Tehnike i metode analize podataka

**Primena SVM za prepoznavanje slova abecede**

Danilo Milošević 1732

**Opis podataka i problema**

Vrši se analiza skupa podataka sa karakteristikama napisanih slova sa ciljem predviđanja koja slova su napisana. Za ovaj problem koristimo SVM odnosno support vector classifier. S obzirom da se vrši klasifikacija sa više klasa moramo da izaberemo da li se radi one-versus-one ili one-versus-rest. Kako skup nije velik, koristićemo one-versus-rest kako bi smanjili broj klasifikatora.

Skup podataka sadrži atribute kao što su širina, visina slova, pozicija gornjeg levog ugla slova i druge vrednosti. Atributi su

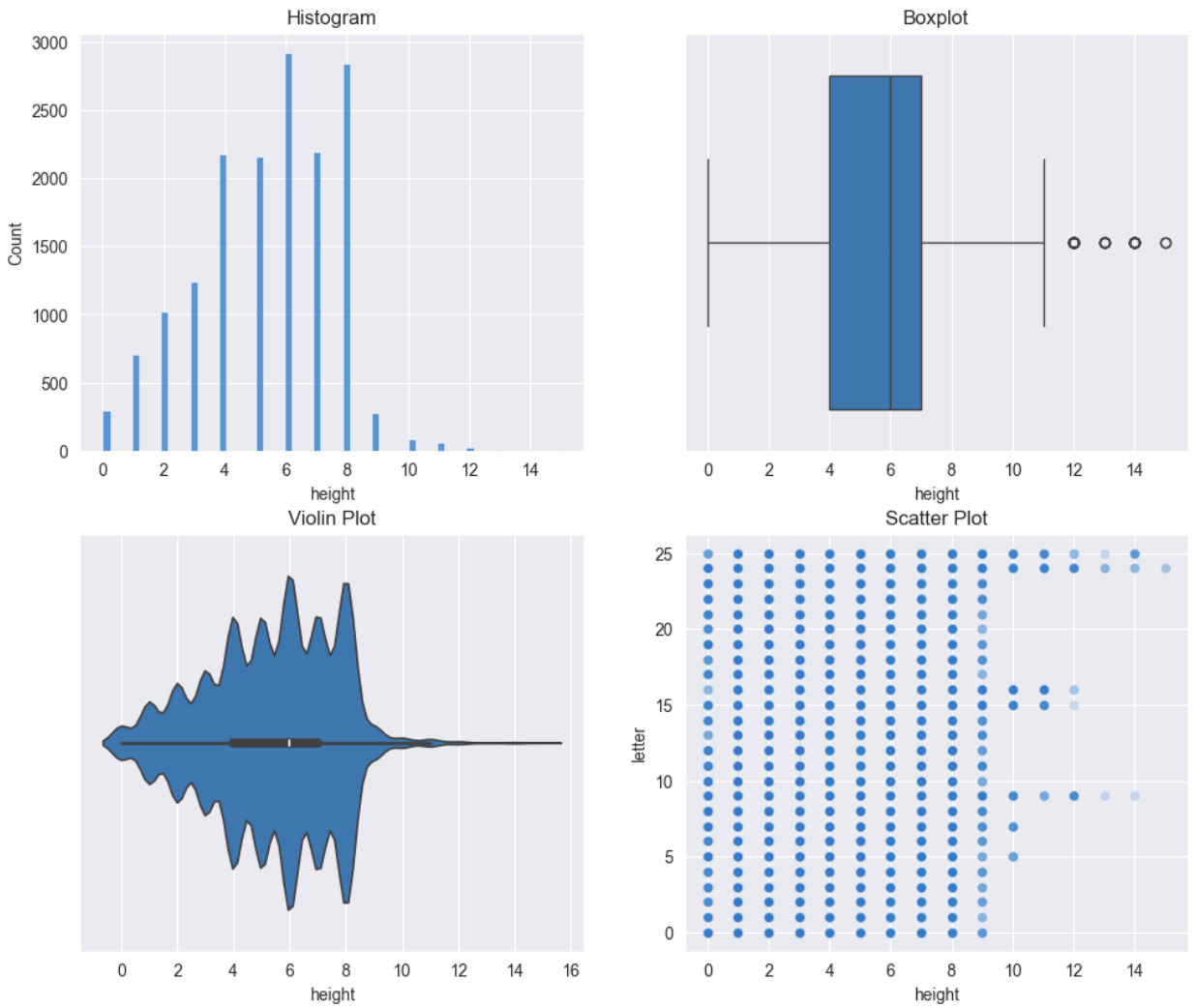
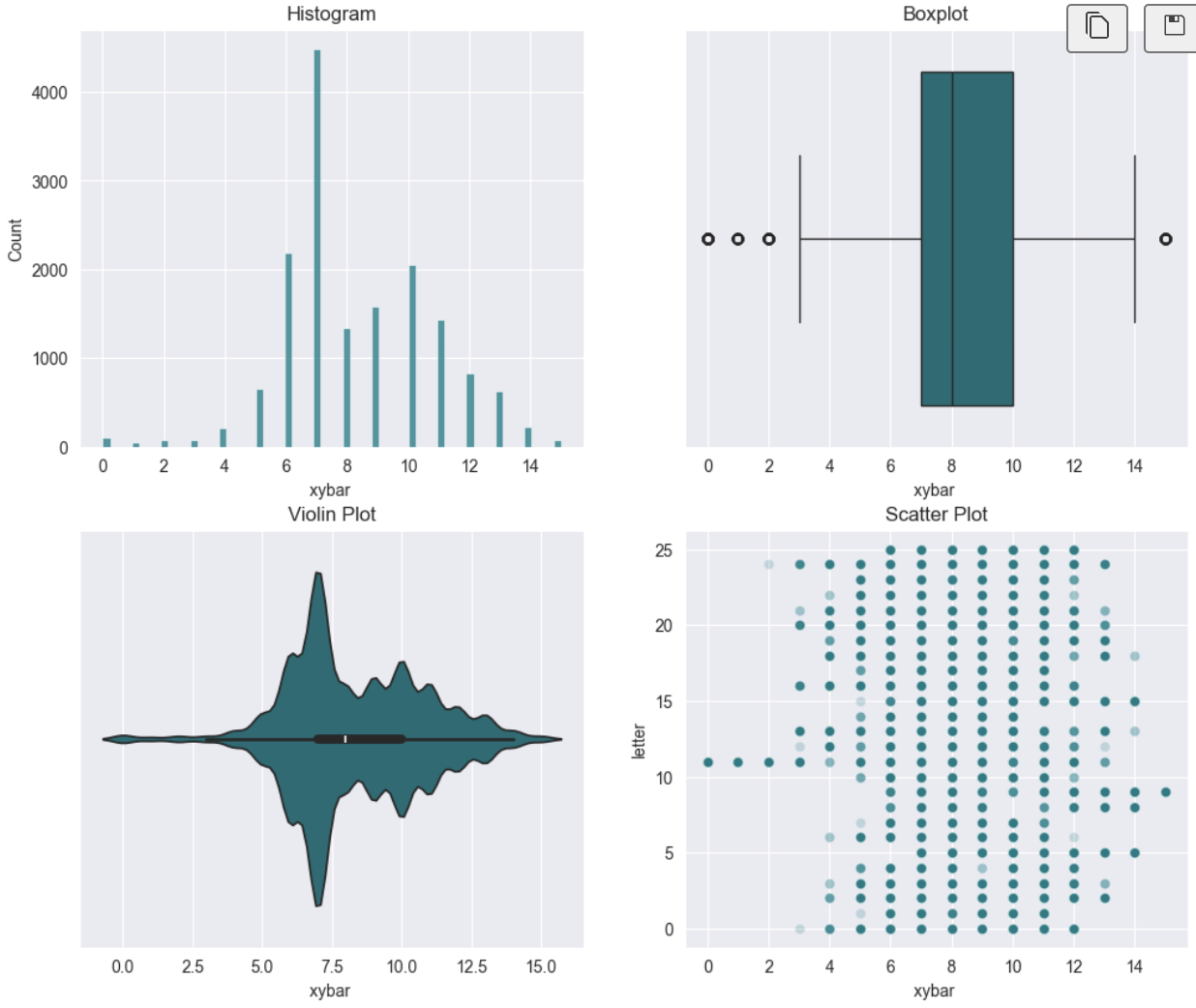
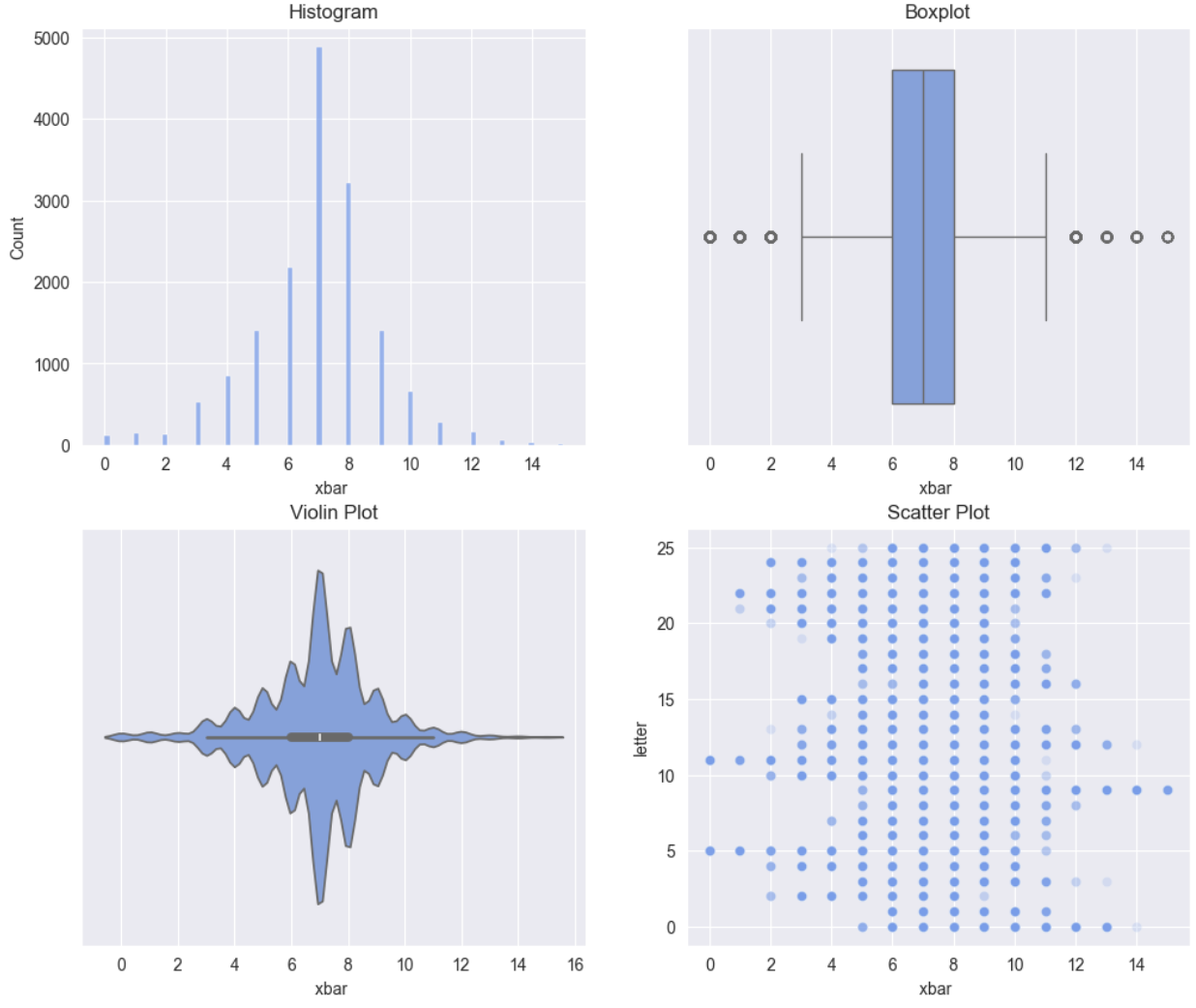
* letter - labela
* xbox
* ybox
* width
* height
* onpix
* xbar
* ybar
* x2bar
* y2bar
* xybar
* x2ybar
* xy2bar
* xedge
* xedgey
* yedge
* yedgex

