Contents

Tehnička Dokumentacija i Analiza: Prediktivno Modeliranje Nivoa Prihoda

Ovaj dokument opisuje razvoj i evaluaciju modela mašinskog učenja za klasifikaciju nivoa prihoda korišćenjem "Adult" skupa podataka. Proces je obuhvatio eksplorativnu analizu podataka, višefazno predprocesiranje, selekciju najvažnijih atributa i uporedno testiranje tri algoritma: Logističke Regresije, Random Forest-a i Gradient Boosting-a. Kao optimalni model identifikovan je **Gradient Boosting Classifier**. Analiza je potvrdila da su demografski i socio-ekonomski faktori, prije svega bračni status i nivo obrazovanja, ključni prediktori visine prihoda.

1. Uvod

1.1. Opis i Značaj Problema

Ovaj projekat se bavi razvojem i evaluacijom modela mašinskog učenja sa ciljem predikcije nivoa prihoda pojedinca, klasifikujući ga u jednu od dve kategorije: prihod do 50.000 dolara godišnje (≤50K) ili iznad tog iznosa (>50K). Sposobnost preciznog predviđanja na osnovu skupa demografskih, obrazovnih i profesionalnih atributa ima direktnu praktičnu primenu u domenima kao što su procena kreditnog rizika, segmentacija tržišta i analiza efikasnosti socijalnih programa.

1.2. Skup Podataka

Osnovu za analizu i modeliranje čini javno dostupan "Adult" skup podataka, izveden iz baze podataka Cenzus Biroa Sjedinjenih Američkih Država. Za potrebe ovog projekta, podaci su podijeljeni u dva odvojena fajla: adult_train.csv za trening modela i adult_test.csv za njegovu finalnu evaluaciju. Svaka instanca u skupu podataka opisuje pojedinca kroz niz atributa, koji obuhvataju numeričke vrijednosti poput starosti (age) i broja radnih sati (hours_per_week), kao i kategorijske podatke kao što su tip zaposlenja (workclass), nivo obrazovanja (education) i bračni status (marital_status). Ciljna promjenljiva, income, je binarnog karaktera i predstavlja ključni ishod koji model treba da nauči da prepozna.

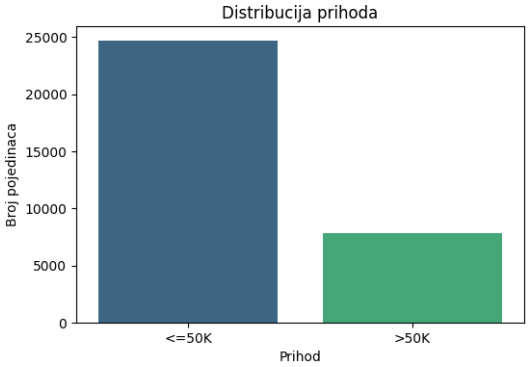
2. Eksplorativna Analiza Podataka (EDA)

EDA je sprovedena sa ciljem razumevanja strukture podataka, identifikacije obrazaca i postavljanja hipoteza za modeliranje.

2.1. Analiza Ciljne Promenljive

Distribucija ciljne varijable income je pokazala izražen disbalans klasa: 76% instanci pripada klasi <=50K, a samo 24% klasi >50K.

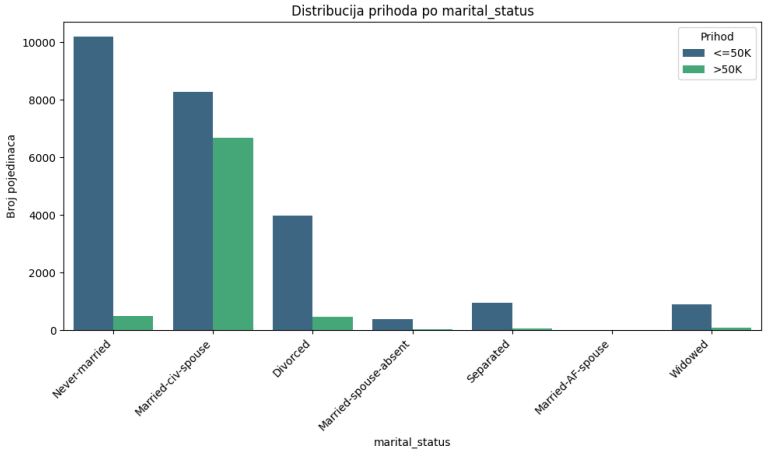
Implikacija: Bez adekvatnog tretmana, modeli će biti pristrasni prema većinskoj klasi, što bi rezultiralo visokom tačnošću, ali lošom sposobnošću prepoznavanja rjeđe klase >50K.



