

Računarstvo usluga i analiza podataka

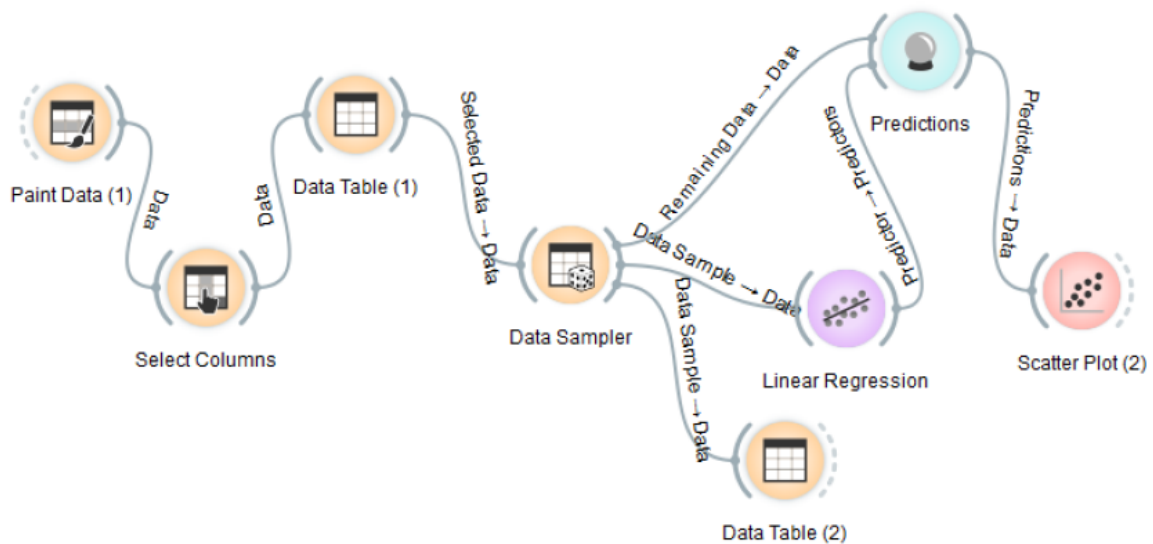
Izveštaj pete laboratorijske vježbe

Marko Ćosić

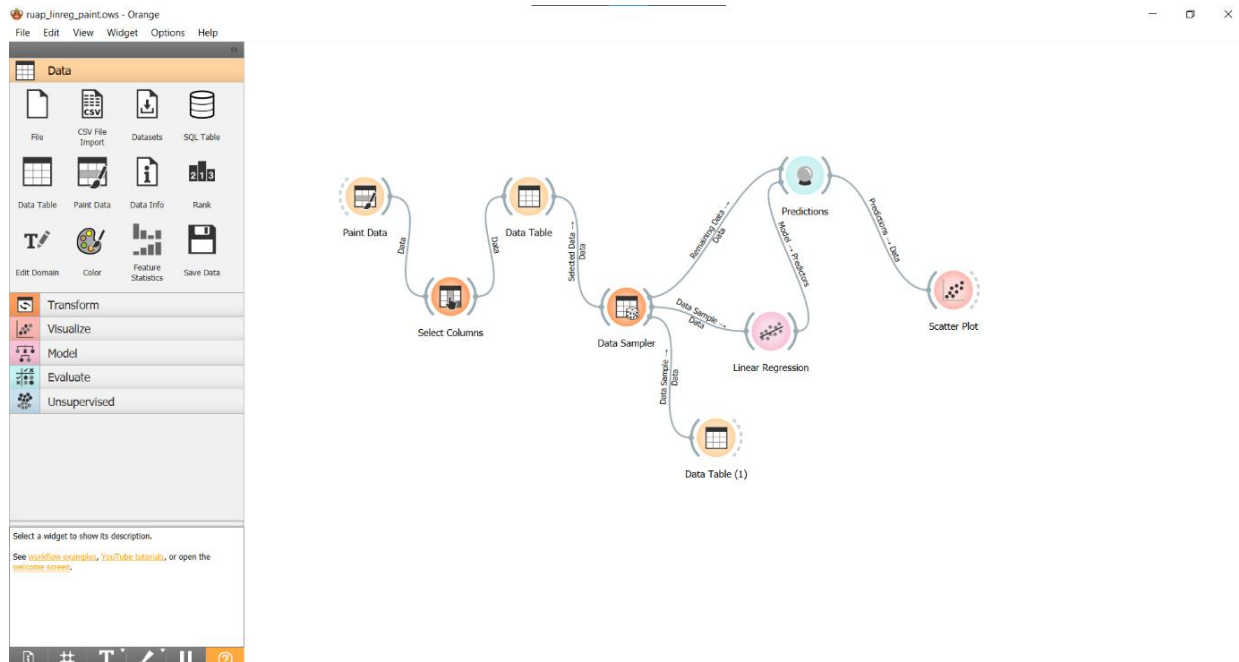
Zadatak 1.

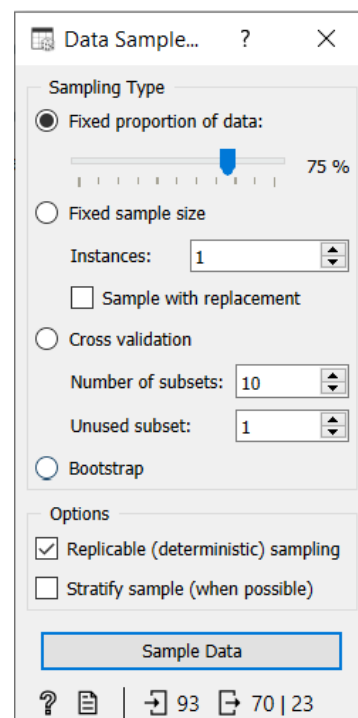
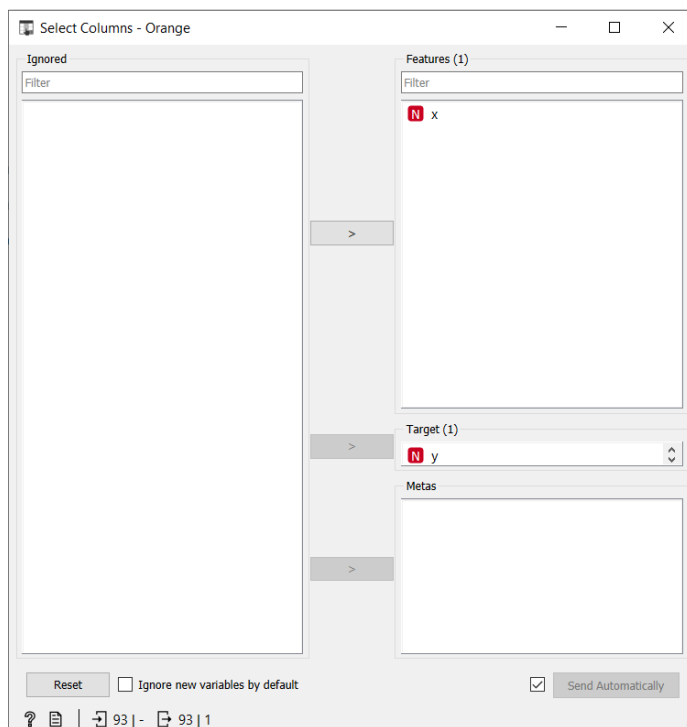
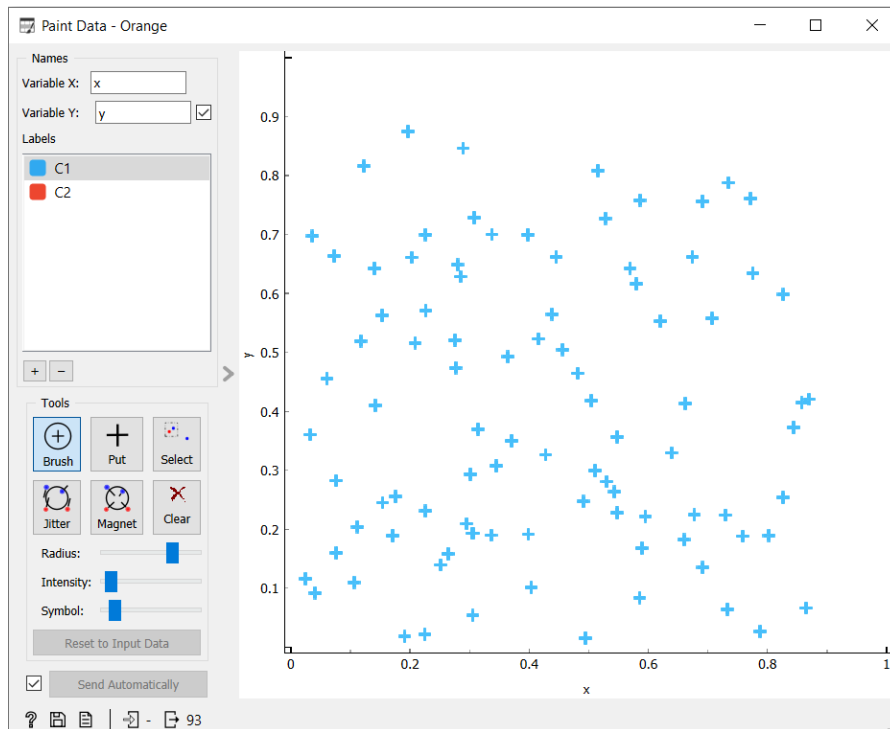
Izraditi linearni regresijski model za predviđanje umjetno generiranih podataka u 2D.

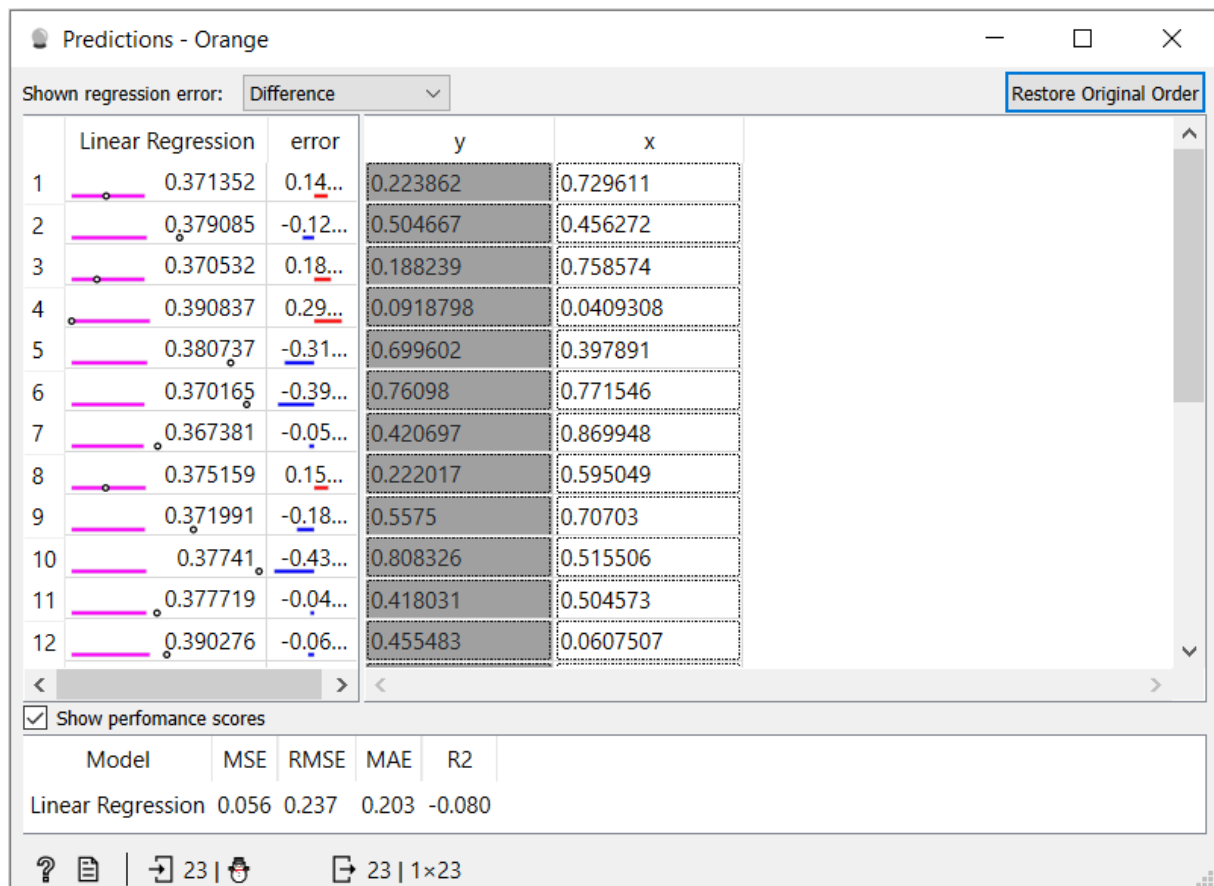
1. Otvoriti Orange alat
2. Odabrati novi dokument i dati mu ime *ruap_linreg_paint*
3. Izraditi eksperiment prema slici
4. "Nacrtati" podatke za jednu klasu
5. Odabrati podatke x-osi kao značajke (na temelju čega se predviđa), a podatke y-osi kao ciljne (vrijednost koju se predviđa)
6. Podijeliti skup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