Svi atributi imaju vrednosti u opsegu 0 - 15, ukupno ima 16 atributa i 20000 uzoraka.

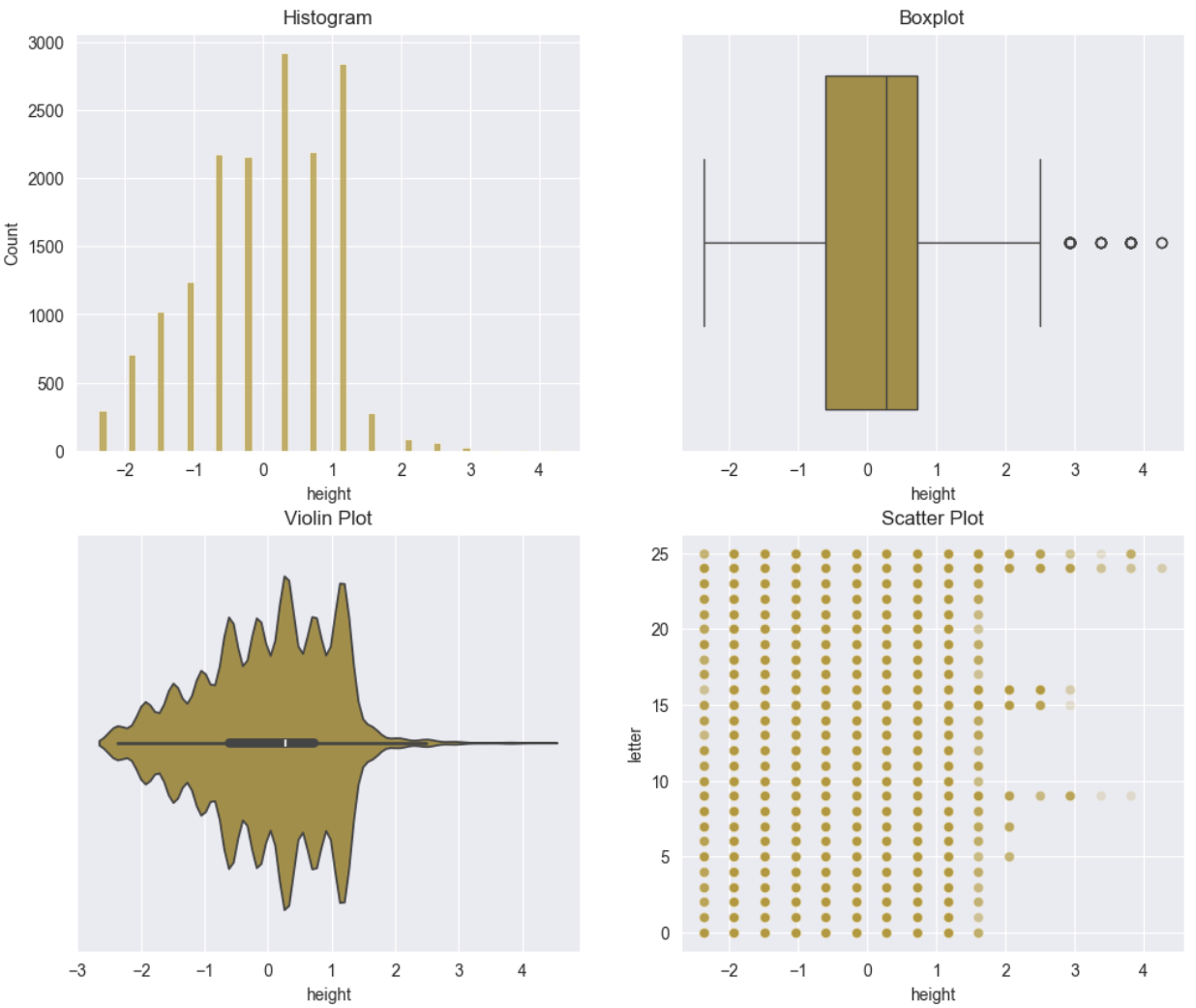
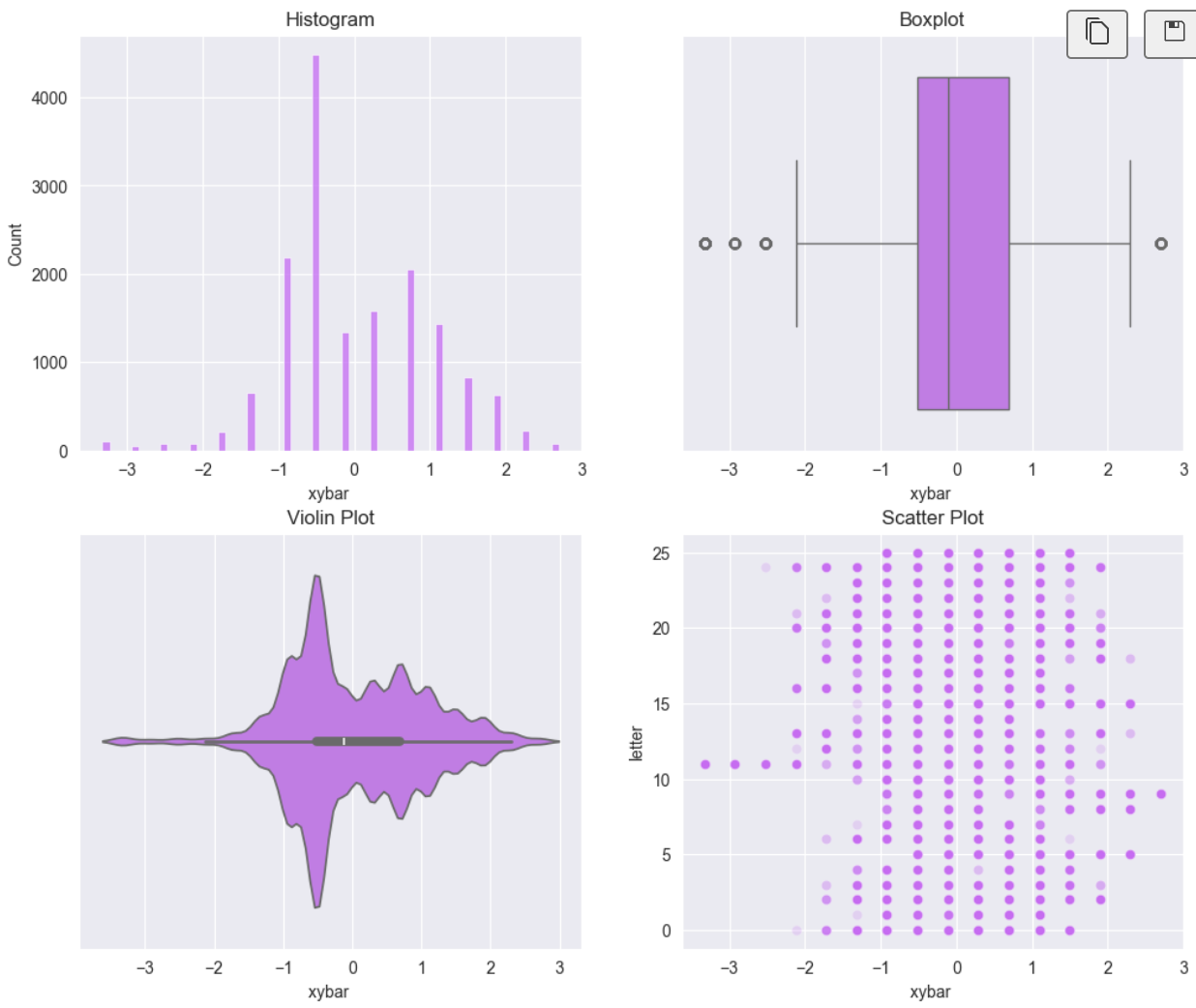
