Redukcija Dimenzionalnosti

Tim [sigma]:

- Marko Vukotić SW-71/2018
- Miloš Popović SW-24/2018

Zadatak:

Na osnovu dostupnih informacija o zaposlenima na istočnoj obali SAD treba izvšiti predikciju njihove rase.

Analiza podataka:

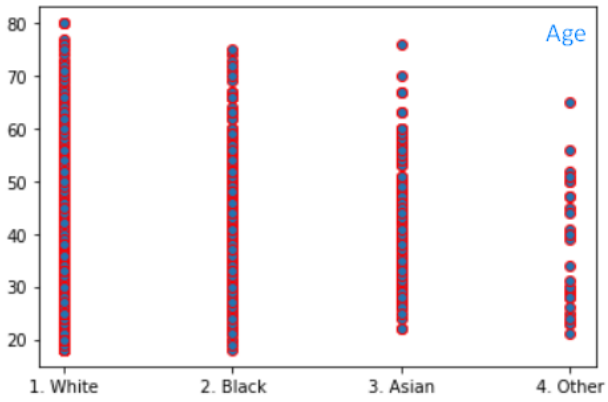
Atributi koji su dati:

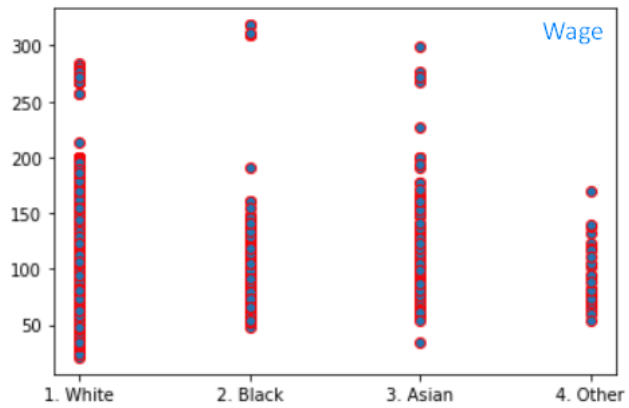
- year - godina prikupljanja informacije
- age - starost zaposlenih
- maritl - bračni status zaposlenih (nikad venčan, venčan, udovica-udovac, razveden, rastavljen)
- education - nivo obrazovanja (bez srednje, srednje, bez fakulteta, fakultet, doktorat)
- jobclass - tip posla (industrijski i informacioni)
- health - zdravstveno stanje (dobro i veoma dobro)
- health_ins - da li zaposlen ima zdravstveno osiguranje (da i ne)
- wage - godisnja plata u hiljadama dolara

Pretprocesiranje podataka:

Primetili smo prazna polja u datim podacima i uklanjali smo te kolone. Takođe kao i u prošlim zadacima je urađen *One Hot Encoding* za polja maritl, education, jobclass, health i health_ins.

Outliere nismo uklanjali jer nismo primetili značajna odstupanja u datim grafovima (year i wage).





Rešenje

Koristili smo PCA algoritam za redukovanje dimenzionalnosti. Takođe smo i DecisionTree koristili a i probali Random Forest i ADAboost

Rezultati

Kao meru smo koristili f1 macro.

DecisionTree:

Rezultat nad validacionim za decision RandomForestClassifier je bio 0.2583189608684731.

Na platformi nakon izvršavanja rezultat je bio 0.34318895108834.

Parametri:
PCA - n_components=11
DecisionTreeClassifier - max_depth=100

Random Forest:

Rezultat nad validacionom: 0.2546918351989355