UNIVERZITET U NOVOM SADU
PRIRODNO-MATEMATIČKI FAKULTET
DEPARTMAN ZA
MATEMATIKU I INFORMATIKU

Bojana Simić

Analiza velikih podataka

- seminarski rad -

Novi Sad, *2021*.

Sadržaj

1.Predgovor

Svake godine milioni životinja se nađu na ulici. Zbog malog interesovanja za udomljavanje, prihvatilišta se nalaze u problemu kako zbrinuti sve te životinje. Većina prihvatilišta su primorana da ubijaju životinje koje ne pronađu dom, kako bi mogli da sklone druge životinje sa ulice. Međutim, ova praksa ne važi za sve. Prihvatilište u Ostinu, Teksas, Sjedinjene Američke Države, je jedno od njih. Pored toga što je najveće prihvatilište koje ne ubija svoje štićenike, on dostavlja podatke o njima. U ovom radu je odrađena analiza skupa podataka tog prihvatilišta u Ostinu eng. Austin Animal Center Shelter Outcomes preuzeta sa kaggle.com. Udomljavanje mačaka je uspešnije i ima više podataka o njima, što me je navelo da odradim analizu samo za pse. Cilj ove analize je da na osnovu prikupljenih podataka predvidimo da li će pas biti usvojen ili ne.

Novi Sad, *22.09.2021.* *Bojana Simić 319m/19*

2. Analiza i predprocesiranje podataka

Za realizaciju ovog rada korišćeno je *Google Colab* okruženje i *PySpark* interfejs *Apache Spark* platforme. Korišćen je programski jezik *Python* i njegove biblioteke za obradu i vizuelizaciju podataka: *Pandas, Numpy, Matplotlib* i *Seaborn*. Implementacija ove analize nalazi u .ipynb datoteci, u prilogu ovog rada.

U ovom skupu podataka se nalazi 78256 instanci i 12 atributa, pritom sve vrednosti su kategoričke.

```
root
 |-- age_upon_outcome: string (nullable = true)
 |-- animal_id: string (nullable = true)
 |-- animal_type: string (nullable = true)
 |-- breed: string (nullable = true)
 |-- color: string (nullable = true)
 |-- date_of_birth: string (nullable = true)
 |-- datetime: string (nullable = true)
 |-- monthyear: string (nullable = true)
 |-- name: string (nullable = true)
 |-- outcome_subtype: string (nullable = true)
 |-- outcome_type: string (nullable = true)
 |-- sex_upon_outcome: string (nullable = true)
```

Slika1. Prikaz skupa atributa

Pre nego što nastavimo sa obradom podataka potrebno je izvršiti analizu i predprocesiranje podataka. Želimo da analiziramo udomljavanje pasa, tako da ćemo najpre iz kolone `animal_type` izdvojiti samo pse. Zatim, kolone `animal_id` i `name` ne sadrže podatke potrebne za ovu analizu. Takođe, kolone `datetime` i `monthyear` su identične, pritom ukoliko oduzmemo vrednosti iz ovih kolona i kolone `date_of_birth` dobijamo starost pasa koja je već predstavljena u koloni `age_upon_outcome`, tako da možemo ih izbrisati, tačnije selektovati sve ostale. Kolona `outcome_subtype` sadrži veliki bro nedostajućih vrednosti i pritom najčešće za udomljene životinje, što znači da nam nije od koristi za ovo istraživanje, pa ćemo je ukloniti. Ostaju nam atributi:

- `age_upon_outcome` - starost psa u vreme kada je napusto prihvatilište
- `breed` - rasa psa
- `color` - boja krzna psa
- `outcome_type` - ishod
- `sex_upon_outcome` - pol psa u trenutku kad je napustio prihvatilište