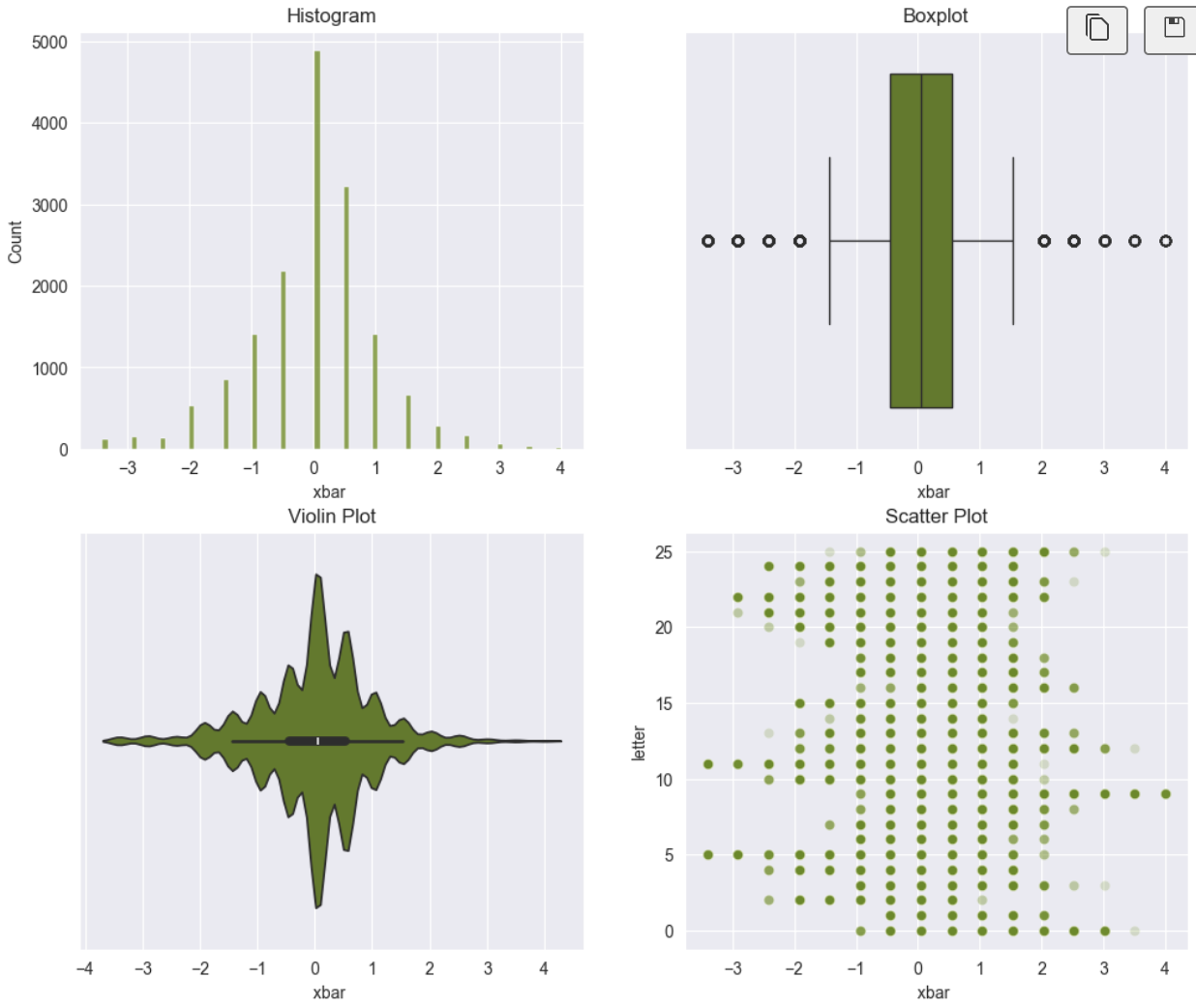
**Obrada atributa**

