Tehnike i metode analize podataka

**Primena stabla odlučivanja (random forests) za predviđanje žanrova pesama**

Danilo Milošević 1732

**Opis podataka i problema**

Vrši se analiza podataka o pesmama, pri čemu atributi predstavljaju različite karakteristike pesme kao što su izvođači, imena albuma i trajanje pesme. Cilj analize je predvideti žanr pesme, pri čemu za klasifikaciju koristimo algoritme sa stablima odlučivanja u osnovi kao što su RandomForest, AdaBoost, HistGradientBoosting i ExtraTrees klasifikatori.

Lista atributa u skupu podataka:

* id - id reda u skupu
* track\_id - jedinstveni id pesme
* artists - izvođači, razdvojeni ';'
* album\_name - ime albuma
* track\_namet - ime pesme
* popularity - vrednost od 0 do 100, bazirana na tome koliko često je pesma puštana na Spotify-u u skorije vreme
* duration\_ms - trajanje pesme u ms
* explicit - da li pesma sadrži eksplicitan tekst
* danceability - koliko je pesma "plesljiva" - od 0 do 1
* energy - koliko je pesma energična (0 do 1), određena na osnovu više odlika pesme (brzina, koliko je pesma glasna...)
* key - u kojoj lestvici je pesma
* loudness - koliko je pesma glasna u dB
* mode - da li je dur ili mol lestvica (1 - dur, 0 - mol)
* speechiness - koliko je zastupljen tekst u pesmi - 0 do 1
* acousticness - verovatnoća da je pesma akustična - 0 do 1
* instrumentalness - verovatnoća da je pesma instrumental - 0 do 1, 1 označava veliku verovatnoću da nema vokala u pesmi
* liveness - verovatnoća da je pesma izvedena u javnosti (koncert) - 0 do 1
* valence - koliko je pesma vesela - 1 predstavlja srećne pesme dok 0 su tužnije pesme
* tempo - brzina pesme u bitovima po minutu
* time\_signature - ritam, takt
* **track\_genre** - žanr pesme, ciljni atribut

Skup podataka sadrži 114000 redova, sa 21 ulaznih promenljivih. Nema null vrednosti u skupu i većina atributa su numerički, izuzev track\_id, artists, album\_name, track\_name i track\_genre koje moramo da enkodiramo.

ID atribute možemo odmah da izbacimo.

**Obrada atributa**

Atribute track\_id i id možemo odmah da izbacimo.

