



# **Komparativna analiza algoritama klasifikacije**

---

Danica Đorđević 1121

# Komparativna analiza algoritama klasifikacije

U ovom radu je vršeno poređenje između sledećih algoritama:

Naivni Bajesov algoritam,

Algoritam K-najbližih suseda,

Algoritam zasnovan na stablima odluke (CART).



# Preprocesiranje

Atribut 'workclass' ima 8 jedinstvenih vrednosti  
Atribut 'education' ima 16 jedinstvenih vrednosti  
Atribut 'marital\_status' ima 7 jedinstvenih vrednosti  
Atribut 'occupation' ima 14 jedinstvenih vrednosti  
Atribut 'relationship' ima 6 jedinstvenih vrednosti  
Atribut 'race' ima 5 jedinstvenih vrednosti  
Atribut 'sex' ima 2 jedinstvenih vrednosti  
Atribut 'native\_country' ima 40 jedinstvenih vrednosti

Potencijalni kandidati za smanjenje broja jedinstvenih vrednosti su kolone *native\_country*, *education* i *occupation*.

United-States	5545
Mexico	138
?	108
Philippines	32
Canada	28
Germany	28
Puerto-Rico	23
Jamaica	22
China	18
South	18

Name: native\_country, dtype: int64

U zavisnosti od distribucija vrednosti, može se izvršiti smanjenje broja jedinstvenih vrednosti za atribut *native\_country* jer 90% ispitanika dolazi iz Amerike. Nove vrednosti za atribut su *United-States* i *Other*.

Prof-specialty	793
Exec-managerial	755
Adm-clerical	751
Craft-repair	741
Sales	717
Other-service	627
Machine-op-inspct	384
?	357
Transport-moving	311
Handlers-cleaners	255

Name: occupation, dtype: int64

HS-grad	2007
Some-college	1451
Bachelors	997
Masters	311
Assoc-voc	263
Assoc-acdm	205
11th	202
10th	187
7th-8th	134
Prof-school	112

Name: education, dtype: int64

Za attribute *education* i *occupation* se ne može izvršiti smanjenje broja jedinstvenih vrednosti jer ne postoji vrednost atributa koja predstavlja većinu.

# Preprocesiranje

```
<=50K    4721  
>50K     1483  
Name: income, dtype: int64
```

Klasni atribut prevesti u numericki oblik, gde >50K predstavlja 1, a <=50K predstavlja 0

```
def _replace_outcome_values(y):  
    for i in range(len(y)):  
        val = y[i].strip()  
        if val == ">50K":  
            y[i] = 1  
        else:  
            y[i] = 0  
    return y
```

Funkcija za zamenu vrednosti klasnog atributa

Kako modeli rade samo sa numeričim vrednostima, neophodno je ove attribute prevesti iz nenumeričkih u numeričke oblike. Biblioteka pandas nudi rešenje za ovaj problem, korišćenjem dummies atributa. Ovaj metod za svaku jedinstvenu vrednost koju kolona poseduje, pravi nove kolone, a staru kolonu briše. Tako, na primer, ako se primeni ova metoda na kolonu race, dobiće se 5 novih kolona sa binarnim vrednostima, dok će stara race kolona biti izbrisana.

	Amer-Indian-Eskimo	Asian-Pac-Islander	Black	Other	White
0	0	0	0	0	1
1	0	0	0	0	1
2	0	0	0	0	1
3	0	0	0	0	1
4	0	0	0	0	1

# Preprocesiranje

```
def _impute_missing_values(X):  
    si = SimpleImputer(missing_values=np.nan, strategy='mean')  
    si.fit(X)  
    X = pd.DataFrame(data=si.transform(X), columns=X.columns)  
    return X
```

