Predlog projekta iz SIAP-a

Predikcija usvajanja životinja iz *Austin Animal Shelter*-a

DEFINICIJA PROJEKTA

Osnovni zadatak projekta jeste analiza podataka o prihvaćenim i udomljenim životinjama iz *Autstin Animal Shelter*-a u Teksasu. Analizom istih, želimo da izvršimo predikciju „udomljivosti“ pristiglih životinja u budućem periodu, koji će biti kasnije identifikovan praćenjem trendova u aktuelnim podacima. Naime, cilj je da za pristigle životinje na osnovu njihovih karakteristika kao što su rasa, veličina, fizičke karakteristike, karakter, starost predvidimo u kom će vremenskom periodu biti udomljena (0-6 meseci, 6-12 meseci, i slično).

Dodatno, želimo da eksplorativnom analizom utvrdimo koje to karakteristike igraju najveću ulogu u povećanju šanse da životinja pronađe svoj stalni dom. Dobijene rezultate potom bismo grafički prikazali.

Nakon eksplorativne analize, kreiranja i treniranja modela iterativno bismo evaluirali tačnost i pokušali da unapredimo isti fokusirajući se na precision, recall, i F1 score.

MOTIVACIJA ZA REŠAVANJE PROBLEMA

Većina azila za životinje suočava se sa problemom prevelike popunjenosti. Kao rezultat toga azili često nemaju mogućnost da prime nezbrinute životinje, ili čak vrše eutanaziju zdravih kako bi oslobodili prostor. Ovo je nešto što bi trebalo svesti na minimum, ili idelano, u potpunosti izbeći. Glavni cilj ovog rada jeste da poveća šanse za udomljavanje životinja iz azila, kroz predikciju dužine njihovog boravka i analizu karakteristika životinja koje dovode do brzog udomljavanja. Sa tim informacijama, moglo bi se postići kreiranje optimalnijeg rasporeda i organizacije prostora u azilima, zatim povećanje šansi slabije udomljivih životinja kroz pojačano promovisanje na društvenim mrežama i slično.

RELEVANTNA LITERATURA

[1] KIM, S.-E., CHOI, J.-H., & KANG, M. (2021). *Adoption Factor Prediction to Prevent Euthanasia Based on Artificial Intelligence*.
Korea Journal of Artificial Intelligence, 9(1), 29–35.
https://doi.org/10.24225/KJAI.2021.9.1.29

Cilj rada bio je vršiti predviđanje ishoda pristiglih životinja u teksaški azil. Metod koji su koristili bila je višeklasna logistička regresija. Skup podataka jeste izveden iz baze Austin Animal Shelter - odnosno, isti skup koji bismo mi koristili. Prilikom istraživanja podaci o novopristiglim životinjama i onim sa ishodom su najpre spojeni u jednu tabelu, a potom podeljeni na trening i test skup (0,7 trening, 0,3 test skup). Kao metod evaluacije korišćeni su accuracy, precision i recall, i pritom su postignuti rezultati veoma uspešni. Average accuracy dostizao je i preko 0,9.

Rezultati su pokazali da su psi generalno dobro udomljivi u svim uzrastima, za razliku od mačaka kod kojih je ovo dominantno u ranijim uzrastima. Takođe jedna od korisnih činjenica na koju rad ukazuje jeste veliki problem izgubljenih životinja koji se odražava i u skupu podataka nad kojim radimo.

Rad će nam biti veoma koristan po pitanju metodologije koja je korišćenja. Pored logističke regresije i metoda evaluacije, smatramo da je pristup spajanja tabela pristiglih životinja sa ishodima (outcome životinje može biti da je vraćena vlasniku, uginula, usvojena, ili prebačena u drugi azil) dobra startna tačka, za koju znamo da donosi poprilično tačne rezultate po pitanju predikcije. Međutim, naša ideja je da se pored baznog skupa takođe fokusiramo na dodatne skupove podataka koji bi proširili opise životinja jnenim fizičkim odlikama kao i karakterom bazirano na njenoj rasi, što bi u teoriji trebalo dati realističniju sliku šta je ono što budući vlasnici najviše vrednuju prilikom odabira svog novog ljubimca.