Rješenje:







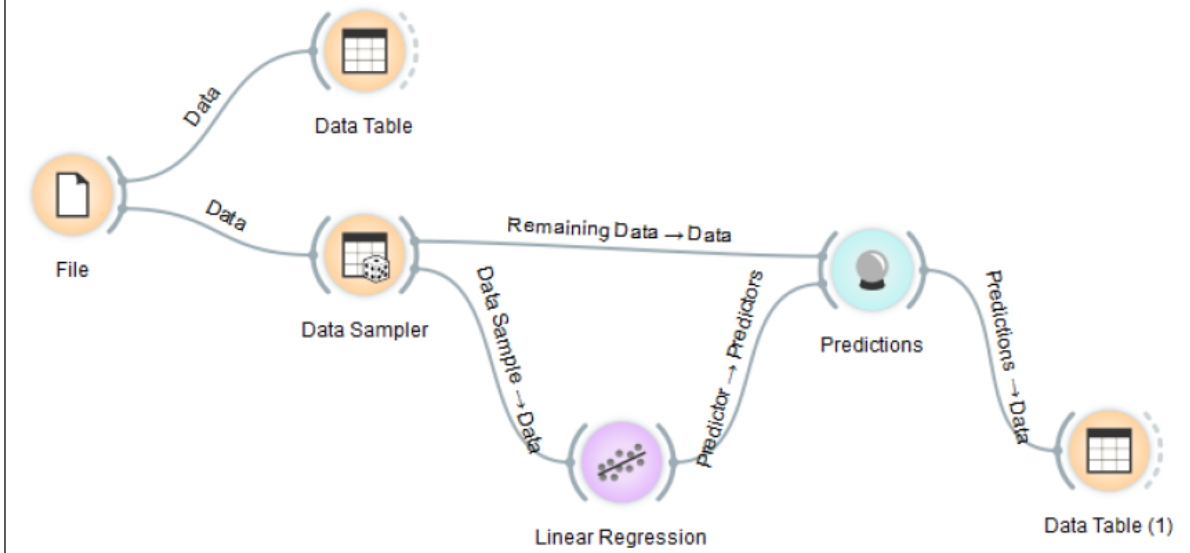
Komentar:

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.

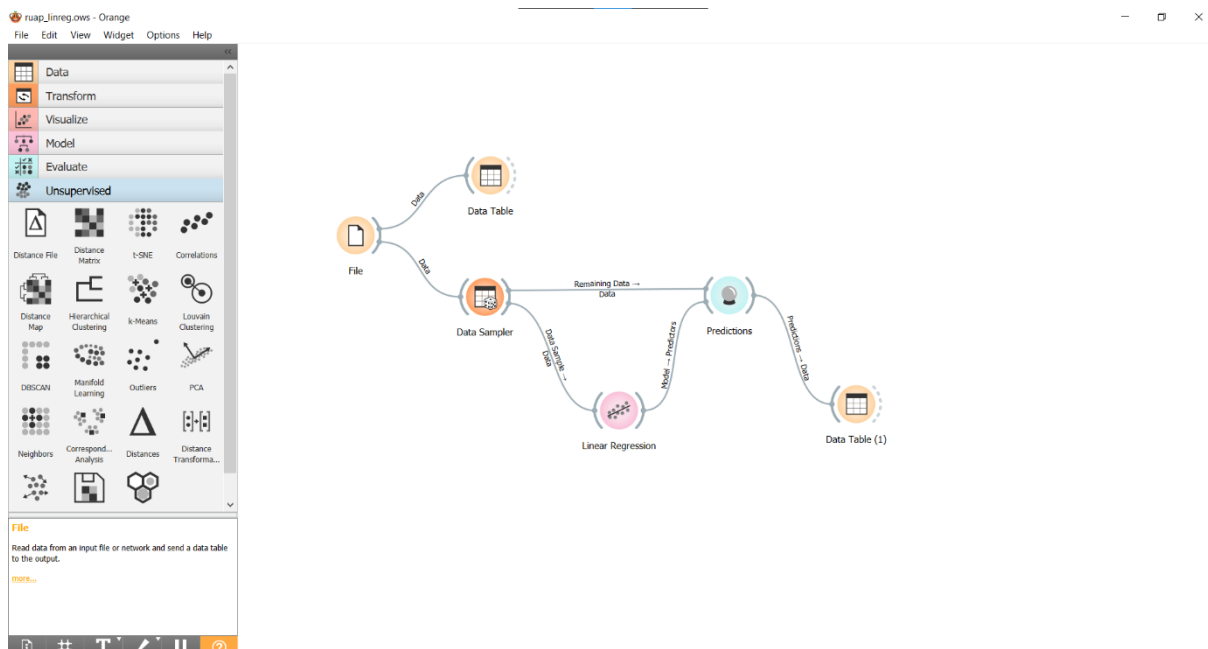
Zadatak 2.

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:



Data Table - Orange

Info
506 instances (no missing data)
13 features
Numeric outcome
No meta attributes.

Variables
☒ Show variable labels (if present)
☐ Visualize numeric values
☒ Color by instance classes

Selection
☒ Select full rows

Restore Original Order

☒ Send Automatically

	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	

506 | 506

Data Table (1) - Orange

Info
126 instances (no missing data)
13 features
Numeric outcome
1 meta attribute

Variables
☒ Show variable labels (if present)
☐ Visualize numeric values
☒ Color by instance classes

