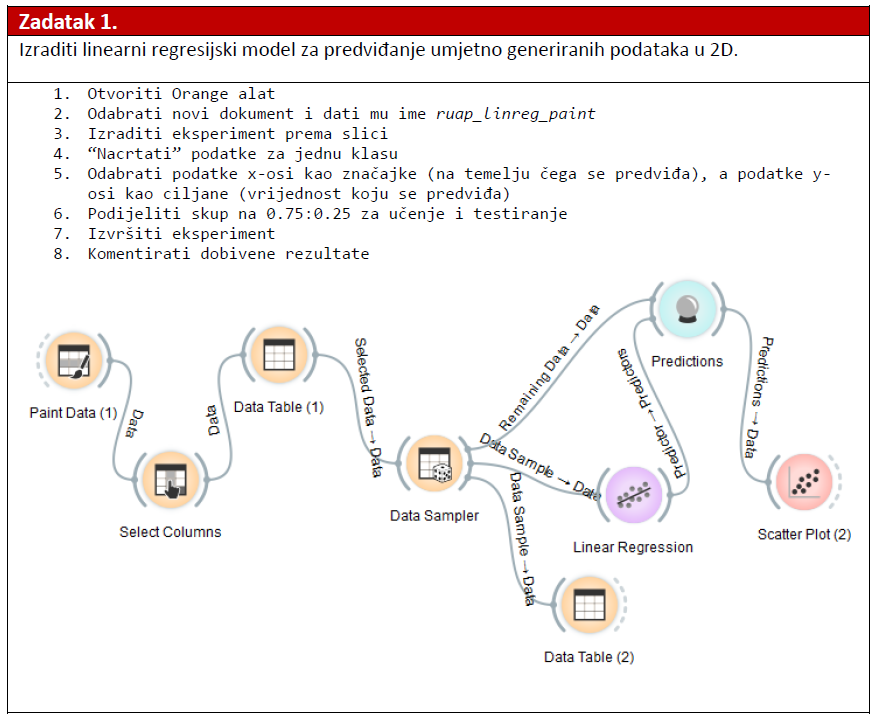
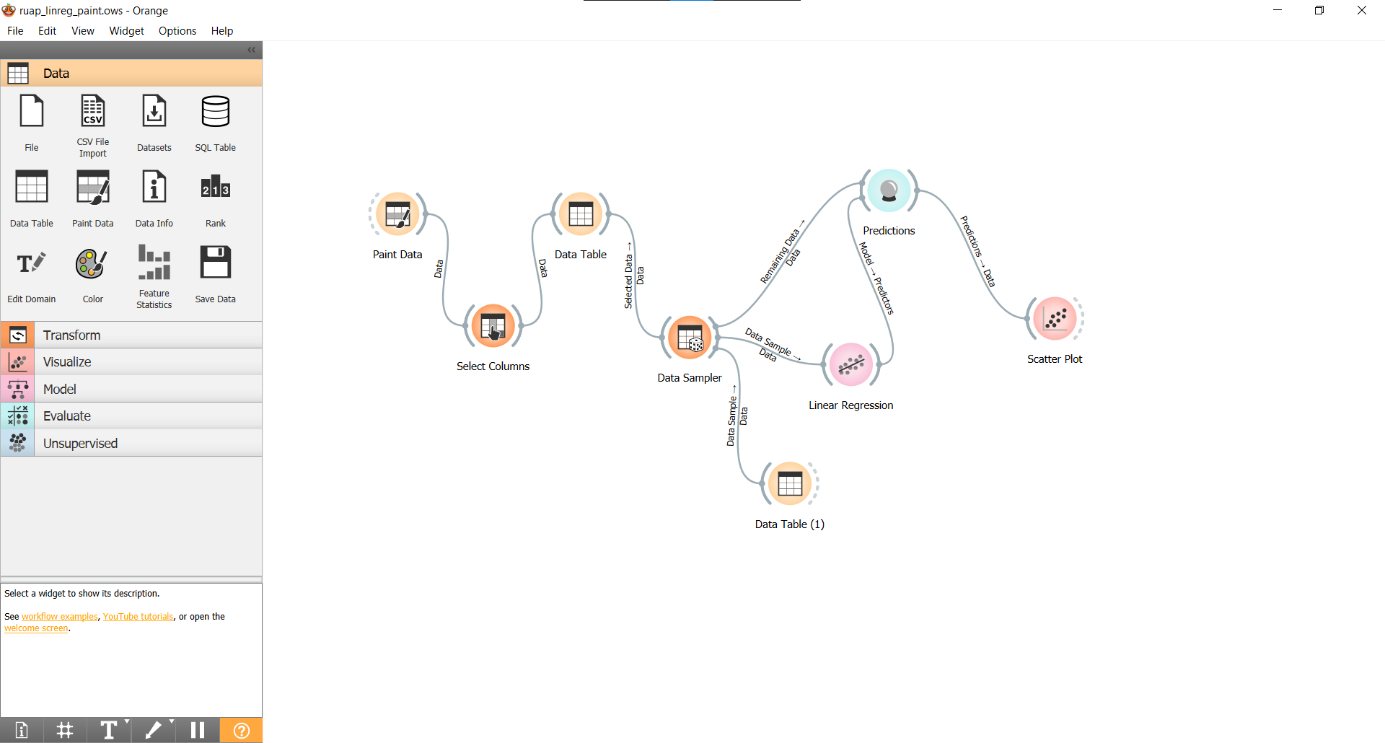
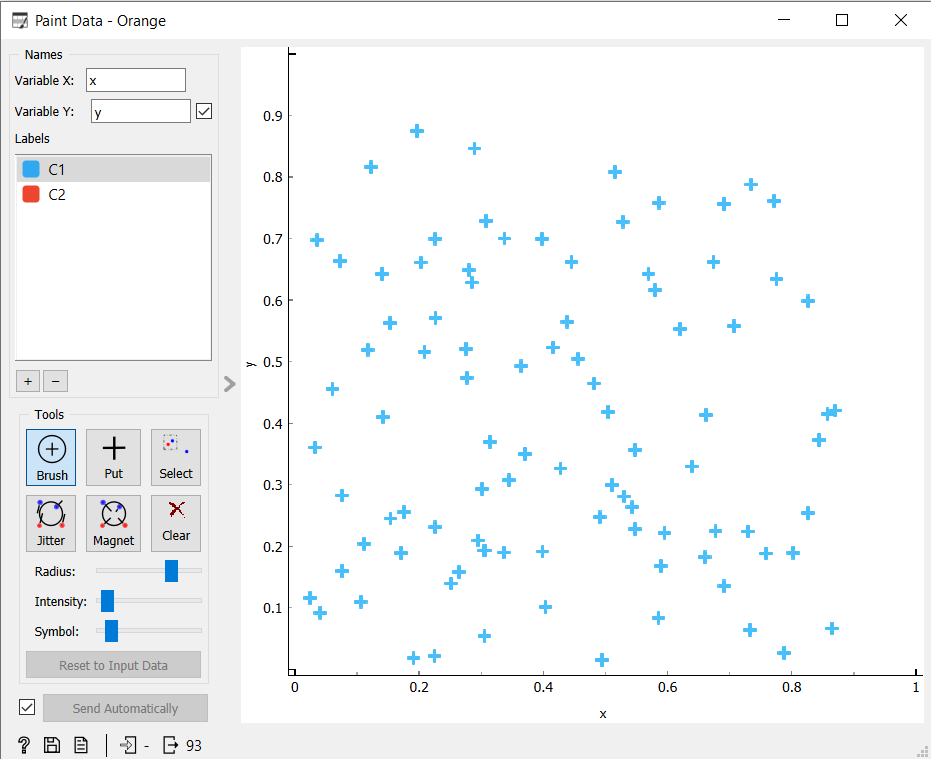
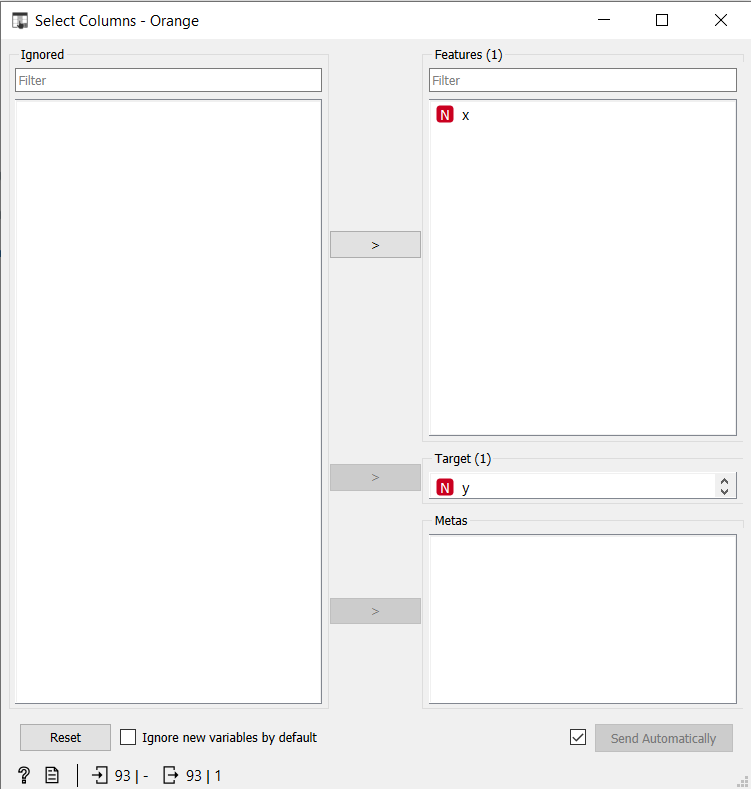
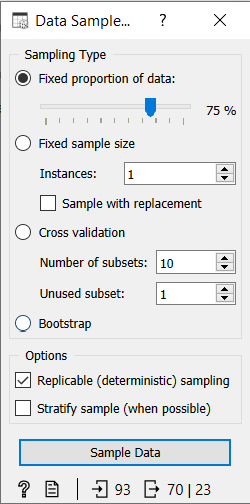
**Računarstvo usluga i analiza podataka**

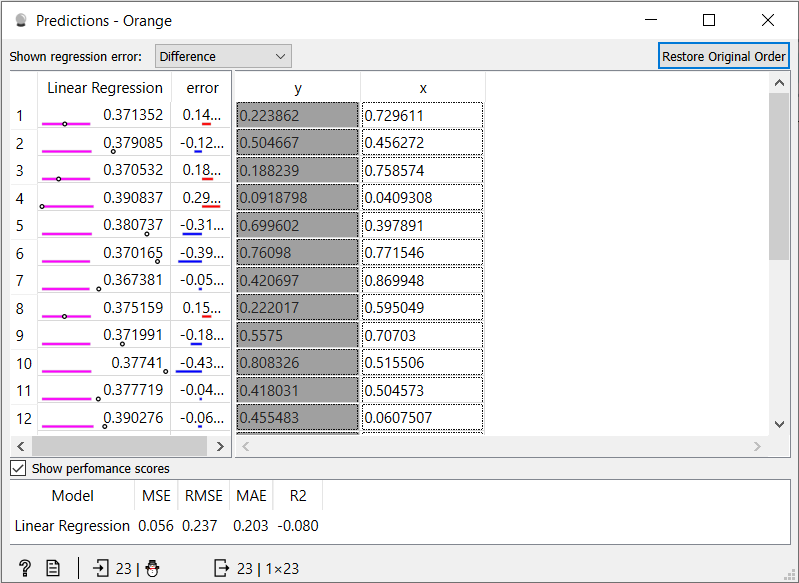
Izvještaj pete laboratorijske vježbe

Marko Ćosić

**Rješenje:**

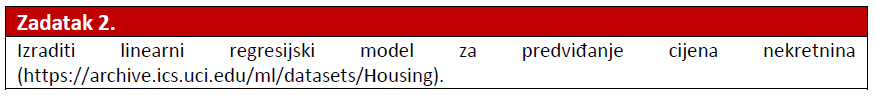
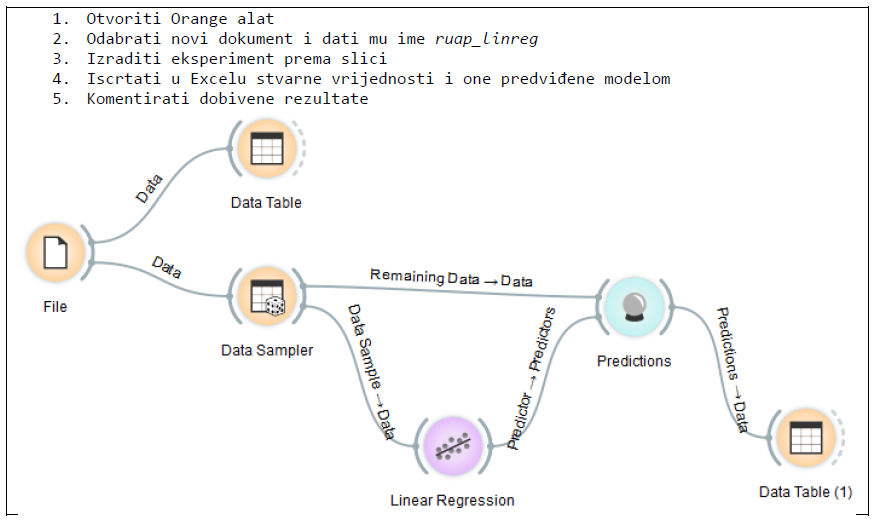