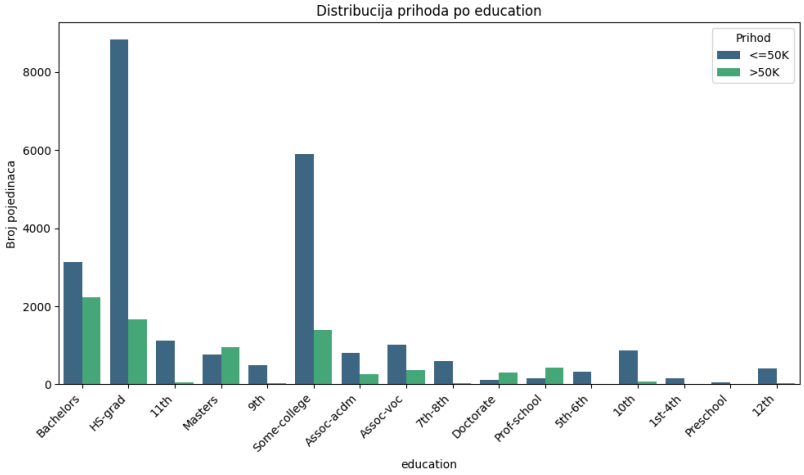
2.2. Analiza atributa

Vizuelizacija odnosa između atributa i prihoda otkrila je sljedeće:

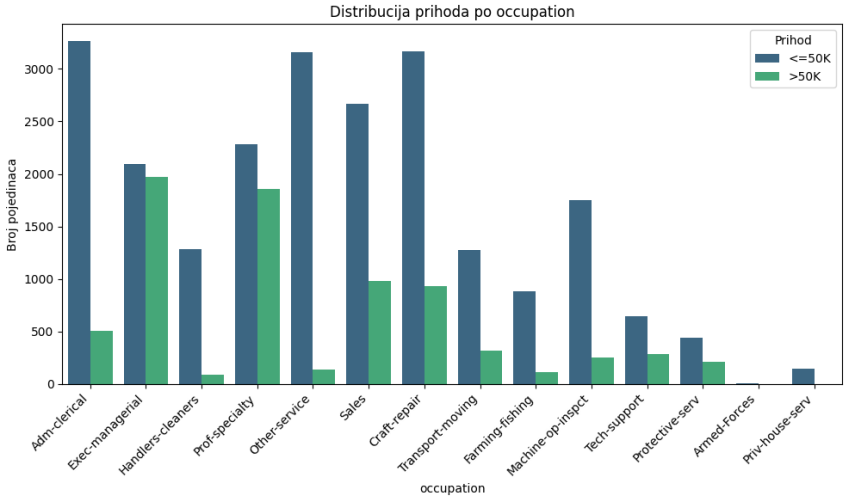
- **Bračni Status (marital_status):** Ovo je vizuelno najjači prediktor. Kategorija Married-civ-spouse je jedina u kojoj je broj osoba sa prihodom >50K značajan, što sugeriše da bračni status (ili faktori povezani s njim, poput stabilnosti i dvojnih prihoda) snažno utiče na zaradu.