Selection
☒ Select full rows

Restore Original Order

☒ Send Automatically

	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	

126 | 126

Predictions - Orange

Shown regression error: Difference

Restore Original Order

	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

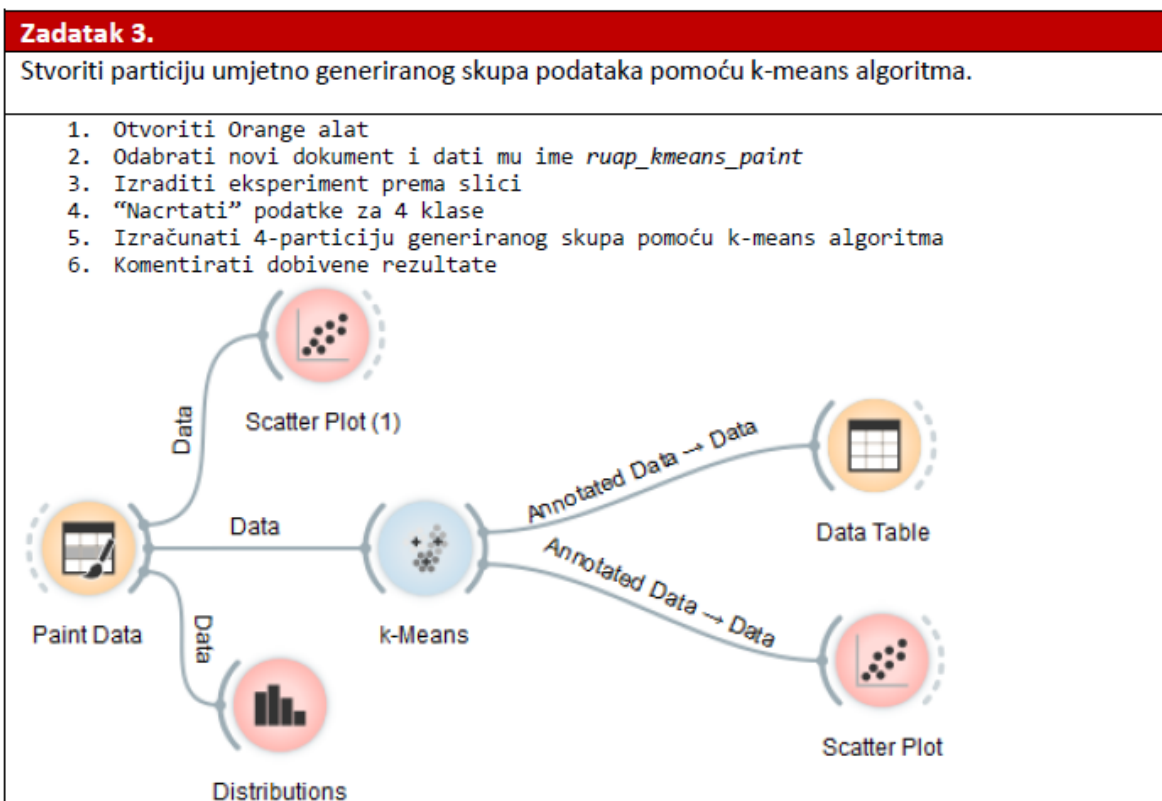
☒ Show performance scores

Model	MSE	RMSE	MAE	R2
Linear Regression	21.096	4.593	3.153	0.685

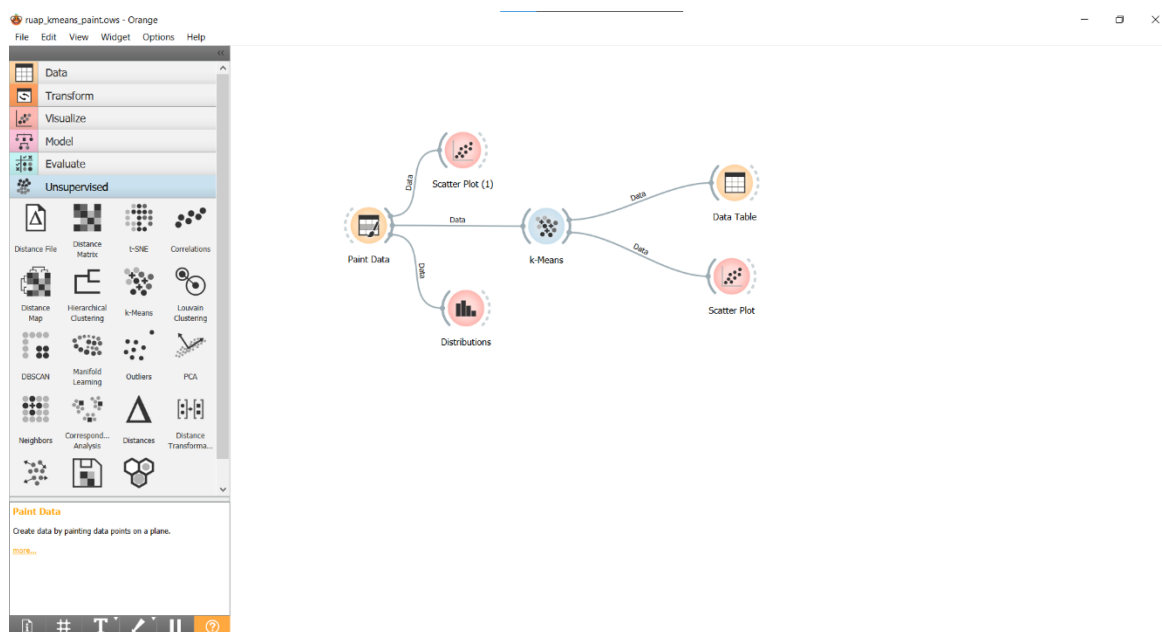
126 | 1x126

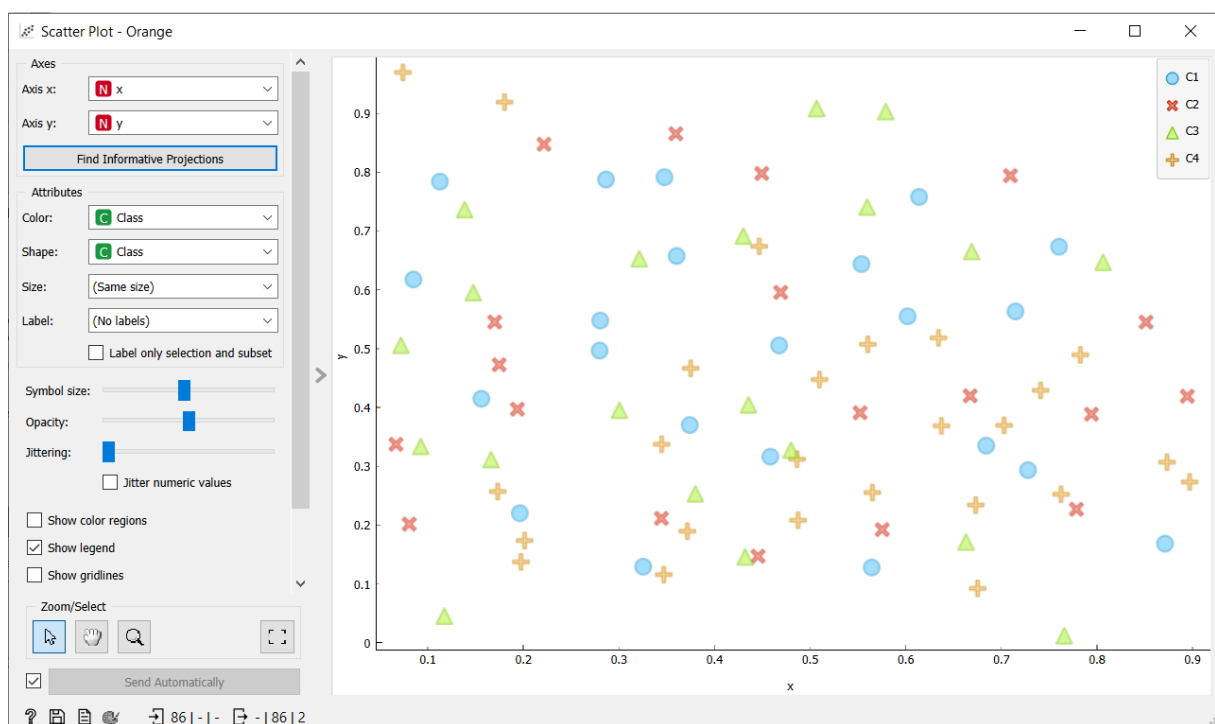
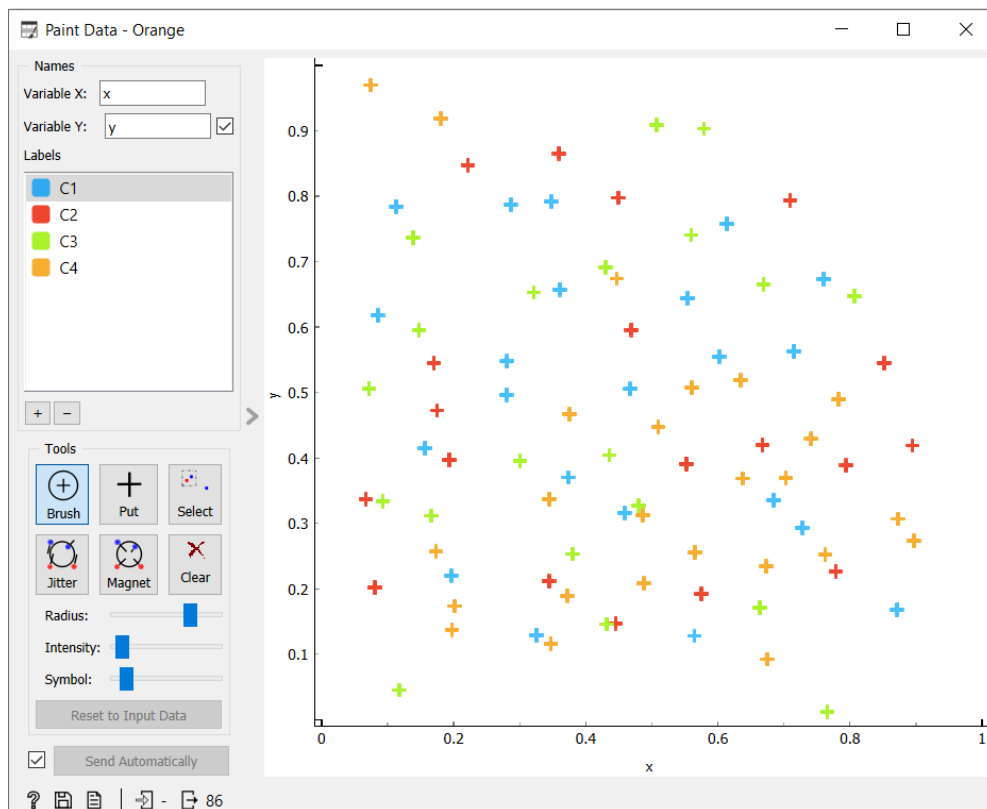
Komentar:

U odnosu 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 usporedbi s prošlim zadatkom, u ovome imamo više značajki na osnovu kojih radimo testiranje, stoga su i rezultati drugačiji.



Rješenje:





Data Table - Orange

Info
86 instances (no missing data)
2 features
Target with 4 values
2 meta attributes

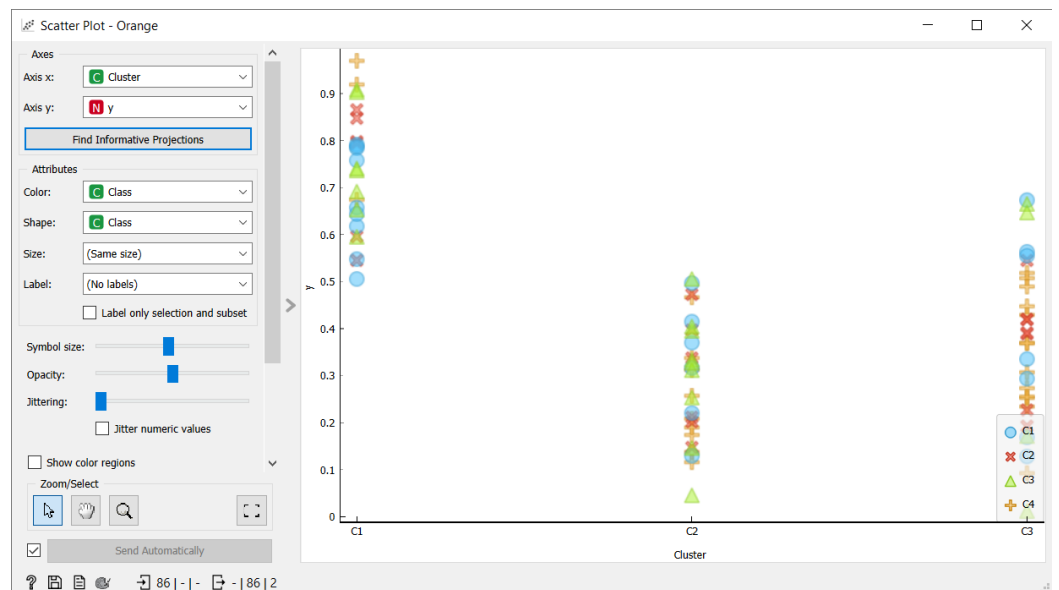
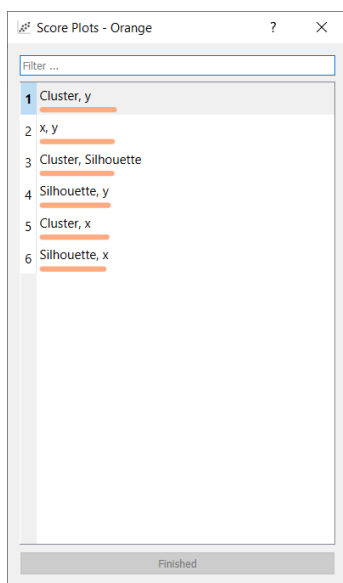
Variables
☒ Show variable labels (if present)
☐ Visualize numeric values
☒ Color by instance classes

Selection
☒ Select full rows

Restore Original Order

☒ Send Automatically

	Class	Cluster	Silhouette	x	y
1	C1	C1	0.670124	0.347684	0.791434
2	C1	C1	0.557691	0.553741	0.643809
3	C1	C3	0.565338	0.601965	0.554963
4	C1	C3	0.619154	0.714852	0.563287
5	C1	C3	0.561955	0.760353	0.673312
6	C1	C1	0.581394	0.614016	0.757927
7	C1	C1	0.638393	0.360395	0.657529
8	C1	C1	0.518843	0.280532	0.547724
9	C1	C2	0.679872	0.196409	0.220218
10	C1	C3	0.646289	0.87119	0.168437
11	C1	C3	0.664231	0.727898	0.293454
12	C1	C3	0.520722	0.564469	0.127946
13	C1	C2	0.652653	0.325429	0.129369
14	C1	C2	0.624807	0.156212	0.41473
15	C1	C2	0.576836	0.458435	0.316315
16	C1	C2	0.541912	0.279881	0.496678
17	C1	C1	0.549766	0.0850957	0.617557
18	C1	C1	0.630163	0.112793	0.78406
19	C1	C1	0.664982	0.286523	0.787473
20	C1	C3	0.659108	0.684239	0.335106
21	C1	C1	0.497753	0.467616	0.505517
22	C1	C2	0.634999	0.374163	0.370107



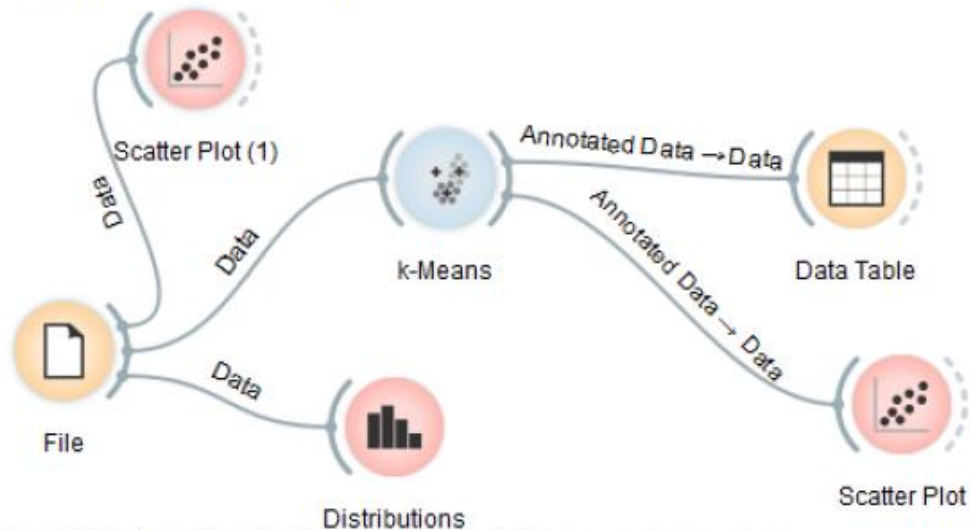
Komentar:

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.