Zamena nedostajućih vrednosti srednjom vrednošću

Dodavanje interakcija

	age	fnlwgt	education_num	capital_gain	...	race_ White_sex_ Male	race_ White_native_country_Other	sex_ Female_native_country_Other	sex_ Male_native_country_Other
0	38.0	234962.0	9.0	2829.0	...	1.0	1.0	0.0	1.0
1	72.0	177226.0	9.0	0.0	...	1.0	1.0	0.0	1.0
2	31.0	259931.0	4.0	0.0	...	1.0	1.0	0.0	1.0
3	55.0	189528.0	4.0	0.0	...	1.0	1.0	0.0	1.0
4	38.0	34996.0	10.0	0.0	...	0.0	1.0	1.0	0.0

[5 rows x 1630 columns]

# Preprocesiranje

```
----- TREE -----
'AUC modela sa preprocesiranjem: 0.7283756413718606'
'AUC modela bez preprocesiranja: 0.6694099378881988'
'Poboljsanje modela: 8.808608917531478%'
('----- BAYES Multinomial -----')
'AUC modela sa preprocesiranjem: 0.3620375371320551'
'AUC modela bez preprocesiranja: 0.5723737510126924'
'Poboljsanje modela: -36.74805378627034%'
('----- K NEIGHBORS -----')
'AUC modela sa preprocesiranjem: 0.8094652984066972'
'AUC modela bez preprocesiranja: 0.5988320280853363'
'Poboljsanje modela: 35.17401549059175%'
('----- TREE PCA -----')
'AUC modela sa preprocesiranjem: 0.7408385093167702'
'AUC modela bez preprocesiranja: 0.6777950310559007'
'Poboljsanje modela: 9.301260022909503%'
('----- BAYES Multinomial PCA -----')
'AUC modela sa preprocesiranjem: 0.833918444504456'
'AUC modela bez preprocesiranja: 0.6021874156089656'
'Poboljsanje modela: 38.48154625767976%'
('----- K NEIGHBORS PCA -----')
'AUC modela sa preprocesiranjem: 0.6005941128814475'
'AUC modela bez preprocesiranja: 0.5988320280853363'
'Poboljsanje modela: 0.2942535992513862%'
```

Često ulazne karakteristike interaguju na neočekivan i nelinearni način prilikom prediktivnog modeliranja. Te interakcije se mogu identifikovati i modelirati pomoću algoritma za učenje. Drugi pristup je osmišljavanje novih karakteristika koje uočavaju ove interakcije i utvrđivanje da li poboljšavaju performanse modela.

Performanse klasifikatora su znatno gore kod naivnog Bajesovog algoritma, ako se ne obavi korak dodavanja interakcija.



# Preprocesiranje – Redukcija dimanzionalnosti k=10

```
def _pca(X):  
    pca = PCA(n_components=10)  
    X_pca = pd.DataFrame(pca.fit_transform(X))  
    return X_pca
```

Korišćenje PCA metode

	0	1	2	3	4	5	6	7	8	9
0	4.603680e+08	-1.565683e+07	7.088114e+05	1.704761e+06	1.473788e+05	-52838.823539	233399.450121	153546.197217	167612.864361	-86417.009491
1	-2.043396e+08	-1.651429e+07	-5.935961e+05	8.347809e+06	1.072941e+05	-685.779770	162372.075717	86929.540199	-19285.642779	-80394.262033
2	-2.043383e+08	-1.648310e+07	2.642779e+06	-1.177344e+06	-1.268122e+06	-27744.038930	209406.504532	124026.132287	48292.490478	96700.176617
3	-2.043374e+08	-1.646142e+07	4.896327e+06	-9.190717e+04	-1.858992e+06	-32092.377953	93546.539915	42950.529469	-38964.830372	68647.632896
4	-2.043425e+08	-1.659134e+07	-8.516450e+06	1.705205e+04	-2.587862e+05	-45291.892752	1504.131647	-54915.983056	3047.100899	24799.014780
...	...	...	...	...	...	...	...	...	...	...
6199	-2.043395e+08	-1.651589e+07	-7.595636e+05	-3.905519e+06	-2.964177e+05	-35100.408105	-190262.990397	-66513.436771	128721.817277	-60992.357319
6200	-2.043389e+08	-1.649903e+07	9.907504e+05	2.198097e+06	-4.030509e+05	-29237.665461	-221283.534397	-33696.082799	74679.528787	-74820.967918
6201	-2.043352e+08	-1.640423e+07	1.075960e+07	2.442868e+06	-9.611259e+05	-12889.197566	-225086.311551	177070.498337	-209992.013130	-88907.992400
6202	-2.043399e+08	-1.652669e+07	-1.881780e+06	-2.694645e+06	1.389111e+05	-33471.930392	-218206.516380	6130.424226	65592.354062	-130637.564220
6203	-2.043373e+08	-1.645879e+07	5.129190e+06	5.330803e+04	-5.904585e+05	-20196.586251	-116003.375779	176838.162313	-13040.662382	-122795.358481

