Dokumentacija o izradi projekta iz predmeta

Softverski alati u sistemima automatskog upravljanja

Adult Census Income

Student:

Minja Drakul RA58/2022

Mentor:

Danilo Kaćanski

Pregled etapa u izradi projekta:

1. Učitavanje podataka i vizualizacija
2. Početno preprocesiranje podataka
3. Eksplorativna analiza skupa
4. Odabir i treniranje modela
5. Podešavanje hiperparametara kreiranog modela
6. Analiza rezultata predikcije
7. Odabir najbitnijih atributa

1. Učitavanje podataka i vizualizacija

Dataset sam dobila u oblika dva .csv fajla:

adult_train.csv

adult_test.csv

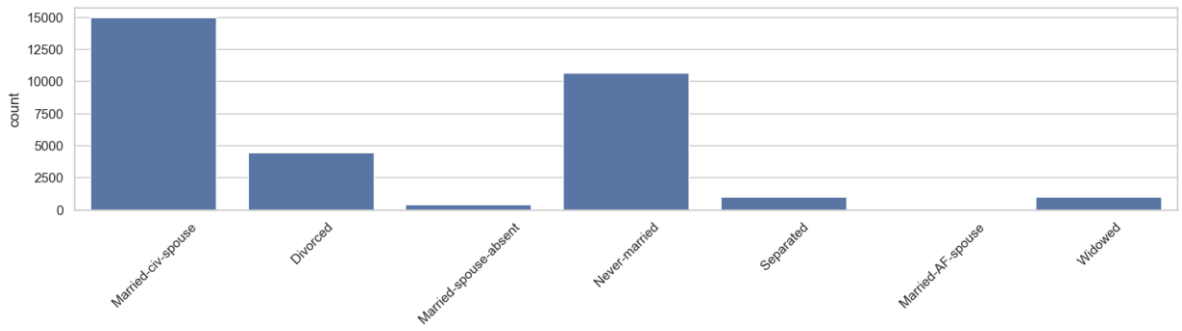
Učitavam ih pomoću *pandas* biblioteke.

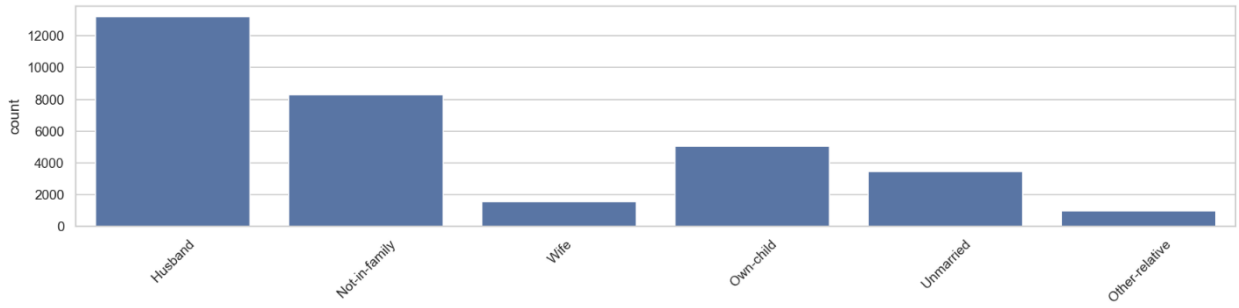
Dataset (*adult_train.csv*) obuhvata sledeće kolone, svaka sa specifičnim značenjem:

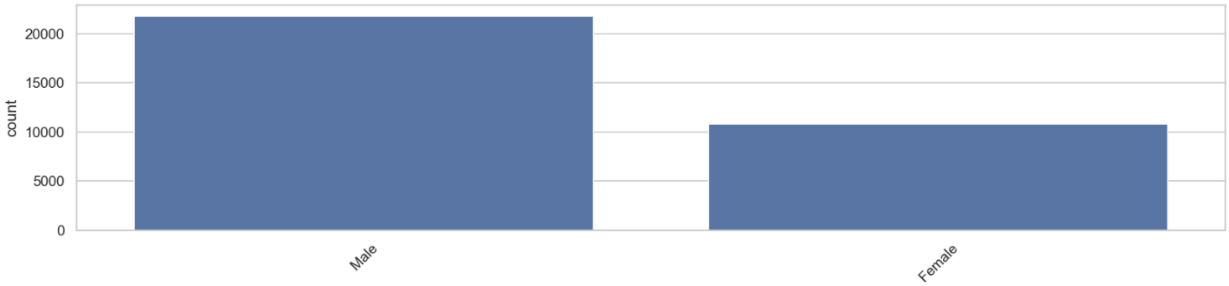
- **age** — starost osobe u godinama.
- **workclass** — tip zaposlenja (npr. privatni sektor, državna služba, samozaposlen).
- **fnlwgt** — finalna težina slučaja koja predstavlja koliko osoba u populaciji odgovara ovom zapisu (statistički ponder).
- **education** — najviši stepen obrazovanja (npr. srednja škola, fakultet).
- **education-num** — numerička vrednost koja odgovara nivou obrazovanja.
- **marital-status** — bračni status (npr. u braku, razveden, nikada u braku).
- **occupation** — zanimanje (npr. tehnička podrška, upravljačke pozicije, zanatske usluge).
- **relationship** — porodična uloga u domaćinstvu (npr. supružnik, dete, ne-porodični član).
- **race** — rasa (npr. White, Black, Asian-Pac-Islander).
- **sex** — pol osobe (muški, ženski).
- **capital-gain** — prihod od kapitalne dobiti (u dolarima).
- **capital-loss** — gubitak kapitala (u dolarima).
- **hours-per-week** — broj radnih sati nedeljno.
- **native-country** — država rođenja.

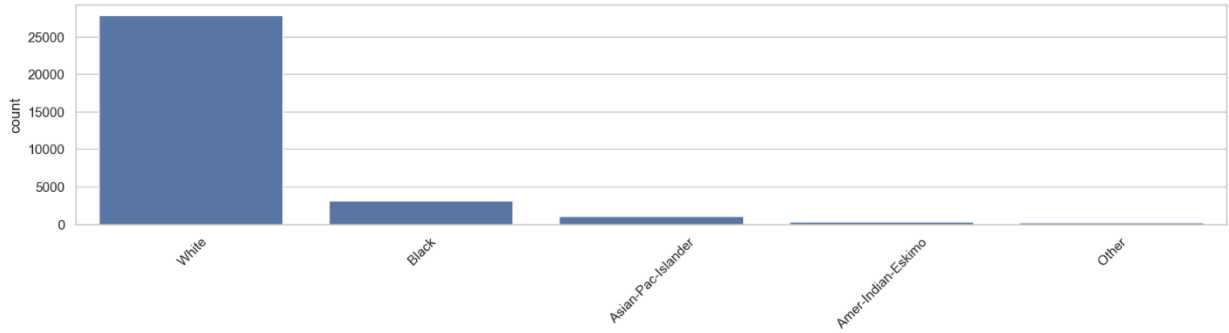
income — **ciljna promenljiva**: <=50K ili >50K.

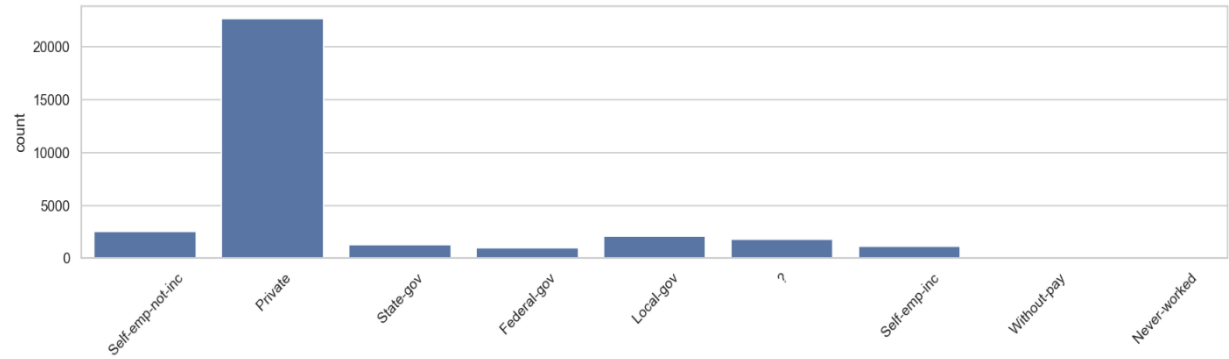
U ovoj fazi još i vizualizujem dataset, i uočavam neke zakonitosti i relacije između kategorija i ciljne promenjive.



