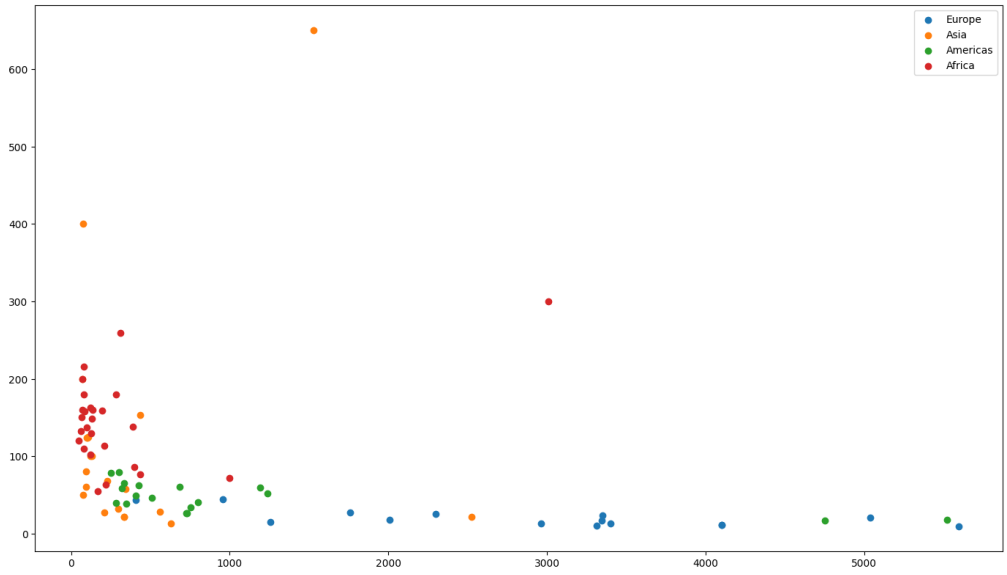
Klasterovanje

Tim Techies

- SW 38/2018, Marko Njegomir
- SW 43/2018, Dušan Erdeljan

Analiza podataka

U trening setu se nalazi 84 podataka. Postoje 4 podatka kojima fali vrednost u infant koloni. Samo 8 zemalja ima naftu, dok ostalih 76 nema naftu, tako da postoji velika nebalansiranost ovog obeležja.



Ilustracija 1 Raspodela klasa u dve dimenzije, gde je na x osi income, a na y osi je infant

Kada se podaci vizualizuju, može se primetiti da postoji određen broj zemalja koje odudaraju od ostalih zemalja sa istog kontinenta. Takođe postoji i preklapanja između nekih klasa, pogotovo kod vrednosti koje su bliže nuli. Može se uočiti da zemlje u Evropi imaju generalno nisku infant vrednost a visoke income vrednosti. Zemlje u Africi generalno imaju imaju više infant vrednosti, a stabilno niske income vrednosti. U proseku su zemlje u Americi bogatije od zemalja u Aziji, i imaju takođe i niže infant vrednosti.

region	income	infant
Africa	307.307692	144.950000
Americas	1087.611111	47.216667
Asia	400.380952	117.090476
Europe	2926.666667	20.260000

Ilustracija 2 Prosečne vrednosti obeležja po kontinentu

Isprobani algoritmi I ostvareni rezultati

Rad sa datasetom

- Model smo trenirali na celom trening skupu podataka
- Izbacili smo sledeće kolone
  - Oil
- Proširili smo dimenzije tako što smo dodavali nelinearne kombinacije kolona
  - Za prvi pokušaj smo dodali sledeće kolone
    - Infant X infant
    - Infant X income
    - Income X income
  - Za drugi pokušaj smo dodali
    - Infant X income
- U drugom pokušaju smo pokušali ručno da otklonimo mnoštvo autlajera tako da napravimo veće I jasnije granice između grupa tačaka koje pripadaju različitim klasama u nadi da će to pomoći algoritmu da bolje uoči grupe.

Gaussian Mixture pristup

- Koristili smo GaussianMixture model sa brojem komponenti 4, I tied tipom kovarijanse I random inicijalnim stanjem.

Ostvareni rezultati nakon evaluacije na test skupu

GaussianMixture je ostvario sledeće rezultate:

- V-measure train: 0.8430025728674821
- V-measure test: 0.814321987395574