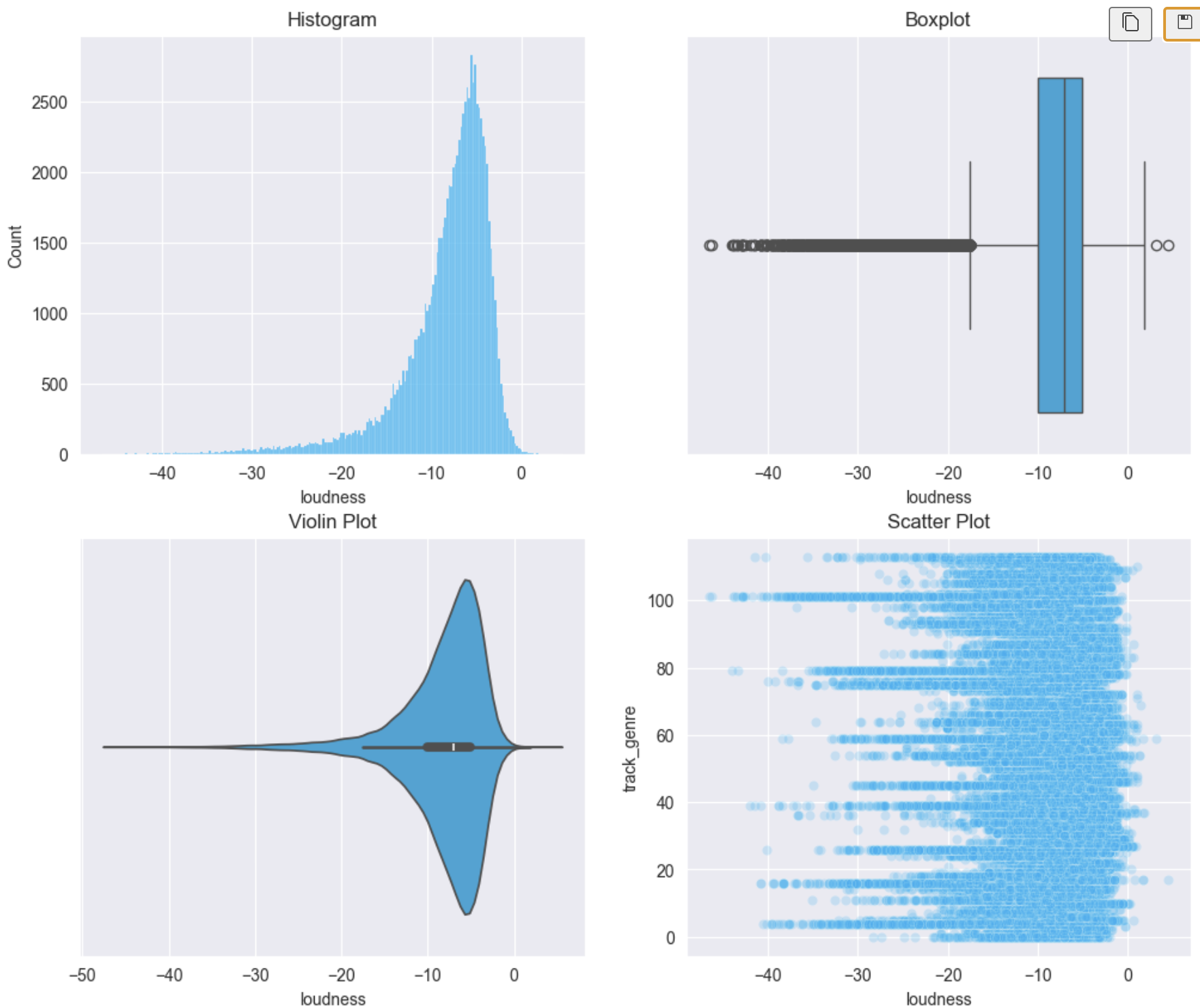
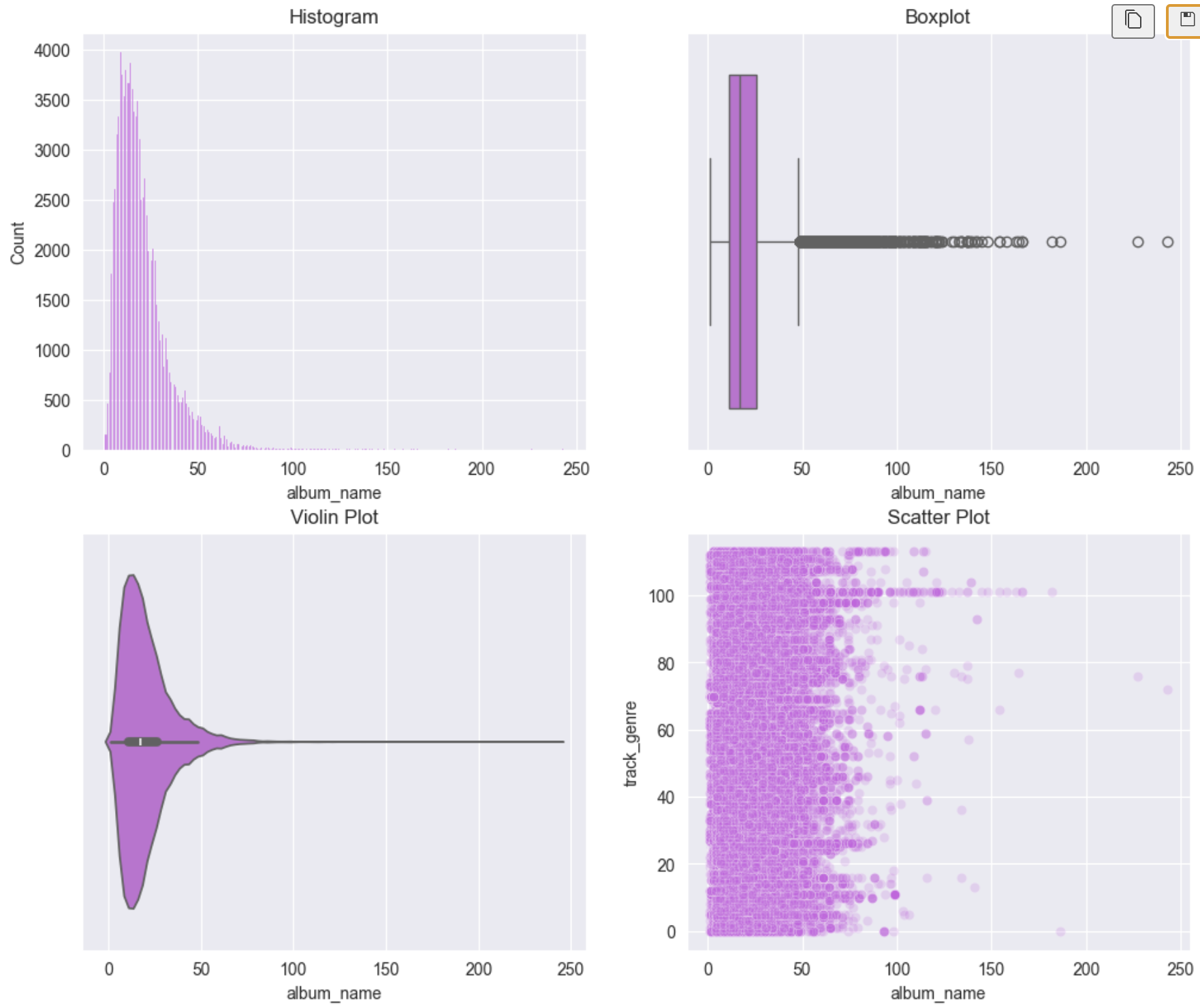
Atribut artists analiziramo na osnovu čega zaključujemo da pesma sa najviše izvođača ima 38 izvođača dok ima 29859 jedinstvenih izvođača. Da bi enkodirali možemo ili da dodamo 38 atributa sa jedinstvenim id-em umetnika ili da atribut artists menjamo ukupnim brojem umetnika koji izvode pesmu. Izabrana je druga metoda.

