Predikcija vakcinacije ispitanika protiv sezonskog gripa i virusa H1N1

Mašinsko učenje (13M051MU)

student: Mihailo Grbić
profesor: doc. dr Predrag Tadić

Zadatak

- Mišljenje ispitanika o vakcinaciji protiv:
  - Sezonskog gripa (binarno)
  - H1N1 virusa (binarno)
- Anketa sa 37 pitanja → podaci
- Postavka
  - Čišćenje i priprema podataka
  - Klasifikacija
    - Predikcija dve labele → dva modela
    - Izostavljena labela se ne koristi kao prediktor
  - Evaluacija

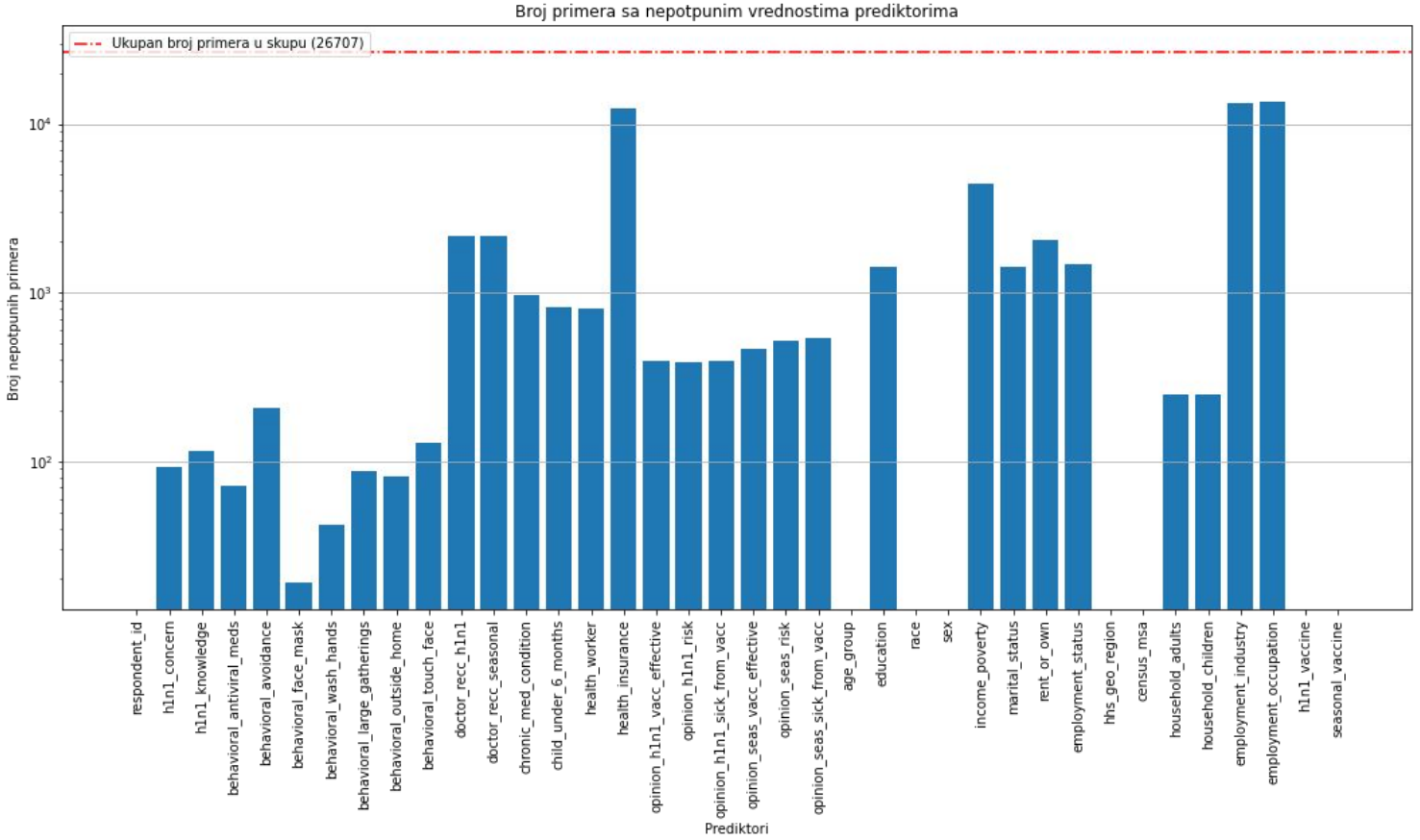
Čišćenje i priprema podataka

Skup podataka

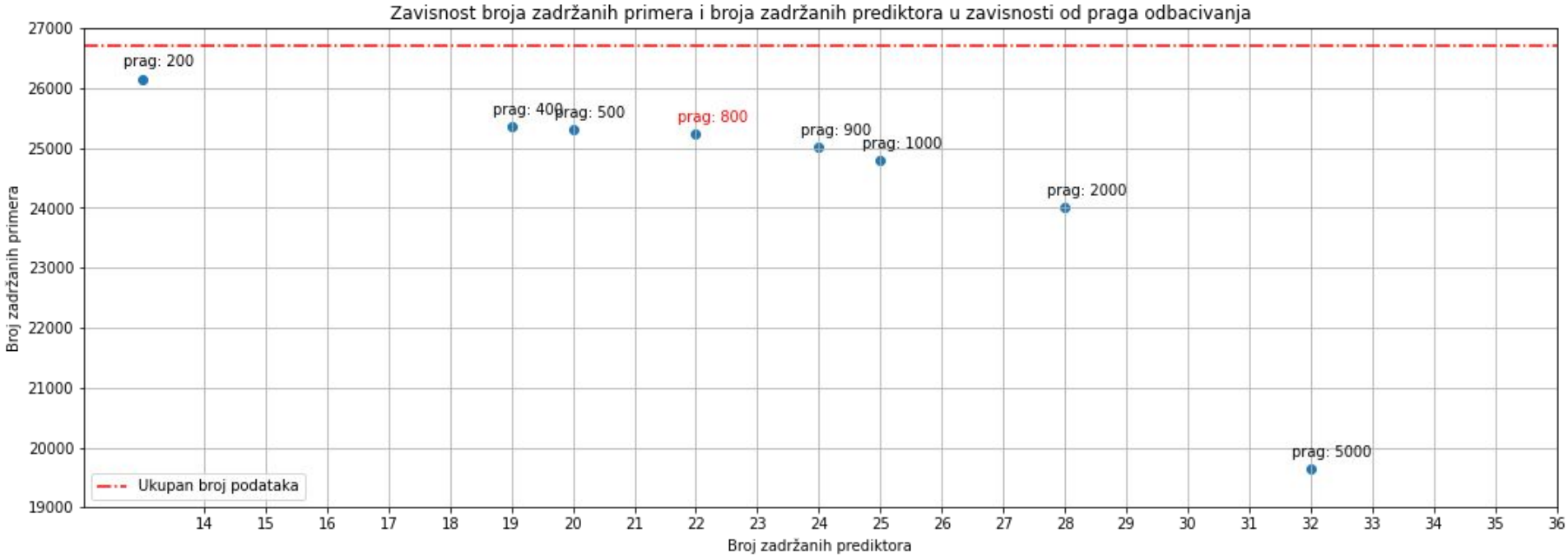
- Osnovne informacije:
  - *training_set_features.csv* : 35 prediktora
    - numeričkih: 23
    - kategoričkih: 12
    - bez indeks ispitanika
  - *training_set_labels.csv* : 2 izlaza
    - *seasonal_vaccine*
    - *h1n1_vaccine*
  - Broj primera: 26707
  - Problemi:
    - Nepotpuni primeri
    - Kategoričke vrednosti

```
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   respondent_id                26707 non-null  int64
 1   h1n1_concern                 26615 non-null  float64
 2   h1n1_knowledge               26591 non-null  float64
 3   behavioral_antiviral_meds    26636 non-null  float64
 4   behavioral_avoidance         26499 non-null  float64
 5   behavioral_face_mask         26688 non-null  float64
 6   behavioral_wash_hands        26665 non-null  float64
 7   behavioral_large_gatherings  26620 non-null  float64
 8   behavioral_outside_home      26625 non-null  float64
 9   behavioral_touch_face        26579 non-null  float64
 10  doctor_recc_h1n1             24547 non-null  float64
 11  doctor_recc_seasonal         24547 non-null  float64
 12  chronic_med_condition        25736 non-null  float64
 13  child_under_6_months         25887 non-null  float64
 14  health_worker                25903 non-null  float64
 15  health_insurance             14433 non-null  float64
 16  opinion_h1n1_vacc_effective  26316 non-null  float64
 17  opinion_h1n1_risk            26319 non-null  float64
 18  opinion_h1n1_sick_from_vacc  26312 non-null  float64
 19  opinion_seas_vacc_effective  26245 non-null  float64
 20  opinion_seas_risk            26193 non-null  float64
 21  opinion_seas_sick_from_vacc  26170 non-null  float64
 22  age_group                    26707 non-null  object
 23  education                    25300 non-null  object
 24  race                         26707 non-null  object
 25  sex                          26707 non-null  object
 26  income_poverty               22284 non-null  object
 27  marital_status               25299 non-null  object
 28  rent_or_own                  24665 non-null  object
 29  employment_status            25244 non-null  object
 30  hhs_geo_region               26707 non-null  object
 31  census_msa                   26707 non-null  object
 32  household_adults             26458 non-null  float64
 33  household_children           26458 non-null  float64
 34  employment_industry          13377 non-null  object
 35  employment_occupation        13237 non-null  object
 36  h1n1_vaccine                 26707 non-null  int64
 37  seasonal_vaccine             26707 non-null  int64
dtypes: float64(23), int64(3), object(12)
```