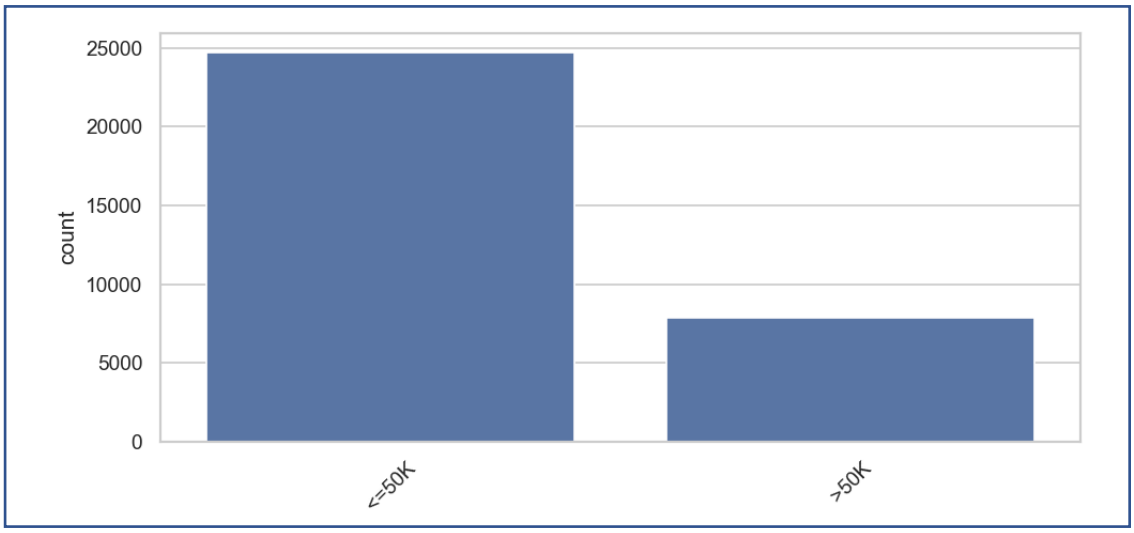








Iz gore prikazanih dijagrama možemo zaključiti da je dataset jako neizbalansiran po mnogim parametrima. Na primer, dvostruko više ispitanika su muškog pola, najviše ljudi je u braku, pa kao posledicu ovoga imamo da je najzastupljenija porodična uloga – otac.



Jedno od najbitniji zapažanja je odnos izlaznih promenjivih. Vidimo da se izlazne promenjive odnose u razmeri (približno) 70-30%.

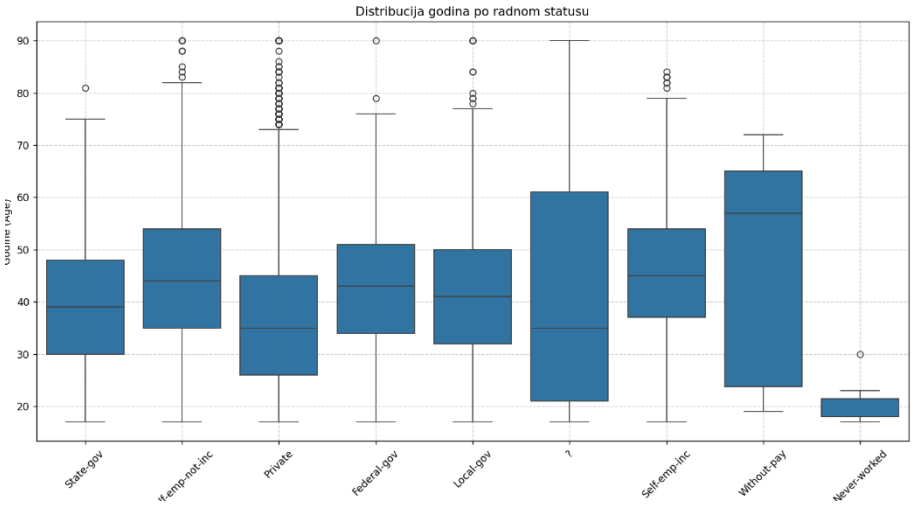
Što nam govori da će model, ako podatke dobro ne procesuiramo, overfitovati, i jako dobro raditi sa kategorijom <=50K, a da će imati problema da pronađe manjinsku klasu >50K.