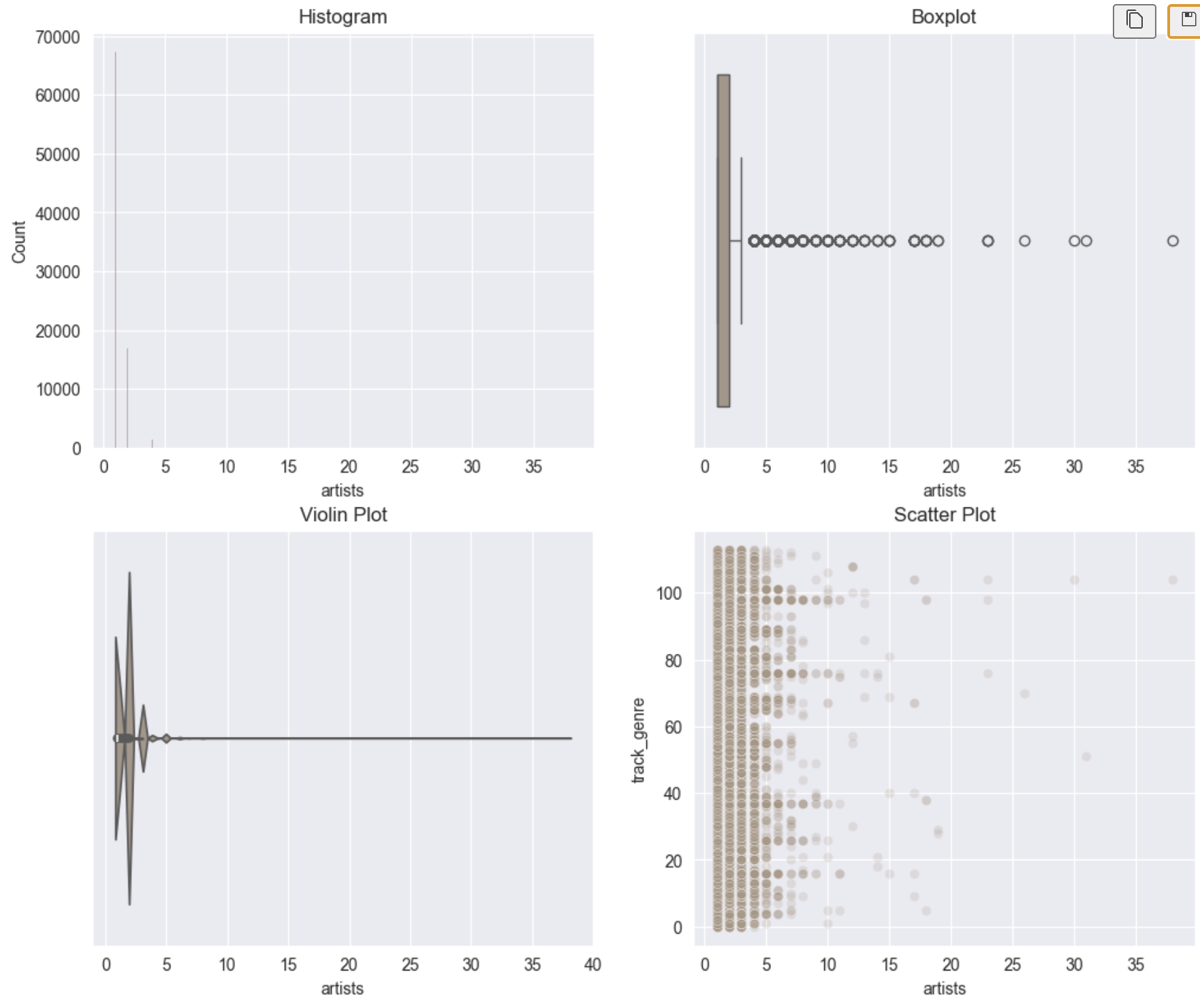
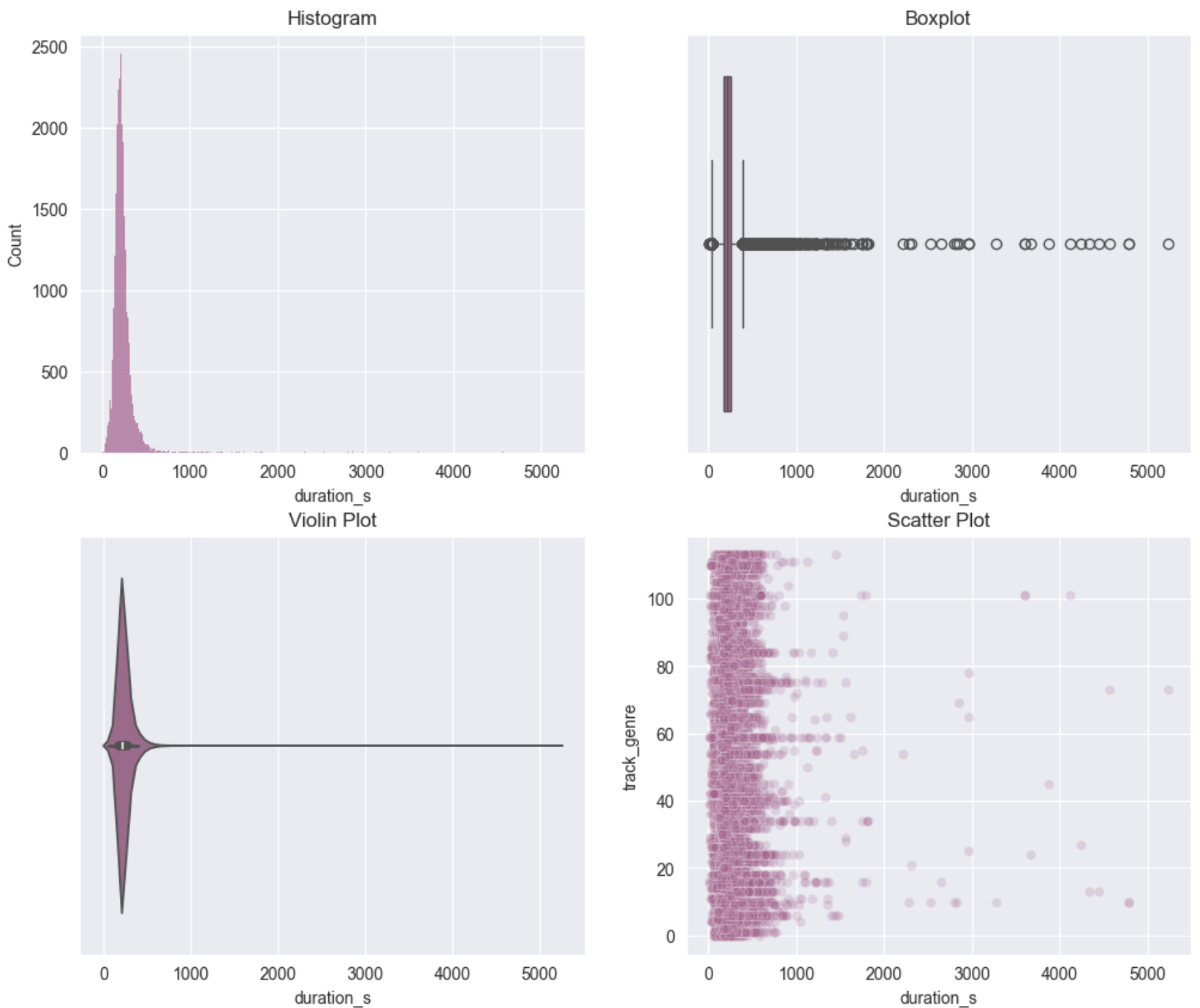
Isti metod primenimo i za track i album name.