Čišćenje podataka



Čišćenje podataka

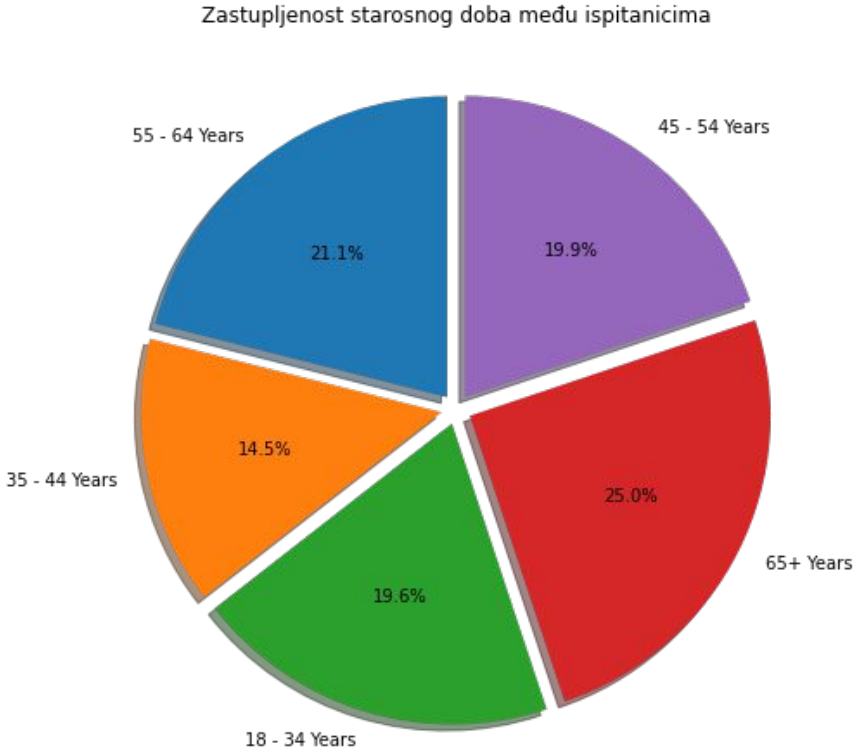


Čišćenje podataka

- Analiza
  - 5/35 “čistih” prediktora
  - 3/35 prediktora kojima nedostaje ~⅓ podataka
- Filtriranje
  - Odbacivanje prediktora koji sadrže više od 800 nepotpunih primera
  - Preostali nepotpuni primeri se odbacuju
- Rezultat
  - Broj zadržanih prediktora: 22/35
  - Broj primera zadržanih primera: 25238
  - ~1500 odbačenih primera (5.61% skupa) ⟶ Prihvatljivo!

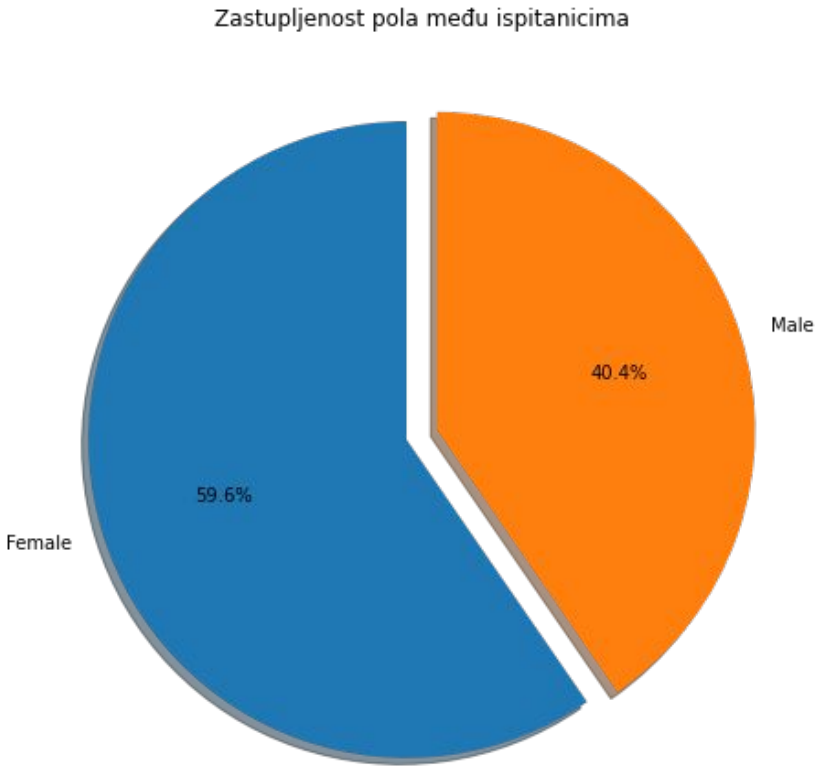
Kategorički prediktori | Starosna grupa

- Punoletni ispitanici
- Pet starosnih grupa
  - Slična zastupljenost
- Gradacioni poredak
  - Kodiranje celobrojnim vrednostima



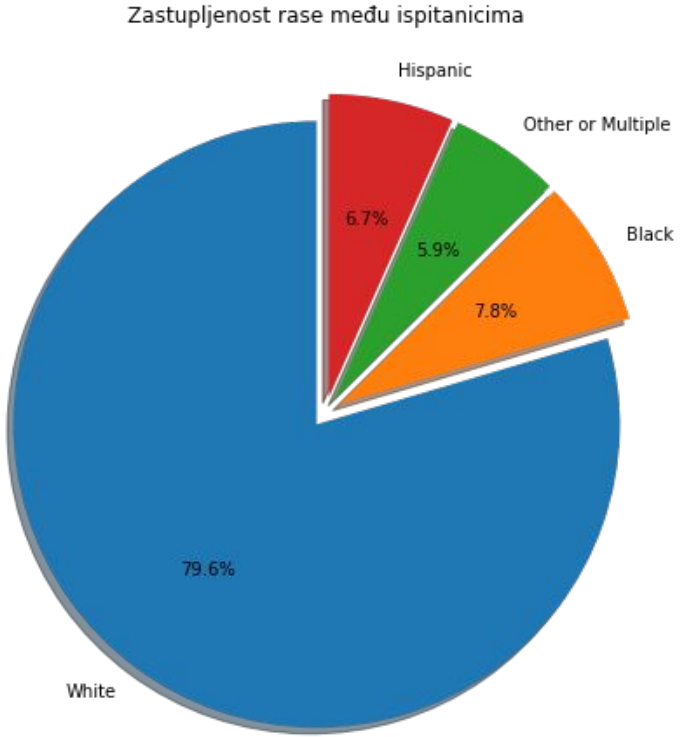
Kategorički prediktori | Pol

- Dva pola (muški/ženski)
  - Nešto veća zastupljenost žena
- Binarno kodiranje
  - 0: Ženski pol
  - 1: Muški pol



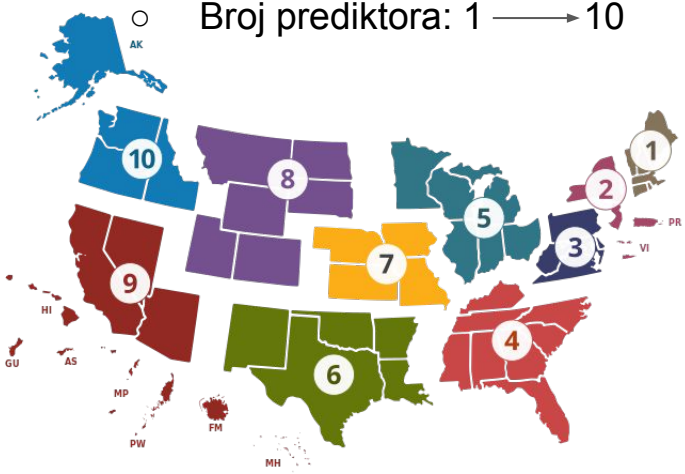
Kategorički prediktori | Rasa

- Podela po boji kože
  - Četiri kategorije
  - Veliku većinu čine ljudi sa belom bojom kože
- Nema smislenog gradacionog poretka
  - *One-Hot Encoding*
  - Broj prediktora: 1 ⟶ 4

