Final Results and Project Report

Animal Shelters

Course: CIND 820

Name: Diego O. Tejada Cardenas

ID: 500669615

Date: 2022-

An animal shelter tends to struggle with resources, each time that an animal stays longer in the shelter the cost increase exponentially. The overcrowding and constant introduction of animals with a high probability of issues (health and behaviour) increase the demand for an already stretch management and resources. Over time this was correlated to an increase in workers turnover, unsanitary conditions, and poorer veterinary care. Moreover, it tends to finish in euthanization for many animals that were not given a real opportunity. For these reasons, I would like to search for anomalies in the data in relation to the characteristics of the adopted and non-adopted animals. Moreover, I would like to search for new insights into the data so I can provide some possible solutions.

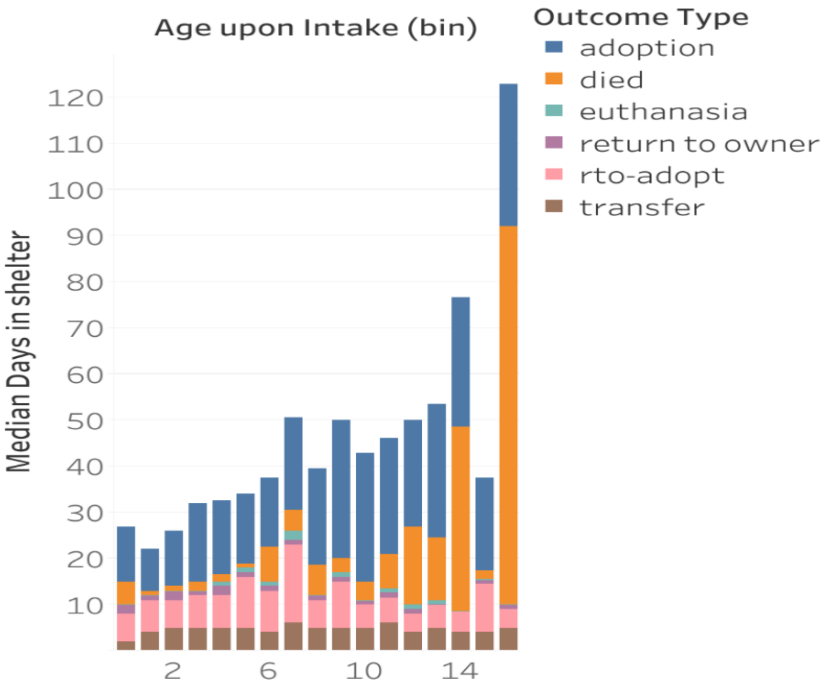
The theme chosen for this project is anomaly detection and data mining & knowledge discovery.

The problem that I trying to solve is the efficiency issues in relation to the resources used for each animal. An animal with characteristics that are in demand should need fewer resources for adoption (promotion, training, etc.). Therefore, the research question would be if there is discrimination again or in favour of certain dogs and/or cats? (Race, size, age).

According to the article by Sloane Hawes, older cats and dogs have been particularly at-risk for euthanasia in animal shelters due to the lower perceived appeal. Her paper found that the condition at intake had the greatest impact on the outcomes for older cats and dogs. Her paper uses the same database as this paper; however, it differs in scope and years. Her paper use data from two years and focus exclusively on dogs over 7 years. On the other hand, this paper uses the data from the last decade and this decade too. Using feature selections techniques, this paper found the same characteristic applies years later and the condition at the moment of intake still is an important feature. The intake condition feature describes the animal at the moment of arrival. In particular, the health, biological characteristic (aged, pregnant, etc.), and behaviour. Due to the traits of this variable, it is assumed that health and behaviour are good predictors of the future outcome of an animal. However, there are some data balancing issues that will have to be in consideration. Moreover, it becomes obvious that the adoption rate and adoption time will vary significantly between an animal that is in treatment (nursing, recovery, etc.) and the ones that are not.

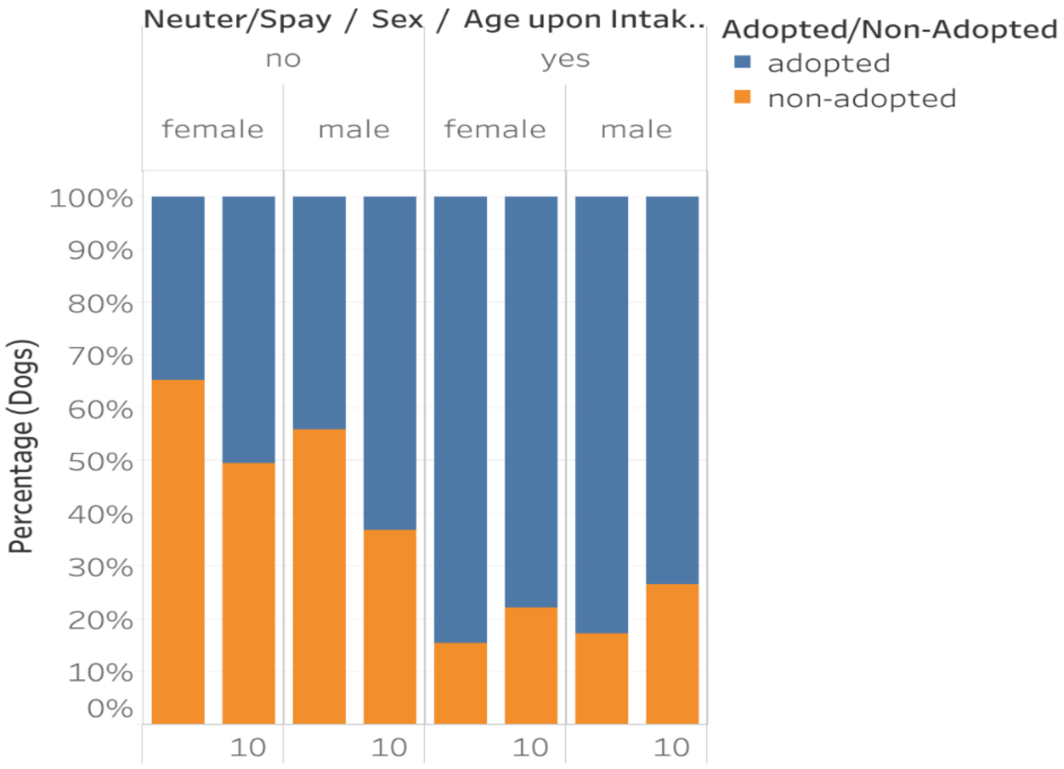
Another paper that evaluates a similar topic is “Why did you choose this Pet”? by Emely Weiss, this study evaluates the reasons for adoptions with a survey of 1600 adopters in animal shelters in New York. Her paper indicated that appearance was the most important reason for adoption. However, it changes between puppies-dogs and kittens-cats. Within species, the age group revealed that appearance was the most important reason for the adoption of a kitten, but for a cat, behaviour with people was the most important reason. In contrast, appearance was the most important reason for the adoption of puppies and dogs. In this dataset, it is not possible to assess the appearance, but it is possible to assess that the majority of euthanasia cases are in large dog breeds and/or guard dogs were described with behavioural issues. Usually, these breed are most intimidating breeds.

A constant feature, that is mentioned in many papers in this essay, is the age variable. A paper by William Brown in ‘Effects of Phenotypic Characteristics” study this variable in more detail and more variables. According to his paper, there is a preference for younger dogs over older ones among all the No-kill Animal shelters in New York. In comparison, this data set suggests the same trend. In the next graph, it is possible to observe the correlation between days in shelter, age of arrival, and the outcome.