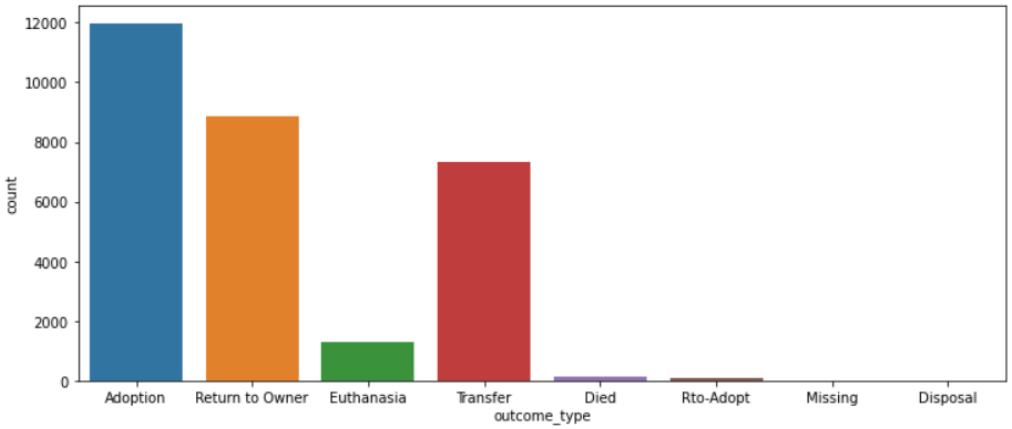
Nakon što smo suzili broj atributa primećeno je da ima duplikata. Najpre smo uklnonili duplikate,a zatim uklonili nedostajuće vrednosti kojih ima svega 4.

```
Broj instanci u skupu podataka:  44242
Broj instanci nakon uklanjanja duplikata:  29798
Broj instanci nakon uklanjanja NULL vrednosti:  29794
```

Slika 2. Prikaz duplikata i nedostajućih vrednosti

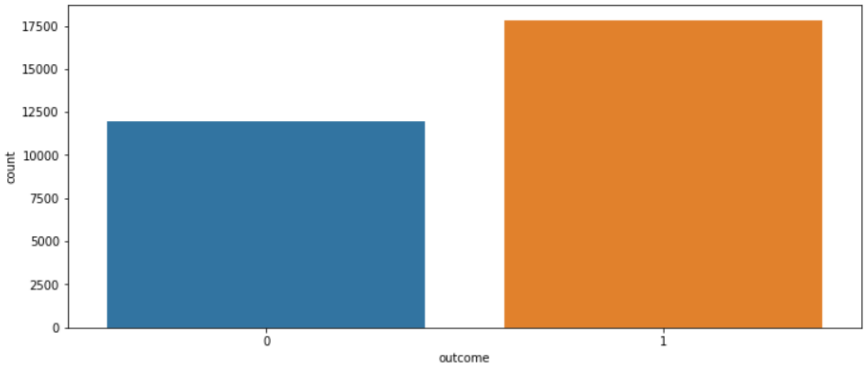
3.Vizuelizacija podataka

Za dalju analizu podataka koristićemo vizelizaciju. Počećemo od kolone za ishod.