# Preprocesiranje – Redukcija dimanzionalnosti k=10

Selekcija K najrelevantnijih atributa

```
select = SelectKBest(f_classif, k=10)
selected_features = select.fit(X_train, y_train)
indices_selected = selected_features.get_support(indices=True)
colnames_selected = [X.columns[i] for i in indices_selected]
```

```
X_train_selected = X_train[colnames_selected] 'Selected columns: 10'
X_test_selected = X_test[colnames_selected] ['marital_status_Married-civ-spouse',
'age_marital_status_Married-civ-spouse',
'education_num_marital_status_Married-civ-spouse',
'education_num_relationship_Husband',
'hours_per_week_marital_status_Married-civ-spouse',
'hours_per_week_relationship_Husband',
'marital_status_Married-civ-spouse_relationship_Husband',
'marital_status_Married-civ-spouse_race_White',
'marital_status_Married-civ-spouse_native_country_Other',
'relationship_Husband_sex_Male']
```




# Preprocesiranje

```
X_train, X_test, y_train, y_test = train_test_split(  
    X, y, train_size=0.9, random_state=1)
```

← Kros validacija

Priprema neobrađenog skupa podataka, na osnovu koga će se vršiti poređenje. Iz skupa neprocesiranih podataka su izbačene kolone sa nenumeričkim podacima, kao i redovi sa nedostajućim vrednostima. Nenumeričke podatke je neophodno izbaciti, da bi primenjivanje bilo kog algoritma bilo moguće.



	age	fnlwgt	education_num	capital_gain	capital_loss	hours_per_week
0	38	234962	9	2829	0	30
1	72	177226	9	0	0	8
2	31	259931	4	0	0	40
3	55	189528	4	0	0	60
4	38	34996	10	0	0	40



# Algoritmi klasifikacije – sklearn biblioteka

```
def _find_model_performance_naive_bayes(X_train, y_train, X_test, y_test):  
    classifier = MultinomialNB() ←  
    classifier.fit(X_train, y_train)  
    y_hat = [x[1] for x in classifier.predict_proba(X_test)]  
    auc = roc_auc_score(y_test, y_hat)  
    return auc
```

```
def _find_model_performance_tree(X_train, y_train, X_test, y_test):  
    classifier = DecisionTreeClassifier() ←  
    classifier.fit(X_train, y_train)  
    y_hat = [x[1] for x in classifier.predict_proba(X_test)]  
    auc = roc_auc_score(y_test, y_hat)  
  
    return auc
```

```
def _find_model_performance_k_neighbors(X_train, y_train, X_test, y_test):  
    classifier = KNeighborsClassifier() ←  
    classifier.fit(X_train, y_train)  
    y_hat = [x[1] for x in classifier.predict_proba(X_test)]  
    auc = roc_auc_score(y_test, y_hat)  
  
    return auc
```