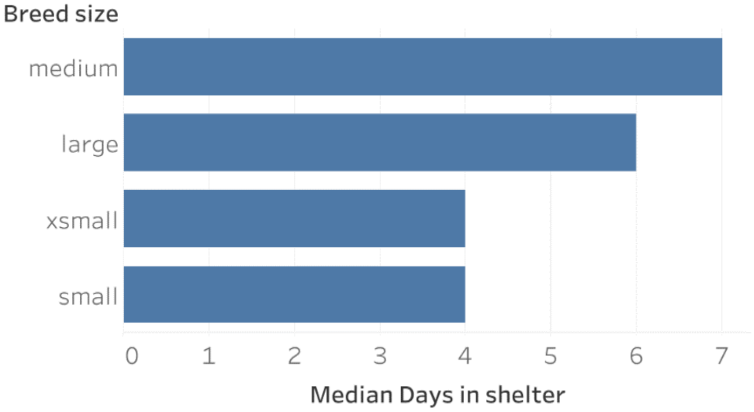


Moreover, his paper also compares the information extracted from other papers which focus on No-Kill animal shelters in the U.S, Australia, Brazil, Ireland, UK, Italy, and the Czech Republic. The combination of the information suggests new sub-topics for observation, the first sub-topic is sex. at first, there are no apparent differences overall between older dogs, assuming that they still can reproduce, but this metric changes when is combine with the information of neutering and spraying. Suggesting that females are preferred everywhere except for Brazil, where males are preferred, based on a survey. The survey indicated that the most common response for this preference contains a version of “females may give birth to unwanted litter” and “males are easier to care for” (Soto et al., 2005). This dataset suggests the same trend that in Brazil if only the animals are intact.

In the next graph, it is possible to observe a minimal difference between the sex if the dog is neutered or spayed. If the dog is not neutered and/or spay, it is observable that females dogs are less demanded when they can reproduce. On the left side of the graph, unneutered females are 34% adopted and unneutered males are 44% adopted (all under 10 years). Assuming that dogs after 10 years cannot reproduce because the max standard age is 7-8 years (legally 5-6 years), only a few cases passed the threshold of 10.



In Italy, demographic differences were observed, men adopted more male dogs from a no-kill shelter, but women adopted more dogs overall. (Normando et al., 2006). In relation to breed, the smallest dogs (lapdogs) had a shorter adoption wait time than medium and large dog. The other papers also concur with this statement and this dataset suggests the same. In this dataset, same trend is possible to observe.



Finally, his paper indicated a shorter wait time for giant breed duo to regional differences (rural and urban). This claim is not possible to verify with this dataset, but each animal type, that it is not adopted, has a large percentage that is transfer. Meaning that the percentage of transfer may indicated the higher demand elsewhere, but also, the output vary because many are transfer to sanctuaries too.

The next research question is if the undesirable characteristics are related to biology or socio-cultural beliefs.

Previously in this essay, Soto indicated that males were preferred because 'males are easier to care for". This result partially responds to the question, but a clearer response is offered by Lori Kogan with "Exploring the common Perception". This paper explicitly tries to respond if black cats take longer to adopt. This paper was based on the analysis of two animal shelters in Colorado and more than 30000 animals were studied for this paper. Moreover, it recollected the perception of the social worker concerning adoption rate differences because of colour. The result of this study suggests that the colour of a cat's coat influence the time required for adoption and that black cats require the longest time to adopt, followed by those that have primarily black coats with other colours mixed in. Also, this result supports the reports from workers about a “Black Dog Syndrome” but in cats. In this dataset, it is not possible to distinguish between darker hair animals and the general population, the differences are too small to be able to access this hypothesis.

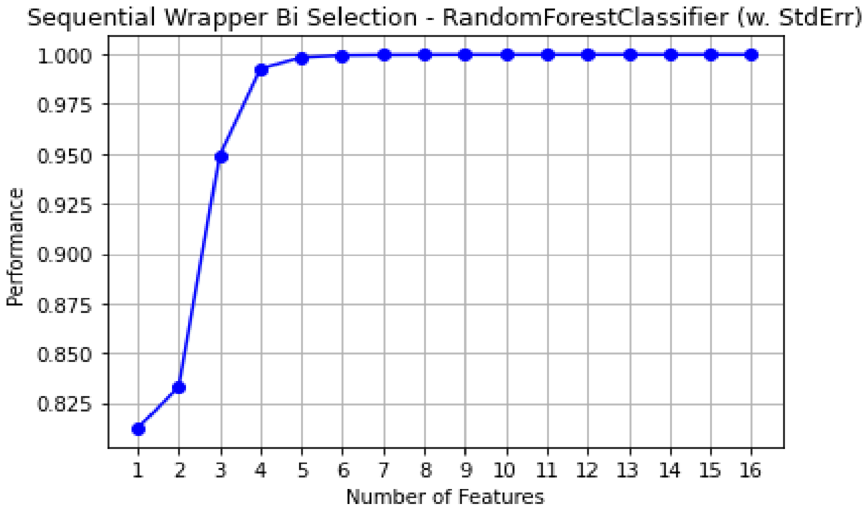
The reason for these questions and analysis of these papers is to discover the less desired characteristics and, in the future, to recommend preventive measures so the least favoured animals are given more resources before they enter the animal shelter, like dealing with stigmas, assumptions or discourage certain actions like buying from puppy mills, certain races, etc.

The data for this project consists of two databases in which the animal shelter gives details of the arrival and departure of animals. The arrival dataset consists of 11 variables where the condition of arrivals is described (colour, breed, age, sex, type, condition, reason, location, month, daytime, name, and ID). It contains information about 124,120 animals. The departure dataset consists of the same variables but with information about 124,491 animals. After the cleaning and merge, the database was reduced to 123,173 animals. Moreover, it was assessed that new variables may benefit the datasets with new approaches. The new variables were created from existing data and American Kennel Club (AKC). The new variables created from existing data were a mix, primary breed, secondary breed, main colour, secondary colour, neuter/spay, adopted/non-adopted, and days in the shelter. On the other hand, the information extracted from the AKC created the variables breed size, trainability, breed characteristic, barking level, activity level, and breed group. The information extracted was only the breed's official names and the breed known categorizations like small, guard dog, easy to train, etc. The official names of the breed will be used from this point forward. Finally, the final data for this study have 27 variables characterized by having 1 Boolean variable, 1 serial variable, 2 text variables, 3 numeric variables, 3 dates variables, and 17 categorical variables. This data was provided by Austin Animal Center on the Texas Official website.

The research question required the determination of the target variable and the variables that affect it the most. For this reason, it was chosen to target the variable Outcome Type, but for simplicity in the visualization, all the outputs in the target variables were compressed to a binary option (adopted or not adopted). The variable outcome type has 5 categories, the 3 positive outcomes are adoption, return to owner, and rto-adopt (return to the first adopter). The rest died (age or medical intervention), euthanasia (behaviour, aggressiveness, and suffering) and transfer (sanctuary or other shelters).