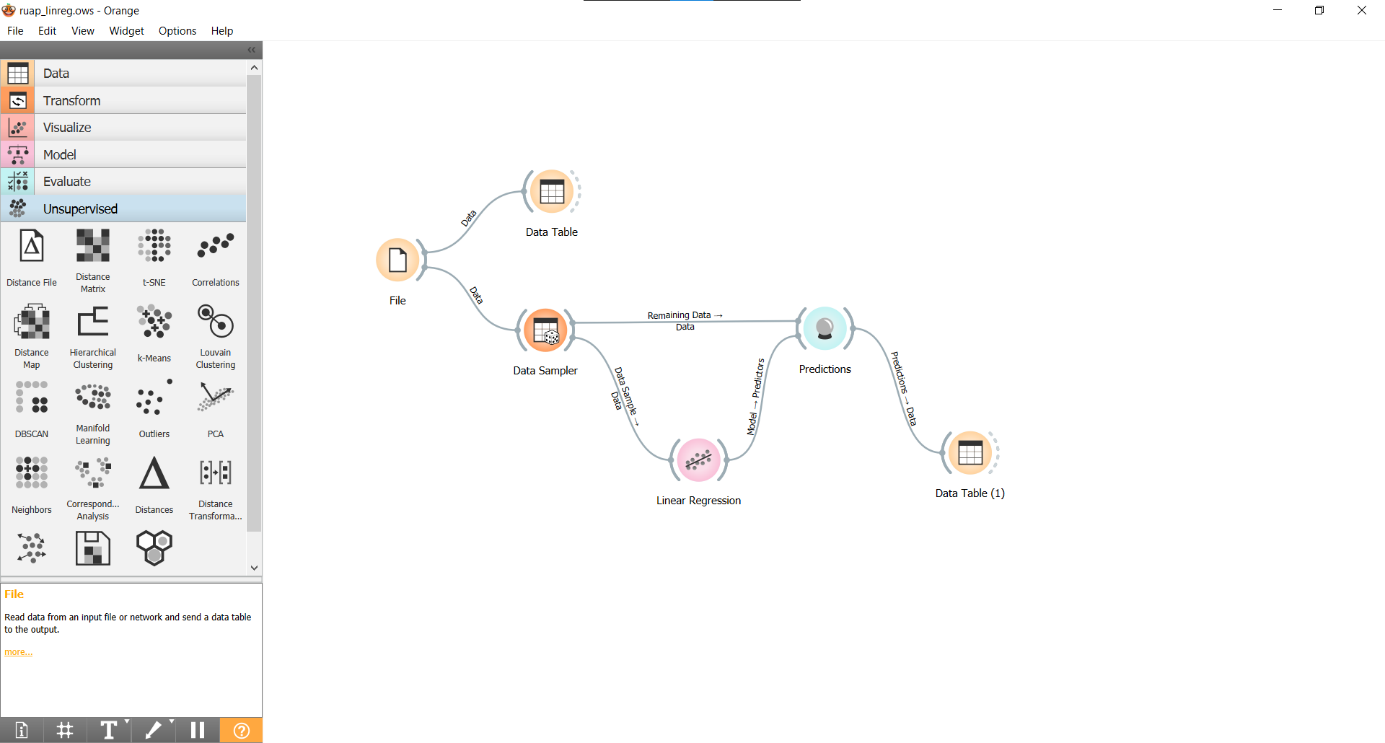


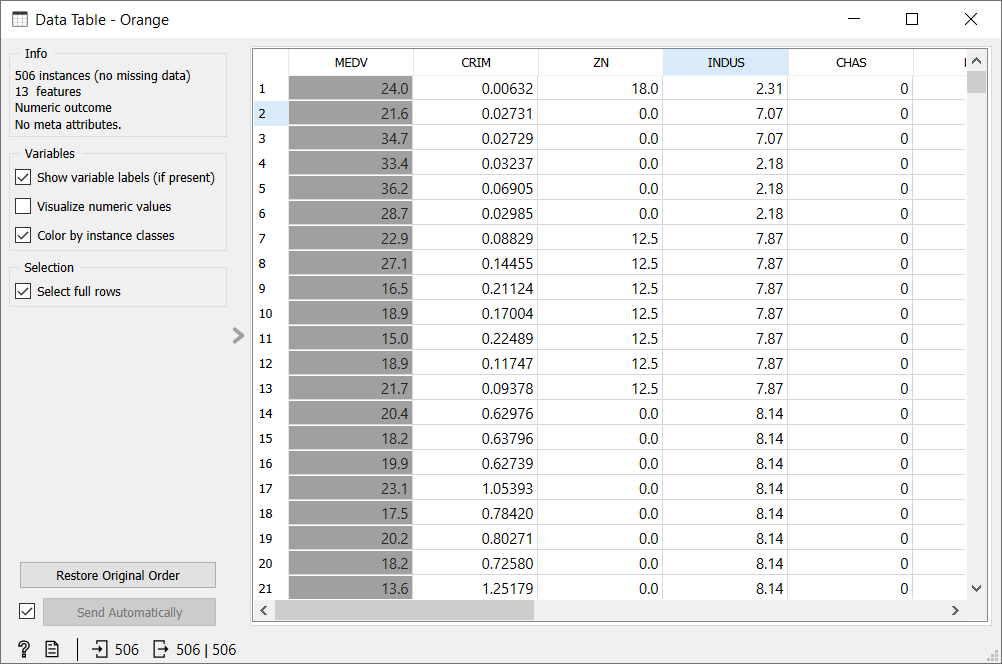


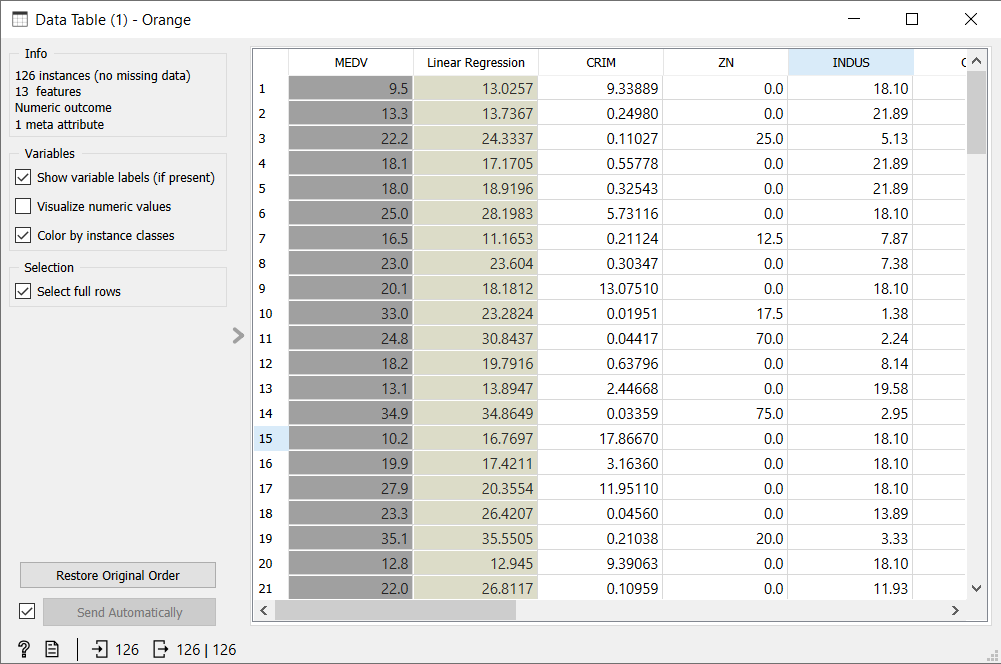
**Komentar:**

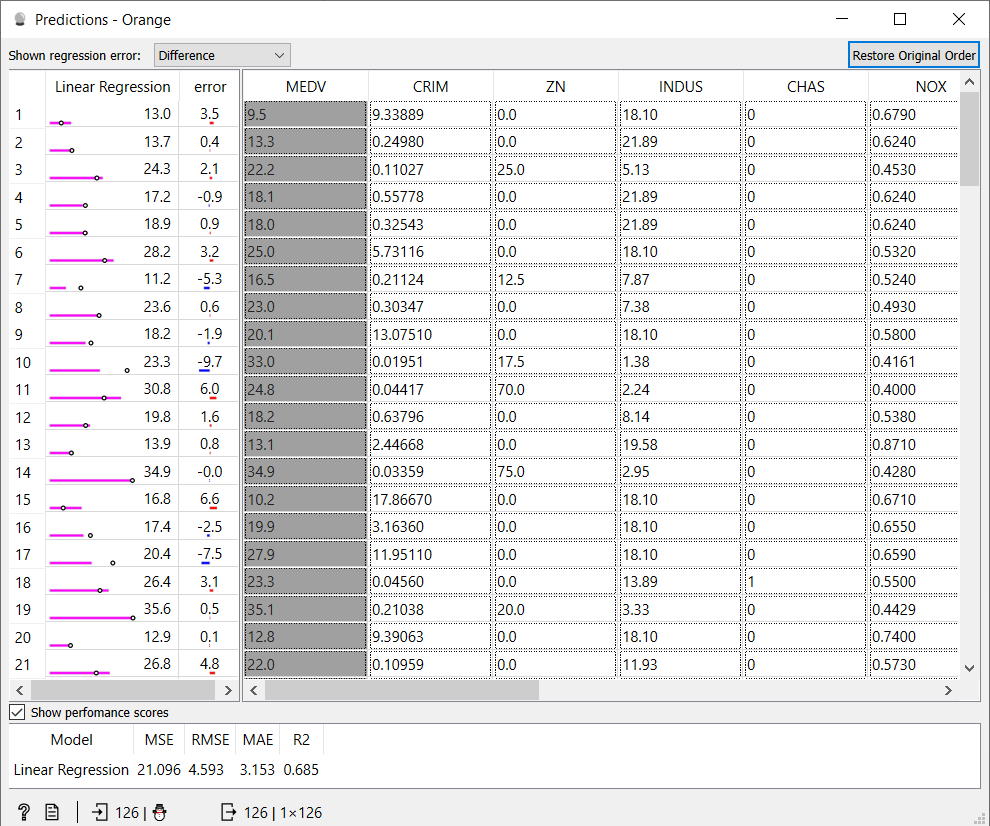
U ovome zadatku nacrtan je dataset, odnosno podatci za jednu klasu. Podatci koji se nalaze na x-osi odabrani su kao značajke (na temelju čega se predviđa), a podatci na y-osi kao ciljane značajke (vrijednosti koje se predviđaju). Može se vidjeti da je nakon podjele skupa 75% : 25% te nakon testiranja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