# Komparativna analiza algoritama k=10

## Selekcija K najrelevantnijih atributa

## PCA metoda

```
'Sa 10 atributa'
'----- TREE -----'
'AUC modela sa preprocesiranjem: 0.6567647853092088'
'AUC modela bez preprocesiranja: 0.6734472049689441'
'Poboljsanje modela: -2.477168148690227%'
'-----'
'----- BAYES Multinomial -----'
'AUC modela sa preprocesiranjem: 0.543532271131515'
'AUC modela bez preprocesiranja: 0.5723737510126924'
'Poboljsanje modela: -5.038924274593053%'
'-----'
'----- K NEIGHBORS -----'
'AUC modela sa preprocesiranjem: 0.7697137456116663'
'AUC modela bez preprocesiranja: 0.5988320280853363'
'Poboljsanje modela: 28.535834676828365%'
'-----'
```

```
'----- TREE PCA -----'
'AUC modela sa preprocesiranjem: 0.6970496894409938'
'AUC modela bez preprocesiranja: 0.671583850931677'
'Poboljsanje modela: 3.791907514450859%'
'-----'
'----- BAYES Multinomial PCA -----'
'AUC modela sa preprocesiranjem: 0.7812179314069674'
'AUC modela bez preprocesiranja: 0.6021874156089656'
'Poboljsanje modela: 29.730032736894064%'
'-----'
'----- K NEIGHBORS PCA -----'
'AUC modela sa preprocesiranjem: 0.7480556305698083'
'AUC modela bez preprocesiranja: 0.5988320280853363'
'Poboljsanje modela: 24.919108445416505%'
'-----'
```

# Evaluacija algoritama

Za evaluaciju algoritama je korišćena AUC (Area Under the ROC Curve) mera. AUC meri celokupno dvodimenzionalno područje ispod cele ROC krive od (0,0) do (1,1). Model čija su predviđanja 100% pogrešna ima AUC od 0,0. Onaj model čija su predviđanja 100% tačna ima AUC 1,0.

```
def _find_model_performance_k_neighbors(X_train, y_train, X_test, y_test):  
    classifier = KNeighborsClassifier()  
    classifier.fit(X_train, y_train)  
    y_hat = [x[1] for x in classifier.predict_proba(X_test)]  
    auc = roc_auc_score(y_test, y_hat)  
  
    return auc
```

Korišćena je  
*roc\_auc\_score* funkcija  
iz *sklearn* biblioteke

```
performanse_improve = (  
    (performanse_processed - performanse_unprocessed) / performanse_unprocessed) * 100
```

Računjanje procenta  
poboljšanja klasifikatora

# Preprocesiranje – Redukcija dimanzionalnosti k=16

```
def _pca(X):  
    pca = PCA(n_components=16)  
    X_pca = pd.DataFrame(pca.fit_transform(X))  
    return X_pca
```

