Klasterovanje

Izveštaj

tim7_24
Nikolina Trivunić SW/64-2019

Zadatak

Klasterovati države na osnovu njihovih geografskih karakteristika u klastere koji predstavljaju geografske regione (kolona region): europe, asia, africa, americas. Zadatak je uspešno urađen ukoliko se na kompletnom testnom skupu podataka dobije v mera (eng. v measure score) veća od 0.13. Zadatak se rešava upotrebom Modela Gausovih mešavina, tj. algoritmom Očekivanje-maksimizacija.

Acceptance criteria: > 0.13

Priprema podataka

Najpre sam pregledala sirove podatke kako bih identifikovala eventualne nedostajuće vrednosti, izuzetke ili nepravilnosti.

Nakon prvog pregleda, otkrila sam nedostajuće vrednosti u skupu podataka, koje je bilo potrebno obraditi kako bih osigurala tačnost analize. Da bih rešila ovaj problem, koristila sam KNNImputer, metod koji zavisi od susednih tačaka kako bi popunio nedostajuće vrednosti. Ovo je omogućilo da sačuvam što više informacija iz podataka, umesto da ih jednostavno izbacim.

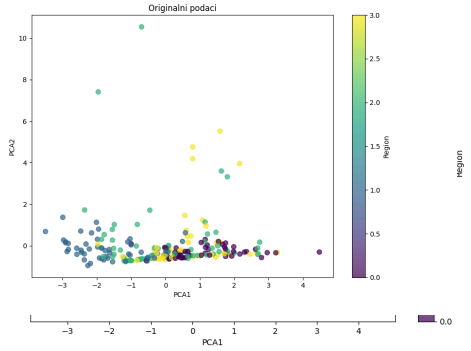
Nakon što sam popunila nedostajuće vrednosti, provela sam normalizaciju podataka koristeći StandardScaler. Ovaj korak je bio važan kako bih osigurala da su sve karakteristike podataka na istoj skali, čime se postiže bolja performansa modela i interpretacija rezultata.

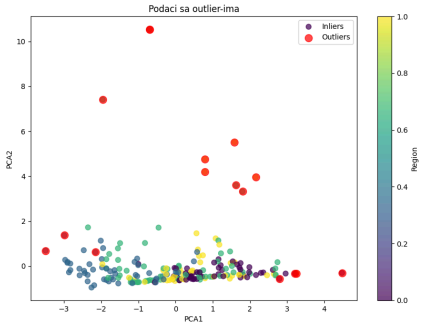
Sledeći korak bio je identifikacija i uklanjanje ekstremnih vrednosti (outliers) u podacima. Primenila sam nekoliko različitih tehnika detekcije izuzetaka, uključujući Local Outlier Factor, DBSCAN i Isolation Forest algoritme. Nakon upoređivanja rezultata, odlučila sam se za Z-score metodu jer je dala najbolje performanse.

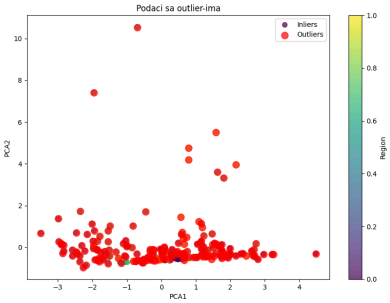
Rezultati sa Local Outlier Factor metodom:
V Measure Score na validacionom skupu: 0.2279391910885644
V Measure Score na test skupu: 0.20493345594289064

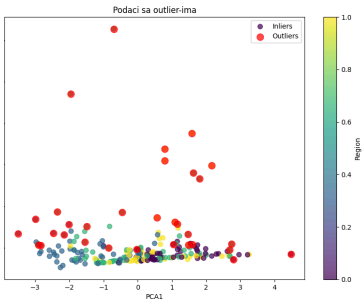
Rezultati sa Isolation Forest metodom:
V Measure Score na validacionom skupu: 0.2720761684522902
V Measure Score na test skupu: 0.14758492656410999

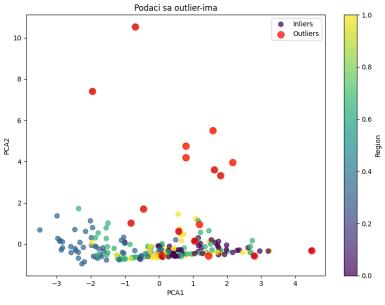
Rezultati sa Z-score metodom:
V Measure Score na validacionom skupu: 0.41394133153366763
V Measure Score na test skupu: 0.5167361522463597

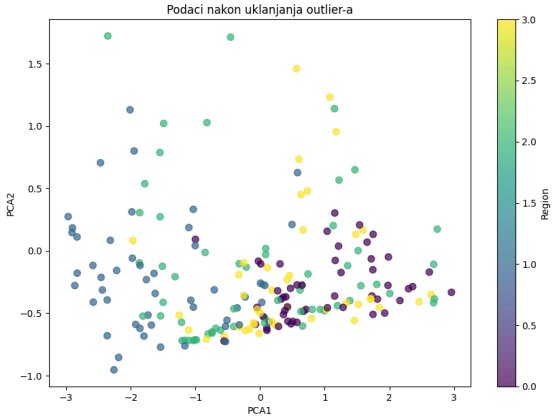












Smanjenje dimenzionalnosti

Nakon pripreme podataka, odlučila sam da smanjim dimenzionalnost podataka kako bih olakšala dalju analizu. Korišćenjem PCA (Principal Component Analysis) algoritma, redukovala sam dimenzije podataka na četiri glavne komponente (PCs). Ovo mi je omogućilo da zadržim najvažnije informacije iz podataka dok istovremeno smanjujem složenost modela.

Modeliranje

Koristila sam GaussianMixture model, dok za optimizaciju hiperparametara zadržala sam se na GridSearchCV kako su rezultati za RandomizedSearchCV bili 0.41745832980843395 na validacionom skupu i 0.36141734339731535 na test podacima. GridSearchCV je na test skupu davao 0.5300105449030372.

Pokušala sam da mapiram podatke kako bih olakšala tumačenje rezultata klasifikacije. Nakon primene mapiranja, V Measure Score na test skupu iznosio je 0.5167361522463597.

```
correlation_matrix = data_cleaned_numeric.corr().abs()

threshold = 0.8
upper_tri = correlation_matrix.where(
    np.triu(np.ones(correlation_matrix.shape), k=1).astype(np.bool_))

to_drop = [column for column in upper_tri.columns if any(upper_tri[column] > threshold)]
data_cleaned.drop(to_drop, axis=1, inplace=True)
```

Analizirala sam korelaciju između atributa u cilju eliminacije visoko koreliranih atributa korišćenjem koda koji sam priložila, postavljen je prag od 0.8 za detekciju visoke korelacije. Međutim, eliminacija koreliranih atributa nije znatno uticala na performanse modela. Ovo je ukazalo na to da visoka korelacija među atributima nije bila ključni faktor u modeliranju klasifikacije regiona.

Zaključak

Nakon detaljne analize podataka i eksperimenata sa različitim tehnikama, zaključujem da je model baziran na GaussianMixture algoritmu uz korišćenje Z-score metode za detekciju outlier-a i GridSearchCV za optimizaciju hiperparametara dao najbolje rezultate u klasifikaciji regiona na osnovu dostupnih atributa. Međutim, moguće je dalje unapređenje modela kroz eksperimentisanje sa drugim metodama i dodatnom obradom podataka.