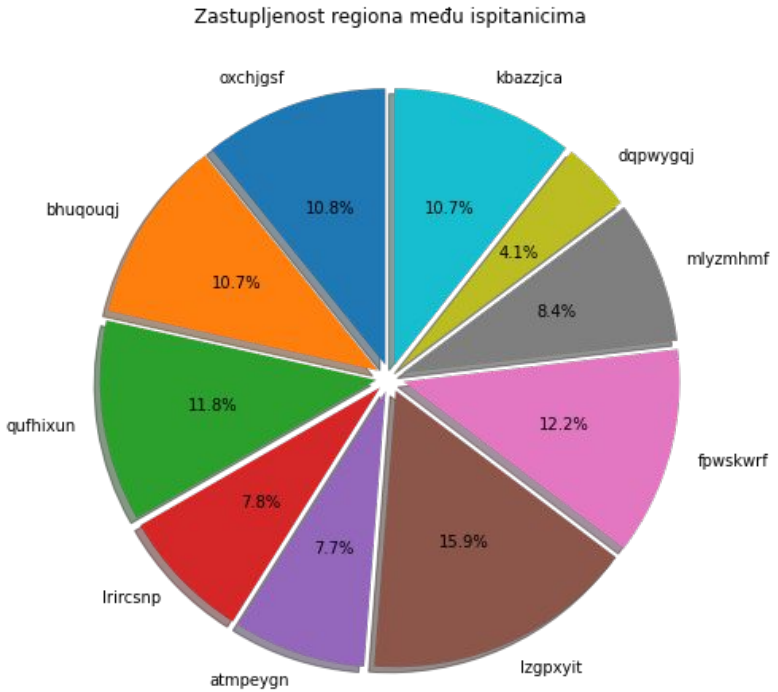


Kategorički prediktori | Geografski region

- Istraživanje je rađeno na području Sjedinjenih Američkih Država
  - Podela na 10 grupa država
  - Ujednačena zastupljenost
- Nema smislenog gradacionog poretka
  - *One-Hot Encoding*
  - Broj prediktora: 1 ⟶ 10

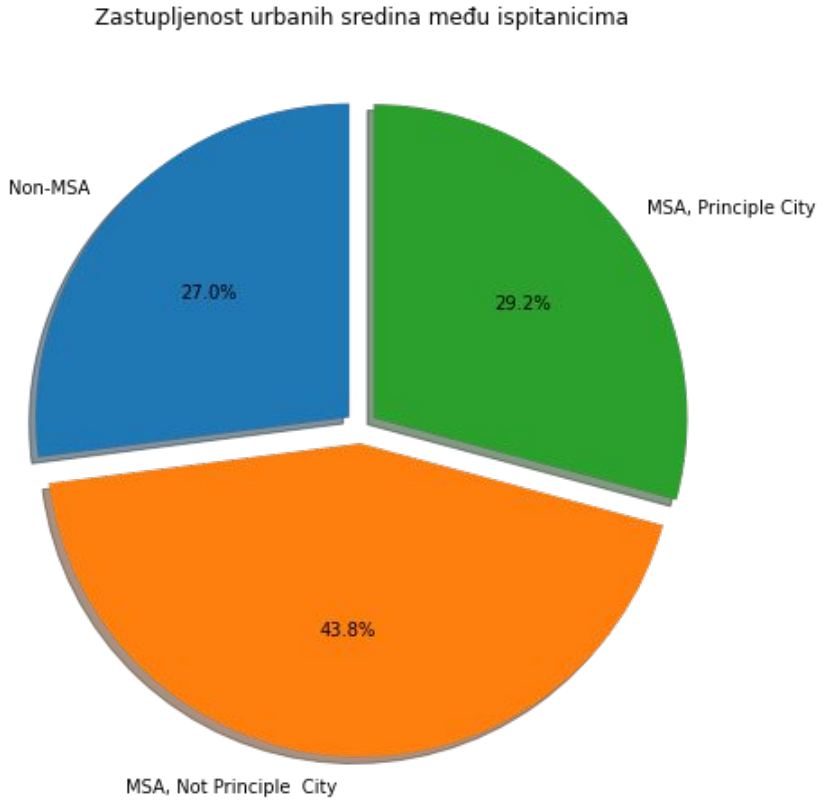


https://www.hhs.gov/about/agencies/iea/regional-offices/index.html



Kategorički prediktori | Urbanost sredine

- Broj stanovnika mesta prebivališta
  - Non-MSA: Populacija < 50 000
  - MSA: Populacija > 50 000
- Najveći udeo ispitanika iz urbanih sredina van glavnog grada
- ~¾ stanovnika su iz urbanih naselja
- Ima smisla gradaciono porediti kategorije
  - *Non-MSA*: 0
  - *MSA, Not Principle City*: 1
  - *MSA, Principle City*: 2