Korišćenje PCA metode

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	4.603680e+08	-1.565683e+07	7.088114e+05	1.704761e+06	1.473788e+05	-52838.823127	233399.496440	153546.708704	167604.980927	-86420.056647	-38942.153503	-43984.884917	12032.053810	35058.607836	-28971.831736	-3417.069975
1	-2.043396e+08	-1.651429e+07	-5.935961e+05	8.347809e+06	1.072941e+05	-685.780065	162372.179690	86931.692867	-19316.273042	-80436.770828	91172.456739	-167399.884897	-61598.178729	-14375.627061	-31402.342408	59575.311803
2	-2.043383e+08	-1.648310e+07	2.642779e+06	-1.177344e+06	-1.268122e+06	-27744.039109	209406.544139	124027.164268	48285.143514	96669.761873	-110149.395997	31937.304111	14709.951132	-92977.475535	-28646.737639	-56047.487296
3	-2.043374e+08	-1.646142e+07	4.896327e+06	-9.190717e+04	-1.858992e+06	-32092.379474	93546.497390	42951.855006	-38962.383915	68600.277555	-50281.223531	38068.110096	9434.404452	-76567.069523	-7843.983161	-75106.540239
4	-2.043425e+08	-1.659134e+07	-8.516450e+06	1.705205e+04	-2.587862e+05	-45291.893158	1504.113576	-54915.379493	3048.554731	24784.844598	-1759.921718	24834.710752	-3744.821129	12172.179321	21431.922709	-7939.900602
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
6199	-2.043395e+08	-1.651589e+07	-7.595636e+05	-3.905519e+06	-2.964177e+05	-35100.409629	-190262.979735	-66509.433353	128709.983198	-61080.424598	-19507.315513	-58770.149717	18653.415971	-60433.382312	19536.669933	-103551.603069
6200	-2.043389e+08	-1.649903e+07	9.907504e+05	2.198097e+06	-4.030509e+05	-29237.665059	-221283.542208	-33696.920410	74682.052125	-74796.775264	-73237.107947	-1893.380768	-711.240868	30913.626292	-95949.855222	2526.792676
6201	-2.043352e+08	-1.640423e+07	1.075960e+07	2.442868e+06	-9.611259e+05	-12889.195594	-225086.102480	177068.683429	-210029.334214	-88864.299892	-225799.892508	176272.148866	-44891.537884	97896.073480	157824.804183	-2705.562305
6202	-2.043399e+08	-1.652669e+07	-1.881780e+06	-2.694645e+06	1.389111e+05	-33471.930169	-218206.461894	6131.672124	65579.249040	-130657.523862	-91270.182650	-29706.883707	14126.221718	-17092.181480	-158174.437808	-91692.467035
6203	-2.043373e+08	-1.645879e+07	5.129190e+06	5.330803e+04	-5.904585e+05	-20196.586339	-116003.378046	176839.196422	-13042.081277	-122807.333868	34272.145529	-16928.148753	-3926.072989	-56226.589396	28965.557837	-6083.871841

# Preprocesiranje – Redukcija dimanzionalnosti k=16

```
select = sklearn.feature_selection.SelectKBest(f_classif,k=16)
selected_features = select.fit(X_train, y_train)
indices_selected = selected_features.get_support(indices=True)
colnames_selected = [X.columns[i] for i in indices_selected]
```

```
X_train_selected = X_train[colnames_selected]
X_test_selected = X_test[colnames_selected]
```

Selekcija K najrelevantnijih atributa

```
'Selected columns: 16'
['marital_status_Married-civ-spouse',
 'relationship_Husband',
 'age_education_num',
 'age_marital_status_Married-civ-spouse',
 'age_relationship_Husband',
 'education_num_marital_status_Married-civ-spouse',
 'education_num_relationship_Husband',
 'hours_per_week_marital_status_Married-civ-spouse',
 'hours_per_week_relationship_Husband',
 'marital_status_Married-civ-spouse_relationship_Husband',
 'marital_status_Married-civ-spouse_race_White',
 'marital_status_Married-civ-spouse_sex_Male',
 'marital_status_Married-civ-spouse_native_country_Other',
 'relationship_Husband_race_White',
 'relationship_Husband_sex_Male',
 'relationship_Husband_native_country_Other']
```

# Komparativna analiza algoritama k=16

## Selekcija K najrelevantnijih atributa

## PCA metoda

```
'Sa 16 atributa'  
'----- TREE -----'  
'AUC modela sa preprocesiranjem: 0.7287537132055091'  
'AUC modela bez preprocesiranja: 0.6798136645962731'  
'Poboljsanje modela: 7.199038671618996%'  
'-----'  
'----- BAYES Multinomial -----'  
'AUC modela sa preprocesiranjem: 0.7244733999459898'  
'AUC modela bez preprocesiranja: 0.5723737510126924'  
'Poboljsanje modela: 26.57348431233783%'  
'-----'  
'----- K NEIGHBORS -----'  
'AUC modela sa preprocesiranjem: 0.8059006211180123'  
'AUC modela bez preprocesiranja: 0.5988320280853363'  
'Poboljsanje modela: 34.57874384153143%'  
'-----'
```

```
'----- TREE PCA -----'  
'AUC modela sa preprocesiranjem: 0.7263975155279503'  
'AUC modela bez preprocesiranja: 0.6746894409937888'  
'Poboljsanje modela: 7.66398158803222%'  
'-----'  
'----- BAYES Multinomial PCA -----'  
'AUC modela sa preprocesiranjem: 0.7942209019713745'  
'AUC modela bez preprocesiranja: 0.6021874156089656'  
'Poboljsanje modela: 31.88932239113863%'  
'-----'  
'----- K NEIGHBORS PCA -----'  
'AUC modela sa preprocesiranjem: 0.7508371590602214'  
'AUC modela bez preprocesiranja: 0.5988320280853363'  
'Poboljsanje modela: 25.383600717031733%'  
'-----'
```