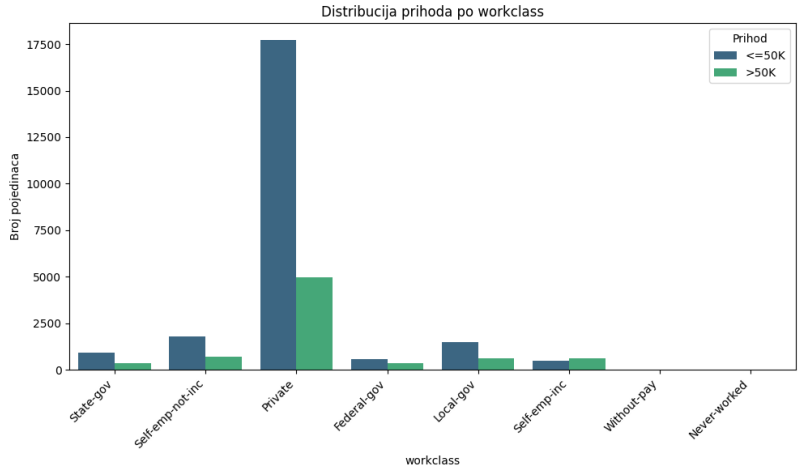
- **Obrazovanje (education i education_num):** Postoji jasna pozitivna korelacija. Viši nivo formalnog obrazovanja (Bachelors, Masters, Doctorate) direktno je povezan sa većom vjerovatnoćom ostvarivanja prihoda iznad 50K.



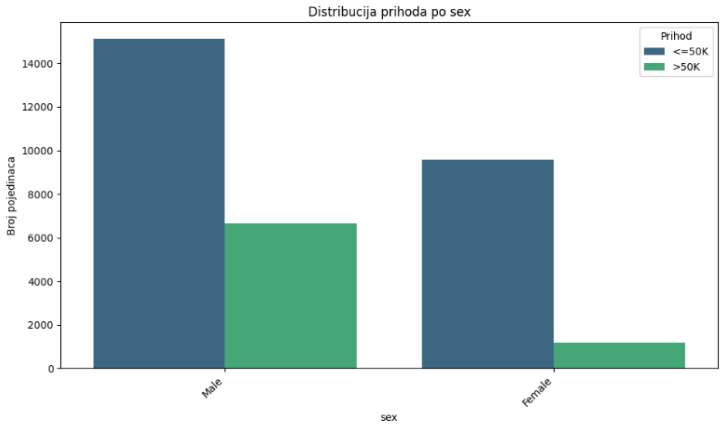
- **Zanimanje (occupation):** Zanimanja poput Exec-managerial i Prof-specialty pokazuju najveći procenat osoba sa visokim primanjima, što je i očekivano.



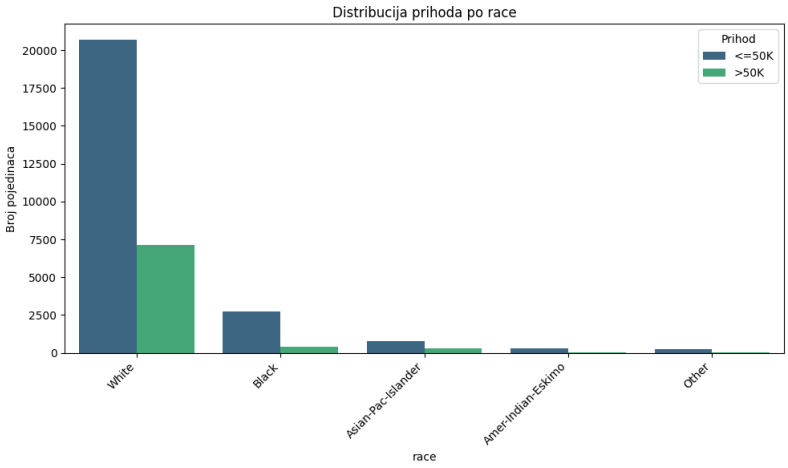
- **Tip Zaposlenja (workclass):** Privatni sektor je najveći poslodavac, ali i druge kategorije kao što su samozaposleni ili rad u državnoj upravi pokazuju različite obrasce prihoda.



