UNIVERSIDADE FEDERAL DE SANTA CATARINA - UFSC
SISTEMAS DE INFORMAÇÃO - INE - CTC
DATA MINING - INE 5644

Trabalho prático de Data Mining

Entrega 2

Bruno Eduardo D'Angelo de Oliveira
Eduardo Demeneck

FLORIANÓPOLIS, 2016

1. Seleção dos dados a serem utilizados para o estudo

No relatório anterior, foi discutido, na seção intitulada “seleção preliminar de uso de atributos”, que deveriam ser utilizadas as seguintes colunas do arquivo de treinamento: *Name, DateTime, OutcomeType, AnimalType, SexuponOutcome, AgeuponOutcome, Breed e Color*.

Contudo, ao se iniciar o processo de pré-processamento dos dados notou-se que a coluna *DateTime* deveria ser descartada. Essa decisão foi baseada em observações feitas durante o processo e ao se pesquisar na plataforma Online Kaggle (ANDRAS ZSOM, 2016), de forma que se constatou que a variável iria interferir negativamente no estudo. Além do fato que em um cenário real não será sabida a hora em que os eventos de desfecho irão ocorrer, fazendo com que o uso da propriedade *DateTime* não seja propício para criação de um modelo prático para predições reais.

Assim sendo, os atributos a serem utilizados ficam sendo os seguintes: *Name, OutcomeType, AnimalType, SexuponOutcome, AgeuponOutcome, Breed e Color*.

2. Transformação e estudo dos atributos

A maioria das colunas selecionadas passou por uma série de tratamentos, para que as informações contidas nelas pudessem ser avaliadas de maneira mais relevante. Abaixo as colunas, e os processos sofridos por elas, são explicados com maior detalhe.

Durante o processo de transformação dos dados, análises exploratórias também foram realizadas, em vista de que é necessário se conhecer bem os dados com que se está trabalhando, para assim tomar decisões conscientes sobre quais variáveis manter ou descartar, e quais podem vir a ser trabalhadas de formas diferentes da que foi aplicada em primeiro momento. Juntamente com as explicações abaixo dos processos de ETL, são exibidos gráficos e constatações realizadas durante esse processo todo.

a. Name

Os dados da coluna *Name* foram tratados de modo a gerar uma coluna nova (chamada *HasName*), que indica se o animal tem ou não um nome. Os valores possíveis para o novo atributo são booleanos, de maneira que *True* indica que o animal tem um nome e *False* o oposto. Após seu uso para gerar o novo atributo, a coluna original foi descartada.

b. AgeuponOutcome

Tendo valores que estavam em escalas diferentes (dias, semanas, meses e anos) esta coluna precisava ser normalizada, pois do contrário não seria possível utilizar as suas informações para geração de modelos preditivos eficientes.

Sendo assim a unidade de medida escolhida para realizar a normalização foi a de menor grão, no caso, **dias**. É importante ressaltar que foi utilizado o valor de **30** dias para a unidade de medida "**meses**", pois é impossível saber ao certo quais meses estão sendo referidos, e o valor de 365 dias foi utilizado para representar anos - assim ignorando anos bissextos.

Após as transformações serem aplicadas os novos valores foram salvos numa nova coluna (a antiga sendo descartada) chamada *DaysUponOutcome*. Abaixo podemos observar a distribuição dos novos valores de idade.

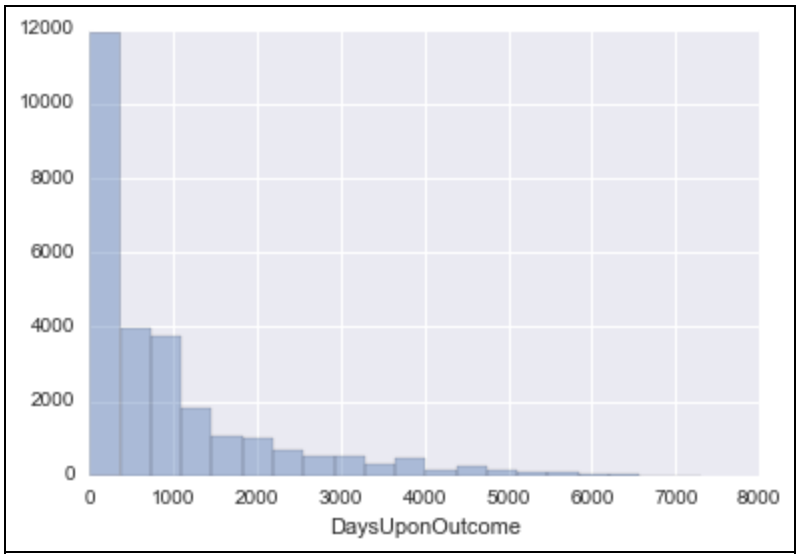

Figura 1: Distribuição dos valores da coluna *DaysUponOutcome.*

É aparente que a grande maioria dos animais é muito jovem, com mais da metade dos registros ficando abaixo da idade de 1000 dias.

