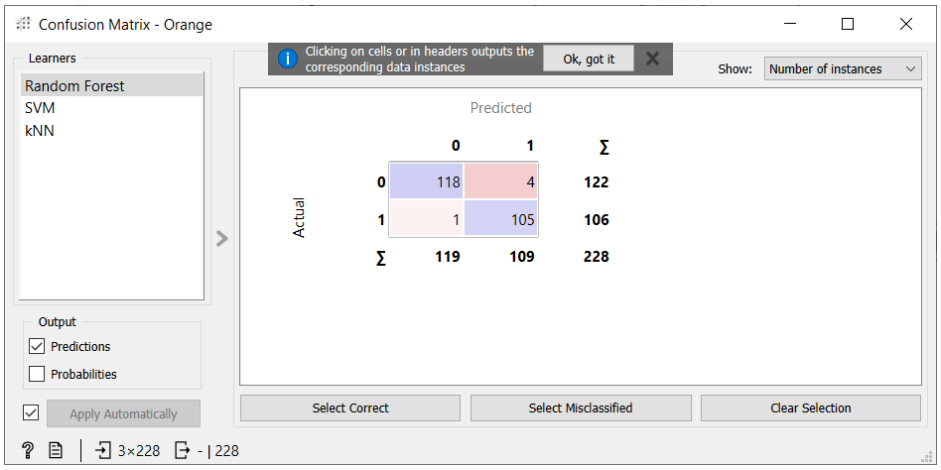
113 | 113 | 3×113

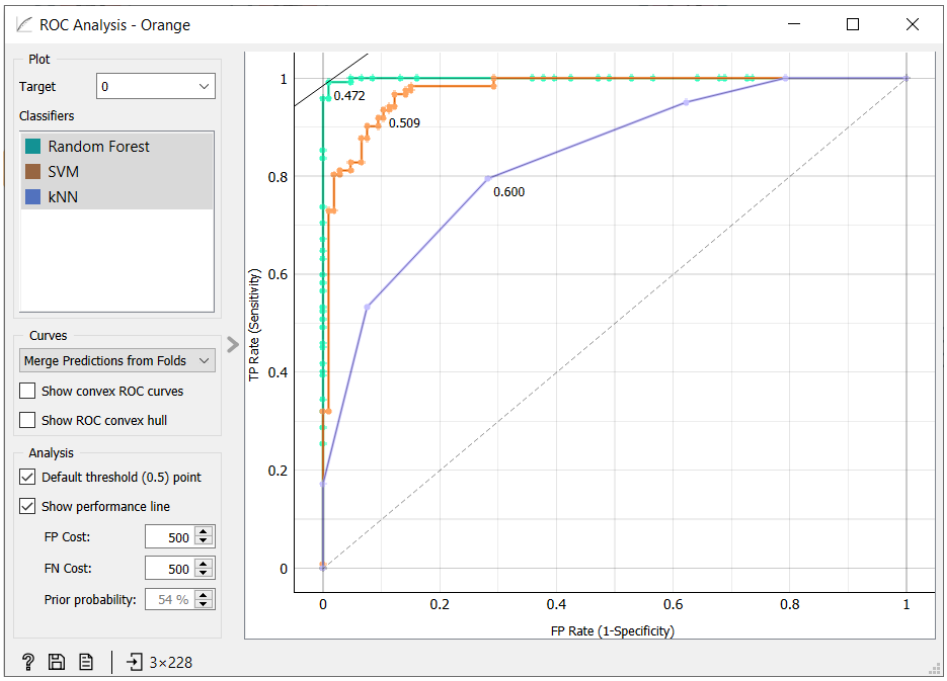




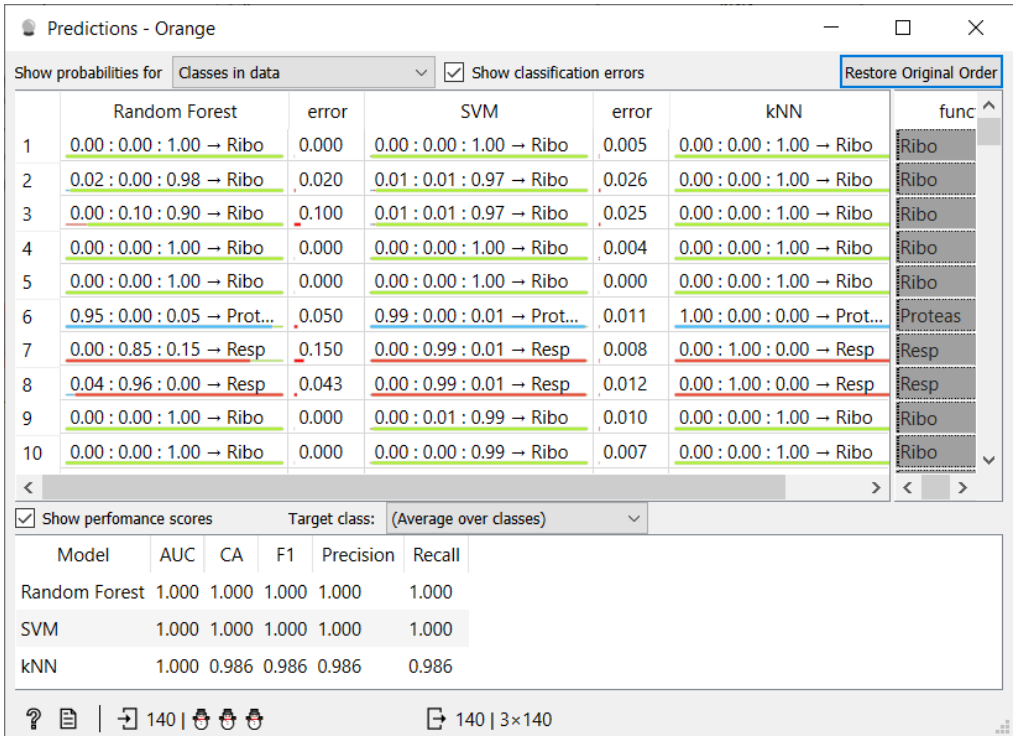
Rezultati za Heart disease dataset

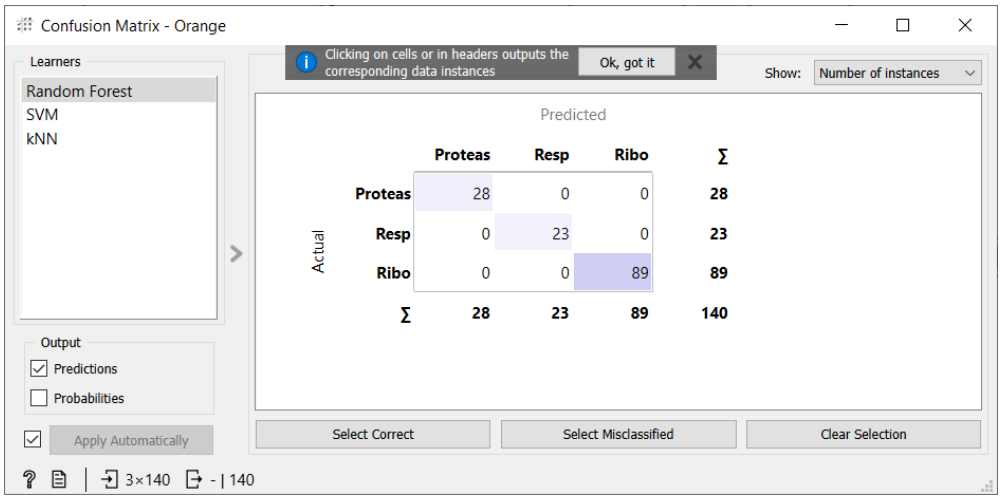


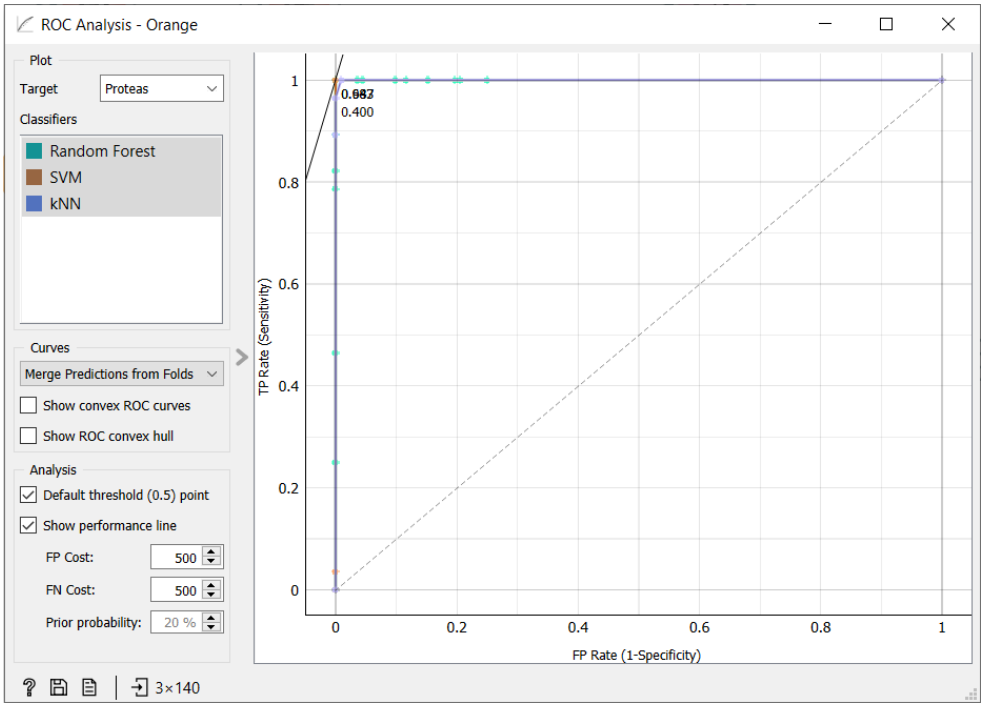




Rezultati za Brown selected dataset







Komentar:

U ovome se zadatku koriste tri dataseta: Iris, Heart disease i Brown selected. U prikazanim rezultatima je vidljivo kako konačni rezultati ovise o količini podataka (feature i target). Gledamo li dataset s manjom količinom podataka, na osnovu rezultata možemo zaključiti da su primjereni algoritmi kNN i Random Forest. Nedostatak navedenih algoritama je taj što nisu brzi kao SVM. Osim toga, SVM i Random Forest imaju vrlo visoku preciznost. Na kraju ako usporedimo preciznosti na sva tri dataseta, kNN ima najmanju preiznost, a nakon njega slijede Random Forest i SVM koji su podjednaki. Za Heart disease je Random Forest precizniji od SVM-a, dok su za druga dva dataseta oba jednaka. kNN je u sva tri slučaja imao najmanju preciznost.