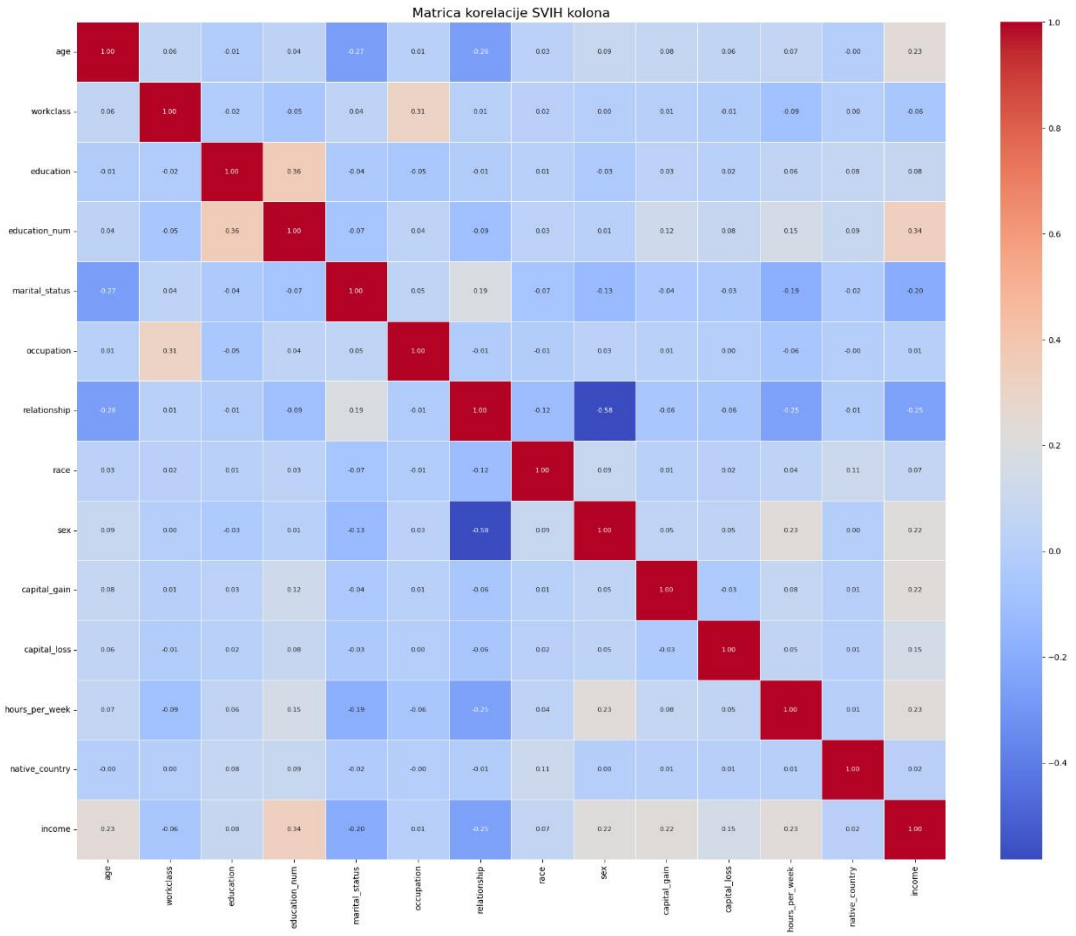
- **Pol (sex):** Grafik jasno pokazuje značajnu razliku u distribuciji prihoda između muškaraca i žena u ovom skupu podataka, gde muškarci čine veći udio u kategoriji sa višim prihodima.



- **Rasa (race) i Zemlja Porekla (native_country):** Pokazano je da postoji disbalans i u ovim kategorijama. Populacija White i osobe iz United-States čine dominantnu većinu u skupu podataka, što može uticati na sposobnost modela da generalizuje na manje zastupljene grupe.



- **Korelaciona Matrica:** Heatmap numeričkih atributa nije pokazao jake linearne korelacije. Ovo je ključan nalaz koji opravdava korišćenje nelinearnih modela (Random Forest, Gradient Boosting) koji mogu da modeliraju složenije interakcije.



3. Predprocesiranje Podataka i Inženjering Atributa

Robustan pipeline za predprocesiranje, definisan u data_preprocessing.py, bio je ključan za pripremu podataka za modeliranje.