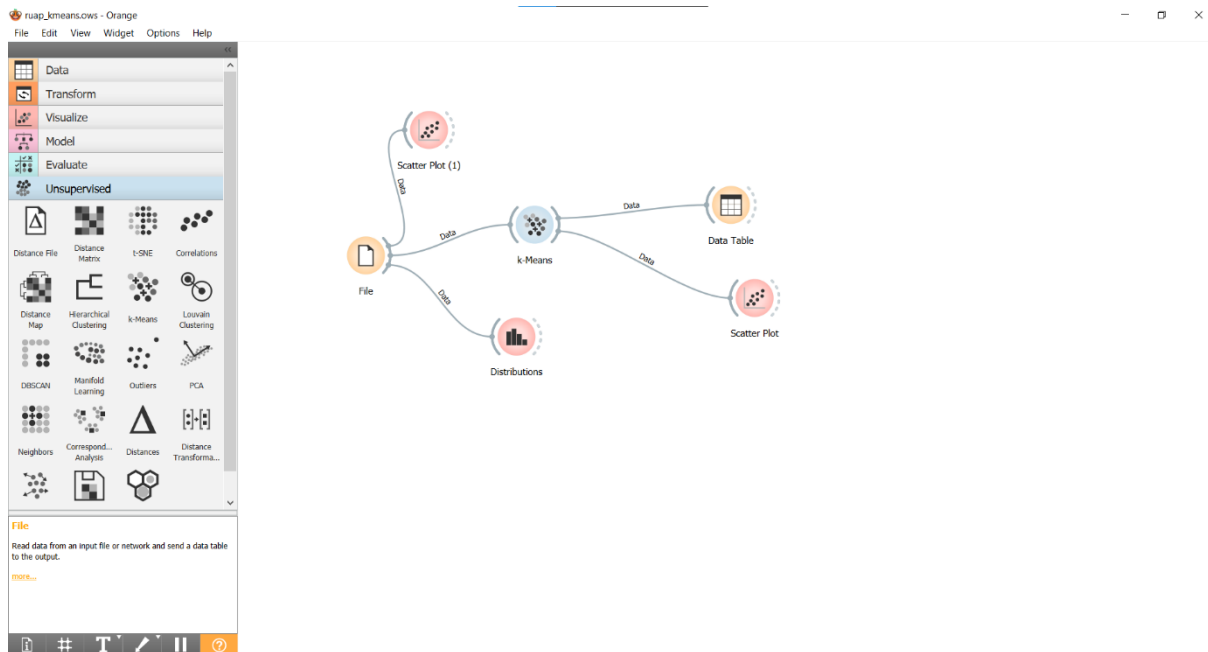
Stvoriti particiju skupa podataka o cvijetu perunike (<https://archive.ics.uci.edu/ml/datasets/Iris>) pomoću k-means algoritma.

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. Pokušati odrediti odgovarajući broj grupa traženjem (optimizacijom) uporabom različitih kriterija vrednovanja

Rješenje:



Rezultati za Iris dataset

Data Table - Orange

Info
150 instances (no missing data)
4 features
Target with 3 values
2 meta attributes

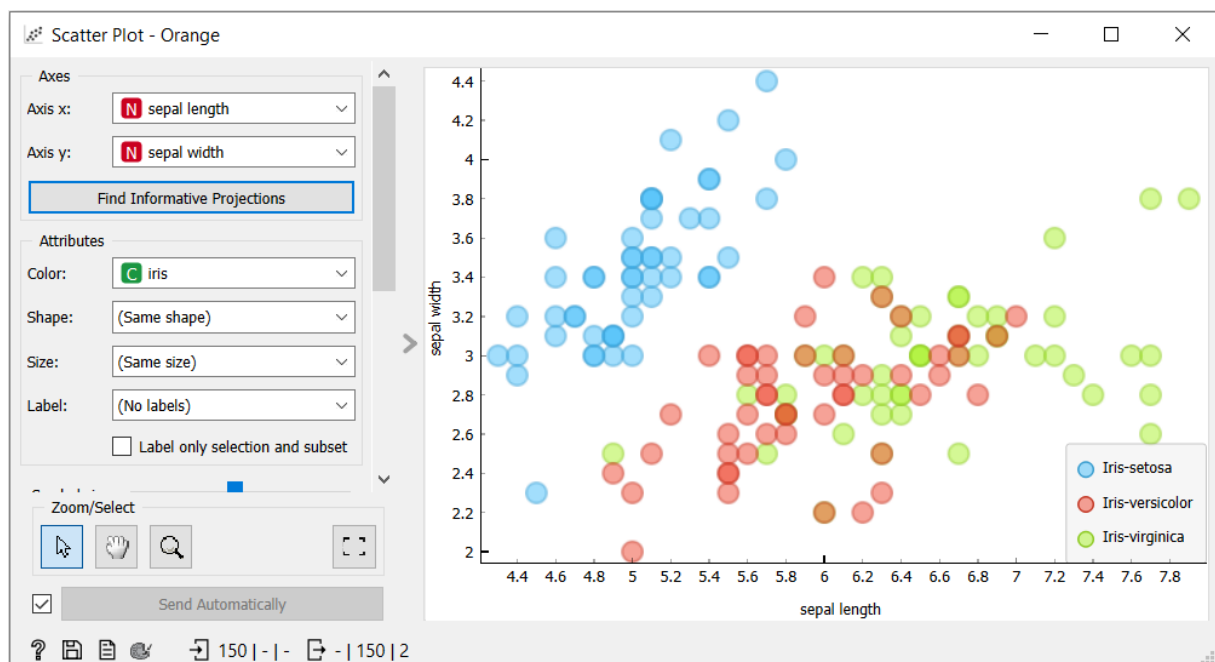
Variables
☒ Show variable labels (if present)
☐ Visualize numeric values
☒ Color by instance classes

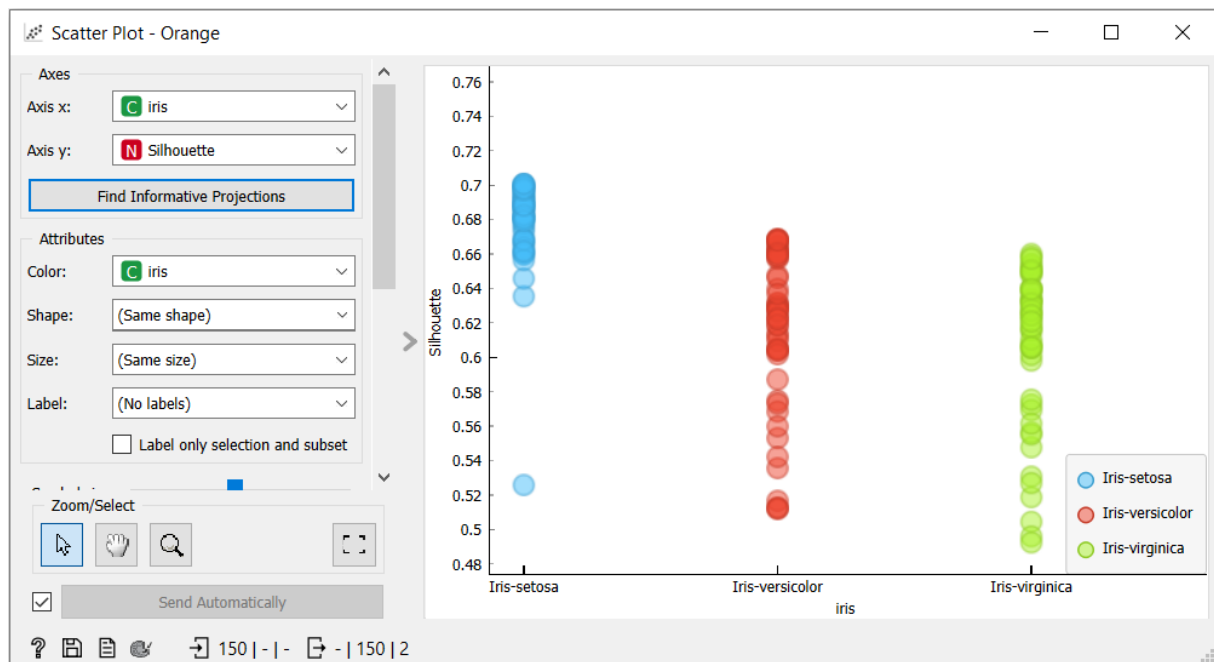
Selection
☒ Select full rows

Restore Original Order

☒ Send Automatically

	iris	Cluster	Silhouette	sepal length	sepal width	petal length	petal width
1	Iris-setosa	C2	0.700676	5.1	3.5	1.4	0.2
2	Iris-setosa	C2	0.665989	4.9	3.0	1.4	0.2
3	Iris-setosa	C2	0.690178	4.7	3.2	1.3	0.2
4	Iris-setosa	C2	0.677784	4.6	3.1	1.5	0.2
5	Iris-setosa	C2	0.699043	5.0	3.6	1.4	0.2
6	Iris-setosa	C2	0.672942	5.4	3.9	1.7	0.4
7	Iris-setosa	C2	0.693669	4.6	3.4	1.4	0.3
8	Iris-setosa	C2	0.700453	5.0	3.4	1.5	0.2
9	Iris-setosa	C2	0.645739	4.4	2.9	1.4	0.2
10	Iris-setosa	C2	0.681067	4.9	3.1	1.5	0.1
11	Iris-setosa	C2	0.687612	5.4	3.7	1.5	0.2
12	Iris-setosa	C2	0.698592	4.8	3.4	1.6	0.2
13	Iris-setosa	C2	0.668474	4.8	3.0	1.4	0.1
14	Iris-setosa	C2	0.66082	4.3	3.0	1.1	0.1
15	Iris-setosa	C2	0.659584	5.8	4.0	1.2	0.2
16	Iris-setosa	C2	0.635396	5.7	4.4	1.5	0.4
17	Iris-setosa	C2	0.675559	5.4	3.9	1.3	0.4
18	Iris-setosa	C2	0.699562	5.1	3.5	1.4	0.3
19	Iris-setosa	C2	0.667565	5.7	3.8	1.7	0.3
20	Iris-setosa	C2	0.688072	5.1	3.8	1.5	0.3
21	Iris-setosa	C2	0.68546	5.4	3.4	1.7	0.2
22	Iris-setosa	C2	0.69094	5.1	3.7	1.5	0.4
23	Iris-setosa	C2	0.692615	4.6	3.6	1.0	0.2





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>

Data Table - Orange

Info

1599 instances (no missing data)

12 features

No target variable.