Potrebno je enkodirati labele za trening. To činimo tako što slovo abecede menjamo rednim brojem u abecedi. Pored toga vrši se prevođenje int64 atributa u float16 kako bi smanjili memorijsku zahtevnost bez gubitka preciznosti.

Ispod imamo vizuelizaciju raspodela nekih od atributa skupa podataka.

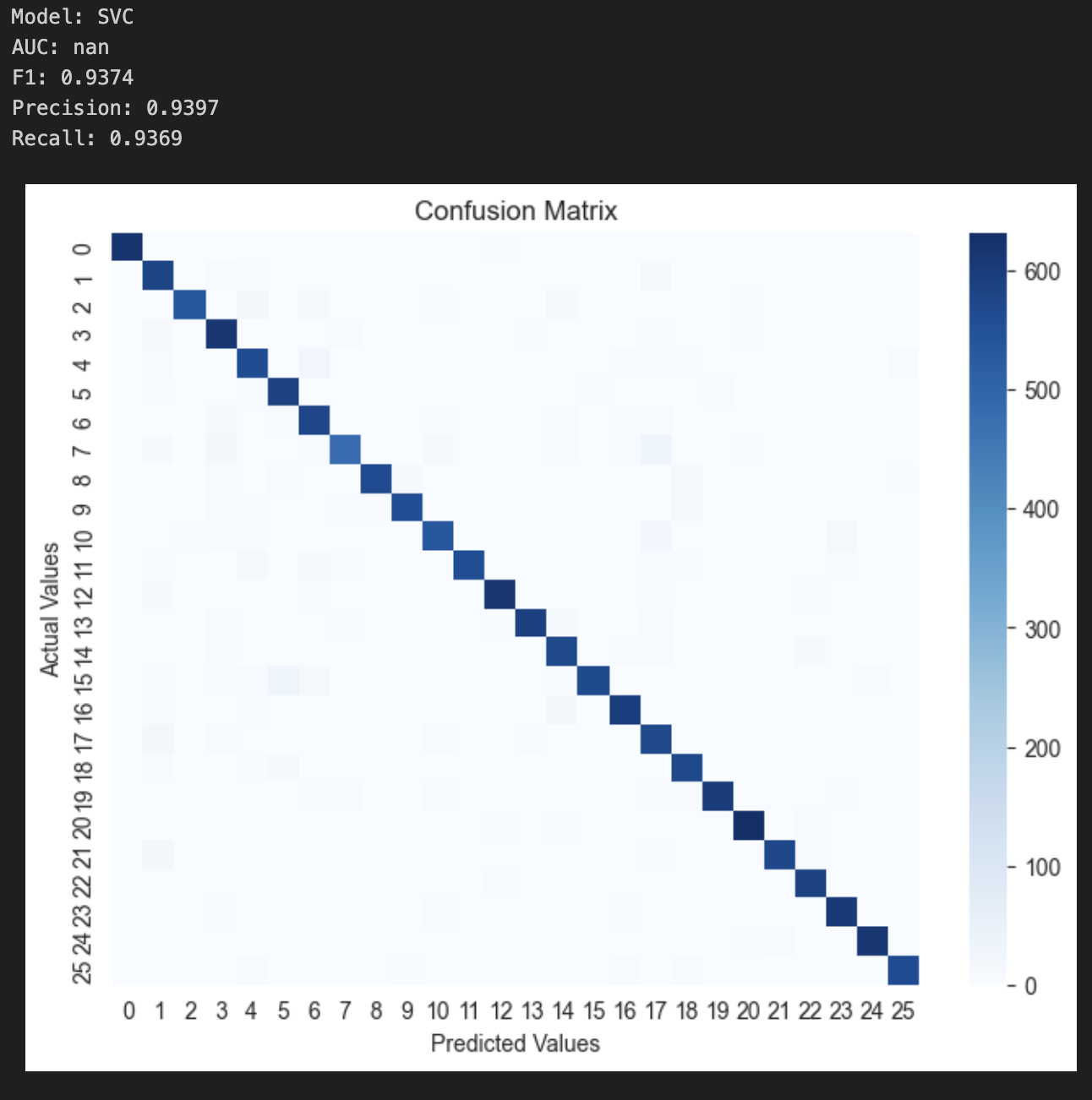


Pre nego što treniramo i tune-ujemo model izvršićemo skaliranje podataka, s obzirom da SVM bolje radi sa podacima koji su skalirani.