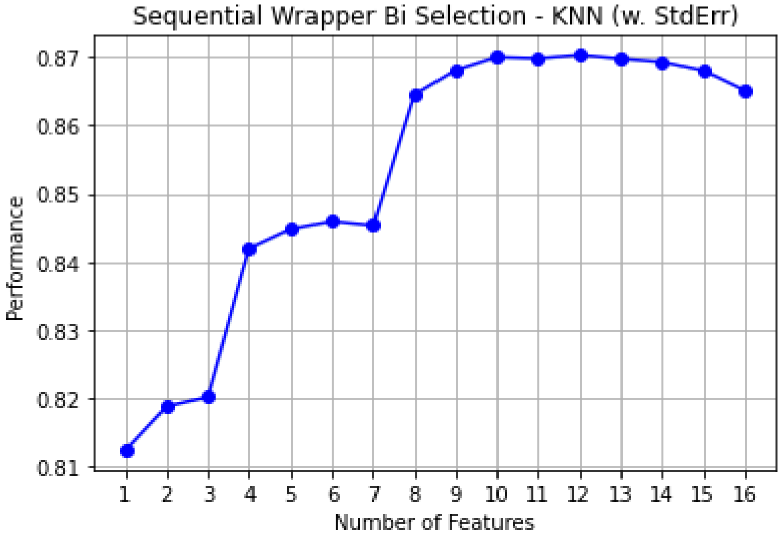
In the next section, this paper explores the techniques used in the selection features and the predictive models. The selection of the most important variables was done by the feature selection techniques such as wrapper method (backward, forward, and both), embedded method and filter method.

The results from the feature selection wrapper (Random Forest) show that 7 variables affect the target variable, the rest is minimal until the point that there is no difference graphically (The forward and both-ways-wrappers were identical). The selection of certain features is unnecessary because the model obtains all the information required from the first variables. The random forest classifier allows an increase of features without affecting its performance of it.

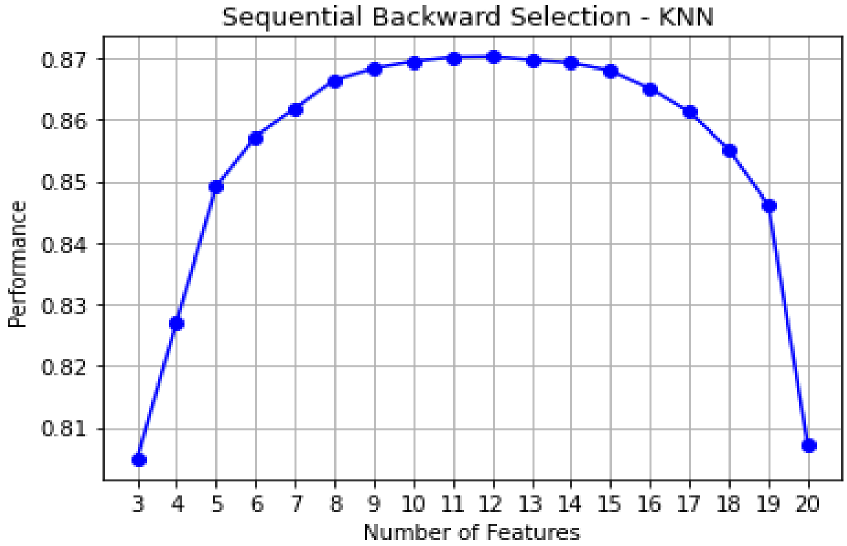


The most important variables according to the wrapper method (random forest) were Intake Type, Intake Condition, Animal Type, Primary Breed, Color, Sex, Breed Size, and neuter/spay.

The results from the feature selection wrapper (KNN) show that there are 9 to 11 variables that affect the target variable, the rest is minimal until a certain point. Then a larger group of variables created noise in the data, reducing its effectiveness.

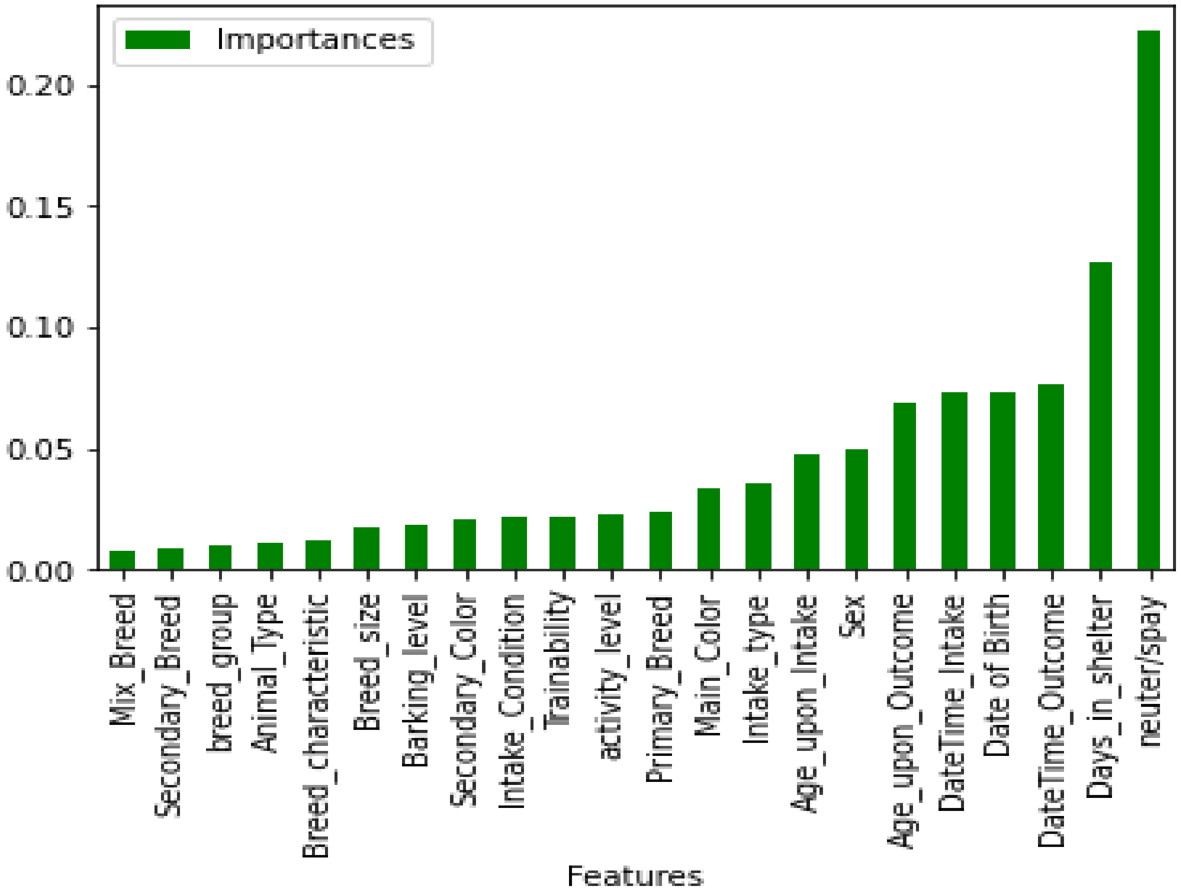


The way that KNN models work allows an increases in variables, but it increase the possible overlapping between clusters. When KNN is allowed to use all the variables, the reduction in performance is significant.