- **Čišćenje:** Nerelevantni atribut fnlwgt je uklonjen. Vrijednosti ? su mapirane u NaN (Not a Number) kako bi ih pandas i scikit-learn mogli programski obraditi. Uklonjen je značajan broj dupliranih redova, čime je osigurana jedinstvenost svake instance.
- **Rukovanje Nedostajućim Vrijednostima:** Primijenjene su standardne strategije imputacije:
    - **Numerički Atributi:** SimpleImputer sa strategijom mean (srednja vrijednost).
    - **Kategorijski Atributi:** SimpleImputer sa strategijom most_frequent (najčešća vrijednost).
- **Upravljanje Ekstremnim Vrijednostima (Outliers):** Korišćenjem "3-Sigma" pravila, vrijednosti koje odstupaju više od tri standardne devijacije od proseka su "zatvorene" (*capping*). Atributi capital_gain i capital_loss su izuzeti iz ovog procesa jer njihove ekstremne vrijednosti (veliki dobici ili gubici) nose izuzetno važan signal i nisu šum.

- **Transformacija Atributa:**
    - **Kategorijski:** Primenjeno je One-Hot Encoding (pd.get_dummies) sa drop_first=True kako bi se izbegla multikolinearnost (redundantnost) među novostvorenim binarnim kolonama.
    - **Numerički:** Primenjen je StandardScaler koji transformiše podatke tako da imaju srednju vrijednost 0 i standardnu devijaciju 1. Ovo je neophodno za optimalan rad modela kao što je Logistička Regresija.
    - **Ciljni:** LabelEncoder je konvertovao <=50K u 0 i >50K u 1.

4. Strategija Modeliranja

4.1. Izbor Modela i Rukovanje Disbalansom

- **Logistička Regresija:** Korišćena kao snažan, interpretibilan *baseline* model. Za disbalans je korišćen class_weight='balanced'.
- **Random Forest Classifier:** Ensemble model baziran na stablima, otporan na *overfitting* i dobar za procijenu važnosti atributa. Takođe je korišćen class_weight='balanced'.
- **Gradient Boosting Classifier:** Napredni ensemble model koji sekvencijalno gradi stabla kako bi ispravio greške prethodnih, često postižući vrhunske performanse. Za disbalans je korišćen SMOTE na trening podacima.

4.3. Optimizacija Hiperparametara

Za svaki model i svaku konfiguraciju atributa, GridSearchCV je korišćen za pronalaženje optimalnih hiperparametara. Kao metrika za optimizaciju odabrana je precision_weighted, signalizirajući da je prioritet bio smanjenje lažno pozitivnih predikcija.

5. Rezultati i Dubinska Analiza

5.1. Uporedni Prikaz Performansi

Tabela ispod sumira ključne metrike za svih šest testiranih konfiguracija.

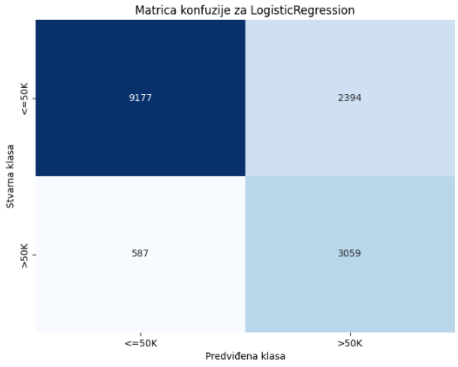
Model	Skup Feature-a	Tačnost (Acc)	F1-Score (W)	Preciznost (W)	Odziv (W)
GradientBoosting	**Selektovani (20)**	**0.8299**	**0.8379**	**0.8597**	**0.8299**
GradientBoosting	Svi	0.8293	0.8372	0.8579	0.8293
RandomForest	Selektovani (20)	0.8228	0.8320	0.8591	0.8228
RandomForest	Svi	0.8110	0.8219	0.8568	0.8110
LogisticRegression	Svi	0.8041	0.8153	0.8491	0.8041
LogisticRegression	Selektovani (20)	0.8017	0.8132	0.8490	0.8017

5.2. Logistička regresija

Logistička regresija postiže solidan odziv za klasu >50K, ali po cijenu veoma niske preciznosti, što generiše veliki broj lažno pozitivnih predikcija. Selekcija feature-a nije donela poboljšanje.