Se criarmos uma nova característica indicando se o animal é jovem ou mais velho, podemos verificar, de forma eficiente, quais faixas de idade tem a maior possibilidade de terem um final positivo, ou as que necessitam de mais atenção e/ou incentivos por parte do canil. As faixas escolhidas foram: Jovem ([0, 3) anos), Jovem Adulto ([3, 5) anos), Adulto ([5, 10) anos) e Idoso ([10, +∞) anos). Abaixo o gráfico gerado nos mostra que animais mais novos, não surpreendentemente, tem taxas de adoção e transferência muito mais altas do que as outras faixas.

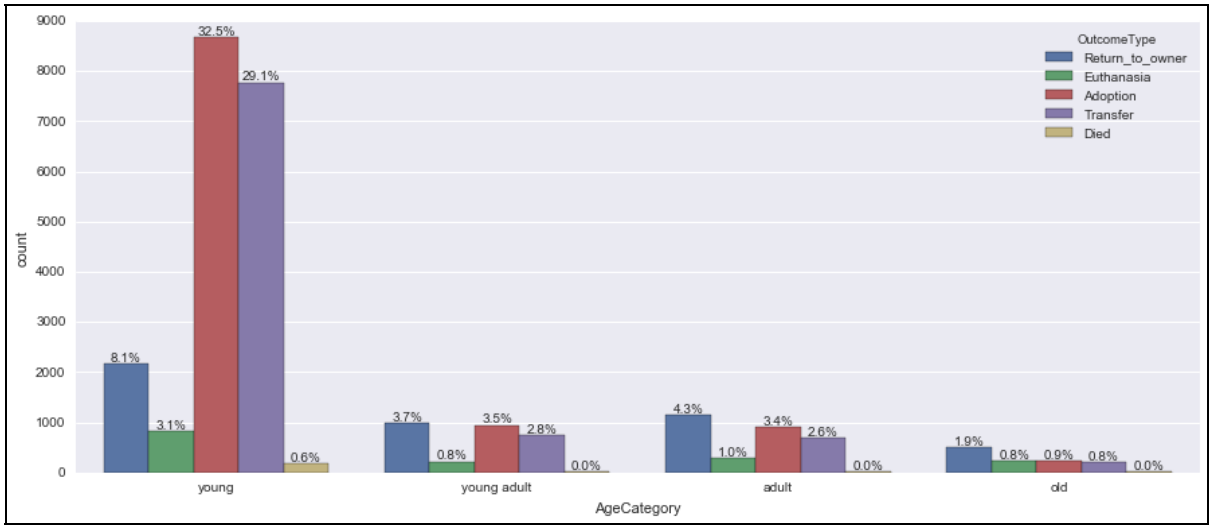

Figura 2: Tipo de desfecho por faixas de idade.

c. SexuponOutcome

Para este atributo foi decidido que seria melhor quebrar as informações contidas em duas novas colunas, que foram chamadas de *Sex* e *Castrated*, que indicam, respectivamente, o sexo do animal e se o mesmo é castrado.

Os valores possíveis são *Male*, *Female* e *Unknown* (para casos onde o valor do campo não foi preenchido apropriadamente) para a coluna *Sex* e, *True* ou *False* para a coluna *Castrated*.

Ao serem criados gráficos com as propriedades novas foi possível notar, como ilustrado pela figura 2, que há um número próximo de machos e fêmeas e que a maioria dos animais foi castrada (Figura 3).

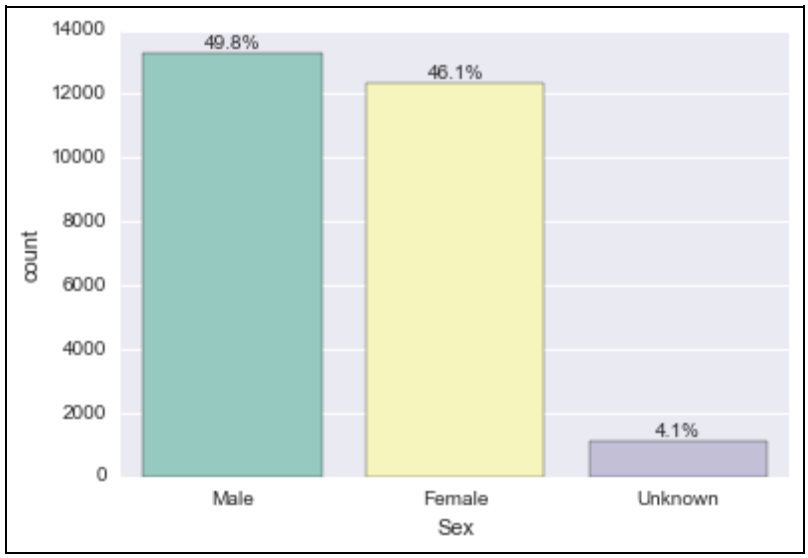

Figura 3: Distribuição dos sexos.