**Rješenje:**



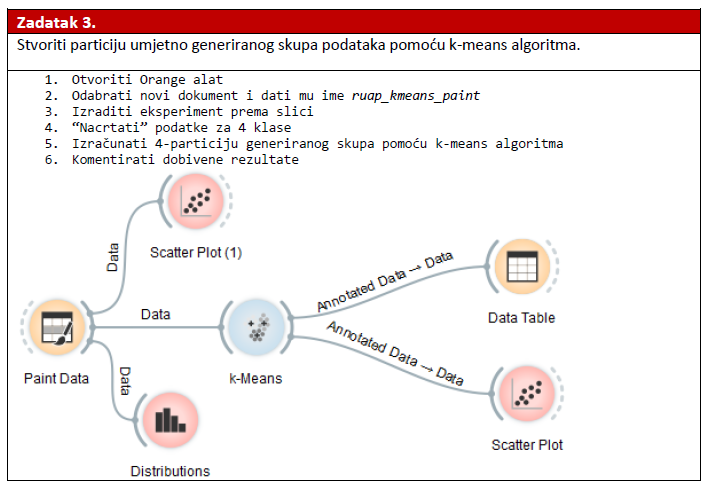




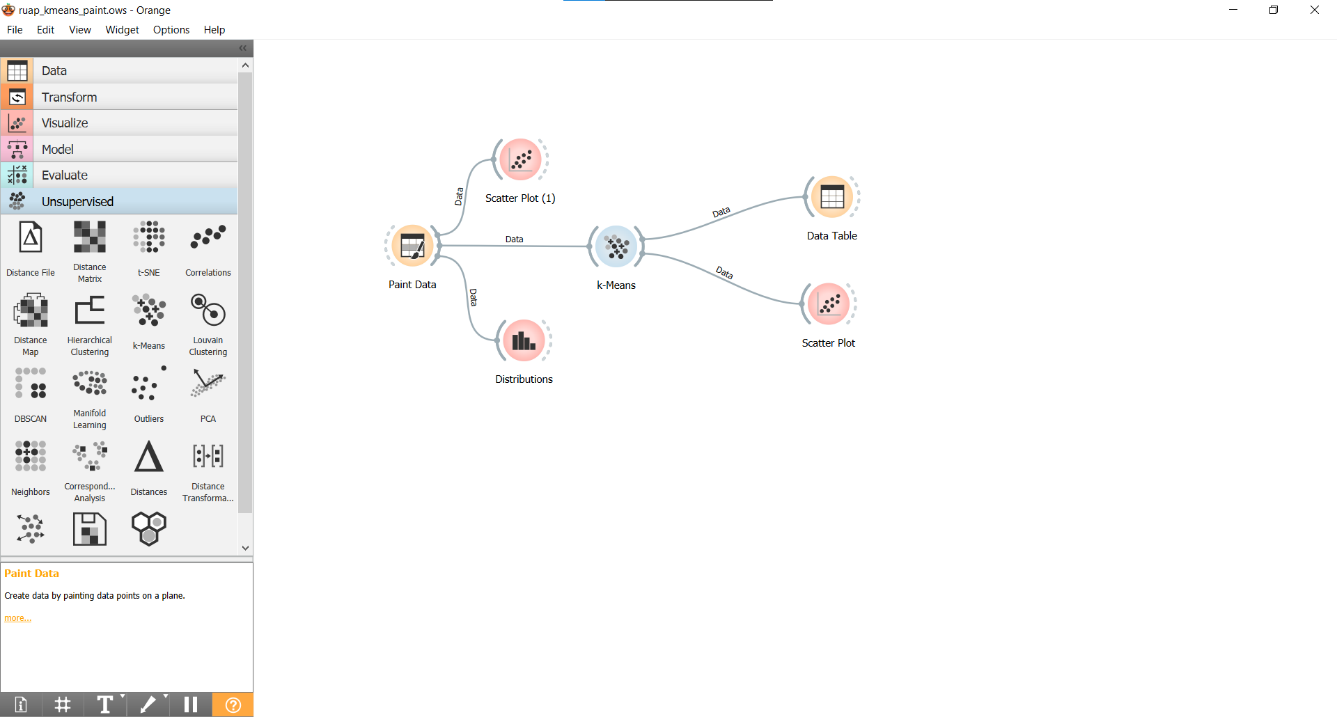
,

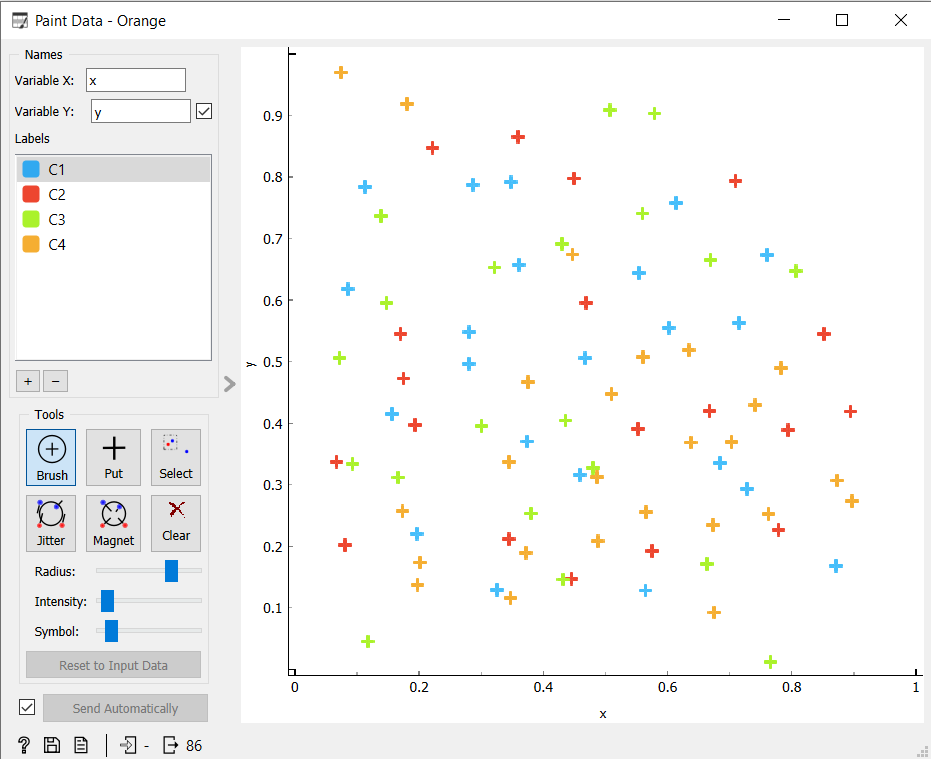
**Komentar:**

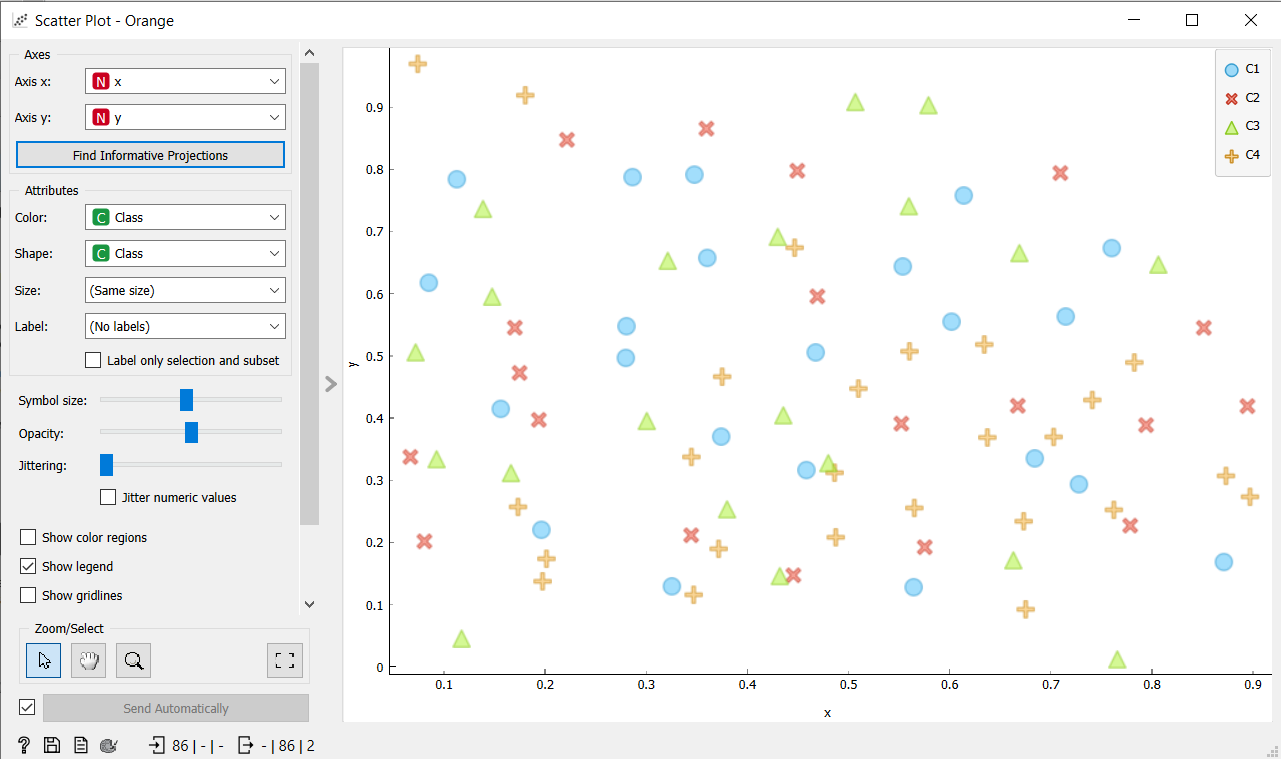
U odnosnu na prethodni zadatak, korišten je Housingdataset koji se učitava prije testiranja. Nakon testiranja može se vidjeti da stvarne vrijednosti iscrtane u Excelu i one predviđene modelom imaju manje razlike (usporedba Data Table i Data Table(1)). Nakon testiranja dobiveni su slijedeći rezultati: 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.