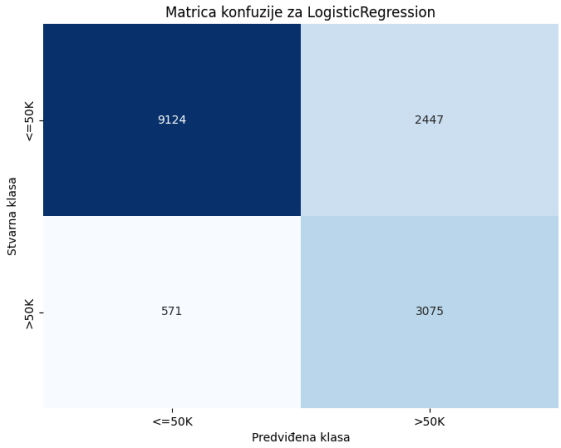
Model sa Svim Feature-ima

```
Ukupna tačnost (Accuracy): 0.8041
Ukupna preciznost (Precision - weighted): 0.8491
Ukupan odziv (Recall - weighted): 0.8041
Ukupan F1-Score (weighted): 0.8153

Detaljan klasifikacioni izvještaj:
              precision    recall  f1-score   support

       <=50K       0.94      0.79      0.86     11571
        >50K       0.56      0.84      0.67      3646

    accuracy                           0.80     15217
   macro avg       0.75      0.82      0.77     15217
weighted avg       0.85      0.80      0.82     15217
```



Model sa Selektovanim Feature-ima

```
Ukupna tačnost (Accuracy): 0.8017
Ukupna preciznost (Precision - weighted): 0.8490
Ukupan odziv (Recall - weighted): 0.8017
Ukupan F1-Score (weighted): 0.8132

Detaljan klasifikacioni izvještaj:
              precision    recall  f1-score   support

       <=50K       0.94      0.79      0.86     11571
        >50K       0.56      0.84      0.67      3646

    accuracy                           0.80     15217
   macro avg       0.75      0.82      0.76     15217
weighted avg       0.85      0.80      0.81     15217
```

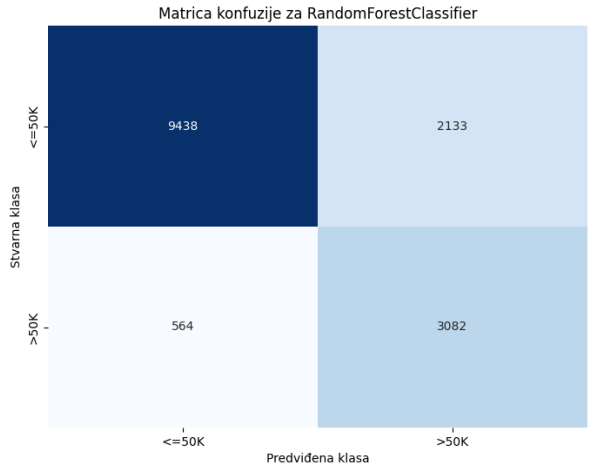


5.3. Random Forest

Random Forest pokazuje bolje performanse od Logističke Regresije. Selekcija feature-a donosi primetno poboljšanje u ukupnoj tačnosti i F1-skoru.

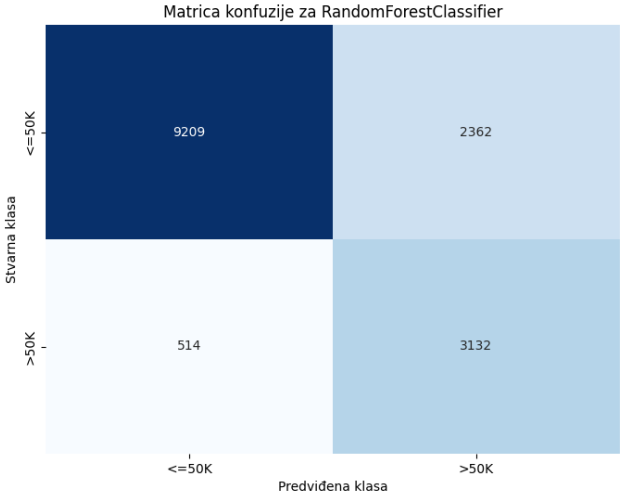
Model sa Svim Feature-ima

```
Ukupna tačnost (Accuracy): 0.8110
Ukupna preciznost (Precision - weighted): 0.8568
Ukupan odziv (Recall - weighted): 0.8110
Ukupan F1-Score (weighted): 0.8219

Detaljan klasifikacioni izvještaj:
              precision    recall  f1-score   support

       <=50K       0.95      0.80      0.86     11571
        >50K       0.57      0.86      0.69      3646

    accuracy                           0.81     15217
   macro avg       0.76      0.83      0.78     15217
weighted avg       0.86      0.81      0.82     15217
```