Atribut track\_genre enkodiramo korišćenjem LabelEncoder dok atribut explicit prevodimo u celobrojne vrednosti (0 ili 1).

Kasnije tokom treninga se pojavio problem gde nije bilo dovoljno memorije, zato su tipovi podataka prevedeni u manje tipove - int32 i int64 su prevedeni u int16 a float64 u float 16.

Ispod možemo da vidimo raspodele pojedinih atributa.





**Trening i tuneovanje**

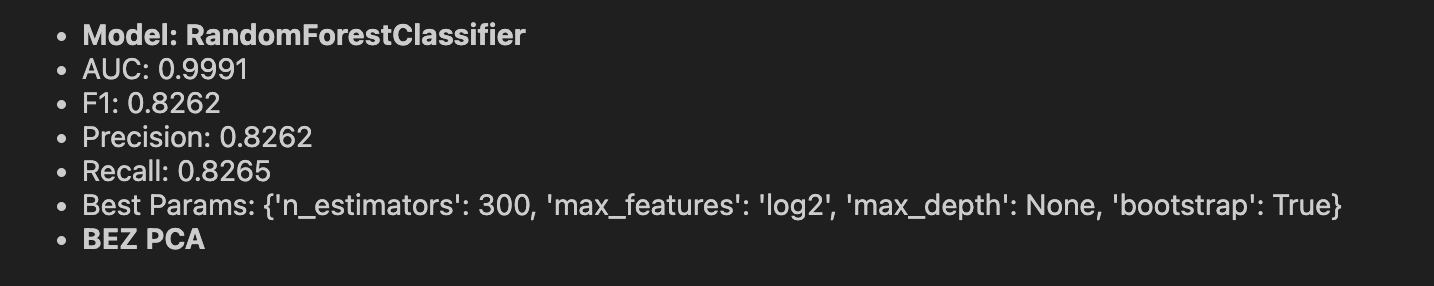
Trening je obavljen na sledeći način

* Inicijalizovani su objekti klasifikatora kao i dictionary koji definiše vrednosti parametara za tuno-vanje
* Definisan je niz transformacija koje mogu da se primene nad podacima - kako stabla odlučivanja nisu osetljiva na scale i da li su podaci normalizovani jedina transformacija koja se primenjuje je PCA
* Funkcija za trening vrši tune-ovanje pomoću RandomizedSearchCV, testira performanse na trening setu - f1 score, auc, precision i recall a zatim iscrtava konfuzionu matricu kao i feature importance

Trenirani su RandomForest, HistGradientBoosting, AdaBoost i ExtraTrees klasifikatori.

Zbog nedostatka memorije poslednje pokretanje je prestalo sa radom ali najbolji rezultat je zabeležen ispod.

Najbolje rezultate je dao RandomForestClassifier sa 300 estimatora i max\_features 'log2', bez primene PCA.



**Testiranje**

Testiranjem modela nad test skupom podataka dobijeni su sledeći rezultati.