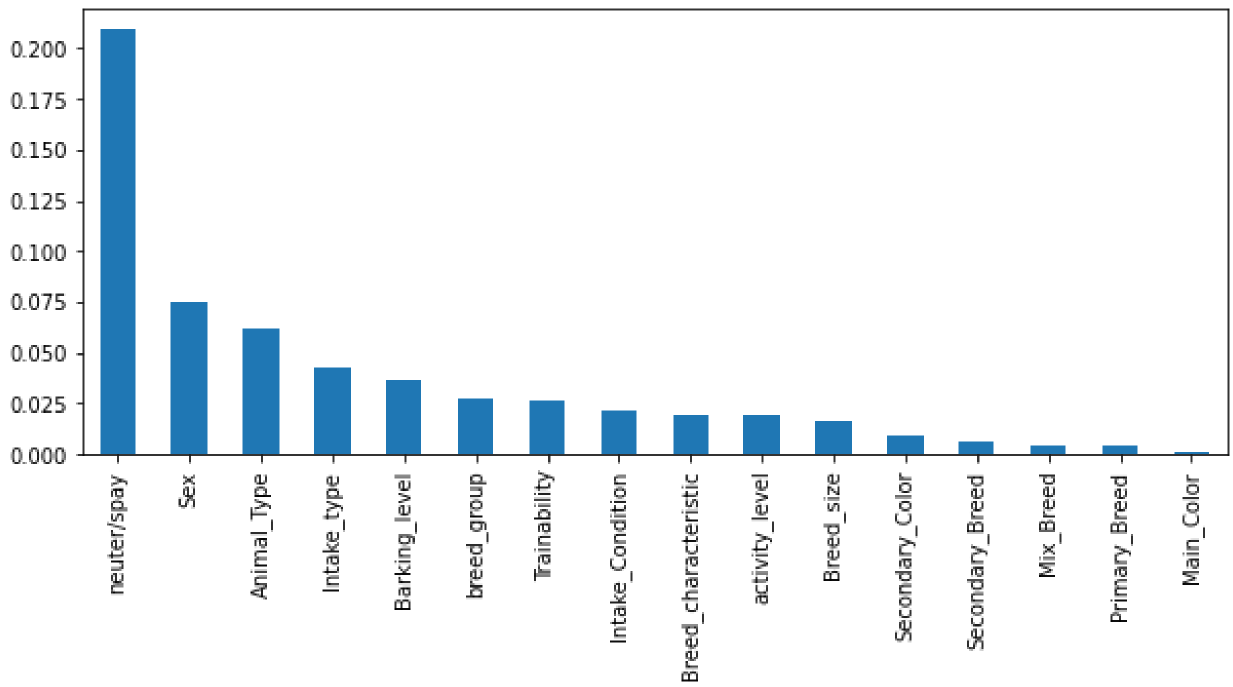


The most important variables according to the wrapper method (KNN) were Intake Type, Intake Condition, Animal Type, Age upon Intake, Mix, Sex, Days in Shelter, Age upon Outcome, and neuter/spay.

The results from the feature selection embedded method (Random Forest) indicated the level of importance according to affect.



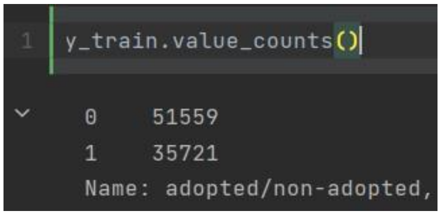
The results from the feature selection filter method indicated the correlation and then the importance after the removal of the highly correlated independent variables.



The research question asked about the existence of certain features in the animals that cause some kind of preferences or discrimination in favour of one and not the other. Therefore, to respond to this, it was decided to create three machine learning models (Random Forest, KNN and SVM) to predict if an animal would be adopted or not.

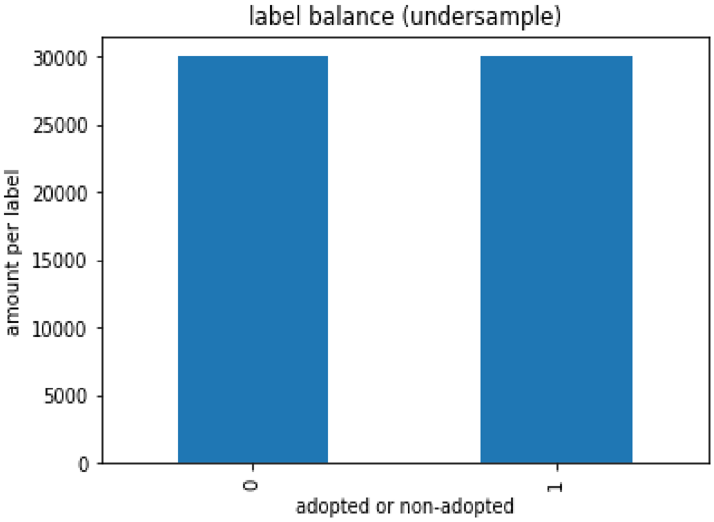
The first model used was Random Forest. The pre modelling procedures were the initial drop of features, cleaning of the target variable (reducing it to a binary option), label encoding of all the categorical variables, split data (20% test and 80% training), balancing of the data (under-sampling method), and standardization (removing the mean and scaling to unit variance).

The initial number of removed features was 7 which consists of identifiers, name (provided by the animal shelter), times of actions, locations, and multicollinear features. Then label encoding was applied for the algorithm to be able to use. Moreover, One-Hot-Encoder was extended to the categorical variables. The target variables, in this case, adoption, was reduced to a binary option from three option (adopted, unknown, non-adopted). The removal of the unknown in the target variable consisted of 25 rows. Then balance methods were applied, but the simpler and better suited for this data was under-sampling because this method works well when the data is substantial and relatively equal in the amount of data from each target class.



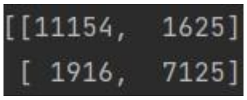
```
y_train.value_counts()

    0    51559
    1    35721
    Name: adopted/non-adopted,
```

After the balancing:



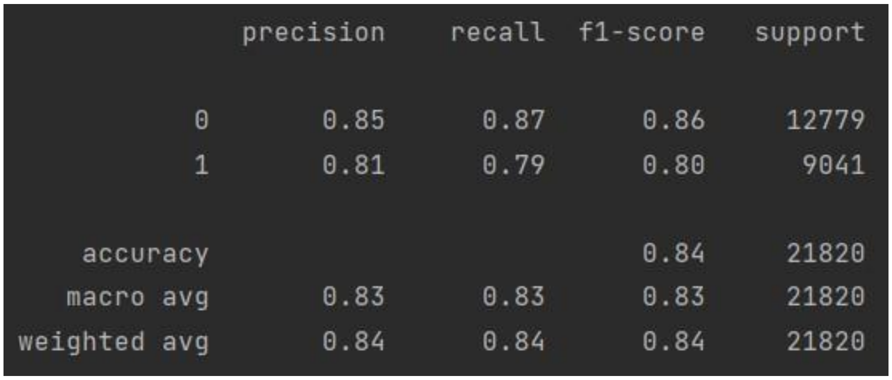
After the application, all the methods previously explained. The model resulted in an accuracy of 83%. To verify and gain a more specific understanding of this output, it was implemented a matrix, classification report, K-fold Cross-validation, log loss, number of trees (performance), and depth effect (performance).

In the next screenshot is possible to observe the matrix with the true positive on the left-top and true negative on the right bottom.



