PCA + klasifikacija

Tim – masinski_ucenjaci:

- SW03/2016 – Nikola Zubić
- SW11/2016 – Filip Mladenović
- SW53/2016 – Mihajlo Kušljić

Zadatak

Na osnovu dostupnih informacija o zaposlenima na istočnoj obali SAD, koristeći *PCA* algoritam, izvršiti predikciju njihove rase (***race***):

- ***1.White***
- ***2.Black***
- ***3.Asian***
- ***4.Other***

Zadatak je uspešno urađen ukoliko se na kompletnom testnom skupu podataka dobije mikro f1 mera (eng. *micro f1 score*) > 0,79.

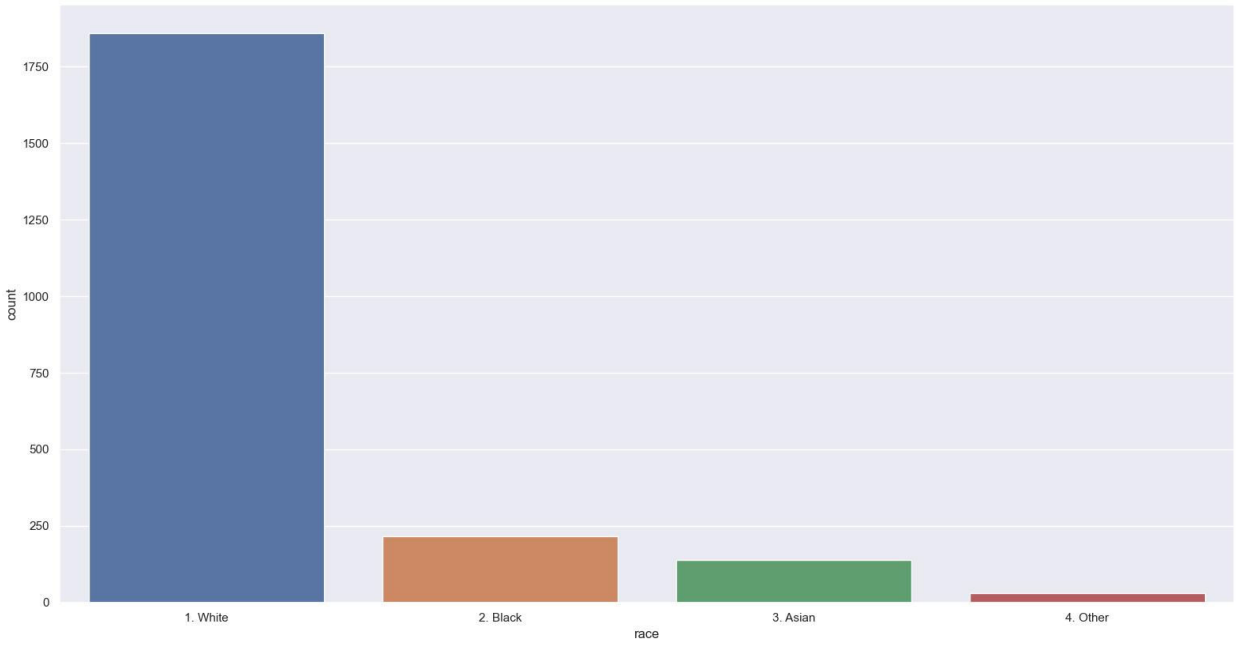
Analiza podataka

Skup podataka sadži opis zaposlenih na istočnoj obali SAD kroz sledeća obeležja:

- ***year*** – godina kada su prikupljane informacije
- ***age*** – broj godina (starost) zaposlenih
- ***maritl*** - bračni status zaposlenih:
  - ***1. Never Married*** - nikad venčani
  - ***2. Married*** – venčani
  - ***3. Widowed*** - udovice/udovci
  - ***4. Divorced*** – razvedeni
  - ***5. Separated*** – rastavljeni
- ***education*** - nivo obrazovanja:
  - ***1. < HS Grad*** - nezavršena srednja škola
  - ***2. HS Grad*** - završena srednja škola
  - ***3. Some College*** - nezavršen fakultet
  - ***4. College Grad*** - završen fakultet
  - ***5. Advanced Degree*** - MSc, PhD
- ***jobclass*** - tip posla:
  - ***1. Industrial*** – industrijski
  - ***2. Information*** – informacioni
- ***health*** - zdravstveno stanje
  - ***1. <= Good*** - dobro ili slabije
  - ***2. >= Very Good*** - veoma dobro ili odlično

- ***health_ins*** - da li zaposleni poseduje zdravstveno osiguranje:
  - ***1. Yes***
  - ***2. No***
- ***wage*** **-** godišnja plata (u hiljadama dolara).