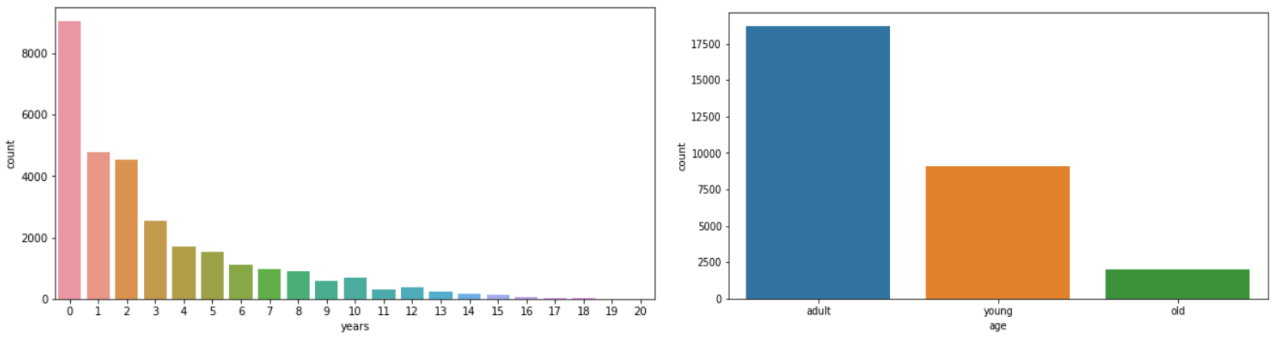
Slika 3. Pregled broja istanci za ishod

Primetimo da pored informacije da je pas usvojen postoje i druge vrednosti kao na primer vraćen vlasniku, nestao, preminuo itd. Sve pse koji su usvojeni postavili smo u jednu kategoriju i dodelili vrednost 0, a ostale smo stavili u drugu kategoriju, vrednost 1.



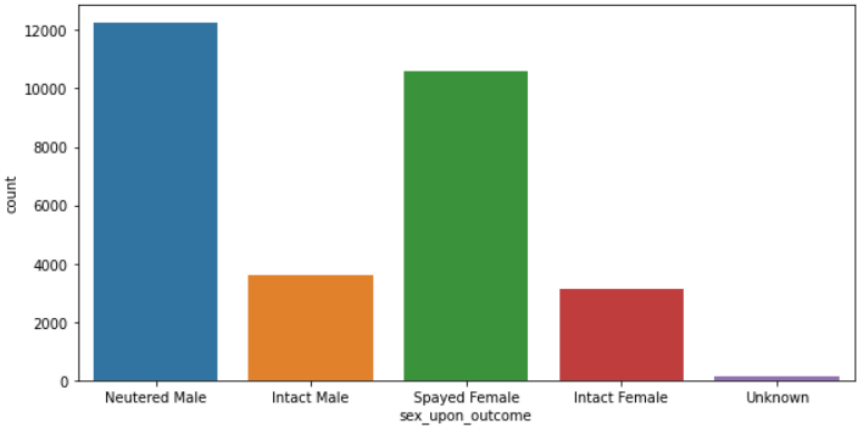
Slika 4. Pregled broja istanci za ishod nakon svrstavanja u kategorije

Dalje, za kolonu koja opisuje starost imamo nekoliko različtih vrednosti, od starih nekoliko dana do starih nekoliko godina. Ukoliko postoji reč years, znači da pas je star jednu ili više godina, tako da ćemo samo skloniti years iz atributa. Ukoliko ne postoji reč years znači da pas nije star ni godinu dana pa ćemo mu dodeliti vrednost 0. Sada, na osnovu dobijenih brojeva, pse prema starosti ubacujemo u kategorije young (mlad pas koji nije star ni godinu dana), adult (odrasli pas, star između 1 i 10 godina) i old (star pas, ima više od 10 godina). Nakon sređivanja podataka, podaci izgledaju kao na slici 5. Pregled broja istanci i vrednosti za pol životinje prikazan je na slici 6.