2 meta attributes

Variables

☒ Show variable labels (if present)

☐ Visualize numeric values

☒ Color by instance classes

Selection

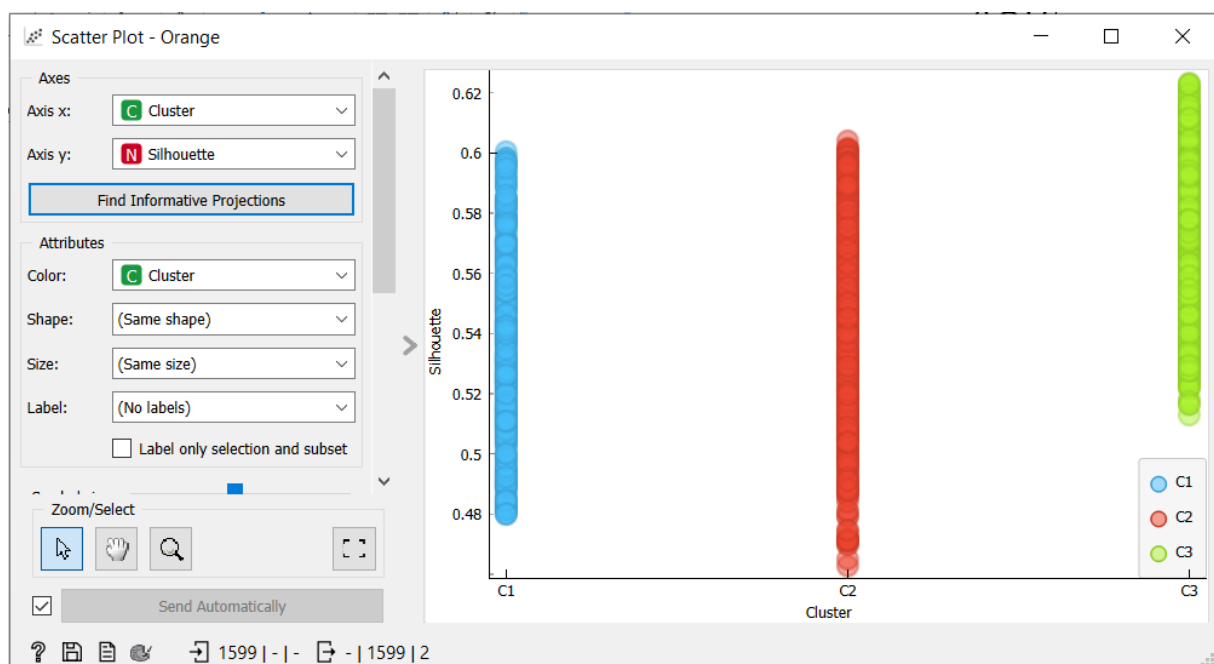
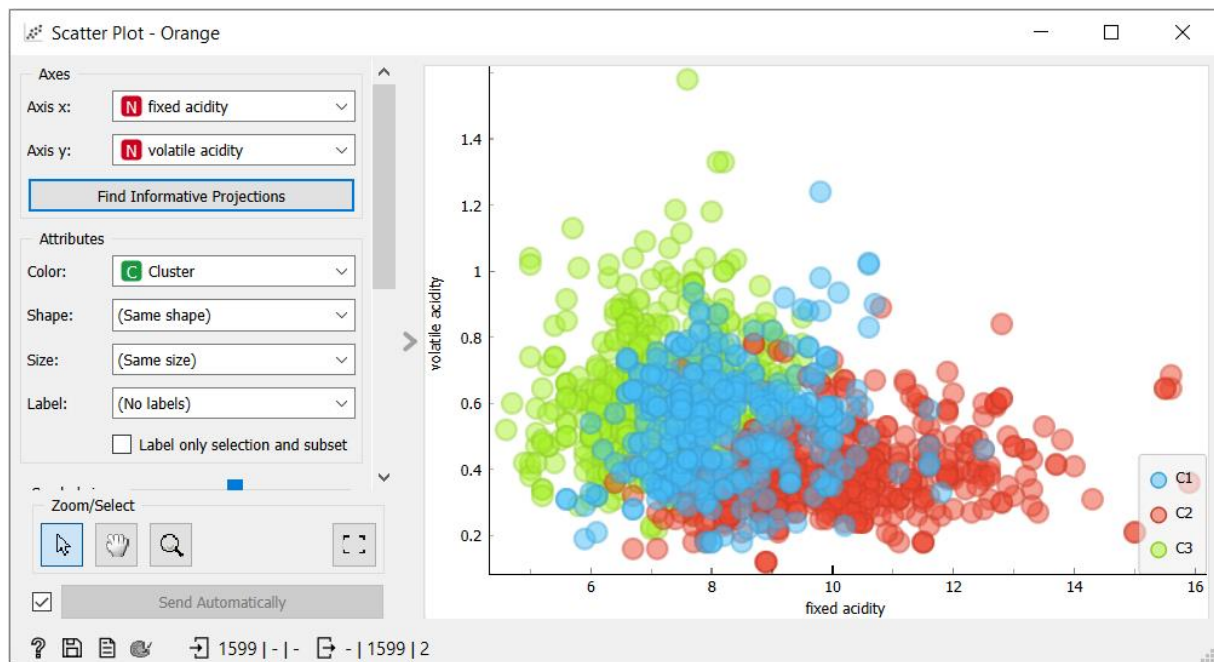
☒ Select full rows

Restore Original Order

☒ Send Automatically

1599 | 1599 | 1599

	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



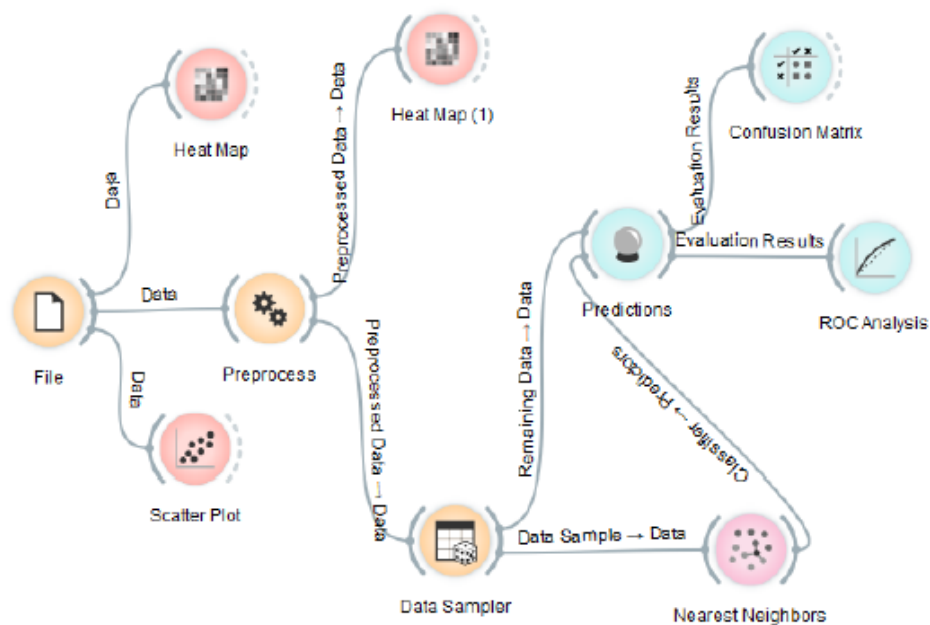
Komentar:

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.

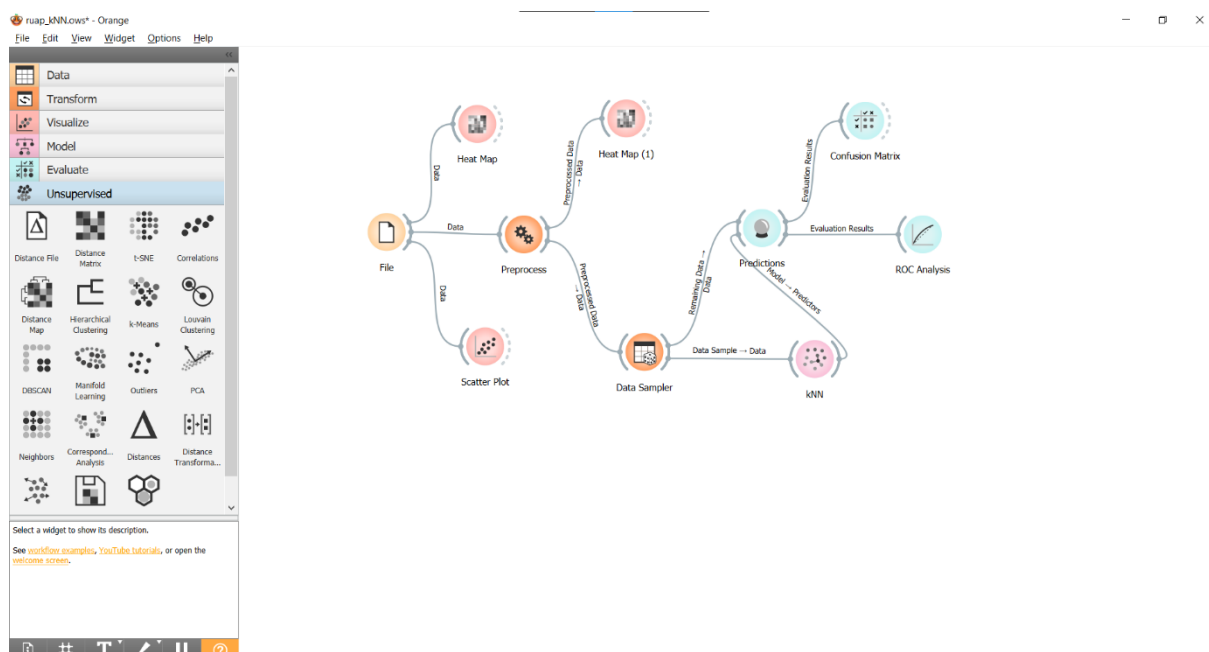
Na primjeru podatkovnog skupa provedite klasifikaciju korištenjem algoritma najbližih susjeda.