# Komparativna analiza algoritama - PCA

16 atributa

```
'----- TREE PCA -----'  
'AUC modela sa preprocesiranjem: 0.7263975155279503'  
'AUC modela bez preprocesiranja: 0.6746894409937888'  
'Poboljsanje modela: 7.66398158803222%'  
'-----'  
'----- BAYES Multinomial PCA -----'  
'AUC modela sa preprocesiranjem: 0.7942209019713745'  
'AUC modela bez preprocesiranja: 0.6021874156089656'  
'Poboljsanje modela: 31.88932239113863%'  
'-----'  
'----- K NEIGHBORS PCA -----'  
'AUC modela sa preprocesiranjem: 0.7508371590602214'  
'AUC modela bez preprocesiranja: 0.5988320280853363'  
'Poboljsanje modela: 25.383600717031733%'  
'-----'
```

10 atributa

```
'----- TREE PCA -----'  
'AUC modela sa preprocesiranjem: 0.6970496894409938'  
'AUC modela bez preprocesiranja: 0.671583850931677'  
'Poboljsanje modela: 3.791907514450859%'  
'-----'  
'----- BAYES Multinomial PCA -----'  
'AUC modela sa preprocesiranjem: 0.7812179314069674'  
'AUC modela bez preprocesiranja: 0.6021874156089656'  
'Poboljsanje modela: 29.730032736894064%'  
'-----'  
'----- K NEIGHBORS PCA -----'  
'AUC modela sa preprocesiranjem: 0.7480556305698083'  
'AUC modela bez preprocesiranja: 0.5988320280853363'  
'Poboljsanje modela: 24.919108445416505%'  
'-----'
```



# Komparativna analiza algoritama – selekcija k

## 16 atributa

```
'Sa 16 atributa'  
'----- TREE -----'  
'AUC modela sa preprocesiranjem: 0.7287537132055091'  
'AUC modela bez preprocesiranja: 0.6798136645962731'  
'Poboljsanje modela: 7.199038671618996%'  
'-----'  
'----- BAYES Multinomial -----'  
'AUC modela sa preprocesiranjem: 0.7244733999459898'  
'AUC modela bez preprocesiranja: 0.5723737510126924'  
'Poboljsanje modela: 26.57348431233783%'  
'-----'  
'----- K NEIGHBORS -----'  
'AUC modela sa preprocesiranjem: 0.8059006211180123'  
'AUC modela bez preprocesiranja: 0.5988320280853363'  
'Poboljsanje modela: 34.57874384153143%'  
'-----'
```

## 10 atributa

```
'Sa 10 atributa'  
'----- TREE -----'  
'AUC modela sa preprocesiranjem: 0.6567647853092088'  
'AUC modela bez preprocesiranja: 0.6734472049689441'  
'Poboljsanje modela: -2.477168148690227%'  
'-----'  
'----- BAYES Multinomial -----'  
'AUC modela sa preprocesiranjem: 0.543532271131515'  
'AUC modela bez preprocesiranja: 0.5723737510126924'  
'Poboljsanje modela: -5.038924274593053%'  
'-----'  
'----- K NEIGHBORS -----'  
'AUC modela sa preprocesiranjem: 0.7697137456116663'  
'AUC modela bez preprocesiranja: 0.5988320280853363'  
'Poboljsanje modela: 28.535834676828365%'  
'-----'
```

# Zaključak



**Hvala na pažnji.**