2. Početno preprocesiranje podataka

Provera postojanja duplikata i uklanjanje

```
Postoje li duplikati? True
Broj redova pre uklanjanja duplikata: 32561
Broj duplikata: 24
Broj redova posle uklanjanja duplikata: 32537
```

Provera nedostajućih vrednosti



Uočila sam nedostajuće vrednosti i među kategorijski i među numeričkim kolonama.

Kako sam ih obradila:

```
# popunjavanje nedostajućih vrednosti kategoričkih kolona
for col in cat_cols:
    most_frequent = X_train[col].mode()[0]
    X_train[col].fillna(most_frequent, inplace=True) #popunjavam sa klasom koja se najvise puta pojavljuje
    X_test[col].fillna(most_frequent, inplace=True)
```

```
#popunjavanje numerickih kolona
for col in num_cols:
    imputer = SimpleImputer(strategy='mean') #popunjavam sa srednjom vrednosti
    X_train[col] = imputer.fit_transform(X_train[[col]])
    X_test[col] = imputer.transform(X_test[[col]])
```

Enkodiranje podataka

Za kodiranje kategorijskih kolona koristila sam **OneHotEncoder,** sa dodatim atributom da ako naiđe na nepoznatu vrednost – ignoriše, odnosno kodira kao 0.

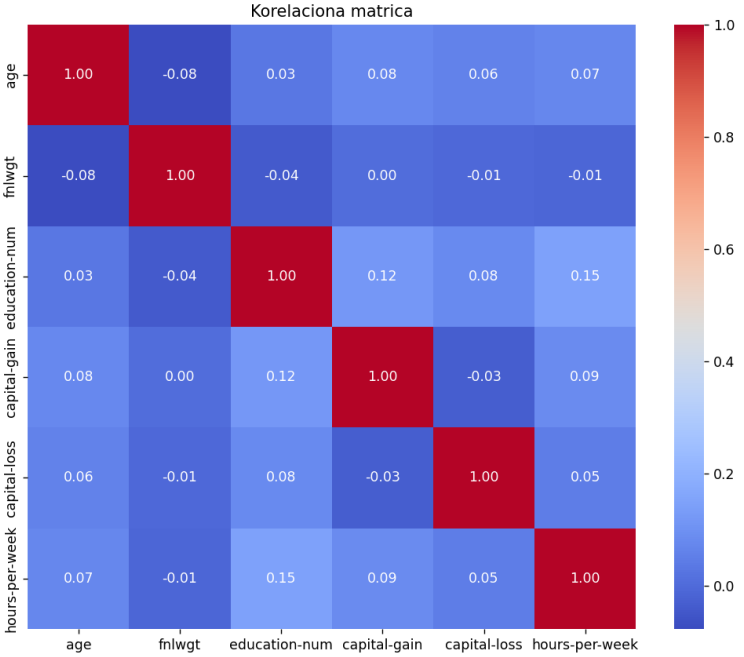
Za kodiranje izlazne promenjive koristila sam **LabelEncoder,** tako da ćemo za enkodovani izlaz imati vrednosti 0 i 1.

Uklanjanje atributa koji ne utiču na izlaz

```
# isključujemo capital-gain i capital-loss zbog prirode podataka
outlier_cols = [col for col in num_cols if col not in ['capital-gain', 'capital-loss']]
```

Isključujem ih iz razloga što skoro 90% ljudi ima capital-gain i capital-loss = 0

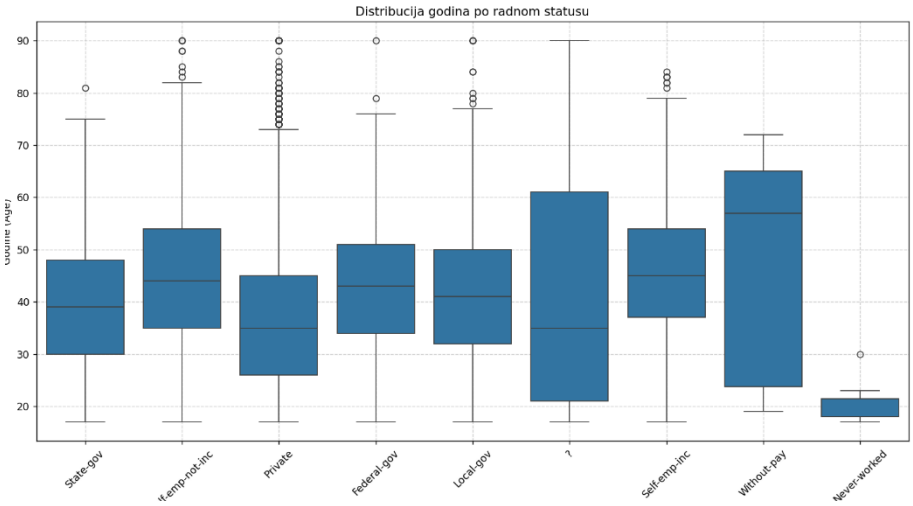
3. Eksplorativna analiza skupa



Gledajući matricu korelacije može se uočiti da najveću korelaciju imaju:

- broj godina obrazovanja sa brojem radnih sati i sa kapitalnom dobiti
- Kapitalna dobit sa godinama i brojem radnih sati

<u>Anomalije i ekstremne vrednosti</u>



Na grafiku možemo uočiti da većina kategorija ima ekstremne vrednosti koje možemo smatrati anomalijama.

Najuočljivija je pojava da u privatnom sektoru rade ljudi koji imaju i do 90 godina.

Dok čak 50% ljudi iz tog sektora ima između 25-45 godina, tako da sve vrednosti preko 73 godine (gornja granica) smatram anomalijama.

Ovdje radim clipping i sve vrednosti preko Q3 lepim na tu gornju granicu, a sve vrednosti ispod Q1 podižem na tu donju granicu. Mada vrednosti ispod ni nemamo.

4. Odabir i treniranje modela

```
--- Rezultati za RandomForest model ---
Tačnost: 0.8220
F1-Score: 0.8173
Preciznost: 0.6485
Model je treniran.
```

```
--- Rezultati za XGBoost model ---
Tačnost: 0.7971
F1-Score: 0.8084
Preciznost: 0.5517
```

```
--- Rezultati za LogisticRegression model ---
Tačnost: 0.8382
F1-Score: 0.8324
Preciznost: 0.6975
Model je treniran.
```

```
--- Rezultati za KNeighbors model ---
Tačnost: 0.8197
F1-Score: 0.8181
Preciznost: 0.6301
Model je treniran.
```

```
--- Rezultati za LightGBM model ---
Tačnost: 0.7976
F1-Score: 0.8096
Preciznost: 0.5505
```

Modele koje biram inicijalno, i za koje ću kasnije izvršiti modifikacije po raznim parametrika da bih dobila željenje metrike, su **LogisticRegression** i **LightGBM** model

5. Podešavanje hiperparametara kreiranog modela

Za optimizaciju hiperparametara koristim RandomizedSearch

```python
rand_search = RandomizedSearchCV(
    estimator=lgbm,
    param_distributions=param_distributions,
    n_iter=100, # Broj nasumičnih kombinacija koje treba isprobati
    cv=3, #cross-validation
    scoring='f1',
    verbose=1,
    # Ovo sprečava grešku kada se kombinuje l1 i lbfgs
    error_score='raise'
)
```

6. Analiza rezultata predikcije

```
--- Rezultati za LogisticRegression model ---
Tačnost: 0.8524
F1-Score: 0.8455
Preciznost: 0.7484
Odziv: 0.5779
```

```
--- Rezultati za LogisticRegression model ---
Tačnost: 0.8055
F1-Score: 0.8168
Preciznost: 0.5622
Odziv: 0.8480
```

Jako dobar odziv

Promjena `class_weight='balanced'`

```
--- Rezultati za LogisticRegression model ---
Tačnost: 0.8551
F1-Score: 0.8506
Preciznost: 0.7344
Odziv: 0.6184
```

Nakon opt. hp ali bez class_weight='balanced'

```
--- Rezultati za LightGBM model ---
Tačnost: 0.8322
F1-Score: 0.8409
Preciznost: 0.6052
Odziv: 0.8608
================================================
```

Sa class_weight = 'balanced'

```
--- Rezultati za LightGBM model ---
Tačnost: 0.8705
F1-Score: 0.8667
Preciznost: 0.7714
Odziv: 0.6526
================================================
```

Bez class_weight = 'balanced'

```
--- Rezultati za LightGBM model ---
Tačnost: 0.8617
F1-Score: 0.8493
Preciznost: 0.8420
Odziv: 0.5200
```