[2] Bradley, J., Rajendran, S. *Increasing adoption rates at animal shelters: a two-phase approach to predict length of stay and optimal shelter allocation*.
BMC Vet Res 17, 70 (2021).
https://doi.org/10.1186/s12917-020-02728-2

Cilj rada je da se povećaju šanse za udomljavanje životinja iz azila kroz dve faze - prva i realizovana se bavi predviđanjem dužine boravka životinja u azilu, dok druga ima u cilju da na osnovu rezultata prve faze i analize utvrdi u koje doba godine i u kojim regionima su veće šanse za udomljavanje radi potencijalne relokacije životinja. Podatke su sakupili sa zvaničnih sajtova azila ili im direktno pisali da bi ih prikupili i svodili su se na podatke o vrsti životinje, rasi, boji, polu, starosti, lokaciji azila, dužini boravka i ishodu. Predikcija je rađena korišćenjem logističke regresije, veštačke neuronske mreže, Random Forest i Gradient Boosting algoritma. Za poređenje modela i evaluaciju korišćeni su precision, recall i F1 mera. Najuspešniji model (*Gradient boost*) dostiže rezultate između 64-70% za precision, recall i F1.

Iako se korišćeni podaci razlikuju, problem koji je rešavan u ovom radu je gotovo isti kao naš, te je ovaj rad koristan kao primer koje metodologije se mogu koristiti i daju dobre rezultate. Tako u našem projektu planiramo da isprobamo metode klasifikacije koje su i ovde korišćene (Random Forest, ANN, ...) s tim što bismo pokušali različite metode obrade podataka i njihove podele pre dovođenja u model. Ovaj rad takođe predstavlja dobar indikator koji atributi će biti od najvećeg značaja za predikciju.

[3] Hawes S, Kerrigan J, Morris K. *Factors Informing Outcomes for Older Cats and Dogs in Animal Shelters*.
Animals. 2018; 8(3):36.
https://doi.org/10.3390/ani8030036

Cilj rada bio je određivanje glavnih indikatora za predviđanje sudbine starijih pasa i mačaka u azilu, za koje postoje veće šanse za eutanaziju zbog predrasuda da su manje pogodni za usvajanje.
Skup podataka nad kojim je vršena analiza sastojao se od 124 mačke i 122 psa, starosti veće od 84 meseci (7 godina). Podatke o životinjama pružio je azil Austin Pets Alive! (APA) specijalizovan za rad sa životinjama sa većom šansom za eutanaziju. Životinje su u APA dovedene iz Austin Animal Centre-a (odgovornog za naš osnovni skup podataka) sa kojim ima saradnju u cilju spašavanja što većeg broja životinja. Atributi prisutni u pomenutom skupu podataka u velikoj meri odgovaraju atributima našeg skupu podataka (datum prihvatanja životinje u APA, tip prihvatanja, procenjena starost životinje, dužina ostanka, bolesti, ishod i brojni drugi...).
Prilikom eksplorativne analize podataka definisane su brojne deskriptivne statistike koje nedvosmisleno prikazuju važnost pojedinih atributa za predviđanje ishoda životinja u azilu. Iskorišćenja je multinomijalna logistička regresija kako bi se na osnovu tipa prihvatanja, stanja životinje prilikom prihvatanja i propisanog lečenja predvideo ishod starijih životinja. Za evaluaciju rezultata korišćen je hi kvardat statistika (chi-square = 60.04 za mačke i chi-square = 62.61 za pse), sa p vrednošću manjom od 0.01 (p value < 0.05 se uzima kao dokaz statistički značajnog rezultata).
Za naš projekat od najveće važnosti je diskusija oko dužine ostanka starijih pasa i mačaka u odnosu na sve mačke i pse koji su prihvaćeni u APA. Zaključak je da stariji psi i mačke u proseku ostaju duže u azilu od svih pasa i mačaka koji su u tom periodu bili u APA. Ovo ukazuje da starost životinje igra bitnu ulogu prilikom procene vremena ostanka životinja u azilu, kao i da pri ovoj proceni potencijalno nisu bitni isti atributi ako su životinje mlađe ili starije.