Slika 5. Pregled broja istanci za starost

Pregled broja istanci i vrednosti za pol životinje prikazan je na sledećoj slici



Slika 6. Pregled broja istanci za pol

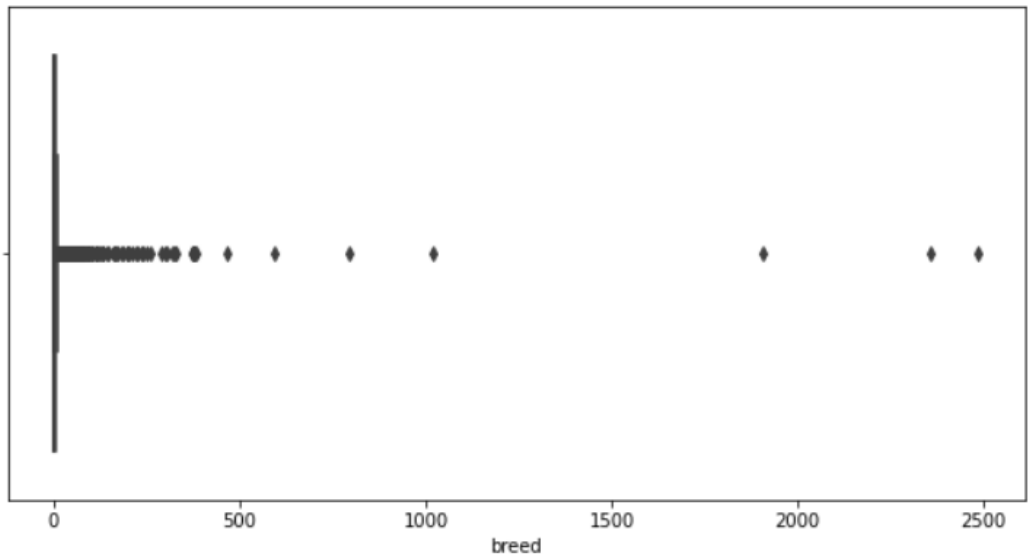
Ostale su još dve kolone za analizu: rasa i boja. Međutim, njihovom analizom smo primetili da postoji veliki broj vrednosti koji one uzimaju.

```
Pit Bull Mix                          2486
Chihuahua Shorthair Mix               2357
Labrador Retriever Mix                1907
German Shepherd Mix                   1021
Australian Cattle Dog Mix              792
                                      ...
Blue Lacy/Australian Cattle Dog         1
Chinese Sharpei/Boxer                   1
Black Mouth Cur/Staffordshire           1
Labrador Retriever/Chihuahua Shorthair  1
Afghan Hound/Labrador Retriever         1
Name: breed, Length: 1893, dtype: int64
```

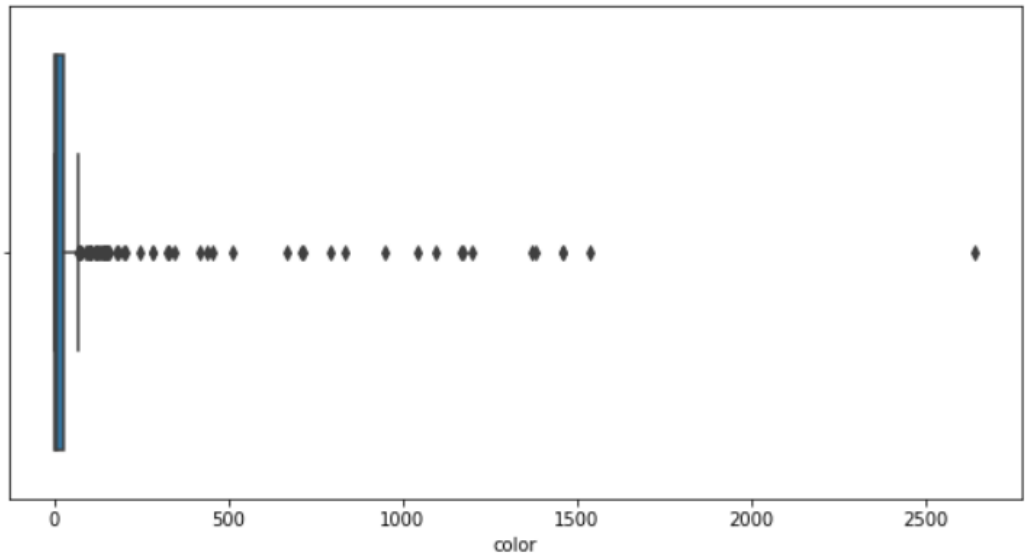
```
Black/White       2644
White             1540
Brown/White       1461
Tan/White         1461
Tricolor          1381
                  ...
Yellow/Yellow        1
Orange               1
Ruddy/Cream          1
Gray/Tricolor        1
Silver/Blue          1
Name: color, Length: 336, dtype: int64
```

Slika 7. Pregled broja istanci za rasu i boju

Preciznije, 1893 vrednosti za rasu i 336 za boju. Ovo znači da postoji mogućnost prisustva *outlier* tj. ekstremnih vrednosti koje moramo rešiti pre nego sto nastavimo sa analizom. Koristeći boxplot grafikon prikazali smo distribuciju vrednosti.