Kategorički prediktori | Rezime

2x Enkodovanje labela

1x Binarizacija

2x One-hot enkodovanje

22 prediktora

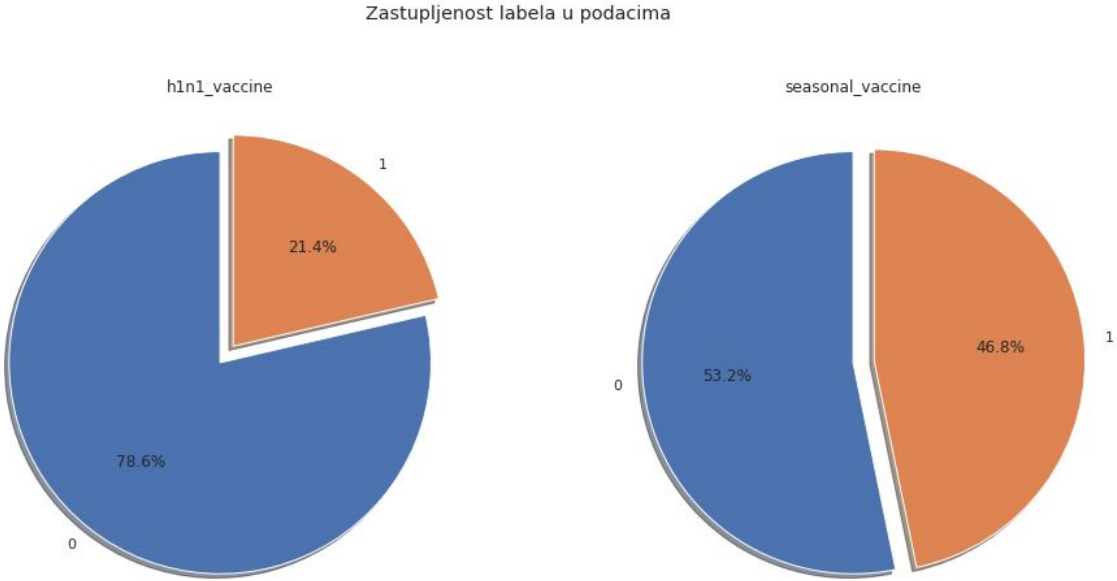
- numeričkih: 17
- kategoričkih: 5

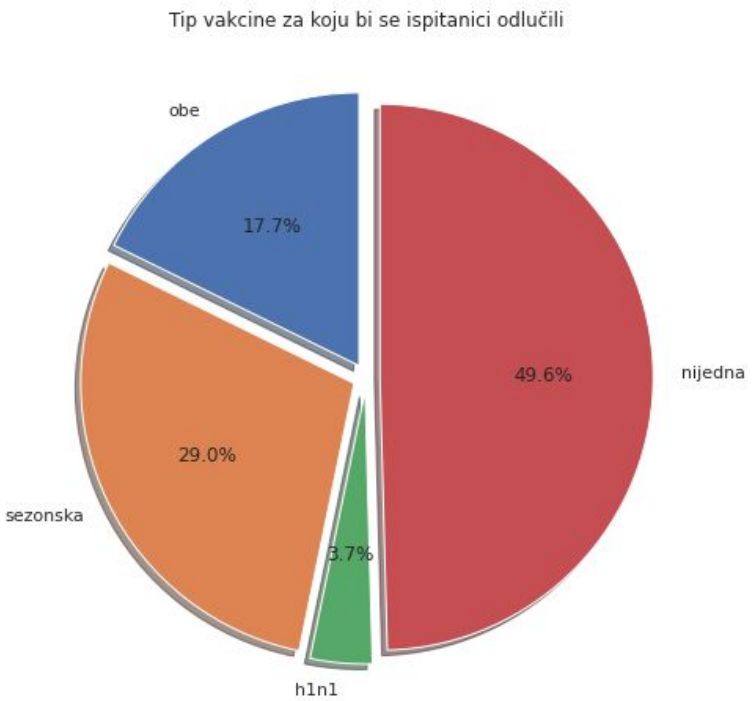
36 prediktora

- numeričkih: 36
- kategoričkih: 0

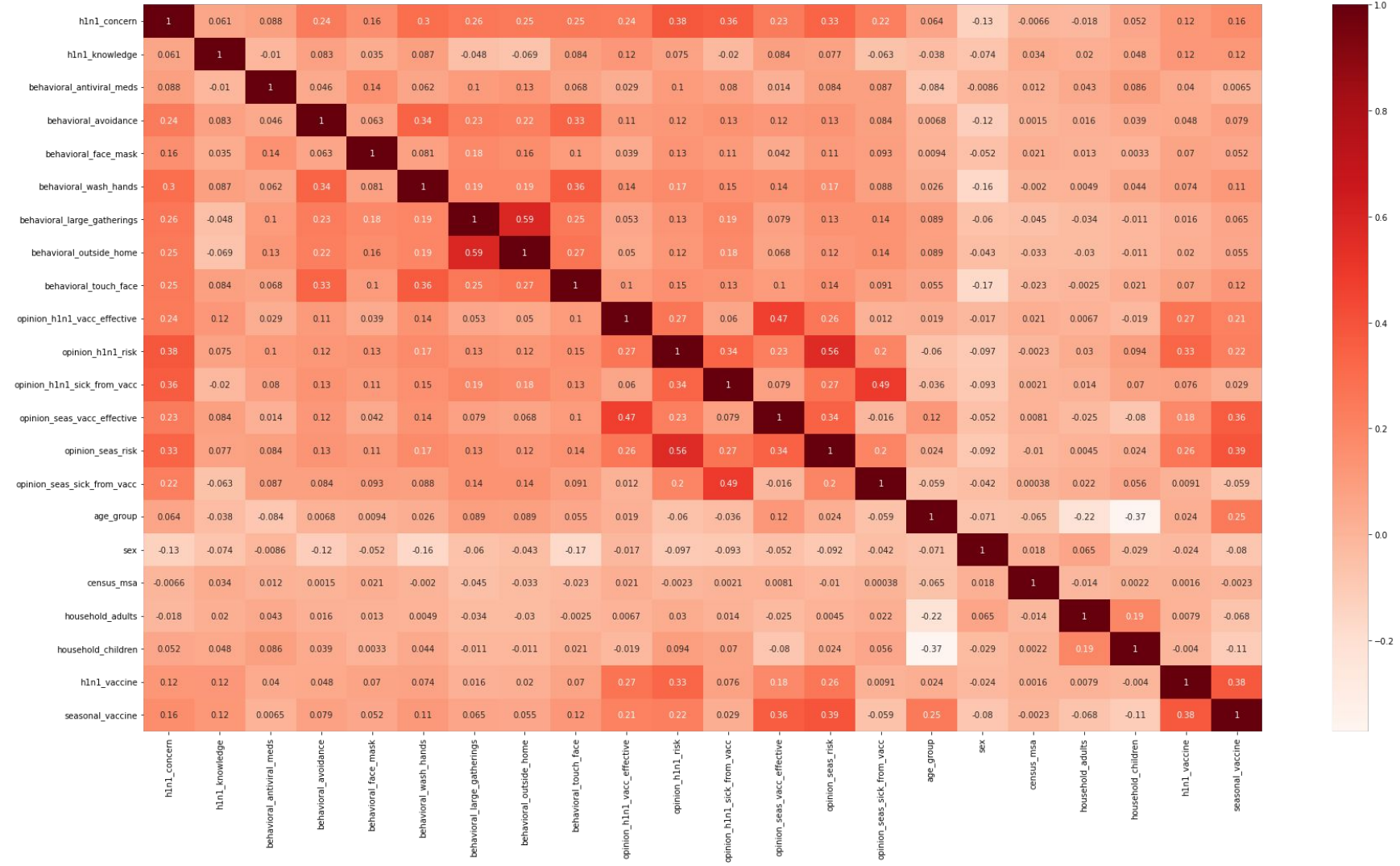
Izlazni podaci

- (Ne)balansiranost klasa
  - 0: “protiv” vakcinacije
  - 1: “za” vakcinaciju

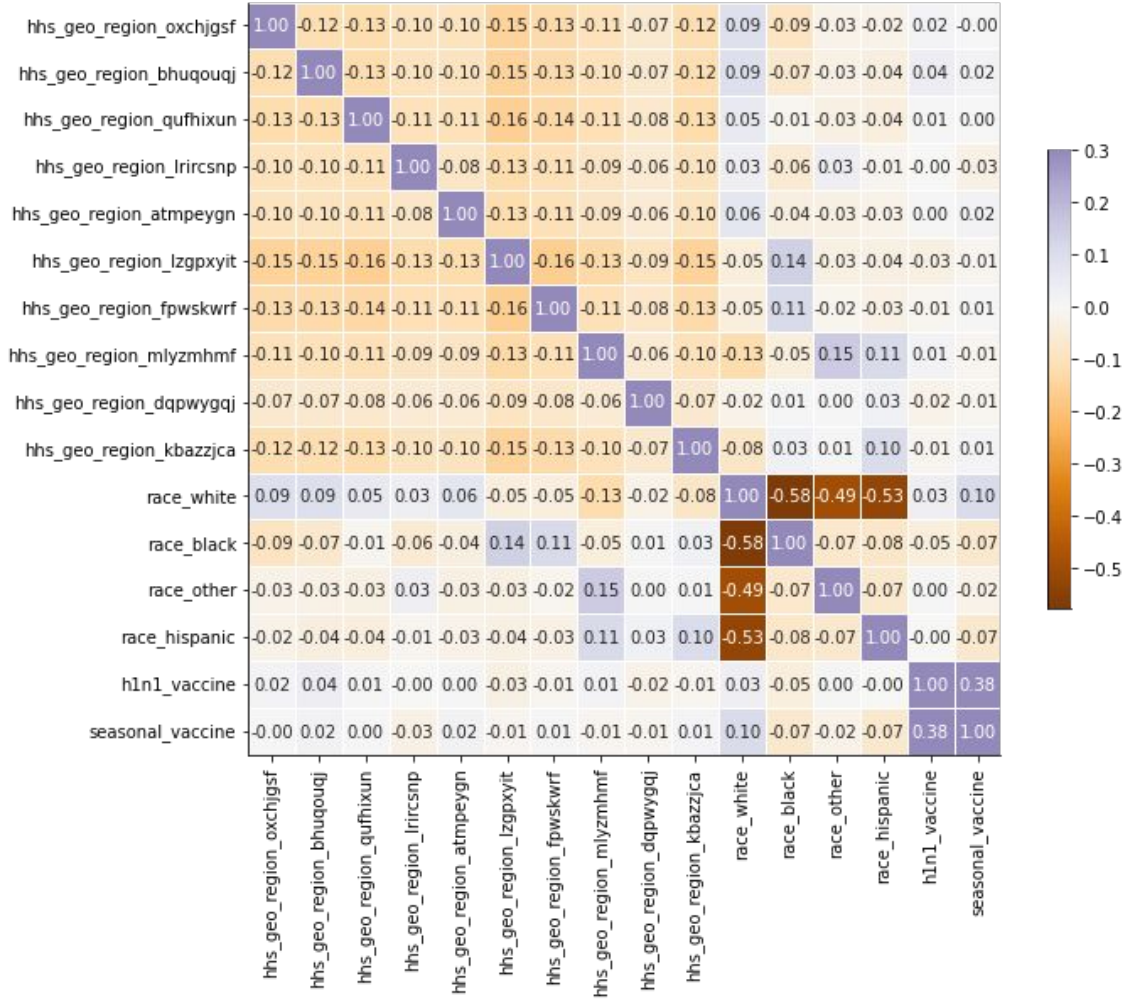




Korelacione mape

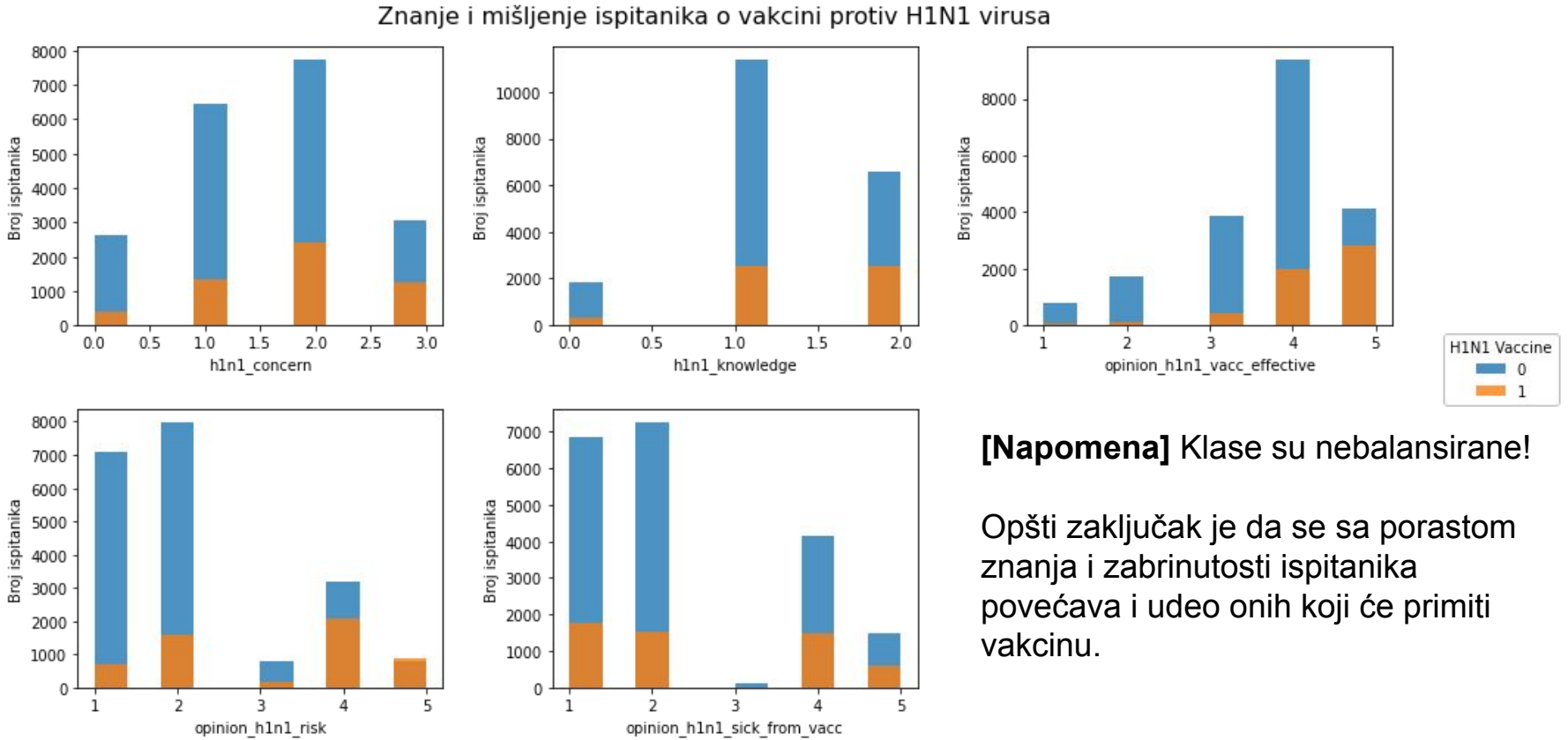

h1n1_concern 1 0.061 0.088 0.24 0.16 0.3 0.26 0.25 0.25 0.24 0.38 0.36 0.23 0.33 0.22 0.064 -0.13 -0.0066 -0.018 0.052 0.12 0.16
h1n1_knowledge 0.061 1 -0.01 0.083 0.035 0.087 -0.048 -0.069 0.084 0.12 0.075 -0.02 0.084 0.077 -0.063 -0.038 -0.074 0.034 0.02 0.048 0.12 0.12
behavioral_antiviral_meds 0.088 -0.01 1 0.046 0.14 0.062 0.1 0.13 0.068 0.029 0.1 0.08 0.014 0.084 0.087 -0.084 -0.0086 0.012 0.043 0.086 0.04 0.0065
behavioral_avoidance 0.24 0.083 0.046 1 0.063 0.34 0.23 0.22 0.33 0.11 0.12 0.13 0.12 0.13 0.084 0.0068 -0.12 0.0015 0.016 0.039 0.048 0.079
behavioral_face_mask 0.16 0.035 0.14 0.063 1 0.081 0.18 0.16 0.1 0.039 0.13 0.11 0.042 0.11 0.093 0.0094 -0.052 0.021 0.013 0.0033 0.07 0.052
behavioral_wash_hands 0.3 0.087 0.062 0.34 0.081 1 0.19 0.19 0.36 0.14 0.17 0.15 0.14 0.17 0.088 0.026 -0.16 -0.002 0.0049 0.044 0.074 0.11
behavioral_large_gatherings 0.26 -0.048 0.1 0.23 0.18 0.19 1 0.59 0.25 0.053 0.13 0.19 0.079 0.13 0.14 0.089 -0.06 -0.045 -0.034 -0.011 0.016 0.065
behavioral_outside_home 0.25 -0.069 0.13 0.22 0.16 0.19 0.59 1 0.27 0.05 0.12 0.18 0.068 0.12 0.14 0.089 -0.043 -0.033 -0.03 -0.011 0.02 0.055
behavioral_touch_face 0.25 0.084 0.068 0.33 0.1 0.36 0.25 0.27 1 0.1 0.15 0.13 0.1 0.14 0.091 0.055 -0.17 -0.023 -0.0025 0.021 0.07 0.12
opinion_h1n1_vacc_effective 0.24 0.12 0.029 0.11 0.039 0.14 0.053 0.05 0.1 1 0.27 0.06 0.47 0.26 0.012 0.019 -0.017 0.021 0.0067 -0.019 0.27 0.21
opinion_h1n1_risk 0.38 0.075 0.1 0.12 0.13 0.17 0.13 0.12 0.15 0.27 1 0.34 0.23 0.56 0.2 -0.06 -0.097 -0.0023 0.03 0.094 0.33 0.22