SKUP PODATAKA

- https://www.kaggle.com/datasets/aaronschlegel/austin-animal-center-shelter-intakes-and-outcomes?select=aac_intakes_outcomes.csv

Osnovni skup podataka nad kojim bismo vršili istraživanje. Podaci su prikupljeni sa javne baze Austin Animal Center-a, koji na godišnjem nivou pruža pomoć preko 18 000 životinja raznih rasa. Ova brojnost čini ga pogodnim kao startnu tačku, kako nećemo brinuti o nedostatku podataka.
Skup podataka razdvaja 2 tabele: *intakes* i *outcomes* – odnosno prihvaćene životinje, i životinje sa krajnjim ishodom (prebačeni u drugi azil, eutanazija (uz razlog), udomljeni, vraćeni vlasniku, uginuli). Za obe tabele imamo veliki broj informacija o životinji kao što su starost, pol, vrsta, boja, ime, kada su pristigli u azil, stanje u kome su pristigli, gde su pronađeni...
Obe tabele sadrže po 72 000 instanci. Ukupan broj kolona je 24 (u svakoj tabeli po 12, s time da su neke od njih poput ID-a životinje, rasa, datum rođenja zajednički). Ciljno obeležje je definisano u *outcomes* tabeli – outcome_type, čija je raspodeljenost 42% za udomljene, 30% za životinje prebačene u drugi azil i 28% za ostale ishode (uginule životinje, uspavane ili izgubljene pa vraćene vlasniku).

- https://datasetsearch.research.google.com/search?query=dog%20breed%20size&docid=L2cvMTFqbl9jbjhjMg%3D%3D

Skup podataka izlistava brojne rase pasa sa njihovim veličinama. Planiramo da pomoću njega uvedemo veličinu životinje kao jedan od faktora udomljivosti, koja se u prethodnim radovima ispostavila kao jedna od ključnih stavki.

- https://www.kaggle.com/datasets/yonkotoshiro/dogs-breeds

U ovom skupu podataka govori se o određenim karakteristikama pasa po njihovim rasama, kao što su osetljivost, koliko su dobri za početnike, adaptabilnost na život u stanu ili na to da budu ostavljeni sami u toku dana. Dodatno, skup podataka sadrži i neke biološke podatke o rasama kao što su visina, težina, prosečan životni vek.

- https://www.kaggle.com/datasets/rturley/pet-breed-characteristics

Skup podataka sadrži informacija o macama i kucama u zavisnosti od njihove rase, gde je osnovni podatak koristan za naš projekat kolona „temperament“. Svaka vrsta saži 5-10 atributa karaktera životinje, što može biti dobar ukazatelj kakvi su temperamenti najtraženiji među udomljenim životinjama.

METODOLOGIJA

Na osnovu rezultata eksplorativne analize podataka grupisaćemo dužinu boravka u kategorije (npr. 0-2 meseca, 2-6 meseci, 6-12 meseci, ...). Na taj način problem predikcije svešćemo na klasifikaciju u neki od definisanih vremenskih razdoblja. Pritom ćemo odrediti koji atributi najviše utiču na dužinu ostanka u azilu i da li se mogu uočiti različiti trendovi kod različitih vrsta životinja. Nakon toga, upotrebićemo različite klasifikacione modele (Random Forest, XGBoost, Gradient Boosting Model, ANN, ...). Ispitati interpretabilnost odabranih modela.

EVALUACIJA

Podelićemo skup podataka na train i test skupove u razmeri 80:20. Kako ćemo treniranje modela vršiti nad train skupom, test skup ćemo upotrebiti za evaluaciju modela tako što ćemo izračunati precision, recall i F1 mere, a zatim uporediti te rezultate kod različitih modela.

PLAN RADA

- Prikupljanje podataka
- Transformisanje prikupljenih podataka
- Kreiranje i treniranje modela
- Evaulacija rezultata
- Vizualizacija dobijenih rezultata

Tim:

1. Ana Grahovac, E2 72/2022
2. Žarko Blagojević, E2 33/2022
3. Tara Pogančev, E2 30/2022