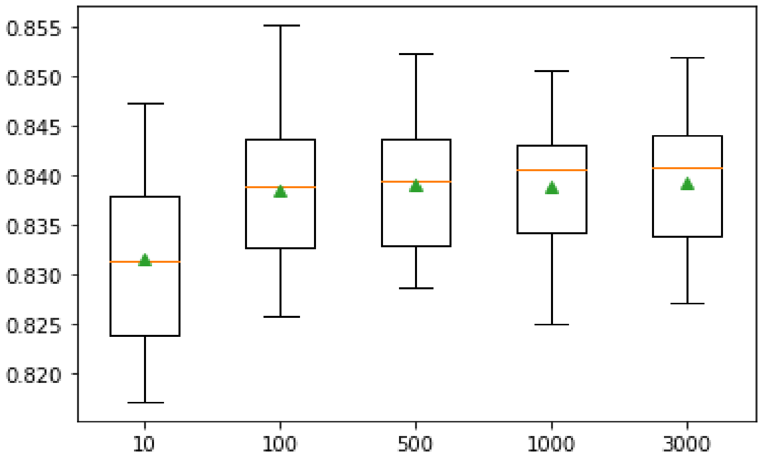
```
[[11154,  1625]
 [ 1916,  7125]]
```

The classification report:

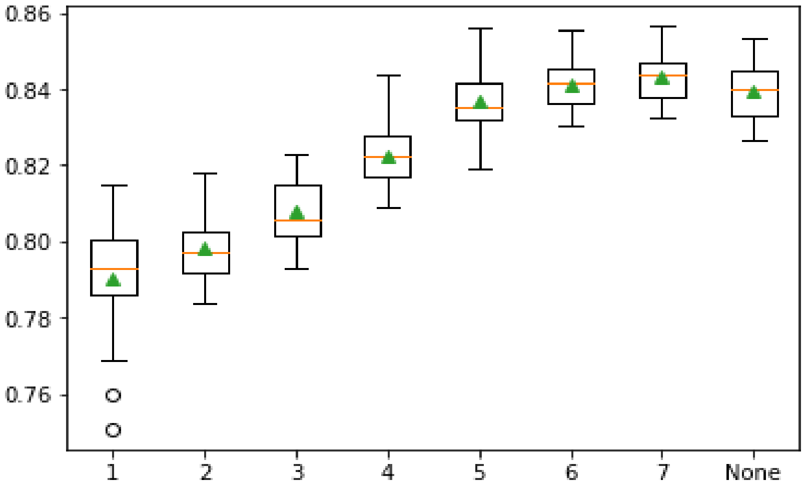


```
              precision    recall  f1-score   support

           0       0.85      0.87      0.86     12779
           1       0.81      0.79      0.80      9041

    accuracy                           0.84     21820
   macro avg       0.83      0.83      0.83     21820
weighted avg       0.84      0.84      0.84     21820
```

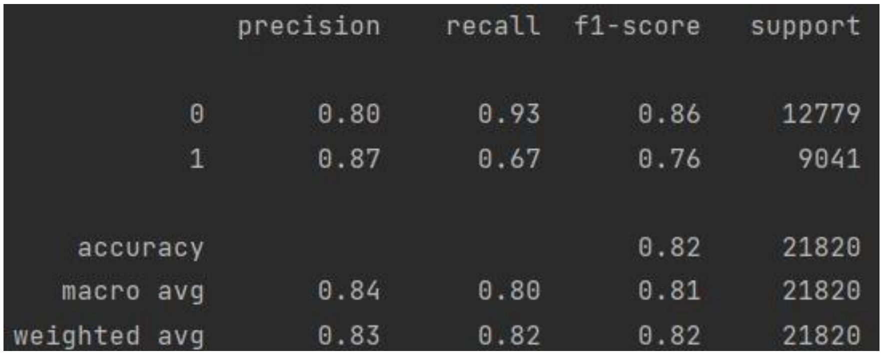
The K-fold Cross-validation had an output of 84% accuracy and the Log loss had a result of -0.456 (the closer to zero the better) with a standard deviation of 0.02 (minimal). The number of trees analysis shows that the performance improves significantly until 10 trees (83.2% accuracy), then the rate of improvement decreases significantly. To achieve the same level of accuracy, obtained in the K-fold cross-validation requires increasing the number of trees. In the next graph, it can be observed the patterns (x is the number of trees and y is performance).



Additionally, it was performed a depth analysis for this model. In the next graph, it can be observed the random forest tree depth effect on performance (x is the level of depth and y is performance). The last x variable is none because no more depth were required and it decreased the performance (84.3 => 83.9)



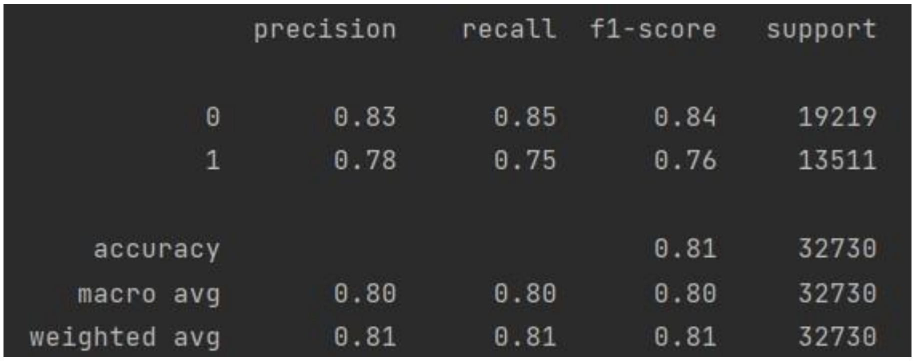
Additionally, it was performed a depth analysis for this model. In the next graph, it can be observed that the random forest tree depth effect on performance (x is the level of depth and y is performance). The last x variable is none because no more depth was required, and it decreased the performance (84.3 => 83.9)



```
              precision    recall  f1-score   support

           0       0.80      0.93      0.86     12779
           1       0.87      0.67      0.76      9041

    accuracy                           0.82     21820
   macro avg       0.84      0.80      0.81     21820
weighted avg       0.83      0.82      0.82     21820
```

The entire KNN model took 6.9 seconds.

The last model used was SVM with the same pre-procedures used in the last models. For practicality the kernel was linear. The initial results were 80% accuracy, 0.77 precision and 0.75 recall. After K-fold cross validation, the accuracy increases to an accuracy of 81.5% and the log loss is 0.02. The classification report is in the next graph.



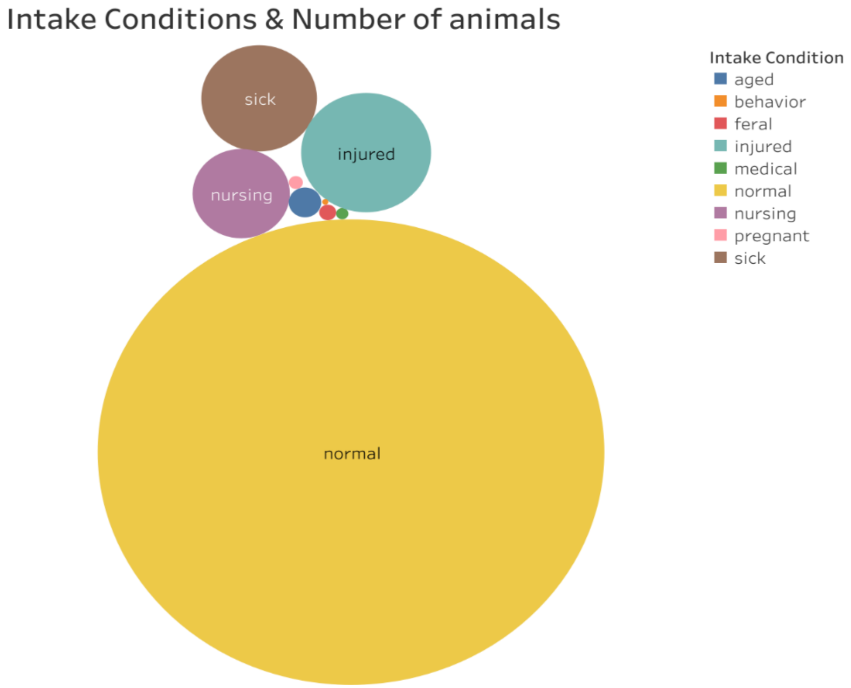
```
              precision    recall  f1-score   support

           0       0.83      0.85      0.84     19219
           1       0.78      0.75      0.76     13511

    accuracy                           0.81     32730
   macro avg       0.80      0.80      0.80     32730
weighted avg       0.81      0.81      0.81     32730
```

The entire SVM model took 30 minutes.