opinion_h1n1_sick_from_vacc 0.36 -0.02 0.08 0.13 0.11 0.15 0.19 0.18 0.13 0.06 0.34 1 0.079 0.27 0.49 -0.036 -0.093 0.0021 0.014 0.07 0.076 0.029
opinion_seas_vacc_effective 0.23 0.084 0.014 0.12 0.042 0.14 0.079 0.068 0.1 0.47 0.23 0.079 1 0.34 -0.016 0.12 -0.052 0.0081 -0.025 -0.08 0.18 0.36
opinion_seas_risk 0.33 0.077 0.084 0.13 0.11 0.17 0.13 0.12 0.14 0.26 0.56 0.27 0.34 1 0.2 0.024 -0.092 -0.01 0.0045 0.024 0.26 0.39
opinion_seas_sick_from_vacc 0.22 -0.063 0.087 0.084 0.093 0.088 0.14 0.14 0.091 0.012 0.2 0.49 -0.016 0.2 1 -0.059 -0.042 0.00038 0.022 0.056 0.0091 -0.059
age_group 0.064 -0.038 -0.084 0.0068 0.0094 0.026 0.089 0.089 0.055 0.019 -0.06 -0.036 0.12 0.024 -0.059 1 -0.071 -0.065 -0.22 -0.37 0.024 0.25
sex -0.13 -0.074 -0.0086 -0.12 -0.052 -0.16 -0.06 -0.043 -0.17 -0.017 -0.097 -0.093 -0.052 -0.092 -0.042 -0.071 1 0.018 0.065 -0.029 -0.024 -0.08
census_msa -0.0066 0.034 0.012 0.0015 0.021 -0.002 -0.045 -0.033 -0.023 0.021 -0.0023 0.0021 0.0081 -0.01 0.00038 -0.065 0.018 1 -0.014 0.0022 0.0016 -0.0023
household_adults -0.018 0.02 0.043 0.016 0.013 0.0049 -0.034 -0.03 -0.0025 0.0067 0.03 0.014 -0.025 0.0045 0.022 -0.22 0.065 -0.014 1 0.19 0.0079 -0.068
household_children 0.052 0.048 0.086 0.039 0.0033 0.044 -0.011 -0.011 0.021 -0.019 0.094 0.07 -0.08 0.024 0.056 -0.37 -0.029 0.0022 0.19 1 -0.004 -0.11
h1n1_vaccine 0.12 0.12 0.04 0.048 0.07 0.074 0.016 0.02 0.07 0.27 0.33 0.076 0.18 0.26 0.0091 0.024 -0.024 0.0016 0.0079 -0.004 1 0.38
seasonal_vaccine 0.16 0.12 0.0065 0.079 0.052 0.11 0.065 0.055 0.12 0.21 0.22 0.029 0.36 0.39 -0.059 0.25 -0.08 -0.0023 -0.068 -0.11 0.38 1
h1n1_concern
h1n1_knowledge
behavioral_antiviral_meds
behavioral_avoidance
behavioral_face_mask
behavioral_wash_hands
behavioral_large_gatherings
behavioral_outside_home
behavioral_touch_face
opinion_h1n1_vacc_effective
opinion_h1n1_risk
opinion_h1n1_sick_from_vacc
opinion_seas_vacc_effective
opinion_seas_risk
opinion_seas_sick_from_vacc
age_group
sex
census_msa
household_adults
household_children
h1n1_vaccine
seasonal_vaccine
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Prediktori transformisani one-hot enkodovanjem