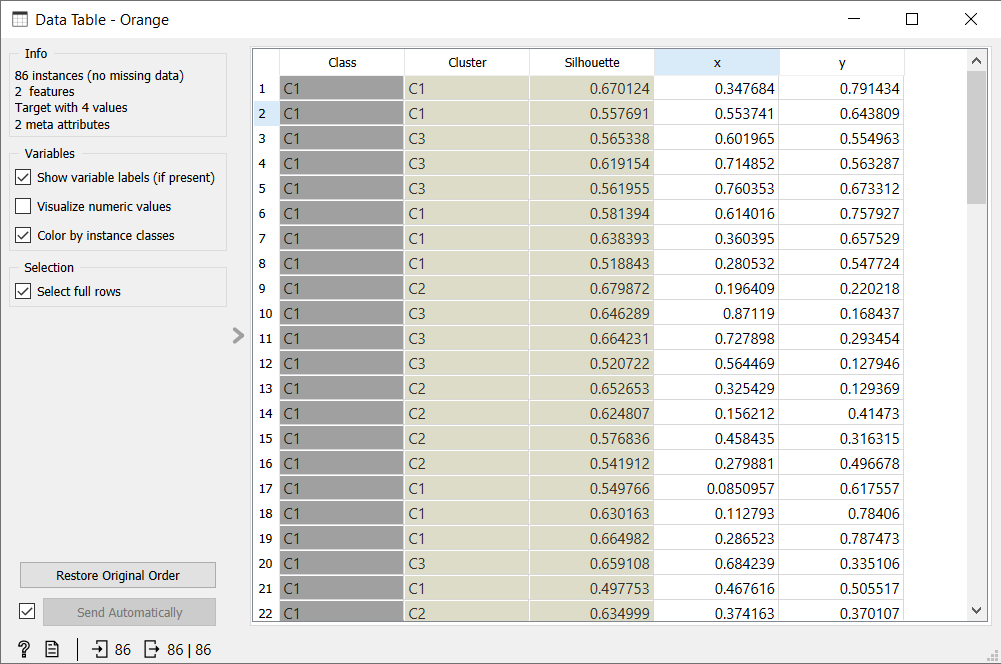


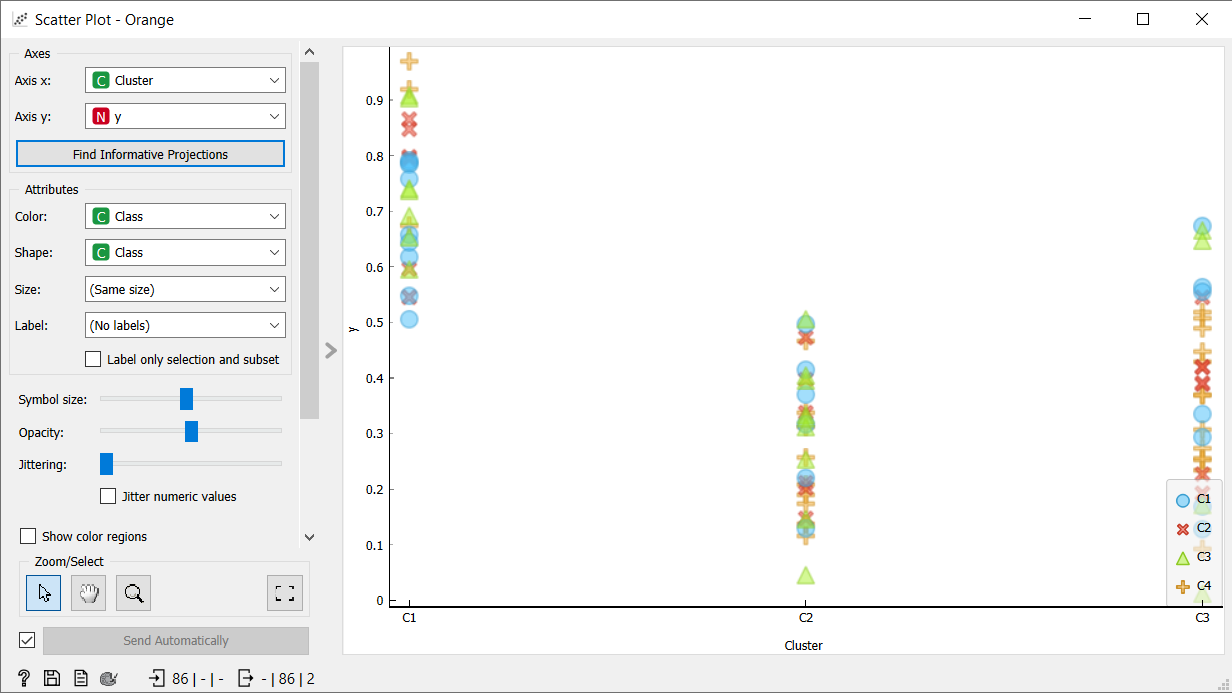
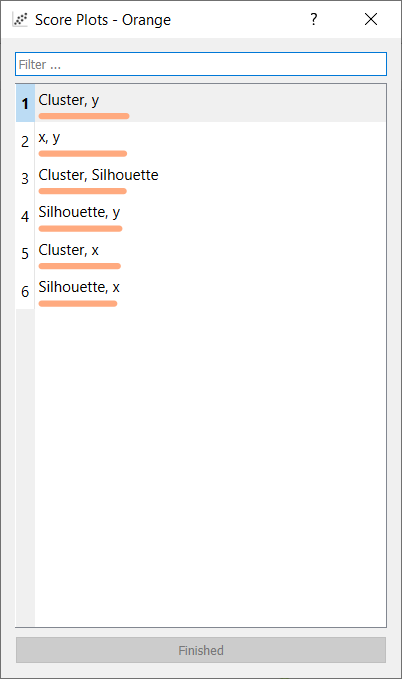
**Rješenje:**





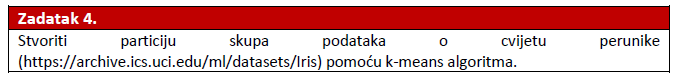
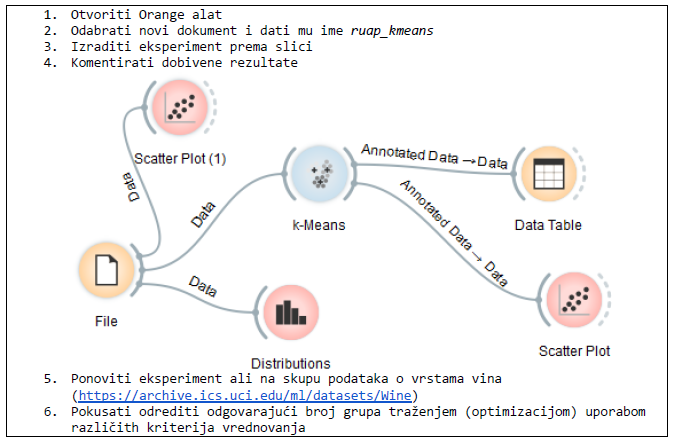


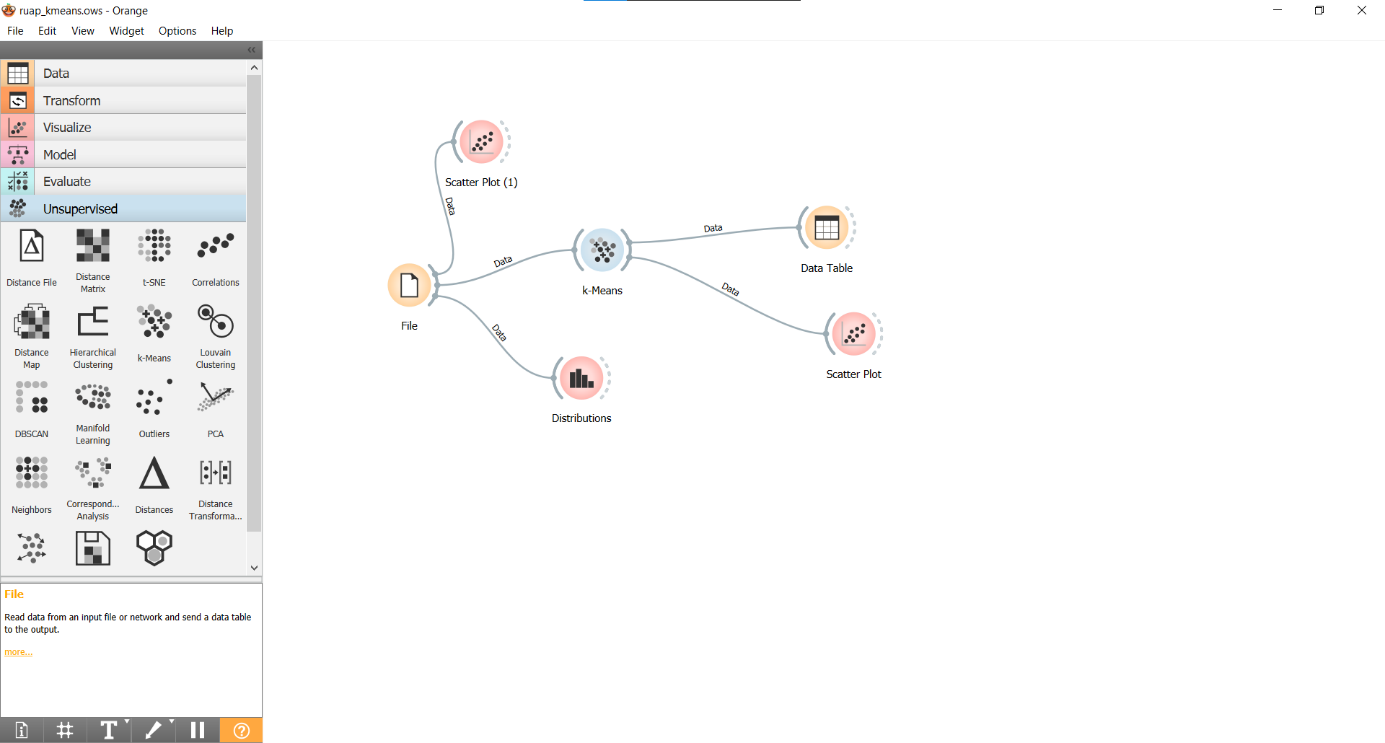


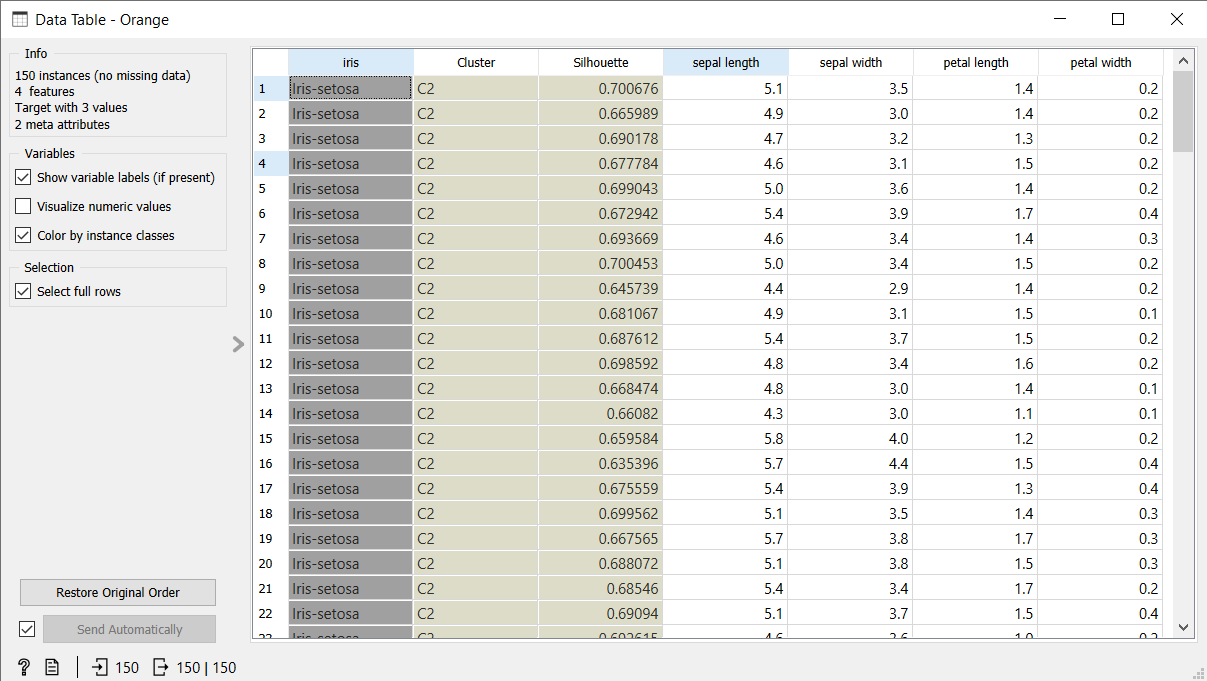


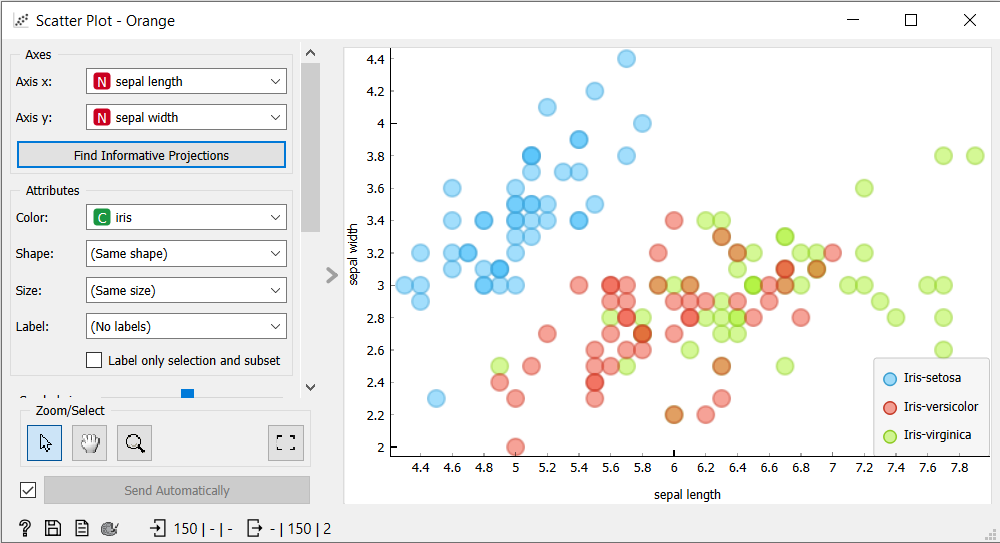
**Komentar:**

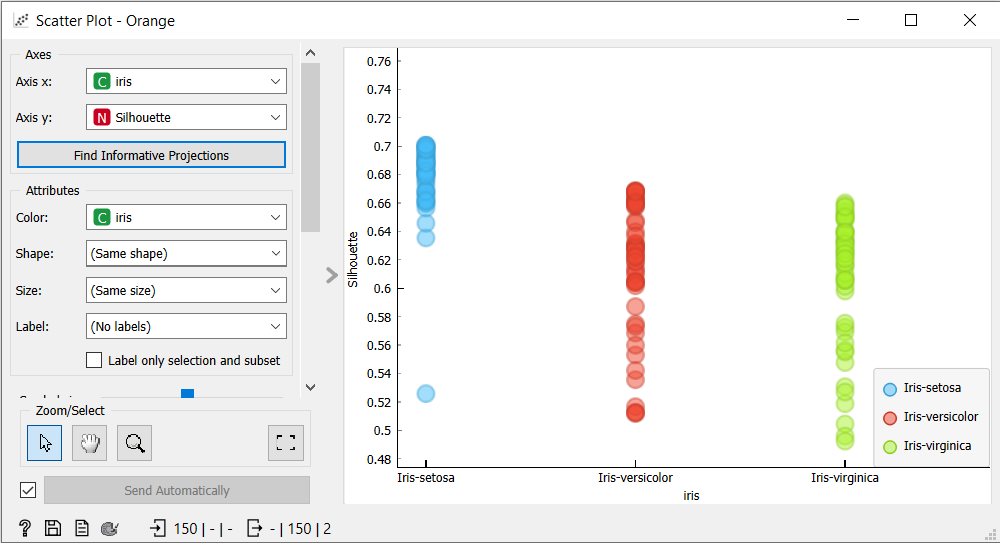
Kmeans je algoritam alternirajuće optimizacije koji na temelju zadanih centara određuje particiju, a na temelju particije centre. Glavni nedostatak Kmeans algoritma, ali i mnogih drugih je što broj grupa mora biti unaprijed poznat. Budući da taj podatak često nije poznat i potrebno je odrediti najprikladniji 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“.