Model sa Selektovanim Feature-ima

```
Ukupna tačnost (Accuracy): 0.8228
Ukupna preciznost (Precision - weighted): 0.8591
Ukupan odziv (Recall - weighted): 0.8228
Ukupan F1-Score (weighted): 0.8320

Detaljan klasifikacioni izvještaj:
              precision    recall  f1-score   support

       <=50K       0.94      0.82      0.87     11571
        >50K       0.59      0.85      0.70      3646

    accuracy                           0.82     15217
   macro avg       0.77      0.83      0.79     15217
weighted avg       0.86      0.82      0.83     15217
```

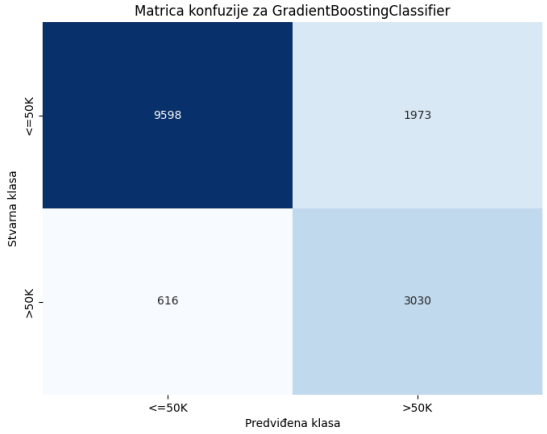


5.4. Gradient Boosting

Gradient Boosting postiže najbolje ukupne rezultate. Model sa selektovanim feature-ima ima neznatno višu tačnost, čineći ga “pobedničkim“ modelom. Nudi najbolji balans između preciznosti i odziva od svih testiranih modela.

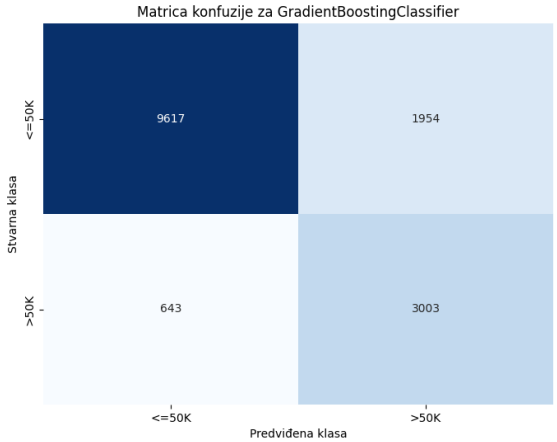
Model sa Svim Feature-ima

```
Ukupna tačnost (Accuracy): 0.8293
Ukupna preciznost (Precision - weighted): 0.8579
Ukupan odziv (Recall - weighted): 0.8293
Ukupan F1-Score (weighted): 0.8372

Detaljan klasifikacioni izvještaj:
              precision    recall  f1-score   support

       <=50K       0.94      0.83      0.88     11571
        >50K       0.61      0.82      0.70      3646

    accuracy                           0.83     15217
   macro avg       0.77      0.83      0.79     15217
weighted avg       0.86      0.83      0.84     15217
```



