Klasterovanje

Tim 6_23

- Anastasija Samčović, SW44/2019
- Strahinja Popović, SW51/2019
- Srđan Đurić, SW63/2019

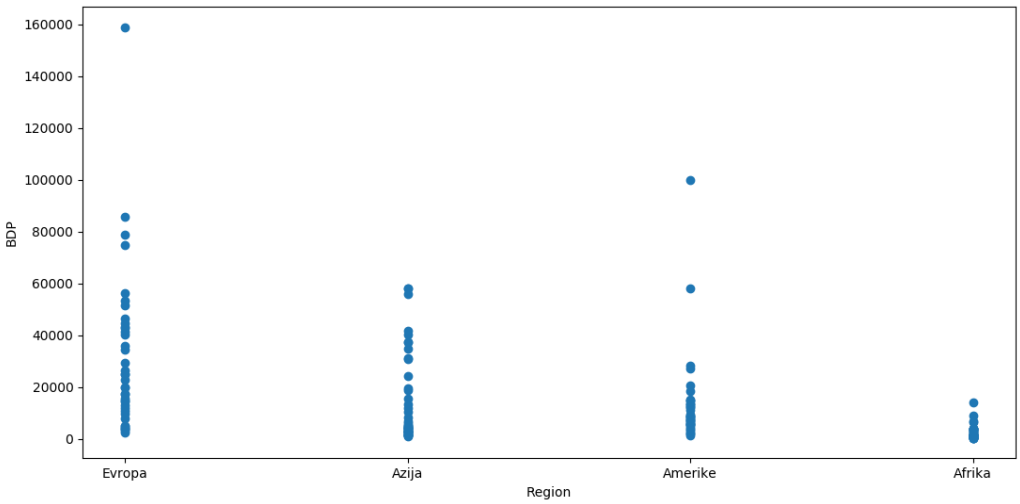
Obrada podataka

Prvi korak u rešavanju datog problema bila je vizualizacija i analiza datog trening skupa podataka po obeležjima. Na taj način smo odredili outliere. Takođe smo primetili postojanje numeričkih obeležja(BDP, Izvoz, Inflacija) i kategoričkih(Region, More, Religija).

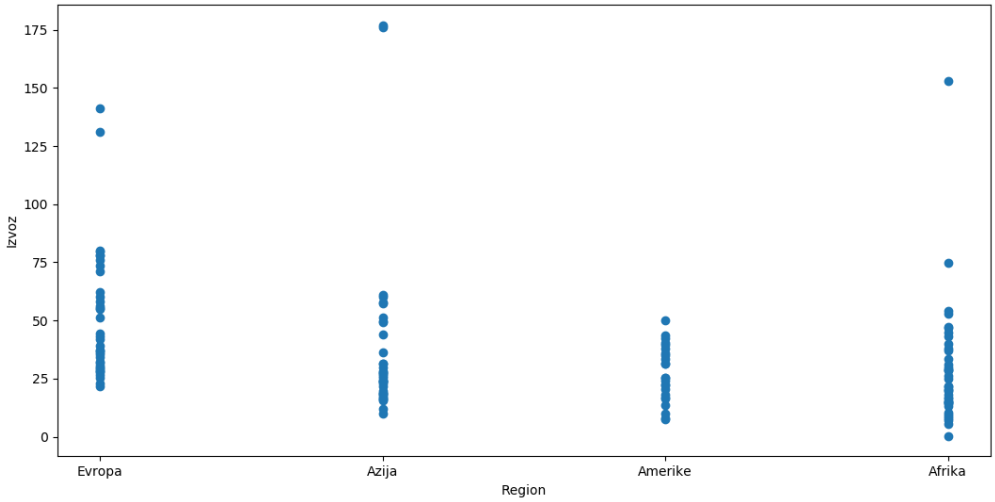
Za obeležje “BDP” smo kao outlier-e posmatrali sve vrednosti veće od 100.000(Slika br. 1), a za “Izvoz” vrednosti veće od 120(Slika br. 2). Kod obeležja izvoz smo prvobitno pokušali da uklonimo samo vrednosti sa vrednošću većom od 150, ali smo bolje rezultate dobili kada smo granicu spustili na 120. Za obeležje “Inflacija” smo zaključili da nam nije relevantno, nakon uklanjanja ovog obeležja dobili smo značajno poboljšanje rezultata. Uklonili smo i obeležje “Region”, jer je ono ciljna varijabla.

Budući da je skup podataka sadržao nedostajuće vrednosti, sve redove koji imaju više od 2 obeležja koja su nedostajuća smo odbacili. Ostatak praznih obeležja smo popunili koristeći “mean” strategiju za numerička i “most frequent” strategije za kategorička obeležja.

Za kategorička obeležja (Religija i More) smo koristili OneHotEncoding, a na numerička obeležja (BDP i Izvoz) smo primenili StandardScaler.



Slika br. 1 Prikaz odnosa BDP-a i Regiona



Slika br. 2 Prikaz odnosa Izvoza i Regiona

Gaussian Mixture Model

Kao što je traženo, primenili smo Gaussian Mixture Model. Koristili smo model iz biblioteke scikit-learn. Za evaluaciju rezultata je korišćena **v mera**.

Poziv:

GaussianMixture(n_components=4,random_state=0,n_init=5,covariance_type='tied')

Parametri:

- n_components = 4 – broj regiona (“Evropa”, “Amerike”, “Azija”, “Afrika”)
- covariance_type=’tied’

Probali smo različite vrednosti za parametar covariance_type i kao što se može videti iz tabele vrednost ‘tied’ je dala najbolji rezultat. U Tabela 1 su prikazani rezultati dobijeni korišćenjem polinoma prvog stepena. Kada smo koristili polinom drugog stepena, rezultati su bili za nijansu bolji, te smo se za krajnje rešenje odlučili da koristimo polinom drugog stepena.

Skup podataka korišćen za testiranje predstavlja 30% datog trening skupa, a skup podataka korišćen za treniranje modela predstavlja 70% datog trening skupa.

covariance_type	rezultat (70/30)
full	0.2943
tied	**0.3267**
diag	0.2943
spherical	0.3162

Tabela 1 Prikaz rezultata