Sa opt. hp i

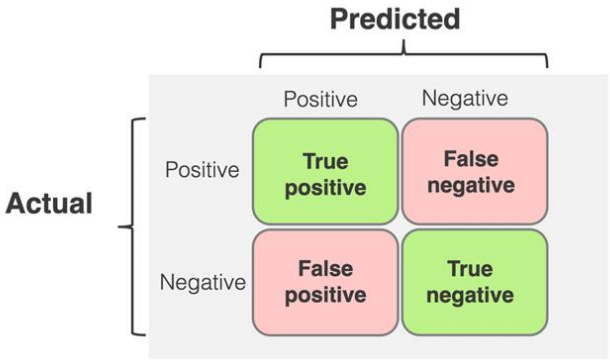
`scoring='precision',`

```
--- Rezultati za LightGBM model ---
Tačnost: 0.8334
F1-Score: 0.8416
Preciznost: 0.6087
Odziv: 0.8511
================================================
```

Sa opt. hp i

`scoring='f1'`

7. Odabir najbitnijih atributa



- **Odziv (Recall)** - Koliki deo stvarno pozitivnih primera je model uspešno identifikovao.

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1-score** - Harmonijska sredina preciznosti i odziva. Korisna je kada postoji neuravnoteženost klasa i potrebna je ravnoteža između preciznosti i odziva.

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

KONTEKST U KOM KORISTIM MODEL

Imajući u vidu da ne postoji najbolja metrika, nego samo najbolja za konkretan problem. Postavila sam model u konkretnu upotrebu.

Recimo da **marketinška agencija** koja prodaje luksuzne proizvode, koristi ovaj model. Agencija želi da cilja potencijalne klijente sa visokim prihodima (>50K) kako bi povećala prodaju.

U tom kontekstu bi bio najbitniji **odziv**, da od svih ljudi koji stvarno zaradjuju >50K, model pronadje što vise takvih.

Jer ako model oznaci da neko zaradjuje >50K, a NE zaradjuje, agencija troši vreme i resurse (novac za oglase, e-mail kampanje, itd.) na klijenta koji nikada neće kupiti luksuzni proizvod.

Ako model predviti da klijent NE zaradjuje vise od 50K, a u realnosti zaradjuje tu sumu, agencija propušta priliku da cilja klijenta koji je spreman da kupi proizvod. Ovo je **propuštena prodaja**. Trošak po grešci je vrlo visok, jer se gubi potencijalni profit.

```
--- Izveštaj o performansama modela: LGBMClassifier ---

Ukupna tačnost (Accuracy): 0.8333
Ukupna preciznost (Precision - weighted): 0.6079
Ukupan odziv (Recall - weighted): 0.8557
Ukupan F1-Score (weighted): 0.7108

Detaljan klasifikacioni izveštaj:
              precision    recall  f1-score   support

       <=50K       0.95      0.83      0.88     11138
        >50K       0.61      0.86      0.71      3506

    accuracy                           0.83     14644
   macro avg       0.78      0.84      0.80     14644
weighted avg       0.87      0.83      0.84     14644
```

```
Matrica konfuzije:
[[9170 1968]
 [ 488 3018]]
================================================

--- Najvažnije obeležje po kategoriji za model 'LGBMClassifier' ---
Najvažnije iz kategorije 'education-num': education-num (Vrednost: 416.0000)
Najvažnije iz kategorije 'capital-gain': capital-gain (Vrednost: 388.0000)
Najvažnije iz kategorije 'capital-loss': capital-loss (Vrednost: 327.0000)
Najvažnije iz kategorije 'hours-per-week': hours-per-week (Vrednost: 499.0000)
Najvažnije iz kategorije 'age': age_23 (Vrednost: 41.0000)
Najvažnije iz kategorije 'workclass': workclass_Private (Vrednost: 89.0000)
Najvažnije iz kategorije 'education': education_Bachelors (Vrednost: 57.0000)
Najvažnije iz kategorije 'marital-status': marital-status_Married-civ-spouse (Vrednost: 105.0000)
Najvažnije iz kategorije 'occupation': occupation_Prof-specialty (Vrednost: 92.0000)
Najvažnije iz kategorije 'relationship': relationship_Not-in-family (Vrednost: 66.0000)
Najvažnije iz kategorije 'race': race_White (Vrednost: 48.0000)
Najvažnije iz kategorije 'sex': sex_Female (Vrednost: 86.0000)
Najvažnije iz kategorije 'native-country': native-country_United-States (Vrednost: 43.0000)
```