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-obrađe (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

Predictions - Orange

Show probabilities for: **Classes in data** ☒ Show classification errors [Restore Original Order](#)

	kNN	error	iris	sepal length	sepal width	petal length	petal width
1	0.00 : 0.00 : 1.00 → Iris-virginica	0.000	Iris-virginica	6.4	3.2	5.3	2.3
2	1.00 : 0.00 : 0.00 → Iris-setosa	0.000	Iris-setosa	5.1	3.8	1.6	0.2
3	0.00 : 0.00 : 1.00 → Iris-virginica	0.000	Iris-virginica	6.9	3.1	5.4	2.1
4	0.00 : 1.00 : 0.00 → Iris-versicolor	0.000	Iris-versicolor	5.9	3.0	4.2	1.5
5	0.00 : 0.20 : 0.80 → Iris-virginica	0.200	Iris-virginica	6.5	3.0	5.2	2.0
6	0.00 : 1.00 : 0.00 → Iris-versicolor	0.000	Iris-versicolor	5.7	2.6	3.5	1.0
7	0.00 : 1.00 : 0.00 → Iris-versicolor	0.000	Iris-versicolor	5.2	2.7	3.9	1.4
8	0.00 : 1.00 : 0.00 → Iris-versicolor	0.000	Iris-versicolor	6.1	3.0	4.6	1.4
9	1.00 : 0.00 : 0.00 → Iris-setosa	0.000	Iris-setosa	4.5	2.3	1.3	0.3
10	0.00 : 1.00 : 0.00 → Iris-versicolor	0.000	Iris-versicolor	6.6	2.9	4.6	1.3

☒ Show performance scores Target class: (Average over classes)

Model	AUC	CA	F1	Precision	Recall
kNN	0.980	0.973	0.973	0.975	0.973

37 | 1×37

Confusion Matrix - Orange

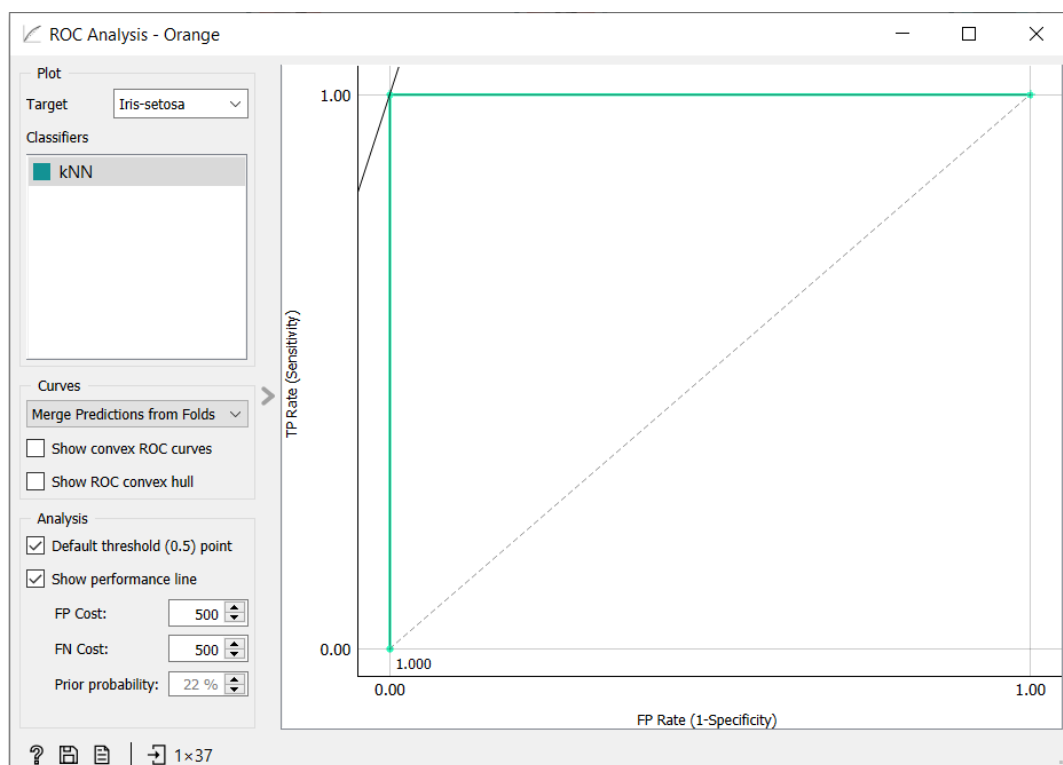
Clicking on cells or in headers outputs the corresponding data instances [Ok, got it](#) Show: **Number of instances**

		Predicted			
		Iris-setosa	Iris-versicolor	Iris-virginica	Σ
Actual	Iris-setosa	8	0	0	8
	Iris-versicolor	0	14	0	14
	Iris-virginica	0	1	14	15
Σ		8	15	14	37

☒ Predictions ☐ Probabilities ☒ Apply Automatically

Select Correct Select Misclassified Clear Selection

1×37 37



Rezultati nakon omogućene normalizacije:

Predictions - Orange

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

	kNN	error	iris	sepal length	sepal width	petal length	petal width
1	0.00 : 0.00 : 1.00 → Iris-virginica	0.000	Iris-virginica	0.1667	0.00	0.4576	0.8333
2	1.00 : 0.00 : 0.00 → Iris-setosa	0.000	Iris-setosa	0.5556	0.50	-0.7966	-0.9167
3	0.00 : 0.00 : 1.00 → Iris-virginica	0.000	Iris-virginica	0.4444	-0.0833	0.4915	0.6667
4	0.00 : 1.00 : 0.00 → Iris-versico...	0.000	Iris-versicolor	-0.1111	-0.1667	0.0847	0.1667
5	0.00 : 0.00 : 1.00 → Iris-virginica	0.000	Iris-virginica	0.2222	-0.1667	0.4237	0.5833
6	0.00 : 1.00 : 0.00 → Iris-versico...	0.000	Iris-versicolor	-0.2222	-0.50	-0.1525	-0.25
7	0.00 : 1.00 : 0.00 → Iris-versico...	0.000	Iris-versicolor	-0.50	-0.4167	-0.0169	0.0833
8	0.00 : 1.00 : 0.00 → Iris-versico...	0.000	Iris-versicolor	0.00	-0.1667	0.2203	0.0833
9	1.00 : 0.00 : 0.00 → Iris-setosa	0.000	Iris-setosa	0.8889	0.75	-0.8983	-0.8333
10	0.00 : 1.00 : 0.00 → Iris-versico...	0.000	Iris-versicolor	0.2778	-0.25	0.2203	0.00

☒ Show performance scores Target class: (Average over classes)

Model	AUC	CA	F1	Precision	Recall
kNN	0.998	0.946	0.946	0.953	0.946

37 | 1×37

Confusion Matrix - Orange

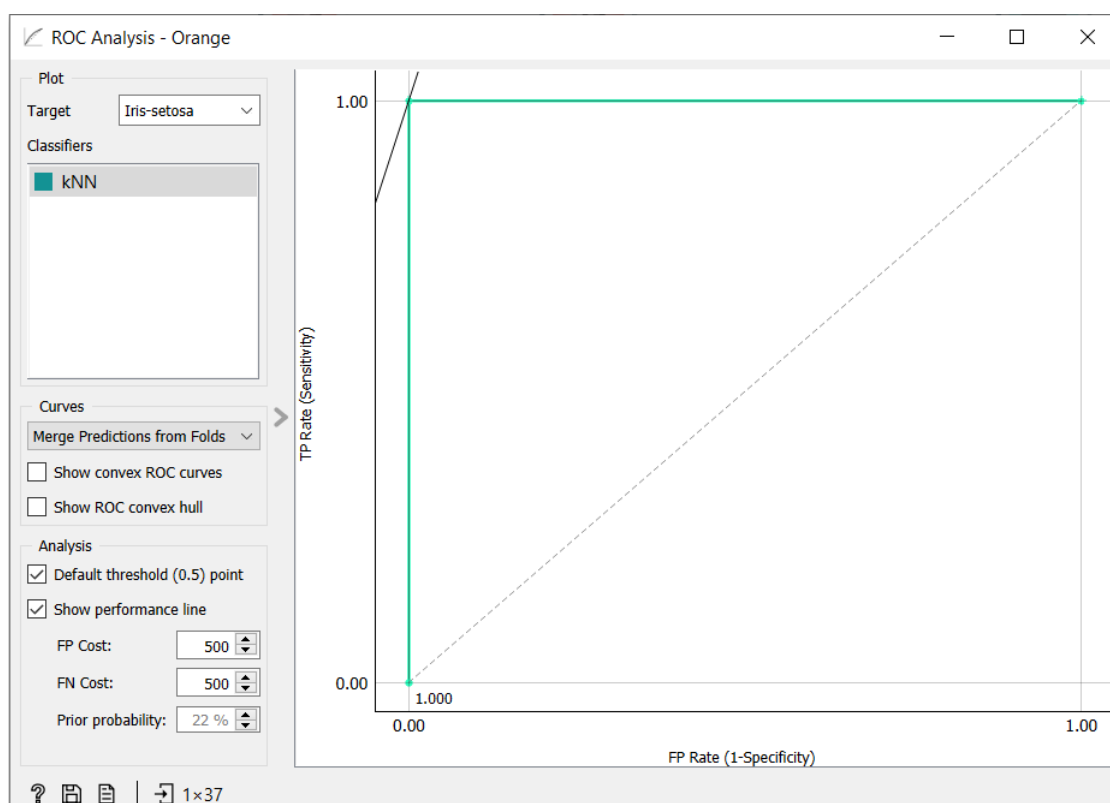
Learners: **kNN**

Clicking on cells or in headers outputs the corresponding data instances **Ok, got it** **Show: Number of instances**

		Predicted			
		Iris-setosa	Iris-versicolor	Iris-virginica	Σ
Actual	Iris-setosa	8	0	0	8
	Iris-versicolor	0	14	0	14
	Iris-virginica	0	2	13	15
Σ		8	16	13	37

Select Correct **Select Misclassified** **Clear Selection**

1×37 | 37



Rezultati nakon omogućene normalizacije i postavljanja parametra $k = 10$:

Predictions - Orange

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

	kNN	error	iris	sepal length	sepal width	petal length	petal width
1	0.00 : 0.00 : 1.00 → Iris-virginica	0.000	Iris-virginica	0.1667	0.00	0.4576	0.8333
2	1.00 : 0.00 : 0.00 → Iris-setosa	0.000	Iris-setosa	-0.5556	0.50	-0.7966	-0.9167
3	0.00 : 0.00 : 1.00 → Iris-virginica	0.000	Iris-virginica	0.4444	0.0833	0.4915	0.6667
4	0.00 : 0.90 : 0.10 → Iris-versico...	0.100	Iris-versicolor	0.1111	0.1667	0.0847	0.1667
5	0.00 : 0.10 : 0.90 → Iris-virginica	0.100	Iris-virginica	0.2222	-0.1667	0.4237	0.5833
6	0.00 : 1.00 : 0.00 → Iris-versico...	0.000	Iris-versicolor	-0.2222	-0.50	-0.1525	-0.25
7	0.00 : 1.00 : 0.00 → Iris-versico...	0.000	Iris-versicolor	-0.50	-0.4167	0.0169	0.0833
8	0.00 : 0.90 : 0.10 → Iris-versico...	0.100	Iris-versicolor	0.00	-0.1667	0.2203	0.0833
9	1.00 : 0.00 : 0.00 → Iris-setosa	0.000	Iris-setosa	-0.8889	-0.75	-0.8983	-0.8333
10	0.00 : 0.90 : 0.10 → Iris-versico...	0.100	Iris-versicolor	0.2778	0.25	0.2203	0.00

☒ Show performance scores Target class: (Average over classes)