**Rješenje:**

Rezultati za Iris dataset

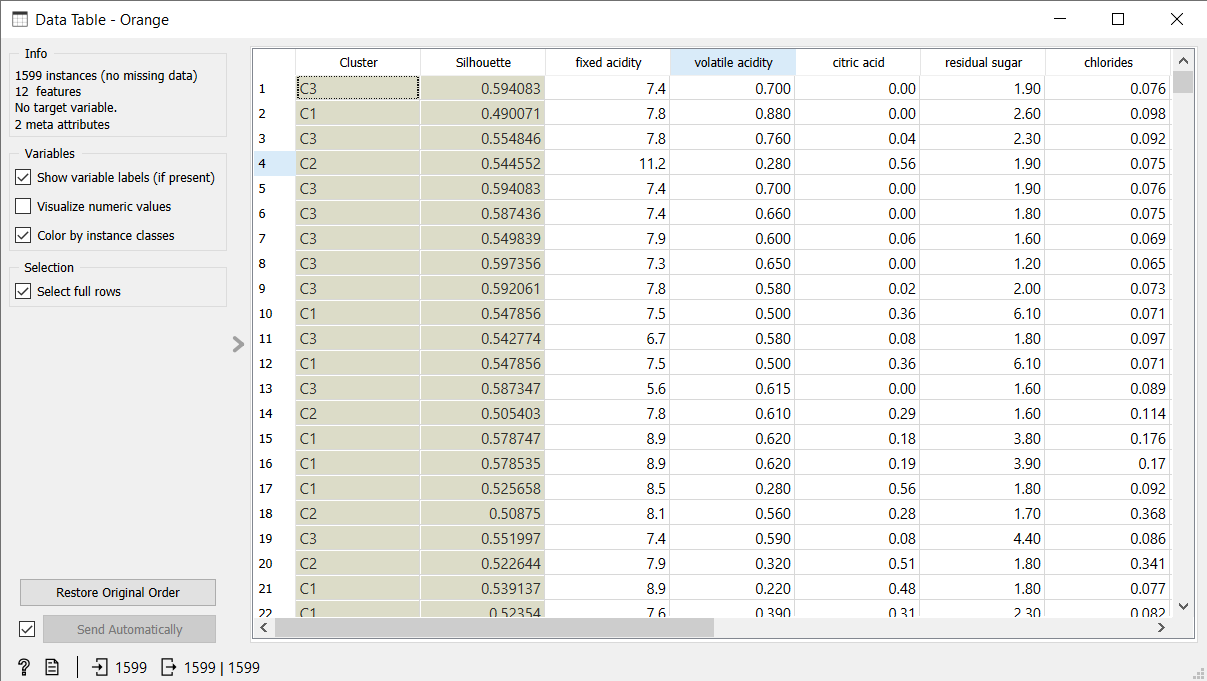


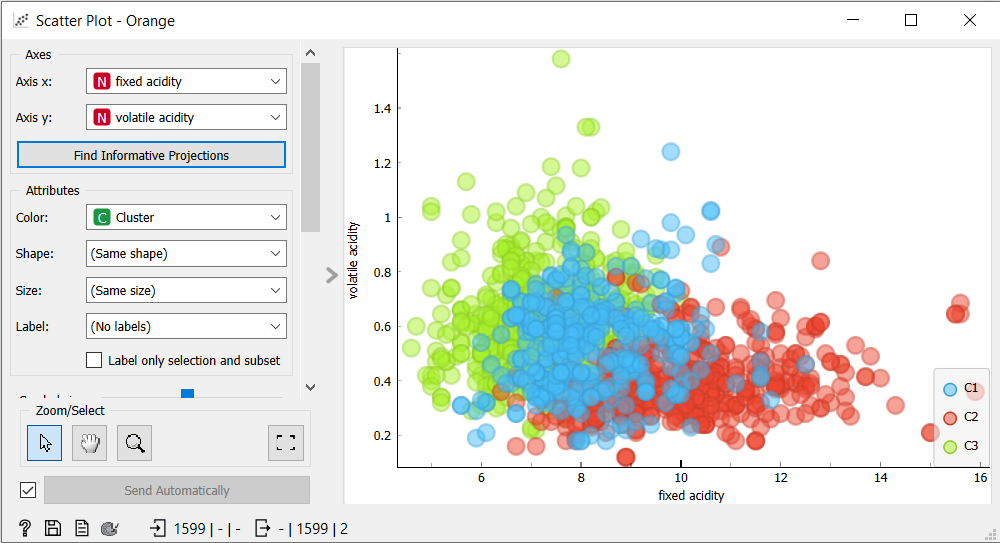


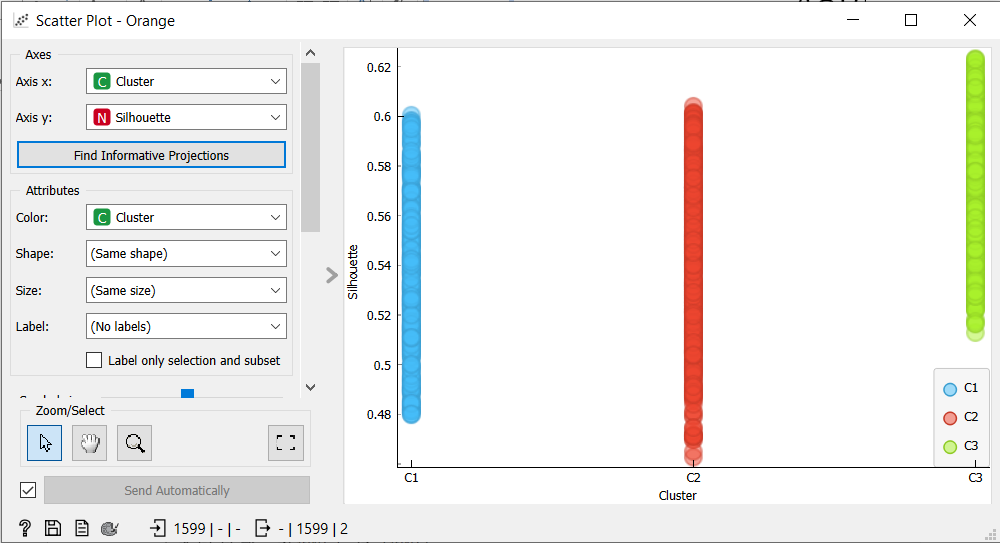
**Komentar:**

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

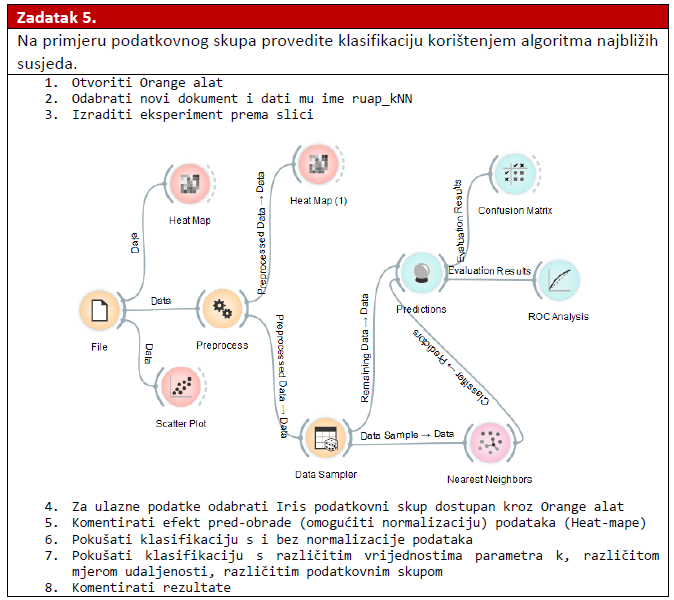
Dataset preuzet sa stranice Kaggle datasets -> <https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009?resource=download>



