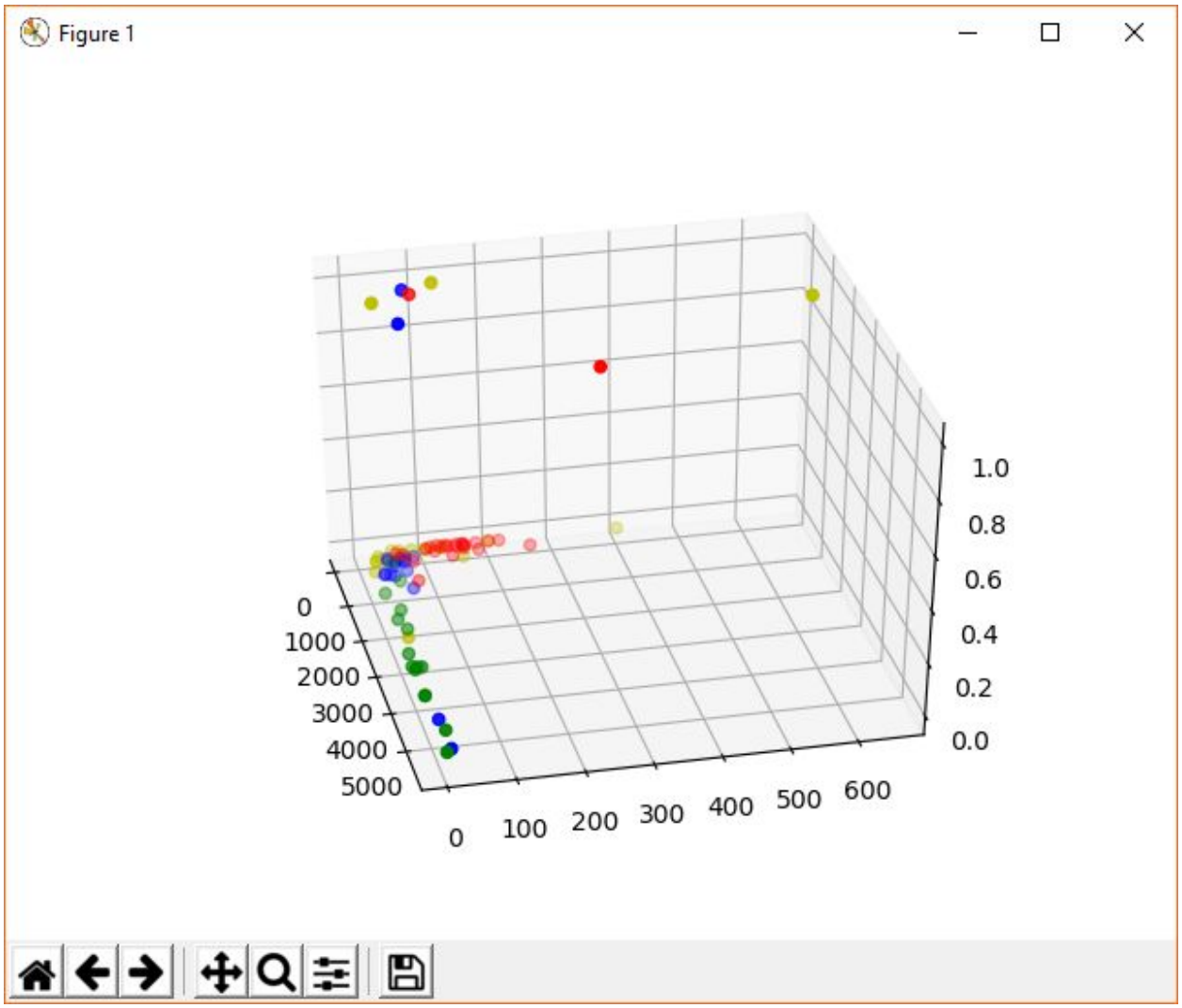
ML18-Z5-code10: Izveštaj

Pristup zadatku

Zadatak je urađen u Python-u uz biblioteke Pandas i Scikit-Learn.

Najpre smo plotovali podatke i primetili da je države veoma teško klasterovati pomoću ovako malog broja obeležja.



Slika 1. Podaci u 3D (income, infant, oil) - Afrika (crvena), Evropa (zelena), Amerike (plava), Azija (žuta).

Tačnost modela utvrđivali smo korišćenjem unakrsne validacije (*RepeatedStratifiedKFold)*, a parametri su optimizovani korišćenjem grid pretrage (*GridSearchCV*).

Isprobani algoritmi

Za klasterovanje koristili smo model Gausovih mešavina (*Gaussian Mixture*).
Za enkodovanje kategoričkih podataka iskoristili smo *label encoding*.
Od metoda za normalizaciju isprobali smo sledeće:

1. Min-max
2. Z-score
3. Bez normalizacije

Za nedostajuće vrednosti:

1. Dopunjavanje korišćenjem *mean* vrednosti
2. Dopunjavanje korišćenjem *median* vrednosti
3. Izbacivanje uzoraka sa nedostajućim vrednostima

Konačno rešenje

Za najbolje rešenje uzorci sa nedostajućim vrednostima su izbačeni (ukupno 4 uzorka) i korišćena je min-max normalizacija. Parametri modela su sledeći:

parametar	Vrednost
n_init	10
covariance_type	diag
init_params	kmeans
max_iter	100
tol	0.1
reg_covar	0.001

S obzirom na osetljivost modela na izbor početnih vrednosti, odlučili smo se da broj inicijalizacija (*n_init*) povećamo sa podrazumevanih 1 na 10.

Ostvareni rezultati

Konačan *V score* na test skupu iznosi 0.620497953934404.