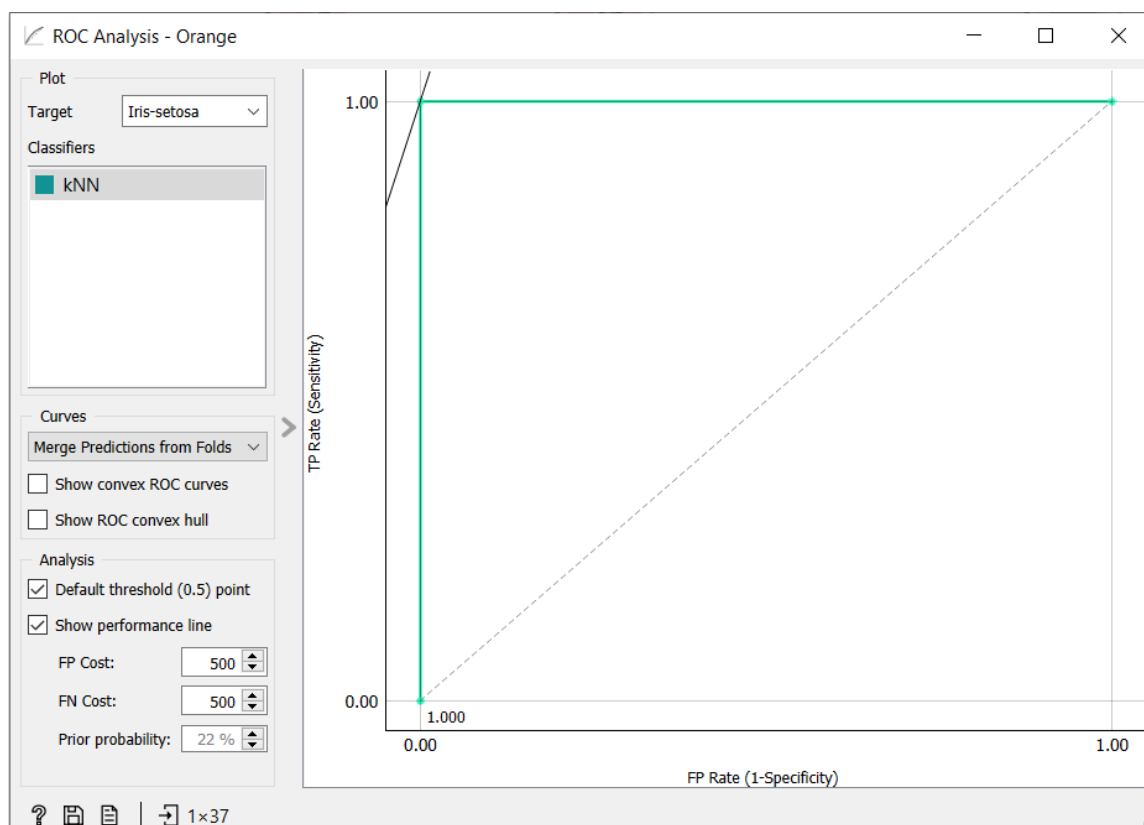
Model	AUC	CA	F1	Precision	Recall
kNN	1.000	0.973	0.973	0.975	0.973

Confusion Matrix - Orange

Clicking on cells or in headers outputs the corresponding data instances **Ok, got it** Show: Number of instances

		Predicted			
		Iris-setosa	Iris-versicolor	Iris-virginica	Σ
Actual	Iris-setosa	8	0	0	8
	Iris-versicolor	0	14	0	14
	Iris-virginica	0	1	14	15
Σ		8	15	14	37

Select Correct Select Misclassified Clear Selection



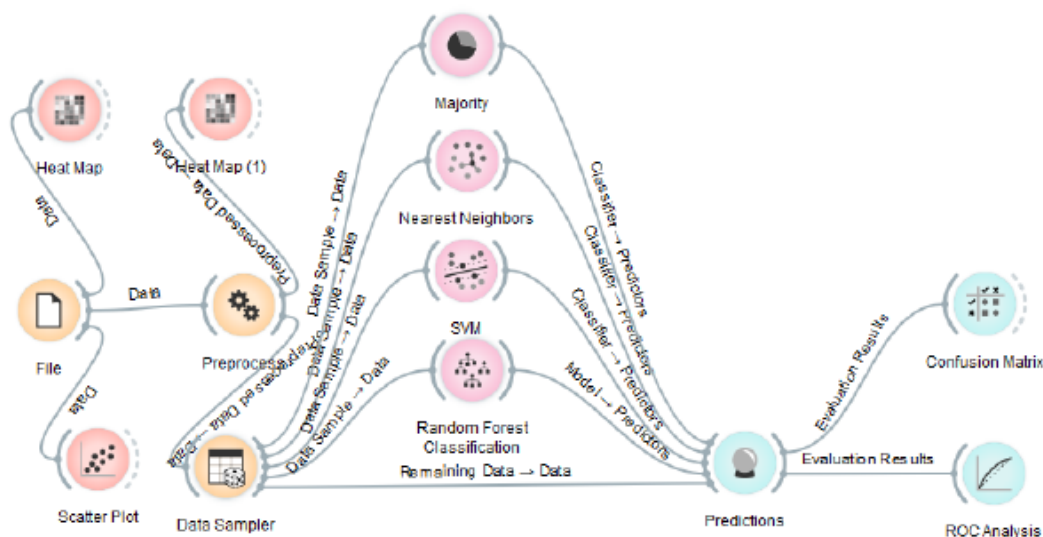
Komentar:

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.

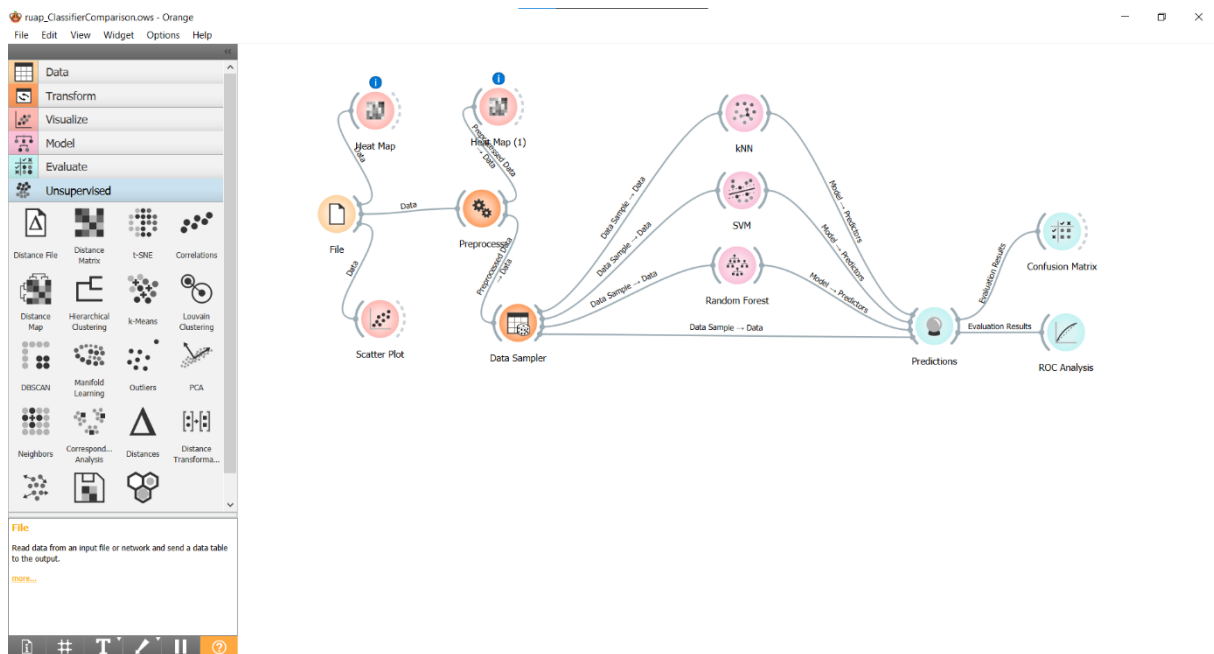
Usporedba nekoliko klasifikatora.

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

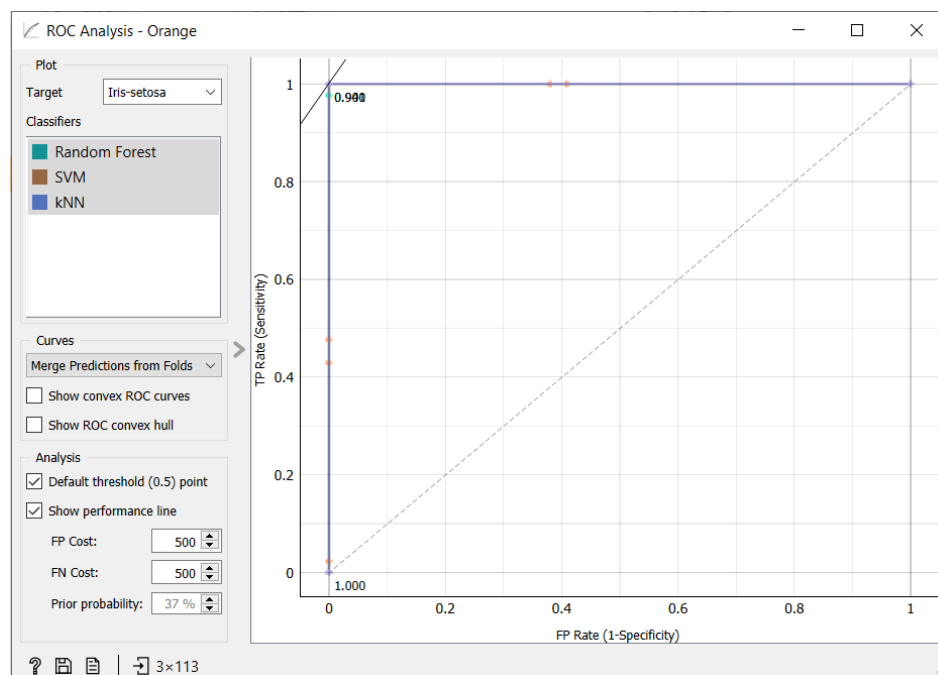
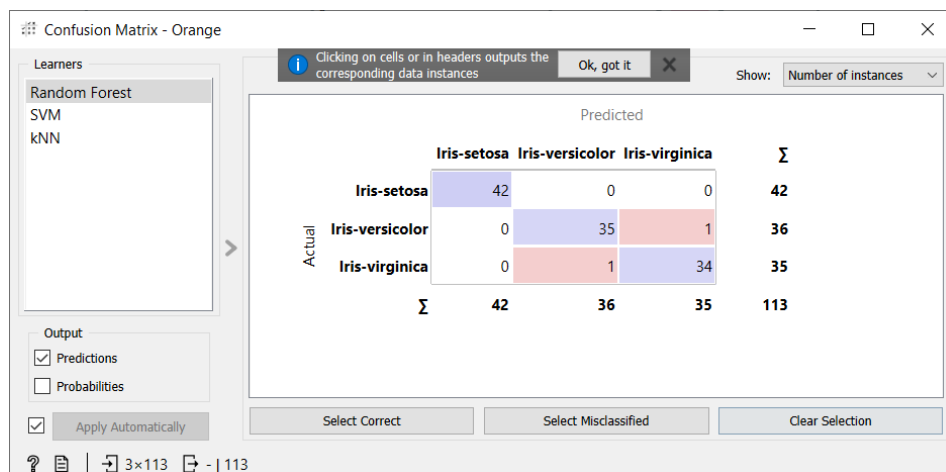
Show probabilities for: Classes in data ☒ Show classification errors Restore Original Order

	Random Forest	error	SVM	error	kNN	
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-versico...
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-setosa
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-virginica
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-versico...
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-versico...
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-setosa
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-versico...
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-virginica
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-versico...
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-versico...