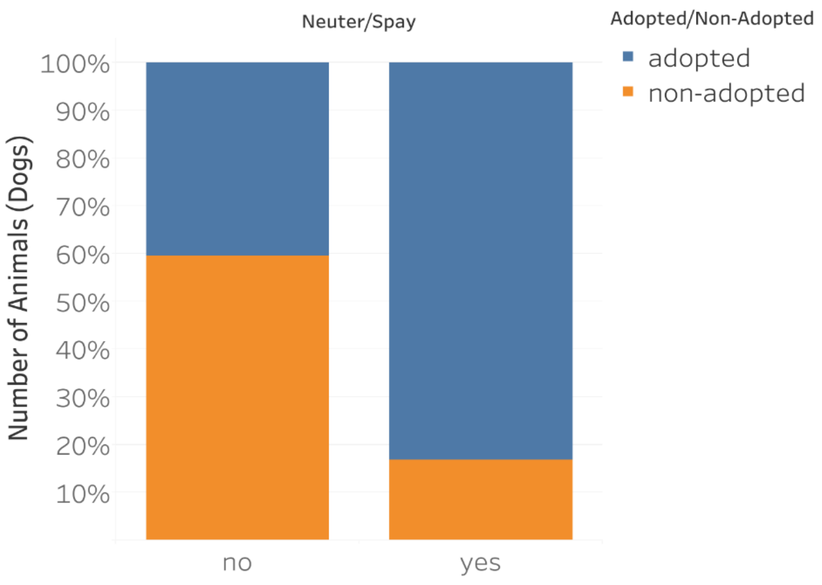
In the next section, this paper will explore the most important features according to features selection and predictive model. The predictive model used the entire set of features and the selected features by the features selector methodologies.

Intake Type is a variable that describes their legal status with humans, meaning that an animal could be considered stray, owner surrender, public assisted, abandoned and euthanasia request. A stray and surrendered dog could easily obtain legal reintegration with a human. the complexity increases in comparison to public assisted dogs that were taken care of by a community (neighbourhood, etc.) but no one took legal possession of. Because in case of medical issues or behavioural issues, it is less likely that a person will come forward. However, almost three-quarters were stray and the owner surrendered.

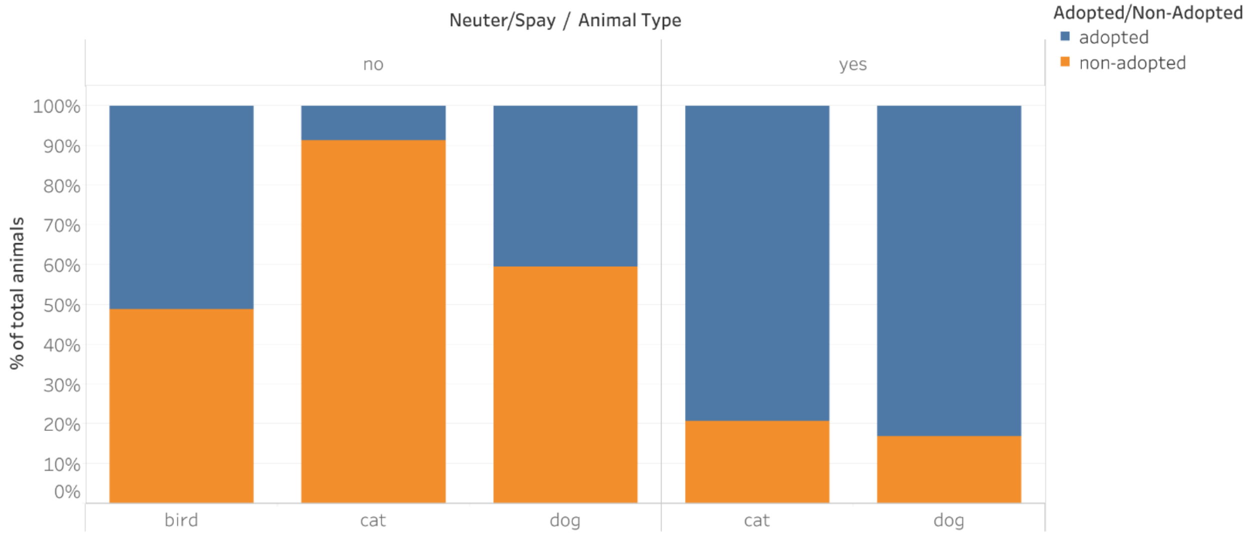
An important feature according to the features selector is the variable Intake Condition. This variable describes the health condition of the animal. Animal health had always been an important reason for adoption, but in this case, almost 90% of dogs arrived in normal conditions.



The variable neuter/spay suggests the existence of large discrepancies in adoption rates between the animals that were neutered or spayed to the ones that did not. In the next graph, it is possible to observe the large discrepancy between groups.

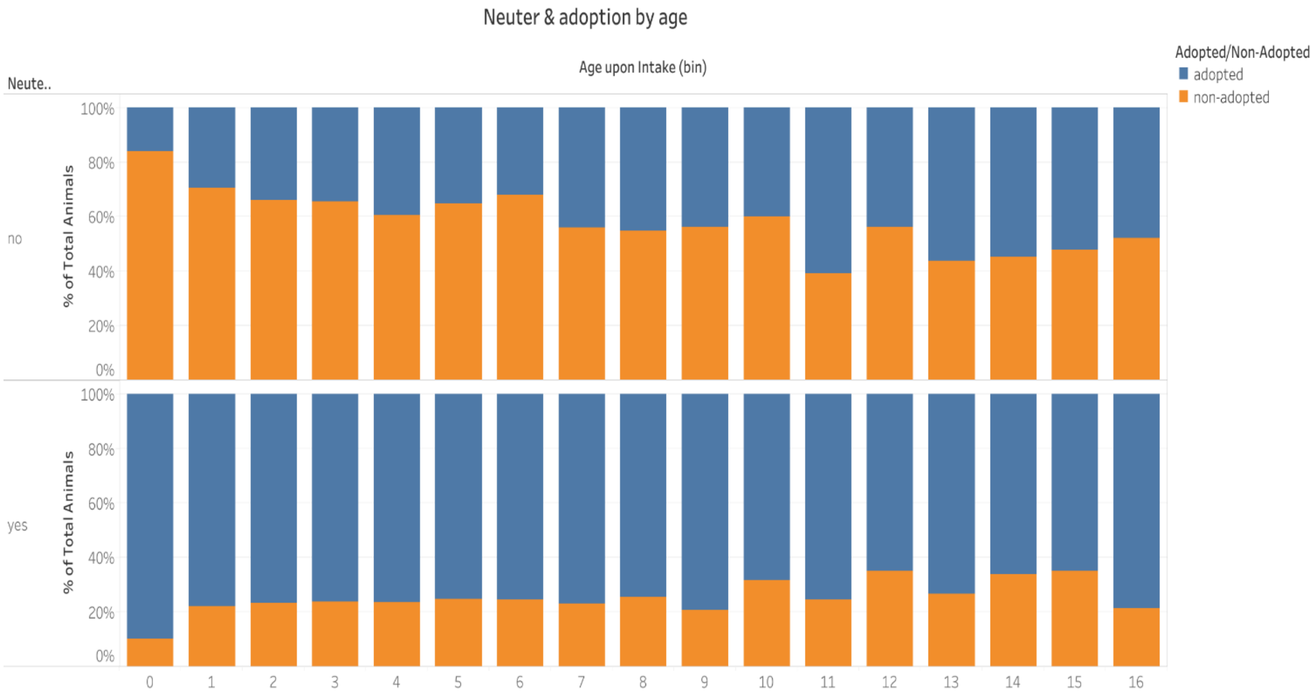


The previous trend extends between species, but with some minor differences like in the cases of birds, it is highly uncommon to have birds neutered because of severe hormonal issues. In the case of dogs and cats, this trend continues.

