Извештај за задатак 5 – кластеровање

Даница Газдић
SV 12/2020

Милош Обрадовић
SV 55/2020

1. Проблем

У овом извештају се решава проблем кластеровања употребом *GaussianMixture* модела.

2. Решење

Недостајуће вредности су израчунате употребом *KNNImputer-a*. Избачене су колоне *region*, *Year* (не носи никакву информацију) и *Literacy Rate* (има јако пуно недостајућих вредности и избацивање је дало боље резултате). Трениран је модел *GaussianMixture* са хиперпараметрима који су оптимизовани употребом *GridSearchCV.*

```
gmm = GaussianMixture()
param_grid = {
    'random_state': [0, 1, 2, 3, 7, 20, 42],
    'init_params': ['kmeans', 'random', 'k-means++', 'random_from_data'],
    'covariance_type': ['full', 'tied', 'diag', 'spherical'],
    'max_iter': [100, 200, 300, 400, 500],
}

grid_search = GridSearchCV(gmm, param_grid, cv=5, scoring='v_measure_score', n_jobs=-1)
grid_search.fit(X_train, y_train)
```

[1] Прво решење је укључивало *Literacy Rate* колону, као и хиперпараметре:
- random_state = 7
- init_params = random
- covariance_type = diag
- max_iter = 100 (default).

Ово решење је дало резултат v_measure_score = 0.21 на локалу и 0.12 на платформи.

[2] Финално решење (0.37 на локалу и 0.35 на платформи) смо добили са избаченом колоном *Literacy Rate* и са хиперпараметрима:
- random_state = 2
- init_params = k-means++
- covariance_type = diag
- max_iter = 100 (default)