Slika 8. Boxplot grafikon za rasu



Slika 8. Boxplot grafikon za boju krzna

Grafikon ukazuje na prisustvo o*utlier* vrednosti, i u praksi se one najčešće uklanjaju. Međutim u našem slučaju uklonili bi ne samo veliku količinu podataka već podatke koje se javljaju najčešće, ostavljajući najređe podatke, tako da mi nećemo ukloniti outlier vrednosti.

Na samom kraju izbacili smo kolonu `age_upon_outcome` i `years` zato što nam nisu više potrebne.

4.Treniranje modela

Kategoričke atribute `age`, `breed`, `color` i `sex_upon_outcome` smo pretvorili u numeričke korišćenjem estimatora *StringIndexer* i *OneHotEncoder*. Zatim uz pomoc transformers *VectorAssembler* sve vrednosti dobijene OneHotEncoder-om smo spojili u jedan vektor predstavljen kolonom features.

Za potrebe mašinskog učenja koristili smo algoritme *RandomForestClassifier*, *LogisticRegression* i *DecisionTreeClassifier*.

Prilikom kreiranja modela koristili smo *pipeline* koji se sastoji od gore pomenutog vektora i algoritma za mašinsko učenje.

Skup podataka smo podelili na trening i test skup u odnosu 80-20. Nakon podele u trening skupu se nalaze 23 631, dok u test skup ima 6131 podataka.

5.Evaluacija modela

Za evaluaciju modela koristili smo *BinaryClassificationEvaluator* i dobili rezultate:

```
Metrika RandomForestClassifier:  areaUnderROC   0.6944031218512405
Metrika LogisticRegression:  areaUnderROC   0.7372135854147365
Metrika DecisionTreeClassifier:  areaUnderROC   0.5431429351252544
```

Model *RandomForestClassifier* pogađa sa tačnošću 69%, bolju tačnost daje *LogisticRegression* sa tačnošću 73% , dok najlošije rezultate daje *DecisionTreeClassifier* svega 54%.

Iz rezultata se može videti da Logistic Regression daje najbolje rezultate, međutim njih smo dobili koristeći default vrednosti. Sledeći korak je da pokušamo da dobijemo bolje rezultate menjajući hiperparametre i koristeći *CrossValidator* sa 5 foldova. Nakon ove evaluacije dobili smo sledeće rezultate:

```
RandomForestClassifier - Cross Validation: areaUnderROC : 0.7430372623684711
LogisticRegression - Cross Validation: areaUnderROC : 0.7381627567526373
DecisionTreeClassifier - Cross Validation: areaUnderROC : 0.5709321495869796
```

Posmatrajući rezultate, poboljšali smo tačnost modela *RandomForestClassifier* za 5% i *DecisionTreeClassifier* za 3%. Model *LogisticRegression* je poboljšan za 0,001%. Sada rezultati pokazuju da je najbolji model *RandomForestClassifier* sa tačnošću 74%, zatim *LogisticRegression* sa tačnošću 73% i na kraju se nalazi *DecisionTreeClassifier* sa tačnošću 57%.

6.Zaključak

U prethodnim sekcijama prikazana je analiza Austin Animal Center Shelter Outcomes skupa podataka. Najpre smo analizom i vizuelizacijom sredili podatke da bi ih iskoristili za kreiranje i treniranje tri modela mašinskog učenja *RandomForestClassifier, LogisticRegression* i *DecisionTreeClassifier.* Zatim smo izvršili evaluaciju modela i došli do zaključka da *RandomForestClassifier* daje najbolje rezultate kada menjamo hiperparametre, dok *LogisticRegression* daje najbolje rezultate za default vrednosti. U oba slučaja *DecisionTreeClassifier* daje najlošije rezultate. Važno je napomenuti da default vrednosti su veoma dobri za naš model *LogisticRegression*, menjajući hiperparametre smo poboljšali rezultate tek za 0,001%. Na samom kraju, model i pipeline je moguće sačuvati za potencijalnu upotrebu u budućnosti, što smo i uradili.