Na osnovu datih obeležja potrebno je odrediti ciljnu labelu – rasu zaposlenog (***race***). Obučavajući skup sadrži 2250 primera. Uočeno je da je skup loše balansiran s obzirom na ciljnu labelu – dato je mnogo više primera za belu rasu u odnosu na ostale (Slika 1):

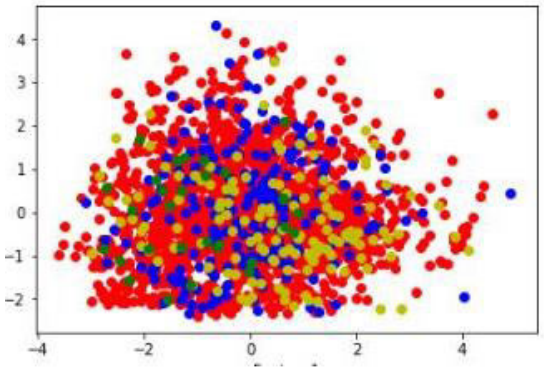


Slika 1 – Broj primera u obučavajućem skupu po rasama

Pretprocesiranje podataka

Za kodiranje kategoričkih obeležja isproban je *Label Encoder* i *One hot encoding*, pri čemu je utvrđeno da *Label Encoder* daje bolje rezultate. Uočeno je da se u obučavajućem skupu za svako obeležje pojavljuju nedostajuće vrednosti. Zato primeri sa nedostajućim vrednostima nisu odbačeni, već su nedostajuće vrednosti zamenjene srednjom vrednošću na obučavajućem skupu za odgovarajuće obeležje.

Preduslov za korišćenje *PCA* algoritma jeste da su podaci centrirani. Stoga su vrednosti obeležja normalizovane korišćenjem *Standard Scaler*-a, koji vrši centriranje podataka. Kako bismo, dodatno, ispitali podatke, iskoristili smo *PCA* da nam konvertuje problem u dve dimenzije. To nam je omogućilo da vizualizujemo skup (Slika 2).



Slika 2 – Vizualizacija obučavajućeg skupa u dve dimenzije

Kako bi se ublažila loša balansiranost obučavajućeg skupa isproban je pristup sa kloniranjem primera za manje zastupljene rase (eng. *upsampling*) kao i ignorisanjem nekih primera bele rase (eng. *downsampling*), pri čemu je bolje rezultate dao *upsampling* pristup.

Isprobani algoritmi

Za validaciju performansi različitih algoritama korišćen je *RepeatedStratifiedKFold* metod sa parametrima: *n_splits* = 10, *n_repeats* = 2 i *random_state* = 36851234. Za redukciju dimenzionalnosti isprobani su algoritmi *PCA* i *Kernel PCA*. Za klasifikaciju isprobani su algoritmi *AdaBoost*, *Bagging*, *Extra-trees*, *Gradient Boosting*, *Random Forest*, *Voting Classifier* sa više različitih modela (*Ada Boost*, *Gradient Boosting*, *Decision Tree* i *Gaussian Naive Bayes*). Takođe je implementiran *Stacking Classifier* koji koristi više modela ansambla. Isproban je *upsampling* primera manjinskih rasa, takođe i *downsampling* primera u obučavajućem skupu čija je klasa bela. Takođe je isproban i *SMOTE* algoritam koji ima za cilj da sintetički generiše nove podatke, međutim on se nije pokazao kao dobar pristup (dobijena je veoma loša mikro f1 mera). Probali smo obučiti i tri zasebna klasifikatora kod kojih je jedan učen da prepoznaje samo trening primere kod kojih *race* ima vrednost *White*, drugi koji prepoznaje samo trening primere kod kojih *race* ima vrednost *Black* i treći kod kojih je vrednost *race Asian* uz *upsampling,* te smo na kraju radili *Count Voting* između ta tri klasifikatora, međutim ovaj pristup se pokazao kao i većina drugih i nije odabran kao najbolji.

Odabrani model

Pošto je obučavajući skup neizbalansiran, povećan je broj primera za klase ***Black*** i ***Asian*** koristeći *upsampling* gde su parametri birani koristeći *gridSearchCV*. Za redukciju dimenzionalnosti korišćen je *Kernel PCA* koji se pokazao boljim od običnog *PCA* jer predstavlja proširenje *PCA* algoritma koje koristi tehnike kernel metoda gde se linearne operacije mogu izvoditi u kernelovskim Hilbertovim prostorima, pa samim tim možemo ih transformisati u više dimenzija što nam daje više informacija. Korišćen je *poly kernel* sa stepenom 14 i brojem komponenti 5. Za klasifikator odabran je *RandomForestClassifier* sa 5 estimatora i maksimalnom dubinom 6. Dobijena mikro f1 mera na testnom skupu na platformi iznosi: **0.825333333333333.**