Analiza veze pojedinačnih prediktora i izlaza

Uticaj znanja i mišljenja ispitanika | H1N1 virus

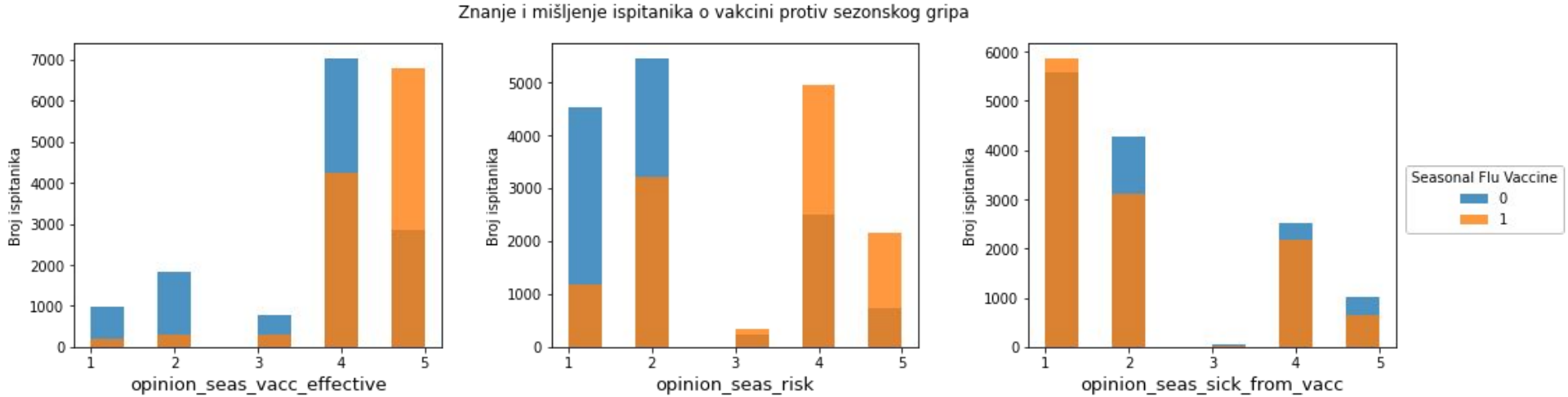


[Napomena] Klase su nebalansirane!

Opšti zaključak je da se sa porastom znanja i zabrinutosti ispitanika povećava i udeo onih koji će primiti vakcinu.

Uticaj znanja i mišljenja ispitanika | Sezonski grip

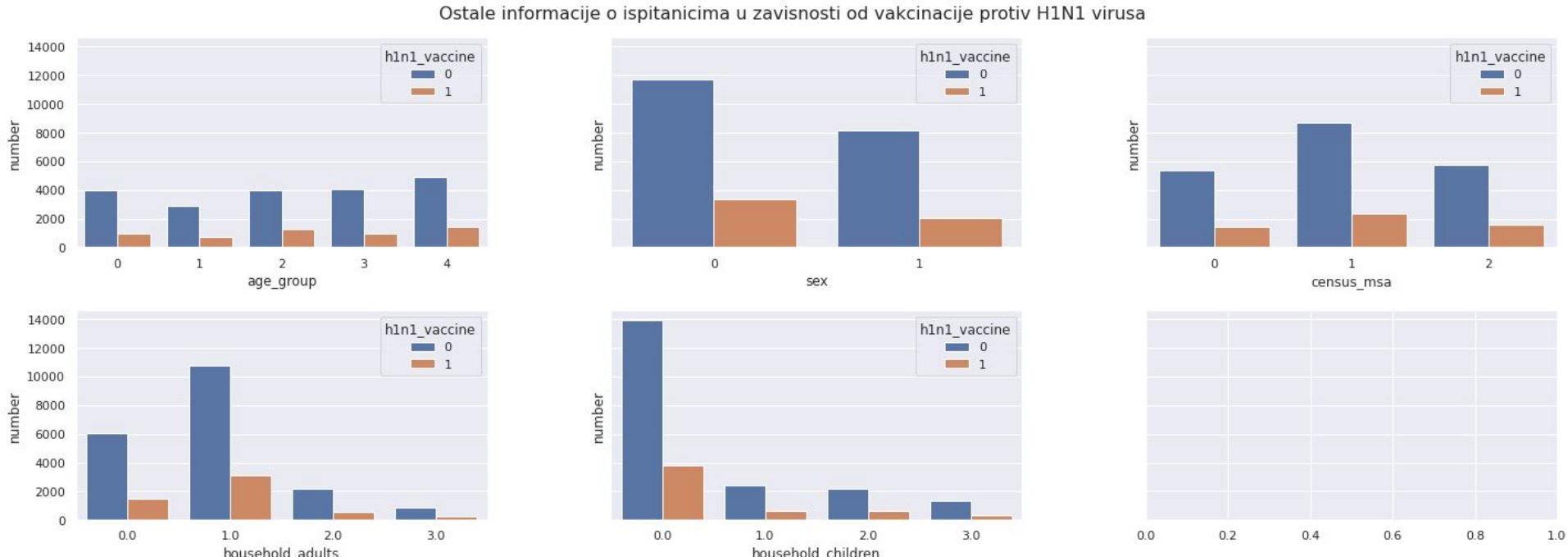
- Rizična grupa i/ili mišljenje da je vakcina efikasna ⟶ veći udeo vakcinisanih
- Učestalo razboljevanje ispitanika nije pokazalo uticaj na opredeljenje za vakcinu



[Napomena] Klase su nebalansirane!