☒ Show performance 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 | 3x113



Rezultati za Heart disease dataset

Predictions - Orange

Show probabilities for: Classes in data

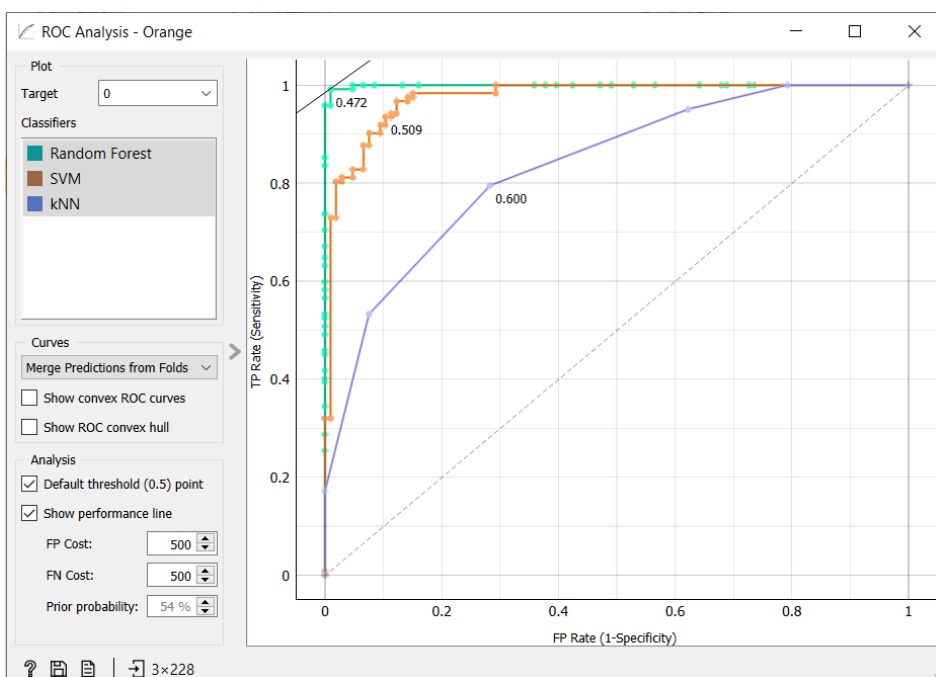
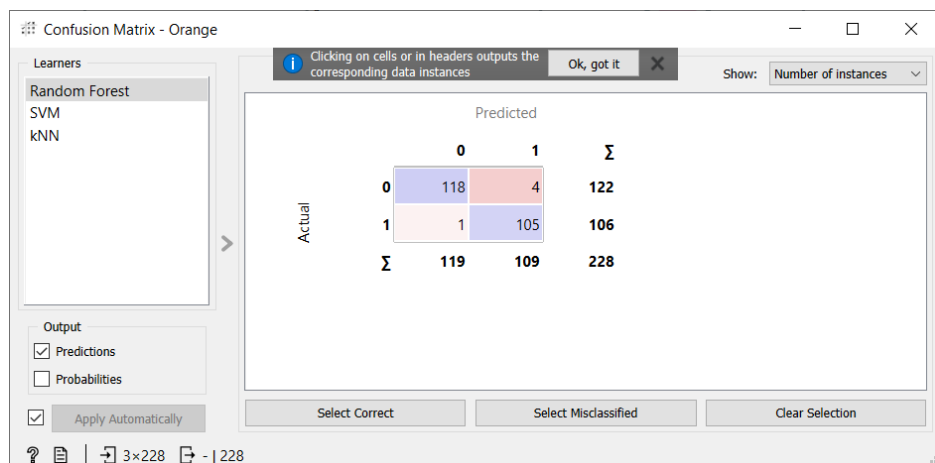
Show classification errors

Restore Original Order

	Random Forest	error	SVM	error	kNN	error	liameter narrowin	age
1	0.89 : 0.11 → 0	0.113	0.85 : 0.15 ...	0.147	1.00 : 0.00 ...	0.000	0	53
2	0.13 : 0.87 → 1	0.128	0.13 : 0.87 ...	0.131	0.60 : 0.40 ...	0.600	1	54
3	0.16 : 0.84 → 1	0.160	0.11 : 0.89 ...	0.109	0.00 : 1.00 ...	0.000	1	56
4	0.21 : 0.79 → 1	0.210	0.29 : 0.71 ...	0.295	0.60 : 0.40 ...	0.600	1	58
5	0.00 : 1.00 → 1	0.000	0.11 : 0.89 ...	0.109	0.40 : 0.60 ...	0.400	1	51
6	0.08 : 0.92 → 1	0.083	0.11 : 0.89 ...	0.111	0.40 : 0.60 ...	0.400	1	53
7	0.02 : 0.98 → 1	0.020	0.04 : 0.96 ...	0.045	0.00 : 1.00 ...	0.000	1	65
8	0.00 : 1.00 → 1	0.000	0.02 : 0.98 ...	0.016	0.00 : 1.00 ...	0.000	1	53
9	0.29 : 0.71 → 1	0.288	0.33 : 0.67 ...	0.332	0.80 : 0.20 ...	0.800	1	40
10	0.62 : 0.38 → 0	0.381	0.34 : 0.66 ...	0.659	0.60 : 0.40 ...	0.400	0	59

Show performance scores Target class: (Average over classes)

Model	AUC	CA	F1	Precision	Recall
Random Forest	0.999	0.978	0.978	0.978	0.978
SVM	0.974	0.917	0.917	0.917	0.917
kNN	0.834	0.759	0.758	0.759	0.759



Rezultati za Brown selected dataset

Predictions - Orange

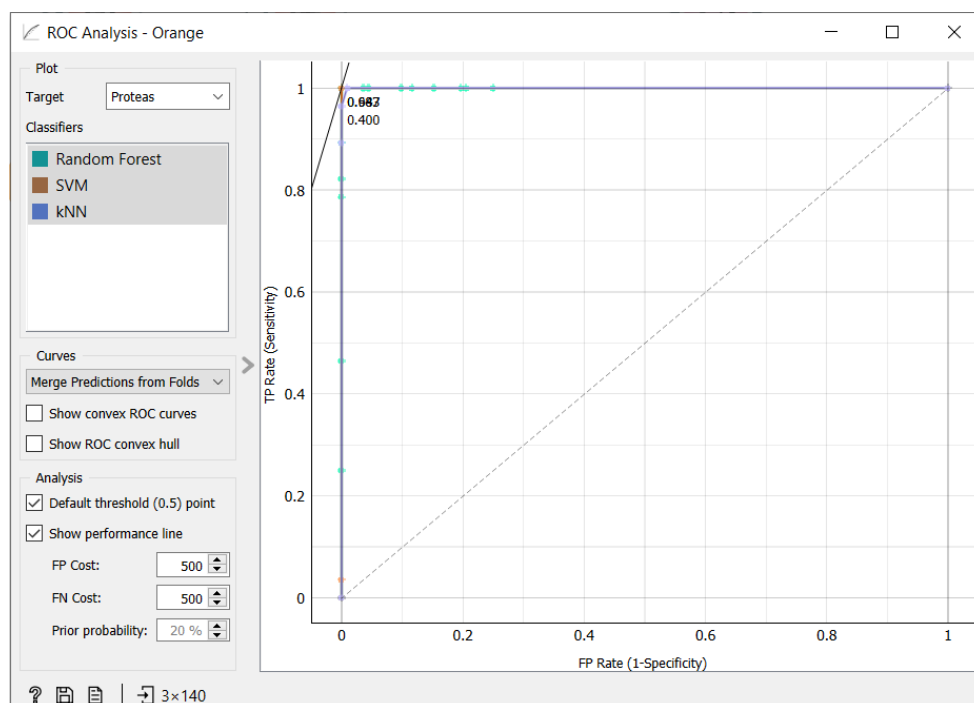
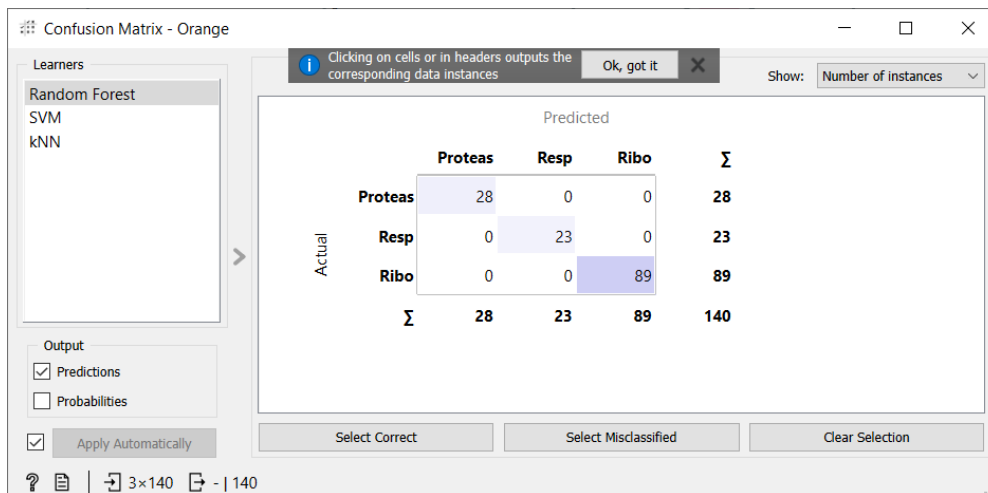
Show probabilities for: Classes in data ☒ Show classification errors [Restore Original Order](#)

	Random Forest	error	SVM	error	kNN	func
1	0.00 : 0.00 : 1.00 → Ribo	0.000	0.00 : 0.00 : 1.00 → Ribo	0.005	0.00 : 0.00 : 1.00 → Ribo	Ribo
2	0.02 : 0.00 : 0.98 → Ribo	0.020	0.01 : 0.01 : 0.97 → Ribo	0.026	0.00 : 0.00 : 1.00 → Ribo	Ribo
3	0.00 : 0.10 : 0.90 → Ribo	0.100	0.01 : 0.01 : 0.97 → Ribo	0.025	0.00 : 0.00 : 1.00 → Ribo	Ribo
4	0.00 : 0.00 : 1.00 → Ribo	0.000	0.00 : 0.00 : 1.00 → Ribo	0.004	0.00 : 0.00 : 1.00 → Ribo	Ribo
5	0.00 : 0.00 : 1.00 → Ribo	0.000	0.00 : 0.00 : 1.00 → Ribo	0.000	0.00 : 0.00 : 1.00 → Ribo	Ribo
6	0.95 : 0.00 : 0.05 → Prot...	0.050	0.99 : 0.00 : 0.01 → Prot...	0.011	1.00 : 0.00 : 0.00 → Prot...	Proteas
7	0.00 : 0.85 : 0.15 → Resp	0.150	0.00 : 0.99 : 0.01 → Resp	0.008	0.00 : 1.00 : 0.00 → Resp	Resp
8	0.04 : 0.96 : 0.00 → Resp	0.043	0.00 : 0.99 : 0.01 → Resp	0.012	0.00 : 1.00 : 0.00 → Resp	Resp
9	0.00 : 0.00 : 1.00 → Ribo	0.000	0.00 : 0.01 : 0.99 → Ribo	0.010	0.00 : 0.00 : 1.00 → Ribo	Ribo
10	0.00 : 0.00 : 1.00 → Ribo	0.000	0.00 : 0.00 : 0.99 → Ribo	0.007	0.00 : 0.00 : 1.00 → Ribo	Ribo

☒ Show performance scores Target class: (Average over classes)

Model	AUC	CA	F1	Precision	Recall
Random Forest	1.000	1.000	1.000	1.000	1.000
SVM	1.000	1.000	1.000	1.000	1.000
kNN	1.000	0.986	0.986	0.986	0.986

140 | 3x140



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.