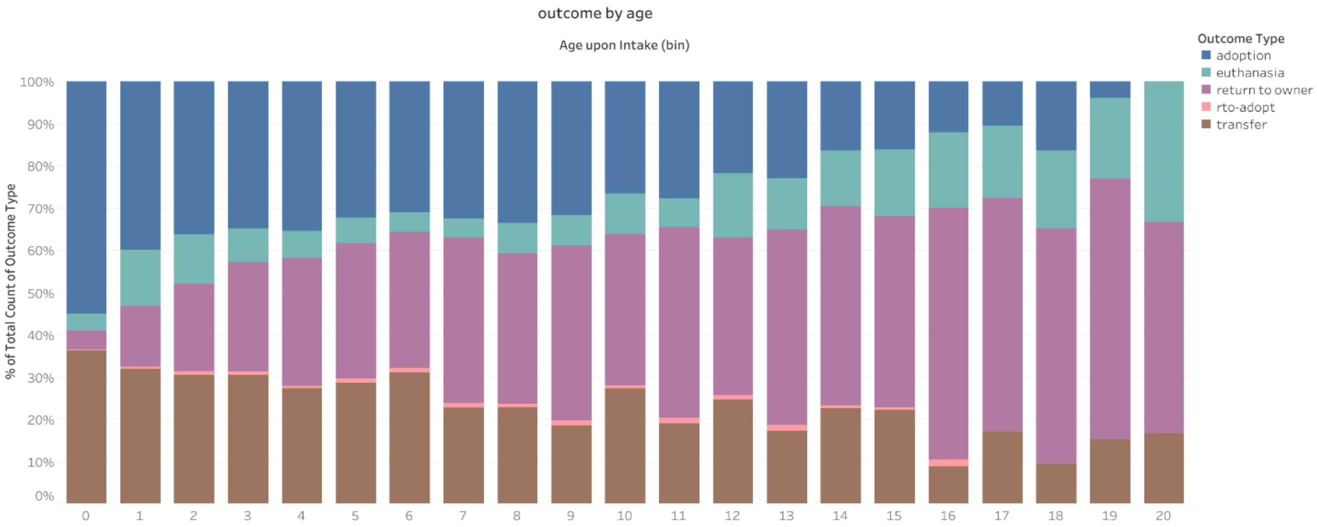


The reasons for these variations are many, but the most common arguments are the possibility of undesirable breeding. Moreover, animal shelters have the incentive to promote neuter dogs and American veterinarian practices do the same. Another argument is the cost related to surgery and the association between un-neuters and future illnesses like mammary tumours, pyometra and prostate cancer. However, this assumption may vary according to the country because veterinaries in Europe and South America recommend based on the individual case and not based on policy. According to the Danish Centre for Bioethics, there is no reason to claim that routine neutering is morally acceptable knowing that there is an increased number of studies that contradict each other and there is not a definitive answer about secondary long-term effects. Suggesting that, Spaying/neutering an animal work for the desire intend of breeding containment but not necessarily for long-term health.

Expanding on the previous trend observe, this trend does not vary much if the age is included as a factor, with the small exception of the first year.

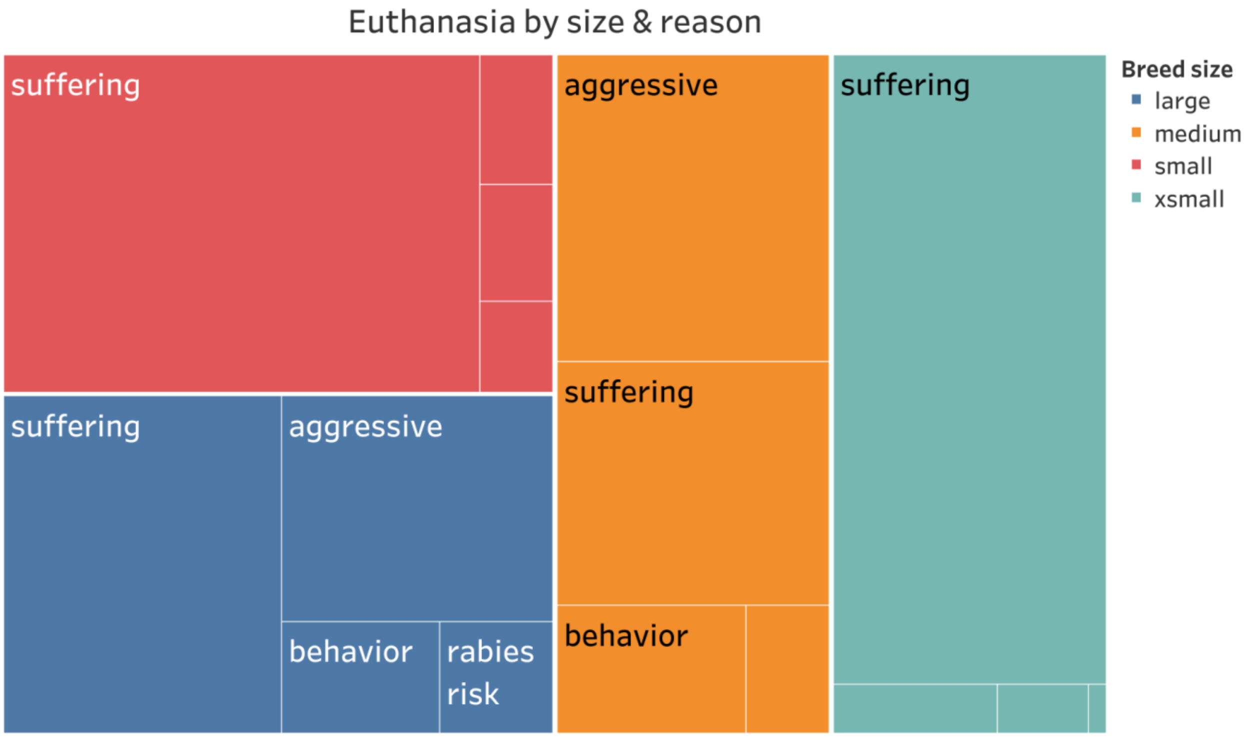


However, the age variables offer a better insight when the scope of specificity changes. In the next graph, it is possible to observe that the adoption rate is the same across ages but the type of adoption changes dramatically. As the dog gets older the adoption by new adopters decreases, at the same time, previous adopters and owners increase the percentage of adoptions. This suggests that as the dogs' age increases the probability of getting adopted by a new person is lower and the only option for these animals is to get adopted by the same person or get transferred to other location. The ones that get transferred may get a better chance of adoption, but some will go to different kinds of sanctuaries. The last detail in the next graph is a higher level of euthanasia as age increases.



The final discovery was in the euthanasia differences between breed sizes and breed types. The dataset suggests that large breeds and medium-size breeds are at risk of a higher level of euthanasia for aggressiveness and behaviourally issues. Moreover, the higher level of euthanasia in the medium and large breeds can be explained by the inclusion of guard dogs and pit bulls.