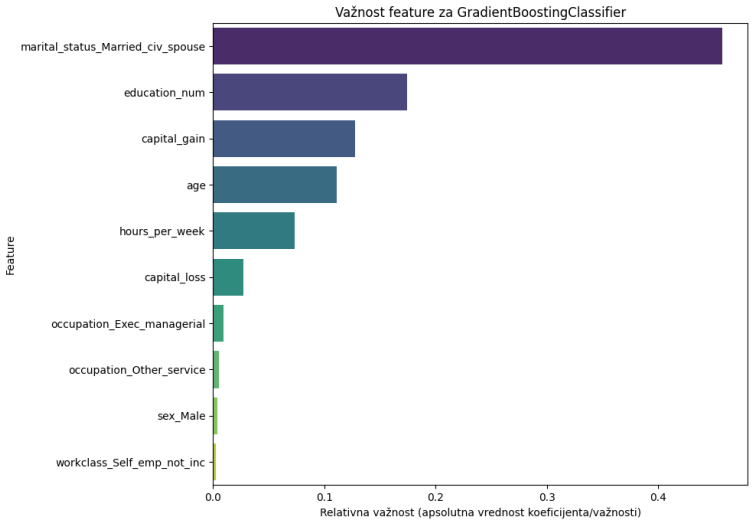
Model sa Selektovanim Feature-ima

```
Ukupna tačnost (Accuracy): 0.8299
Ukupna preciznost (Precision - weighted): 0.8597
Ukupan odziv (Recall - weighted): 0.8299
Ukupan F1-Score (weighted): 0.8379

Detaljan klasifikacioni izvještaj:
              precision    recall  f1-score   support

       <=50K       0.94      0.83      0.88     11571
        >50K       0.61      0.83      0.70      3646

    accuracy                           0.83     15217
   macro avg       0.77      0.83      0.79     15217
weighted avg       0.86      0.83      0.84     15217
```



5.5. Važnost Atributa

Analiza važnosti atributa za pobednički model potvrđuje uvide iz EDA. Pet najuticajnijih prediktora su: bračni status, broj godina obrazovanja, kapitalni dobitak, starost i broj radnih sati nedeljno.



6. Zaključak

6.1. Sumarni Pregled Nalaza

Projekat je uspješno demonstrirao primenu end-to-end procesa mašinskog učenja za rešavanje realnog klasifikacionog problema. Kroz sistematičnu evaluaciju, Gradient Boosting Classifier, treniran na setu od 20 selektovanih atributa(enkodiranih), nedvosmisleno se izdvojio kao pobednički model, postigavši najviše ocjene na svim ključnim metrikama, uključujući tačnost od 83.0% i ponderisani F1-skor od 83.8% na test podacima.

Selekcija atributa se pokazala kao ključna strategija, ne samo za smanjenje kompleksnosti i vremena treninga, već i za poboljšanje performansi kod modela poput Random Forest-a, efektivno smanjujući uticaj šuma i rizik od *overfittinga*.

6.2. Ključni Uvidi i Poslovne Implikacije

Analiza je pružila nekoliko fundamentalnih uvida:

1. **Potvrđena je superiornost nelinearnih modela:** Gradient Boosting i Random Forest su konzistentno nadmašili Logističku Regresiju, potvrđujući hipotezu iz EDA da su veze između atributa i prihoda suviše složene da bi ih linearni model efikasno uhvatio.
2. **Konzistentnost važnosti atributa:** Nezavisno od korišćenog modela, atributi kao što su bračni status (Married-civ-spouse), kapitalni dobici, nivo obrazovanja i starost uvek

su se nalazili na vrhu liste po važnosti. Ovo pruža visok stepen povjerenja da ovi faktori predstavljaju fundamentalne pokretače prihoda u analiziranom skupu podataka.

3. **Praktični značaj kompromisa metrika:** Najvažniji praktični uvid je niska preciznost (61%) za klasu >50K. U poslovnom kontekstu, ovo znači da bi skoro 40% resursa (npr. marketing budžeta, vreme prodajnih agenata) bilo pogrešno usmjereno na pojedince koji ne pripadaju ciljnoj grupi sa visokim prihodima. Svest o ovom kompromisu je ključna za donošenje odluka o primeni modela.

6.3. Ograničenja Projekta

Važno je prepoznati sledeća ograničenja:

1. **Starost podataka:** Podaci potiču iz 1994. godine. Društveno-ekonomska dinamika se značajno promijenila, što znači da bi performanse modela na savremenim podacima vjerovatno bile drugačije.
2. **Performanse na manjinskoj klasi:** Iako je odziv dobar, niska preciznost za klasu >50K ostaje glavno ograničenje modela za primene gde je cena lažno pozitivnih predikcija visoka.
3. **Ograničen skup atributa:** Skup podataka ne sadrži informacije koje bi danas mogle biti veoma relevantne, kao što su industrija zaposlenja, specifična lokacija (grad/država) ili posjedovanje nekretnina.