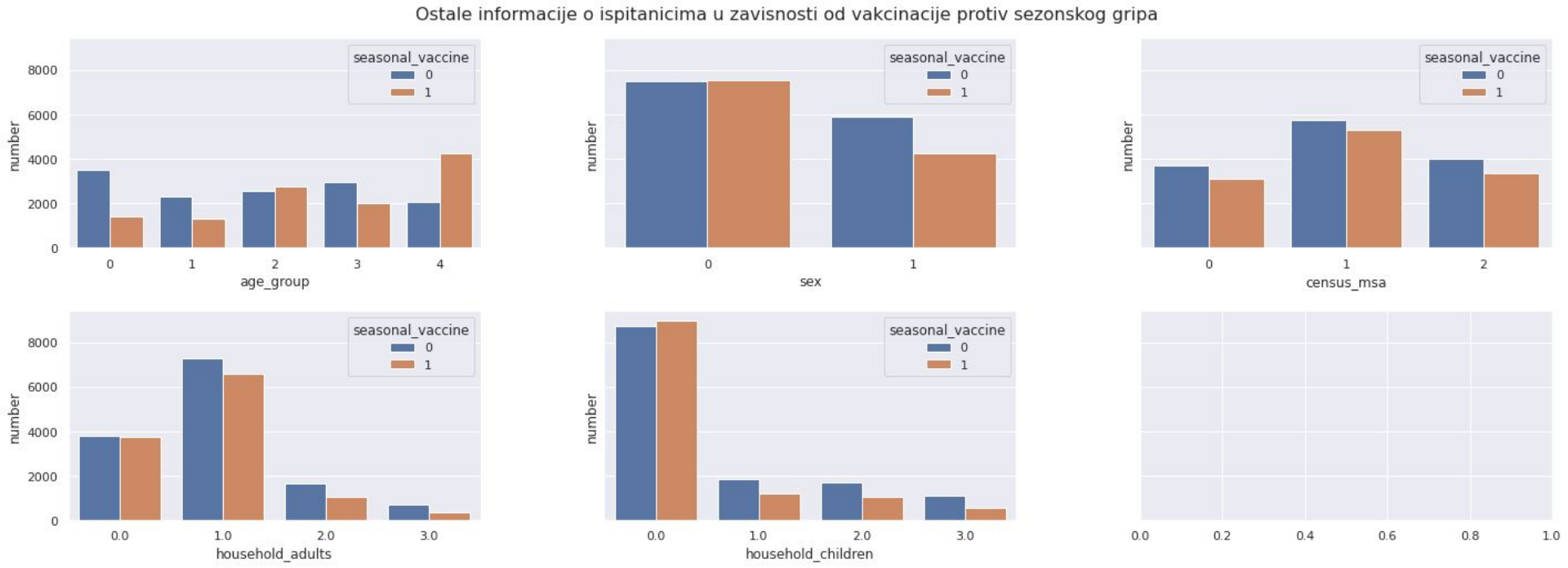
Uticaj ostalih prediktora | H1N1

Teško je uočiti trend u podacima koji bi sugerisao na opredeljenje za vakcinu

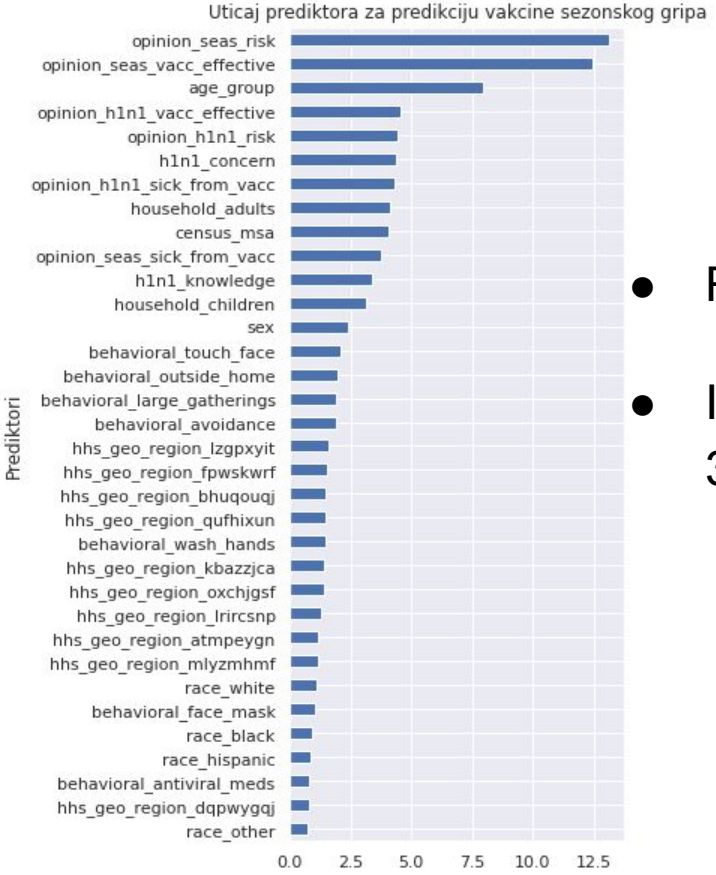


Uticaj ostalih prediktora | Sezonski grip

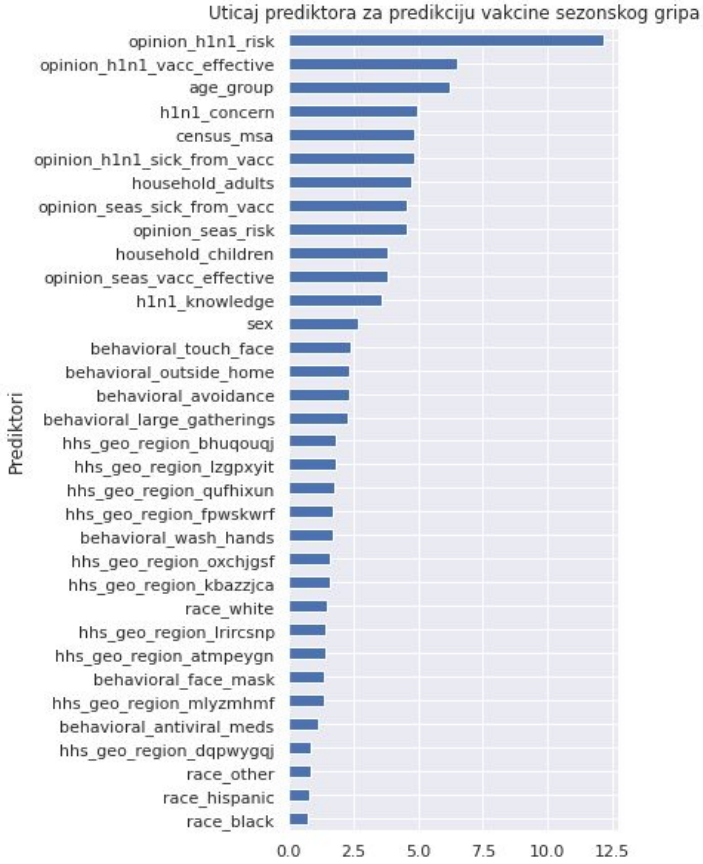
- Ispitanici sa preko 65 godina se najčešće opredeljuju za vakcinu
- Žene se nešto češće opredeljuju za vakcinu od muškaraca, a muškarci češće odbijaju vakcinu
- Ostali prediktori nisu naročito informativni



Značaj prediktora



- Random Forest
  - 200 estimatora
- Isti prediktori su u top 3 najinformativnijih



Izbor modela

Metodologija

- Podela inicijalnog skupa
  - *obučavajući skup : testirajući skup = 90% : 10% = 22714 : 2524*
- Standardizacija podataka
  - Učenje statističkih parametara na obučavajućem skupu
  - Transformacija prediktora obučavajućeg i testirajućeg skupa
- Unakrsna validacija
  - Obučavajući skup ⟶ 4 struka
  - Odnos zastupljenosti klasa kao u inicijalnim podacima
- Pretraga optimalnih hiper-parametara njihovim iscrpnim kombinovanjem
- Izbor na osnovu srednje balansirane tačnosti na validacionom skupu

[Napomena] Metodologija je nezavisno primenjena na obe labele!

Modeli i mogući hiper-parametri

Vakcina protiv sezononskog gripa

- **Logistička regresija**
  - 'penalty': ['l1','l2'],
  - 'C': [0.5, 1, 5, 10, 20],
  - 'solver': ['liblinear'],
  - 'class_weight': ['balanced', None]
- **Metoda nosećih vektora**
  - Linearni / Polinomijalni / Gausovski kernel
    - 'C': [1, 5, 10]
  - Polinomijalni kernel
    - 'degree': [2, 3, 5]
- **Random forest**
  - 'n_estimators': [50, 100, 200]
  - 'criterion': ['gini', 'entropy']
  - 'max_depth': [3, 5, 10]
  - 'max_features': ['sqrt','log2']

Vakcina protiv H1N1 virusa

- **Logistička regresija**
  - 'penalty': ['l1','l2'],
  - 'C': [0.5, 1, 5, 10, 20],
  - 'solver': ['liblinear'],
  - 'class_weight': ['balanced']
- **Metoda nosećih vektora**
  - Linearni / Polinomijalni / Gausovski kernel
    - 'C': [1, 5, 10]
    - 'class_weight': ['balanced']
  - Polinomijalni kernel
    - 'degree': [2, 3, 5]
- **Random forest**
  - 'n_estimators': [50, 100, 200]
  - 'criterion': ['gini', 'entropy']
  - 'max_depth': [3, 5, 10]
  - 'max_features': ['sqrt','log2']
  - 'class_weight': ['balanced']

Modeli i optimalni hiper-parametri

Vakcina protiv sezononskog gripa

- **Logistička regresija**
  - 'penalty': [***'l1'***,'l2'],
  - 'C': [0.5, 1, 5, ***10***, 20],
  - 'solver': [***'liblinear'***],
  - 'class_weight': ['balanced', ***None***]
- **Metoda nosećih vektora**
  - Linearni / Polinomijalni / ***Gausovski kernel***
    - 'C': [***1***, 5, 10]
  - Polinomijalni kernel
    - 'degree': [2, 3, 5]
- **Random forest**
  - 'n_estimators': [50, 100, ***200***]
  - 'criterion': ['gini', ***'entropy'***]
  - 'max_depth': [3, 5, ***10***]
  - 'max_features': ['sqrt',***'log2'***]