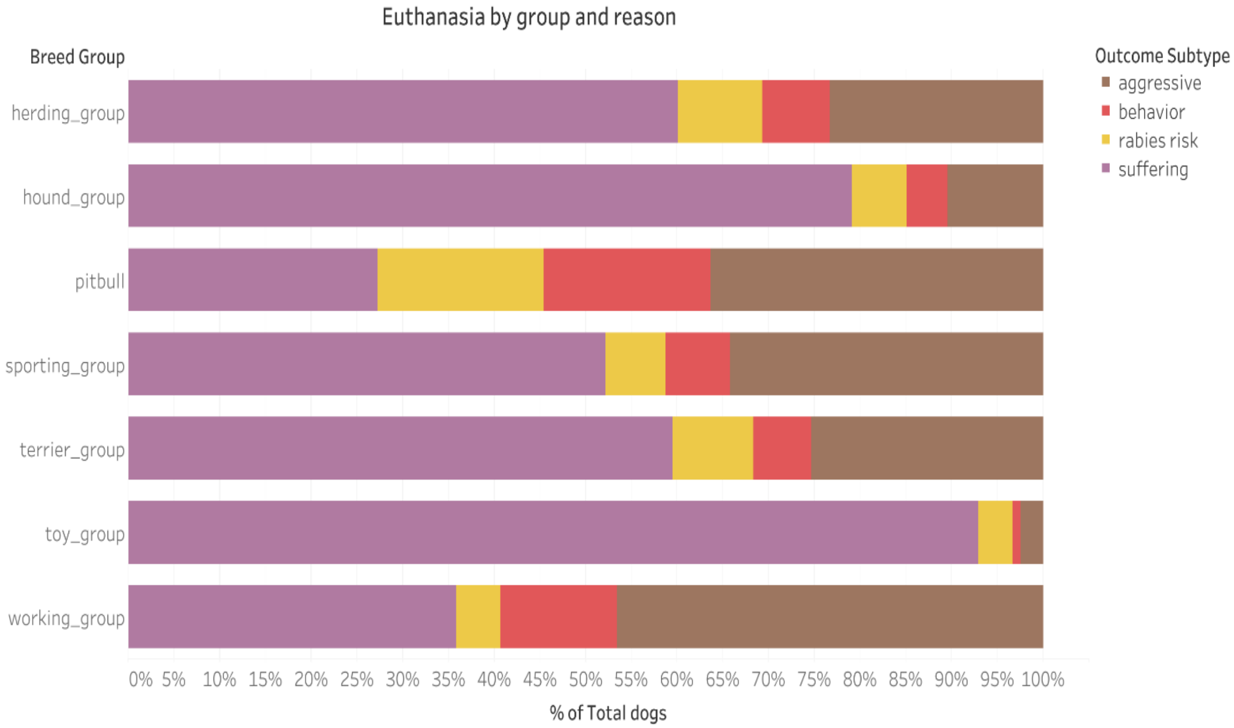
In the next graph, it is possible to observe that a large section of the euthanasia reasons small and very small is suffering. Meaning that euthanasia in a very small or small dog is highly probable due to some severe health issue or because age. However, the other sizes present a higher level of euthanasia for aggressiveness and behaviour. Medium-size breeds have 45% of the euthanasia due to aggressiveness (the biggest in this size category) and large-size breeds have 33% (the second biggest).



The difference previously mention is of concern because aggressiveness and behaviourally issues are not perpetual factors in these breeds and sizes. At some point in the animal life, these issues rise without anyone properly addressing it. Meaning the cost for an untrained or unattended dog is higher according to the size.

In the next graph, this issue is explored in more detail. The euthanasia differences presented in the last graph appear in this one too. The group that constitutes almost exclusively medium and large size breeds are the most disfavored (pit bulls and working breeds). The working group is formed by the combination of breeds that work as a guard on farms or for personal protection and other very specific work. The Pitbulls were considered part of the working group but over time their reputation changes with their inclusion in dog fights.

In the next graph, the pit bulls that are euthanized by suffering is slightly over 25% and for working breeds is slightly over 35%. The rest is euthanized by aggressiveness, behaviour and rabies risk.



In conclusion, it is possible to answer the main question in this paper. Some differences may be considered statistical discrimination in favour of certain animals. The statistical discriminations were found in variables related to age, health, neuter/spay, and breed/size. Many papers, including this paper, indicate that age is one of the main factors for adoption. Age as the only variable fall short of a proper explanation because it fluctuates significantly according to breed and reproducibility. Colour is also frequently appearing as an important variable but according to the models in this paper, they are not the most important as other papers suggest. The most important variable and one of the less written about is neuter/spay in the adoption rate. The large indifference adoption rates between spay/neuter animals and un-neutered are significant. In the research process for this paper, no paper explicitly analyzes this effect. Some had written about it but not with the scope required, only as a correlation found, such as the paper Adoption of Shelter Dogs in a Brazilian Community by Martins Soto.

The euthanasia issues in this dataset present a well-known issue in the animal rescue communities. A topic that had been explore but none combine with the spay/neuter scope.

The final suggestion to address these discoveries would be the creation of a Medicare system for older dogs because one of the main issues in adoption is health, which deteriorates with age. This would decrease the barrier to adoption of dogs over 5 years and their age could become a positive variable because older dogs have the tendency to be calmer and require less exercise than younger dogs. On the other hand, the discrepancy between neuter and un-neuter dogs may require a more focused study. The study to address this discrepancy should focus on humans because the policies, beliefs and assumptions may be the cause of the lower adoption rate. From the behavioural and health perspective, there is no definitive answer.

Finally, the shortcoming of this paper is the lack of analysis of the humans involved. The adopter represents the demand side of these issues and not enough information could be gathered about them. Another shortcoming is the external factors in supply. Meaning how finance, the worker and politics affected the adoption rates of these animals.

Bibliography

Brown, W. P., Davison, J. P., & Zuefle, E. M. (2013). Effects of Phenotypic Characteristics on the Length of Stay of Dogs at Two No Kill Animal Shelters. *Journal of Applied Animal Welfare Science*, 1-16.

Hawes, S. M., Kerrigan, J. M., Hupe, T., & Morris, K. N. (2020). Factors Informing the Return of Adopted Dogs and Cats to an Animal Shelter. *animals - MDPI*.

Hawes, S., Kerrigan, J., & Morris, K. (2018). Factors Informing Outcomes for Olders Cats and Dogs in Animal Shelters. *animals - MDPI*.

Kay, A., Coe, J. B., Young, I., & Pearl, D. (2018). Factors Influencing Time to Adoption for Dogs in a Provincial Shelter System in Canada. *ournal of applied animal welfare science* , 375–388.

Kogan, L. R., Schoenfeld-Tacher, R., & Hellyer, P. W. (2013). Cats in Animal Shelters: Exploring the Common Perception that Black Cats Take Longer to Adopt. *The Open Veterinary Science Journal*, 18-22.

Martins Soto, F. R., Ferreira, F., Regina Pinheiro, S., Nogari, F., Regina Risseto, M., de Souza, O., & Amaku, M. (2005). Adoption of Shelter Dogs in a Brazilian Community: Assessing the Caretaker Profile. *Journal of Applied Animal Welfare Science*, 105-116.

Palmer, C., Corr, S., & Sandoe, P. (2012). Inconvenient Desires: Should We Routinely Neuter Companion Animals? *Anthrozoos* , S153-S172.

Weiss, E., Miller, K., Mohan-Gibbons, H., & Vela, C. (2012). Why Did you Choose This Pet?: Adopters and Pet Selection Preferences in Five Animal Shelters in the United States. *animals - MDPI*.

GitHub:

https://github.com/diego0tc/CIND-840_diego_500669615