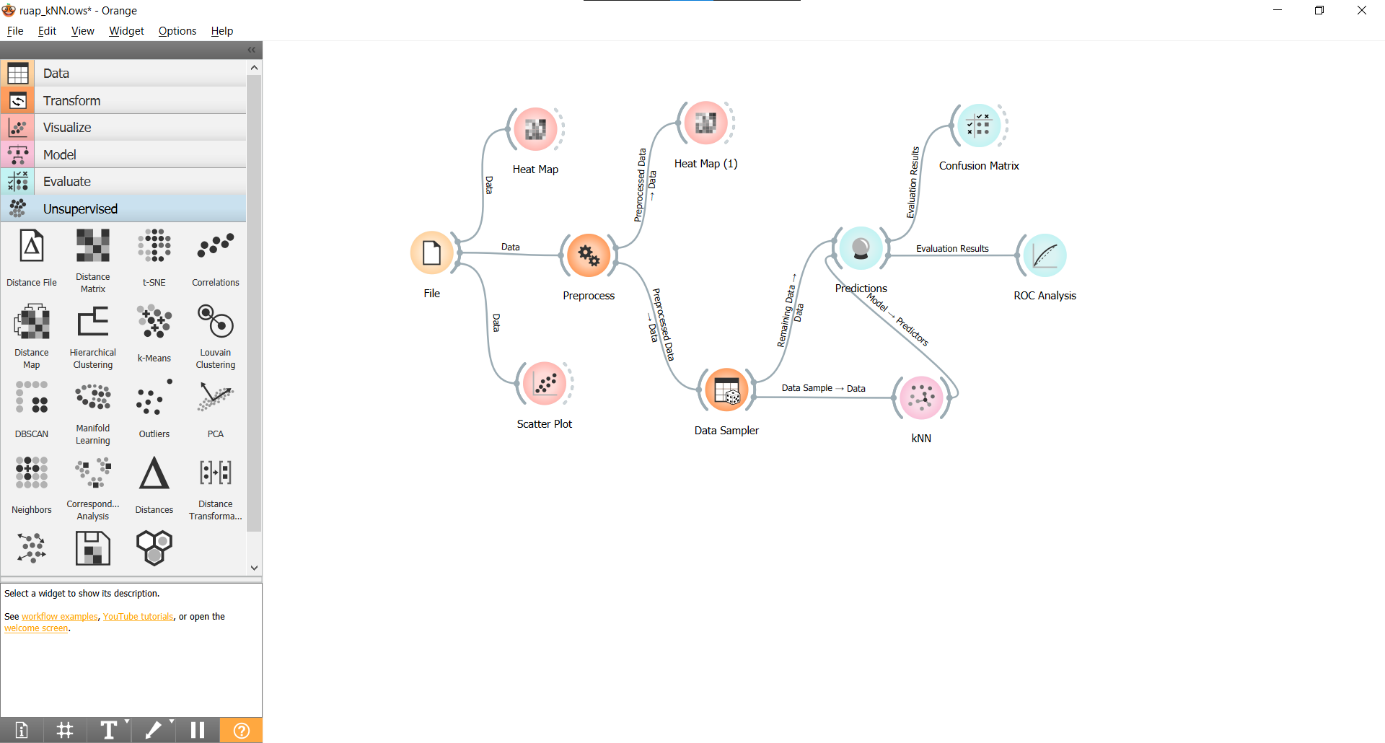


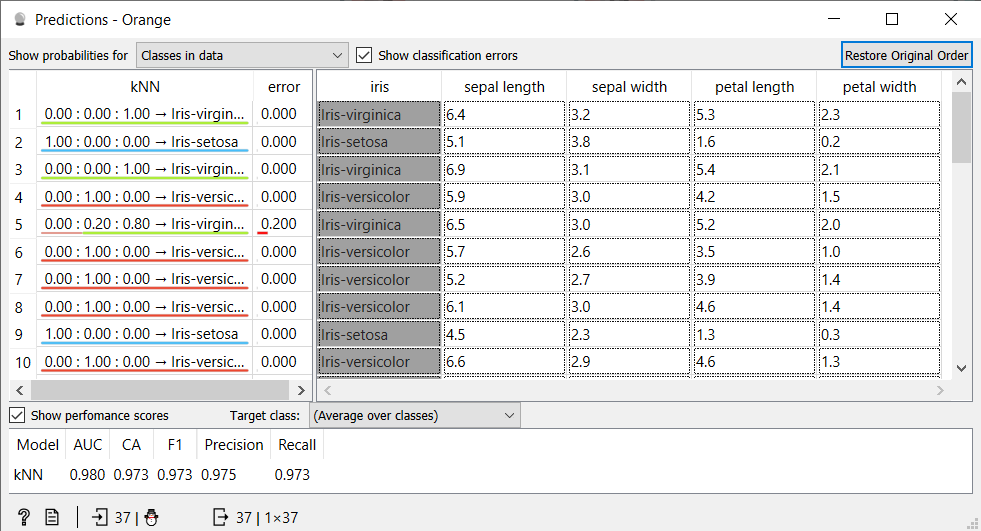
**Komentar:**

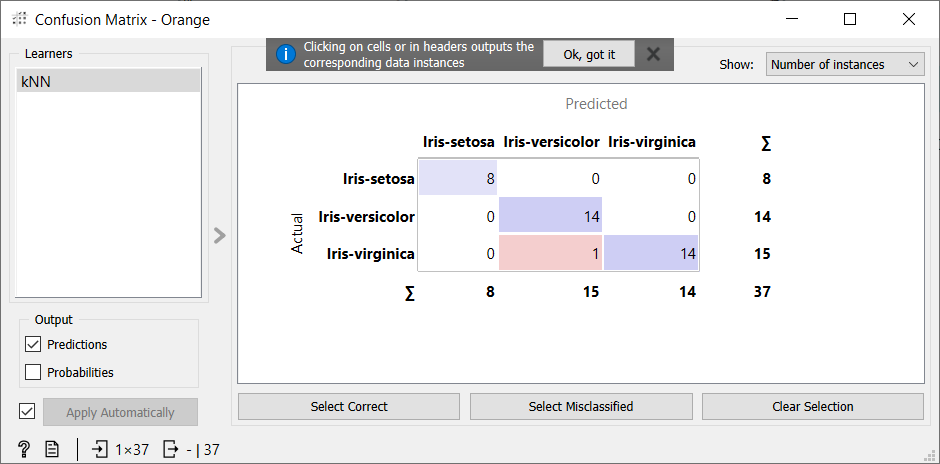
Usporedimo li rezultate s prethodnim datasetom, može se vidjeti da je u ovome primjeru Silhouette vrijednost bila otprilike 0.62, što je manje nego u prethodnom primjeru. Budući da se Silhouette vrijednost nalazi u intervalu [-1, 1], a vrijednosti koje se nalaze bliže vrijednosti 1 ukazuju da je uzorak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

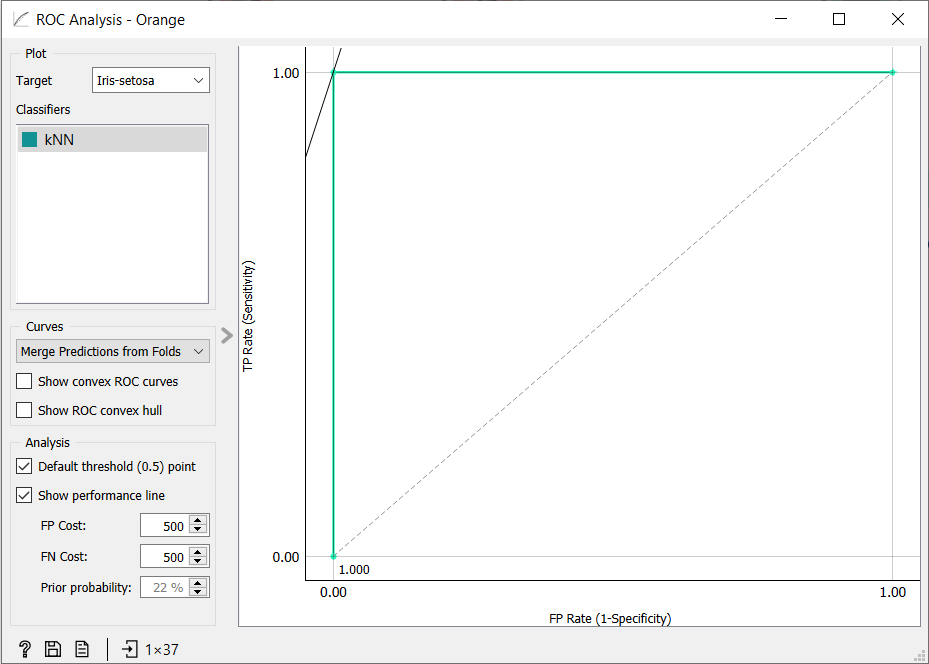


**Rješenje:**

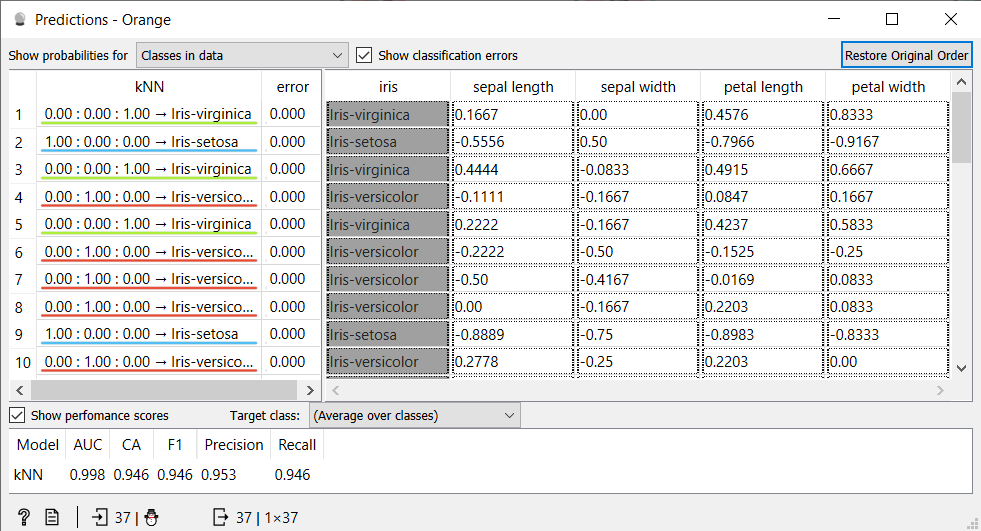


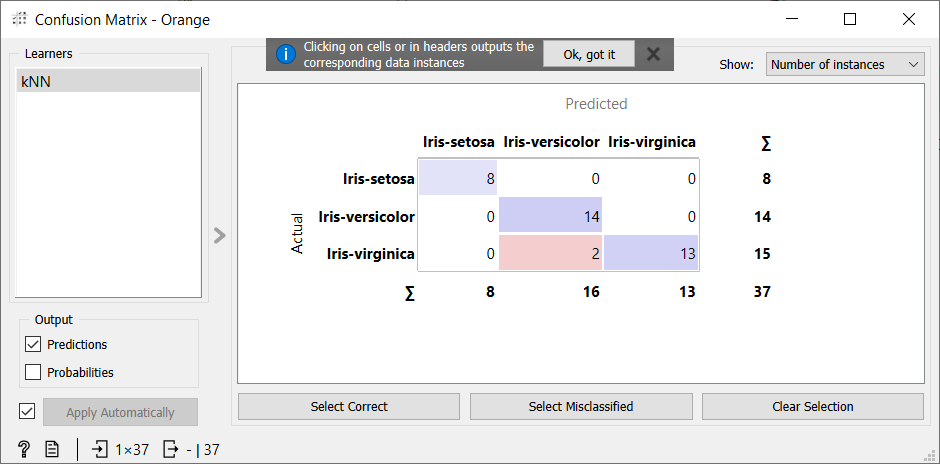
Rezultati bez normalizacije

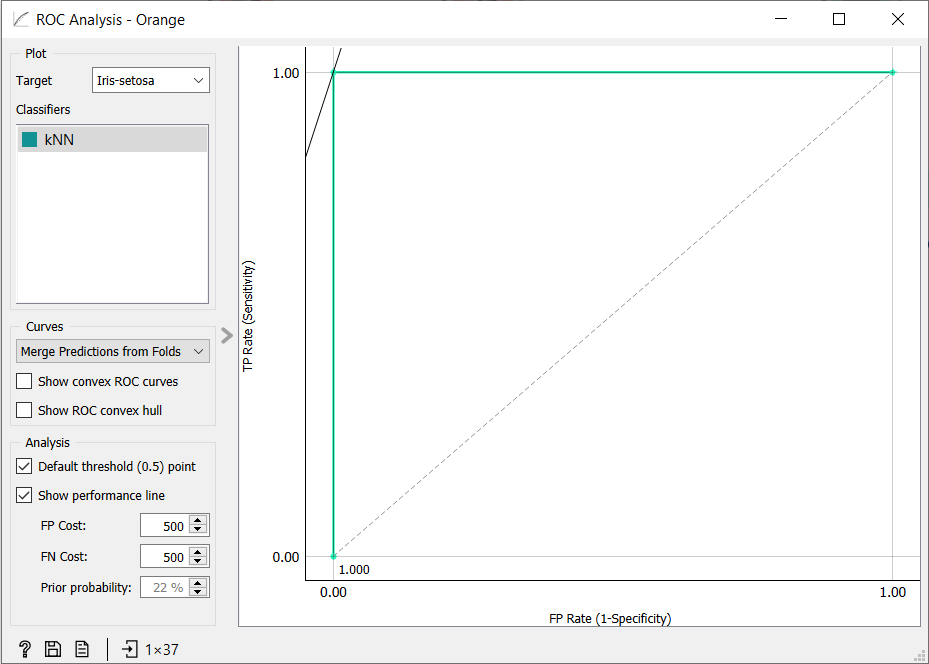




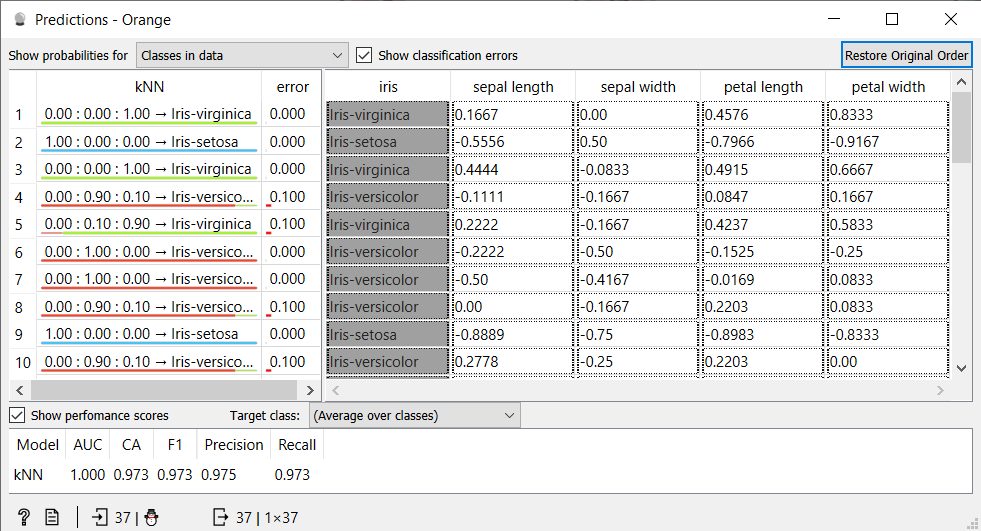
Rezultati nakon omogućene normalizacije:

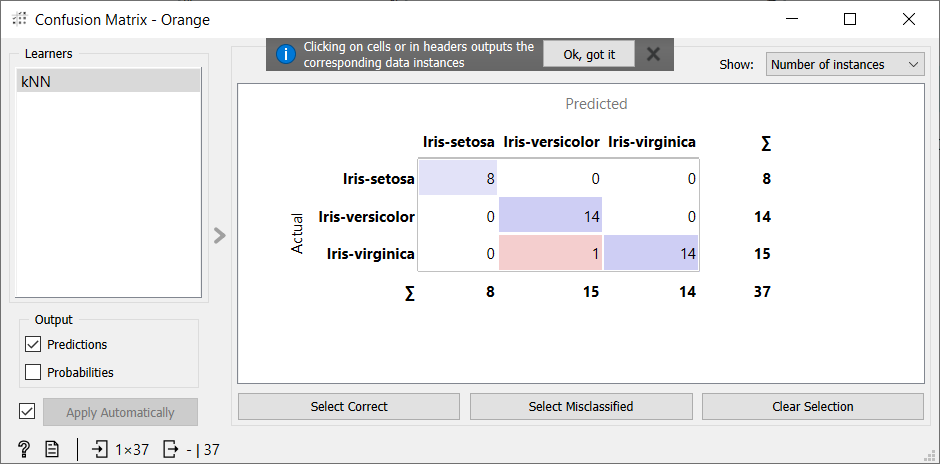


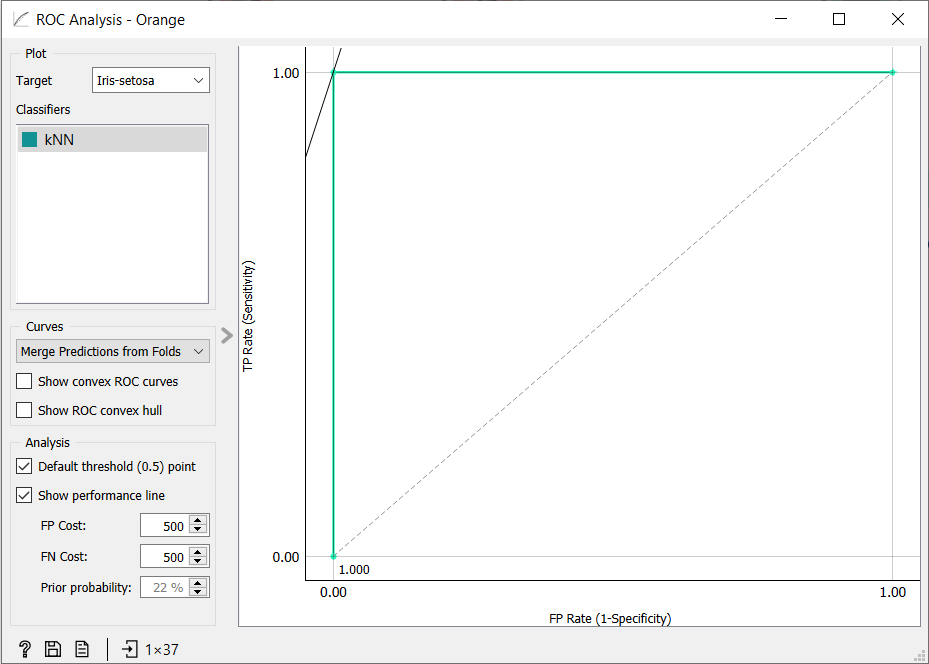




Rezultati nakon omogućene normalizacije i postavljanja parametra k = 10:

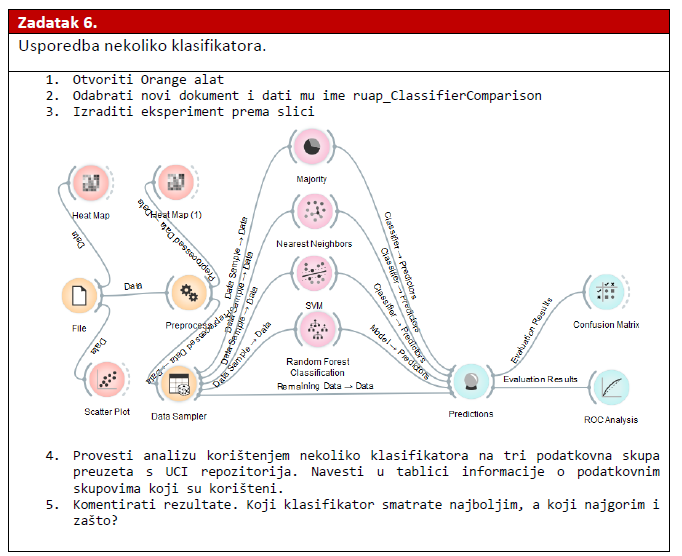


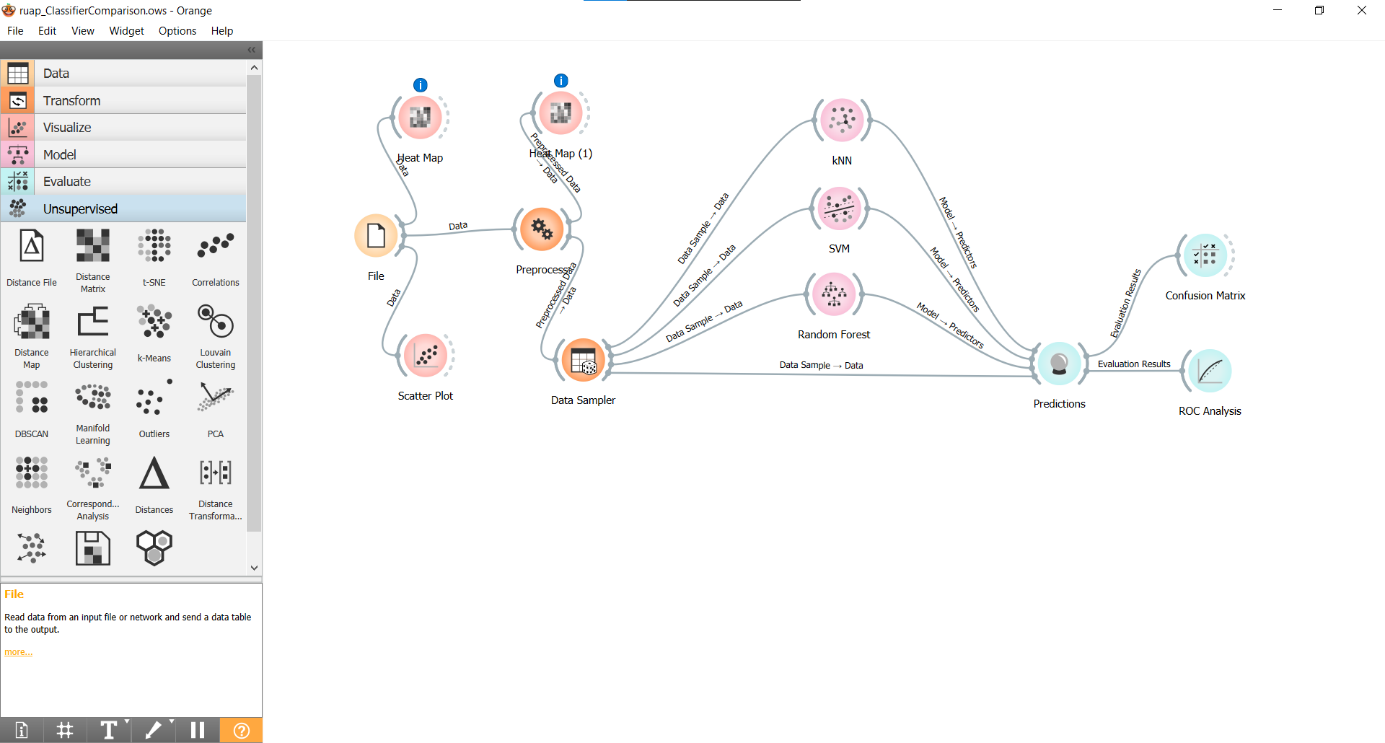


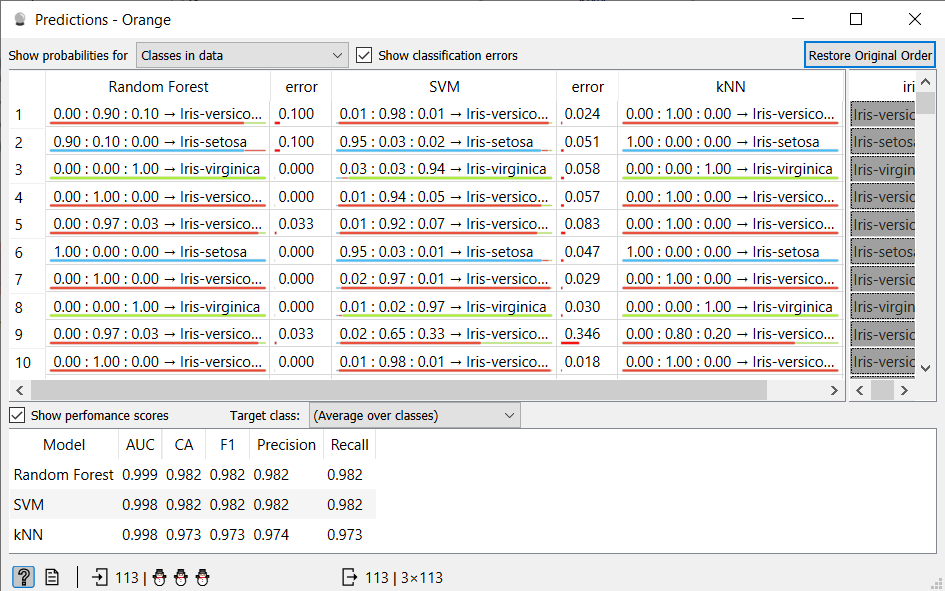


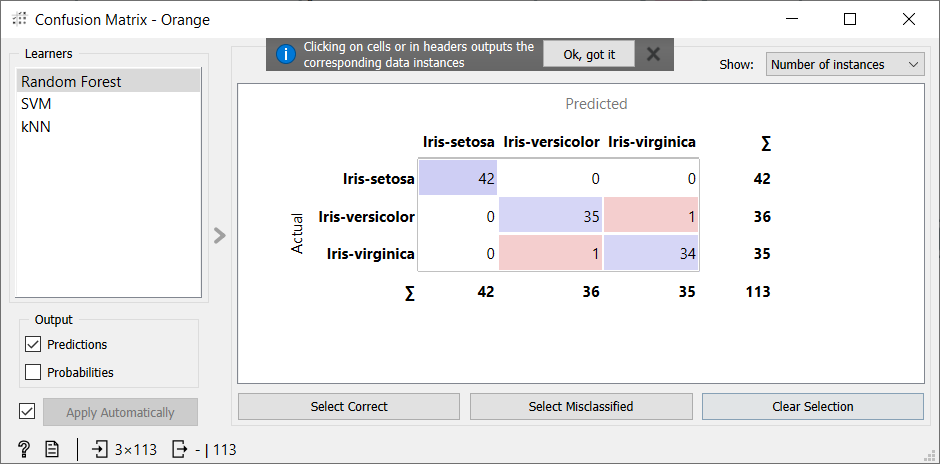
**Komentar:**

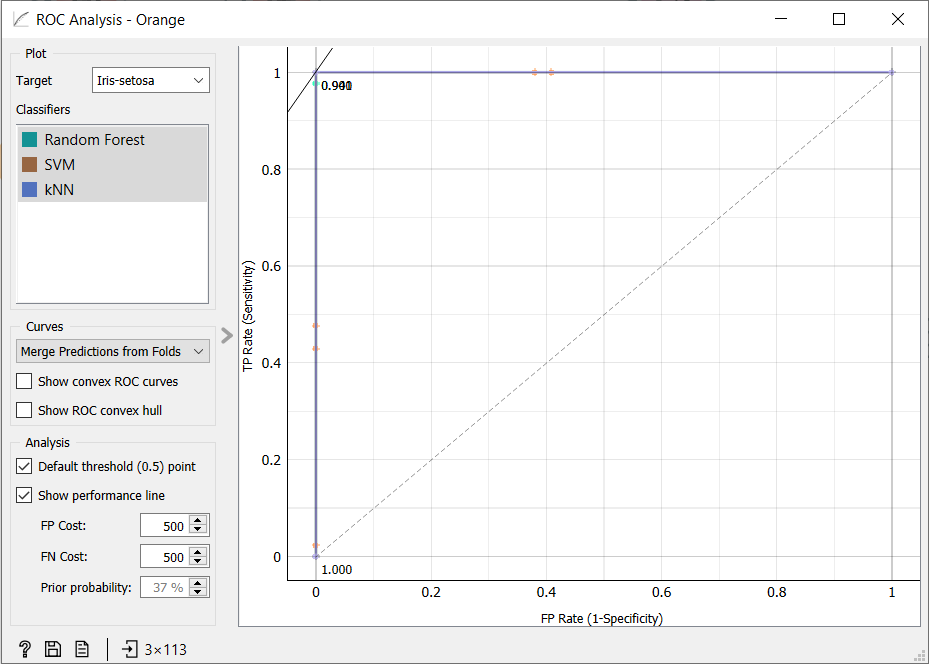
Usporedimo li dobivene rezultate, može se vidjeti da su rezultati nakon omogućene normalizacije nešto lošiji u usporedbi s druga dva rezultata. Uzmemo li slučaj bez normalizacije i slučaj s uključenom normalizacijom i povećanim parametrom k, vidimo da se dobiju bolji, odnosno precizniji rezultati. Analiza 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.

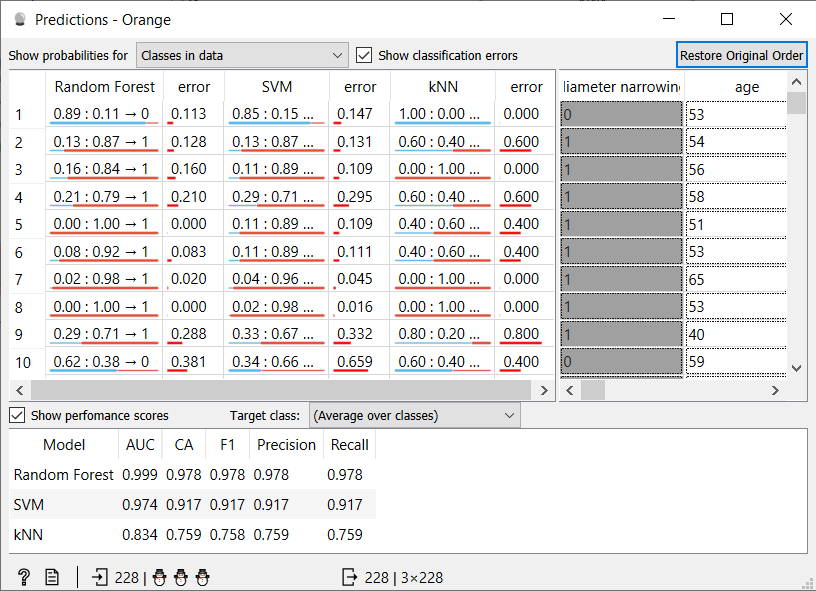


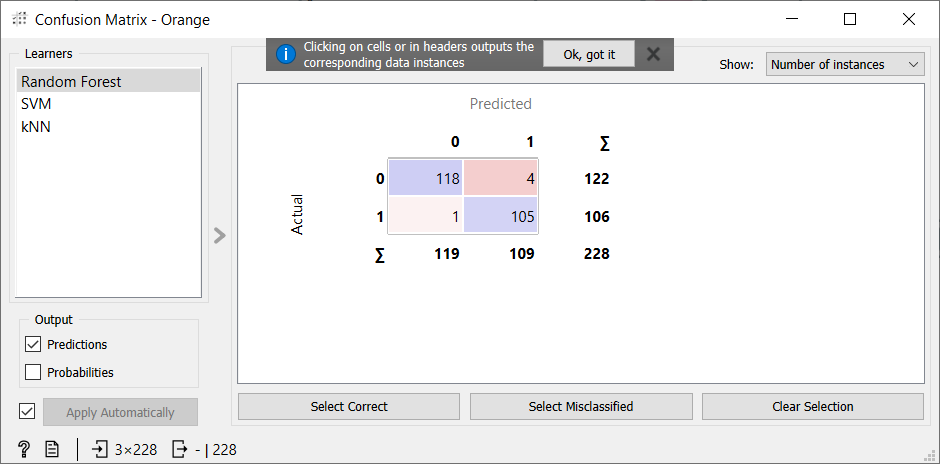
**Rješenje:**

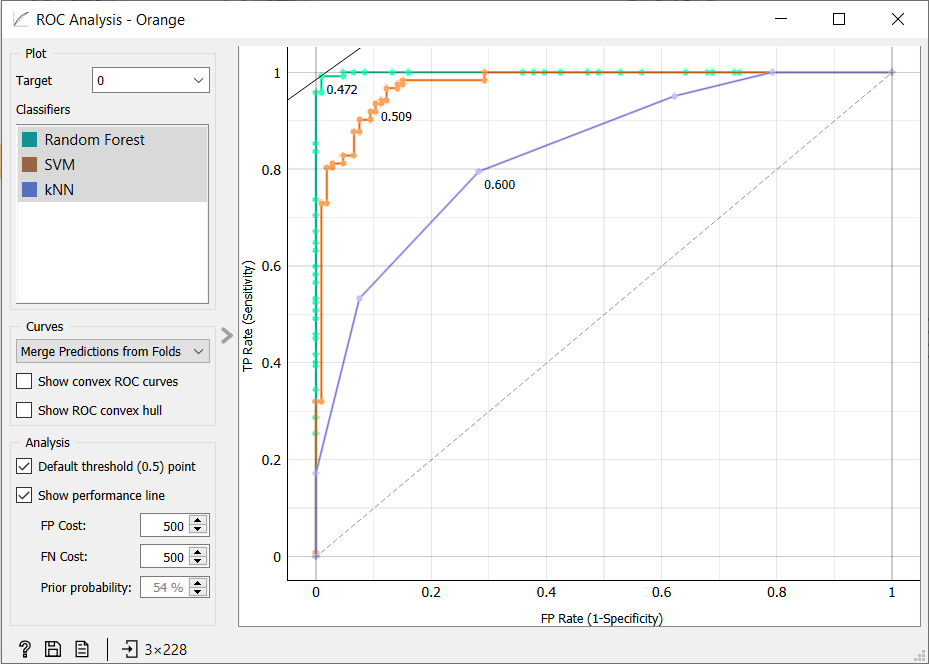
Rezultati za Iris dataset

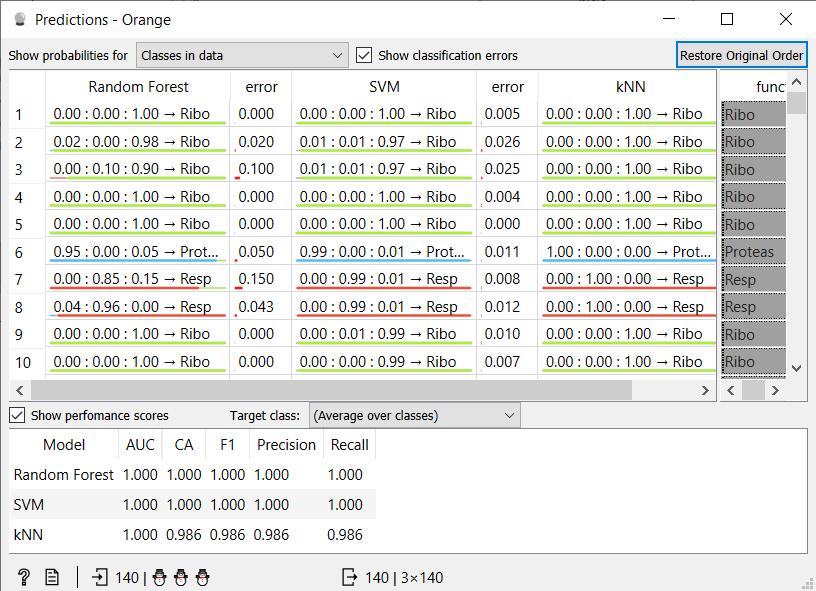


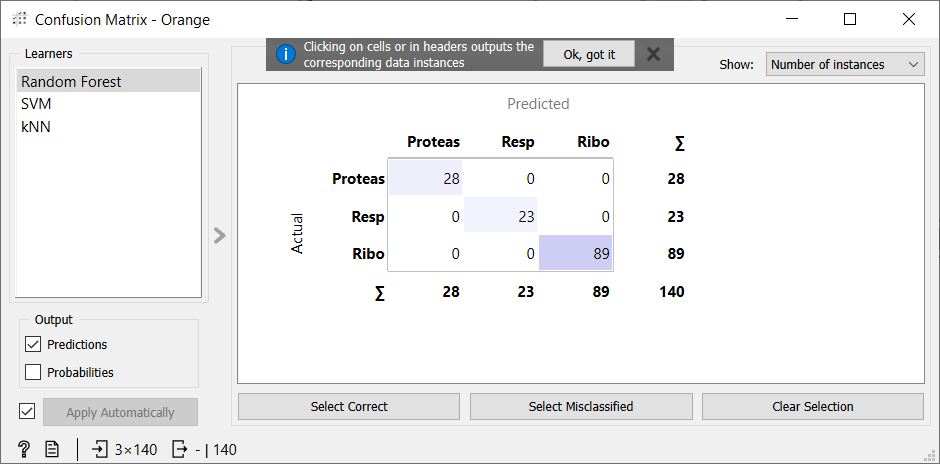


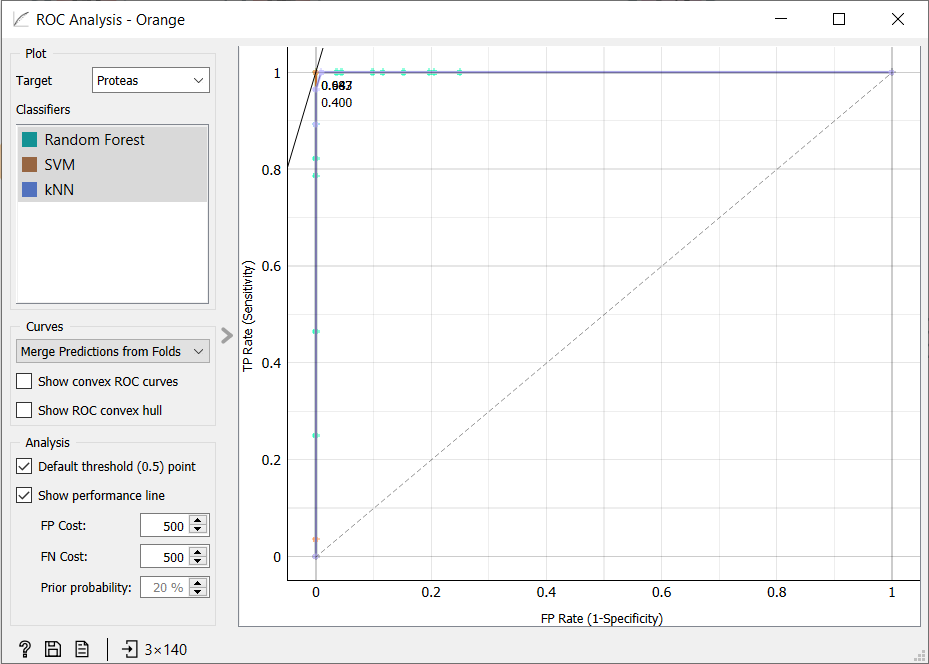
Rezultati za Heart disease dataset





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.