Vakcina protiv H1N1 virusa

- **Logistička regresija**
  - 'penalty': [***'l1'***,'l2'],
  - 'C': [0.5, 1, ***5***, 10, 20],
  - 'solver': [***'liblinear'***],
  - 'class_weight': [***'balanced'***]
- **Metoda nosećih vektora**
  - ***Linearni*** / Polinomijalni / Gausovski kernel
    - 'C': [***1***, 5, 10]
    - 'class_weight': [***'balanced'***]
  - Polinomijalni kernel
    - 'degree': [2, 3, 5]
- **Random forest**
  - 'n_estimators': [***50***, 100, 200]
  - 'criterion': [***'gini'***, 'entropy']
  - 'max_depth': [3, ***5***, 10]
  - 'max_features': ['sqrt',***'log2'***]
  - 'class_weight': [***'balanced'***]

Evaluacija modela sa optimalnim hiper-parametrima

Tumačenje:
precision: *Među svim ljudima za koje konačni model bude predvideo da [će se / se neće] vakcinisati, [<precision 1>/<precision 0>]% će se vakcinisati*
recall: *Među svim ljudima koji [će se / se neće] vakcinisati, model će predvideti [<recall 1>/<recall 0>]%.*

Vakcina protiv sezononskog gripa | *train*

Logistička regresija

	precision	recall	f1-score	support
0	0.76	0.78	0.77	12090
1	0.74	0.72	0.73	10624
accuracy			0.75	22714
macro avg	0.75	0.75	0.75	22714
weighted avg	0.75	0.75	0.75	22714

Metoda nosećih vektora

	precision	recall	f1-score	support
0	0.80	0.83	0.81	12090
1	0.80	0.76	0.78	10624
accuracy			0.80	22714
macro avg	0.80	0.79	0.80	22714
weighted avg	0.80	0.80	0.80	22714

Random forest

	precision	recall	f1-score	support
0	0.80	0.83	0.82	12090
1	0.80	0.76	0.78	10624
accuracy			0.80	22714
macro avg	0.80	0.80	0.80	22714
weighted avg	0.80	0.80	0.80	22714

Vakcina protiv H1N1 virusa | *train*

Logistička regresija

	precision	recall	f1-score	support
0	0.90	0.72	0.80	17858
1	0.41	0.70	0.52	4856
accuracy			0.72	22714
macro avg	0.65	0.71	0.66	22714
weighted avg	0.79	0.72	0.74	22714

Metoda nosećih vektora

	precision	recall	f1-score	support
0	0.89	0.75	0.82	17858
1	0.42	0.67	0.52	4856
accuracy			0.73	22714
macro avg	0.66	0.71	0.67	22714
weighted avg	0.79	0.73	0.75	22714

Random forest

	precision	recall	f1-score	support
0	0.90	0.72	0.80	17858
1	0.41	0.71	0.52	4856
accuracy			0.72	22714
macro avg	0.65	0.71	0.66	22714
weighted avg	0.79	0.72	0.74	22714

Evaluacija modela sa optimalnim hiper-parametrima

Tumačenje:
precision: *Među svim ljudima za koje konačni model bude predvideo da [će se / se neće] vakcinisati, [<precision 1>/<precision 0>]% će se vakcinisati*
recall: *Među svim ljudima koji [će se / se neće] vakcinisati, model će predvideti [<recall 1>/<recall 0>]%.*

Vakcina protiv sezononskog gripa | *test*

Logistička regresija

	precision	recall	f1-score	support
0	0.75	0.77	0.76	1344
1	0.73	0.71	0.72	1180
accuracy			0.74	2524
macro avg	0.74	0.74	0.74	2524
weighted avg	0.74	0.74	0.74	2524

Metoda nosećih vektora

	precision	recall	f1-score	support
0	0.75	0.79	0.77	1344
1	0.74	0.71	0.72	1180
accuracy			0.75	2524
macro avg	0.75	0.75	0.75	2524
weighted avg	0.75	0.75	0.75	2524

Random forest

	precision	recall	f1-score	support
0	0.76	0.78	0.77	1344
1	0.74	0.72	0.73	1180
accuracy			0.75	2524
macro avg	0.75	0.75	0.75	2524
weighted avg	0.75	0.75	0.75	2524

Vakcina protiv H1N1 virusa | *test*

Logistička regresija

	precision	recall	f1-score	support
0	0.89	0.70	0.78	1984
1	0.38	0.69	0.49	540
accuracy			0.70	2524
macro avg	0.64	0.69	0.64	2524
weighted avg	0.78	0.70	0.72	2524

Metoda nosećih vektora

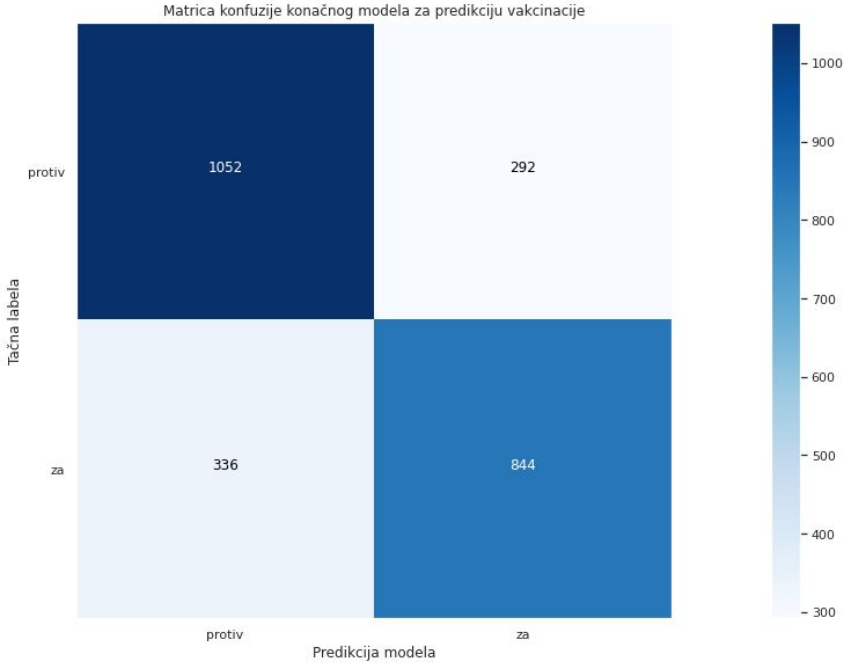
	precision	recall	f1-score	support
0	0.89	0.73	0.80	1984
1	0.40	0.66	0.50	540
accuracy			0.71	2524
macro avg	0.64	0.69	0.65	2524
weighted avg	0.78	0.71	0.73	2524

Random forest

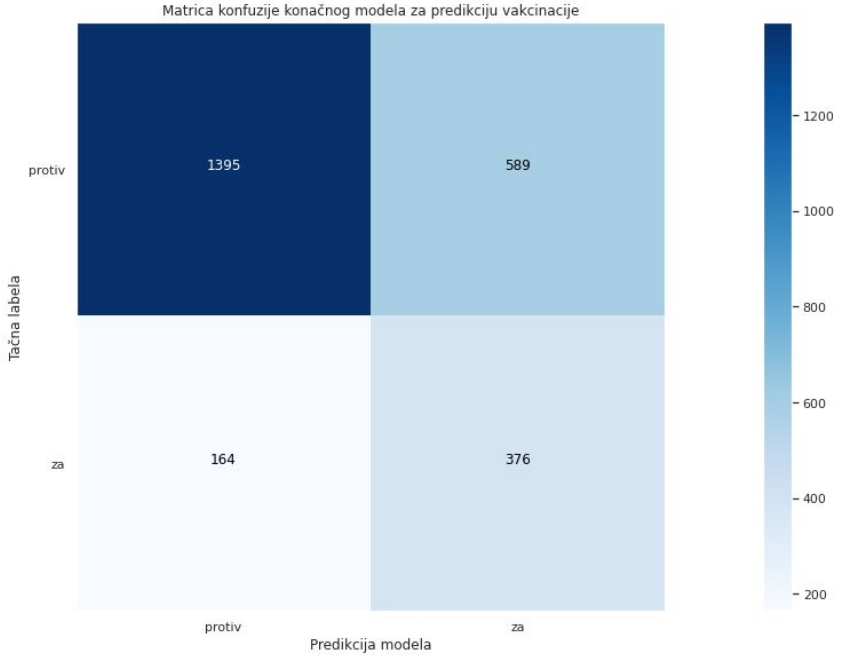
	precision	recall	f1-score	support
0	0.89	0.70	0.79	1984
1	0.39	0.70	0.50	540
accuracy			0.70	2524
macro avg	0.64	0.70	0.64	2524
weighted avg	0.79	0.70	0.73	2524

Matrica konfuzije na testirajućem skupu

Vakcina protiv sezononskog gripa



Vakcina protiv H1N1 virusa



Hvala na pažnji!