**Trening**

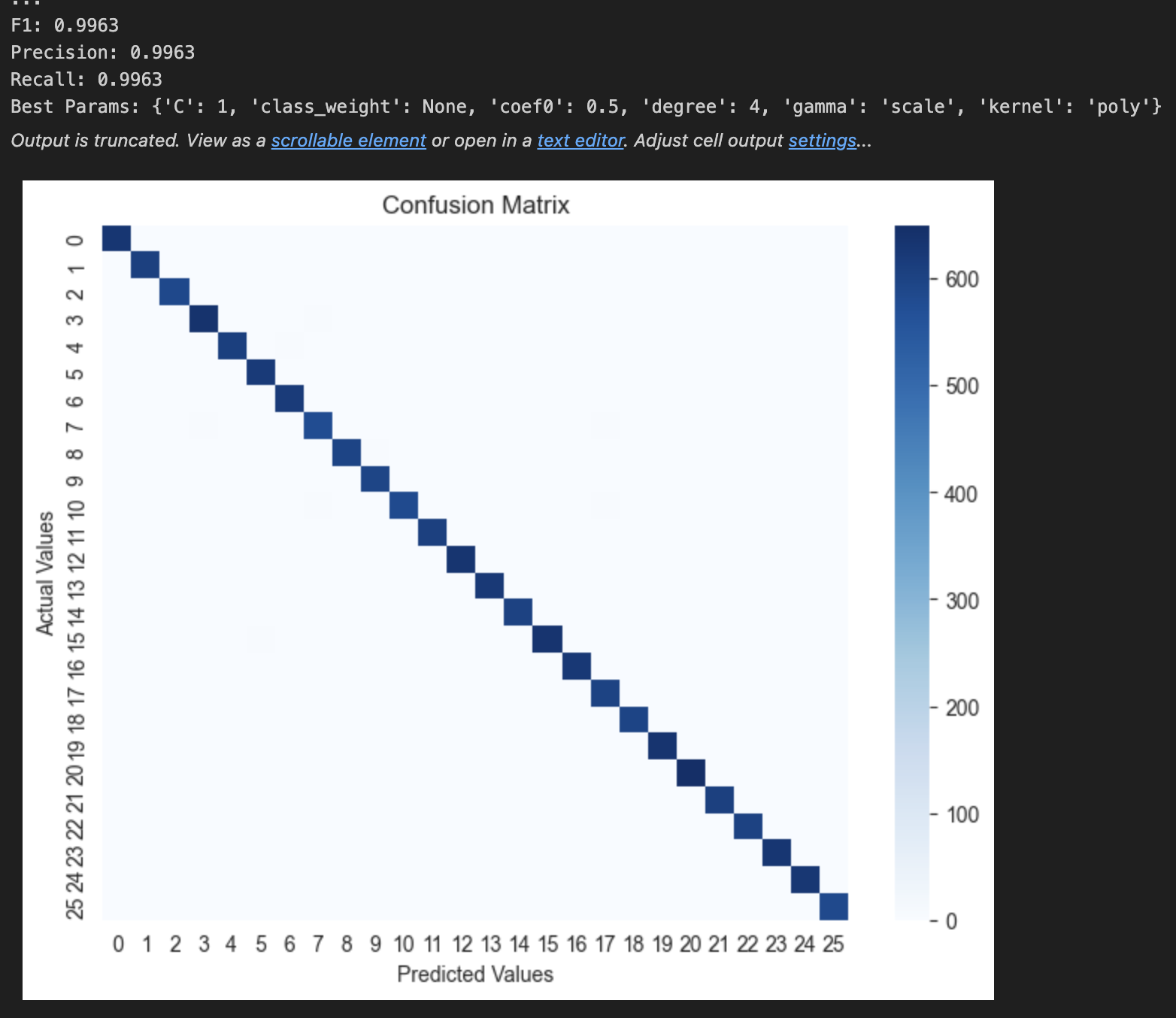
Prvo ćemo trenirati SVC model bez ikakvog tune-ovanja nad 80% početnog skupa podataka, dok će preostalih 20% biti za testiranje. Funkcija za trening će izračunati AUC, F1, Precision i Recall i prikazaće konfuzionu matricu.



Bez ikakvog tuninga model već daje odlične rezultate, gde je preciznost oko 94% a po konfuzionoj matrici vidimo da model retko pravi greške.

Iako daje dobre rezultate, model je moguće poboljšati. Za tune-ovanje koristimo HalvingGridSearchCV koji će izvršiti pretragu za što boljim parametrima tako što trenira više modela nad podskupom podataka, bira najbolje a zatim trenira ih opet nad većim podskupom dok se ne dobije najbolji model. Ovim procesom se dobija model koji koristi koeficijent regularizacije C=1, polinomijalni kernel sa stepenom 4.

Ispod je dat najbolji model kao i njegove performanse. Kao što vidimo model dostiže precision i recall oko 99.6 %. Model je pored toga isproban zajedno sa PCA ali daje gore rezultate - oko 98.5% precision i recall.



**Testiranje**

Testiranje modela je dalo blago gore rezultate nego tokom treninga. Rezultati su svakako odlični, oko 96% za precision i recall. Iako gori tokom treninga, tune-ovani SVC zajedno sa PCA je dao bolje rezultate tokom testiranja - 97%.