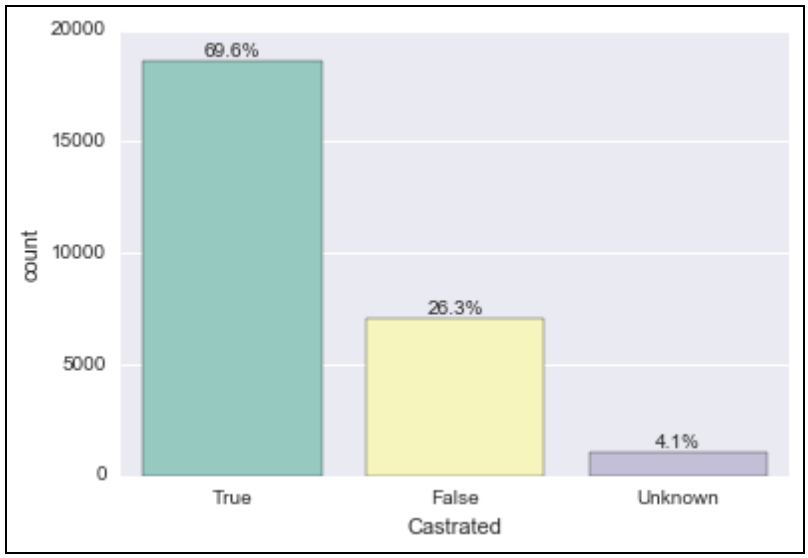

Figura 4: Proporção de animais castrados ou não.

Ao relacionarmos as novas informações com o desfecho de cada indivíduo do dataset, temos o insight de que o sexo do animal parece não influenciar muito em seu destino final, porém se o animal é castrado a taxa de adoção e transferência são muito mais elevadas, o que indica que este atributo vem a ter, possivelmente, uma correlação maior com a classe que se deseja prever.

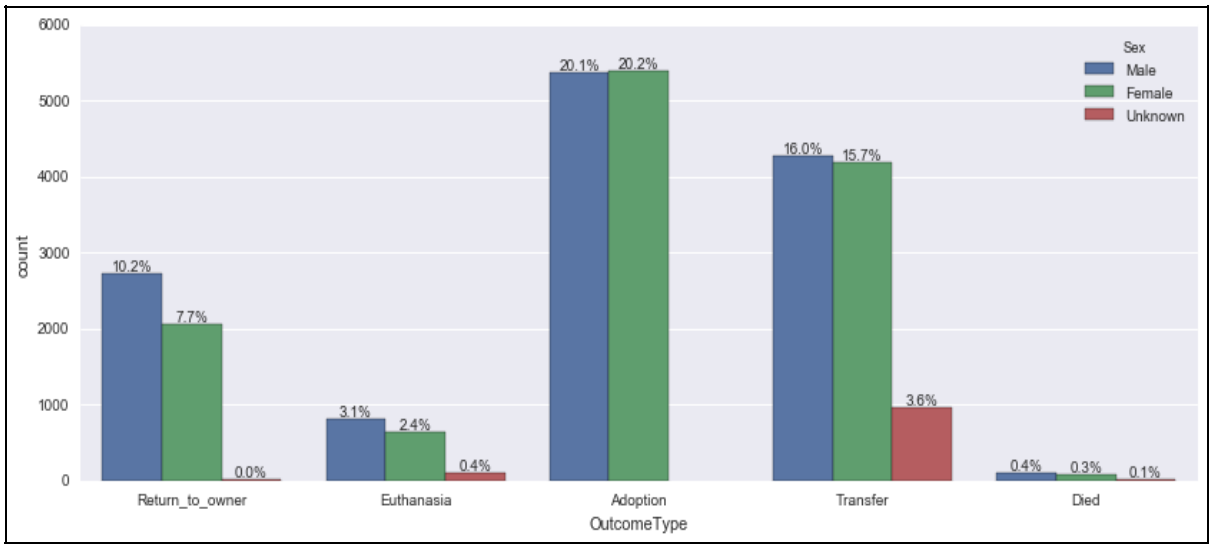

Figura 5: Relação do sexo com o desfecho do animal.

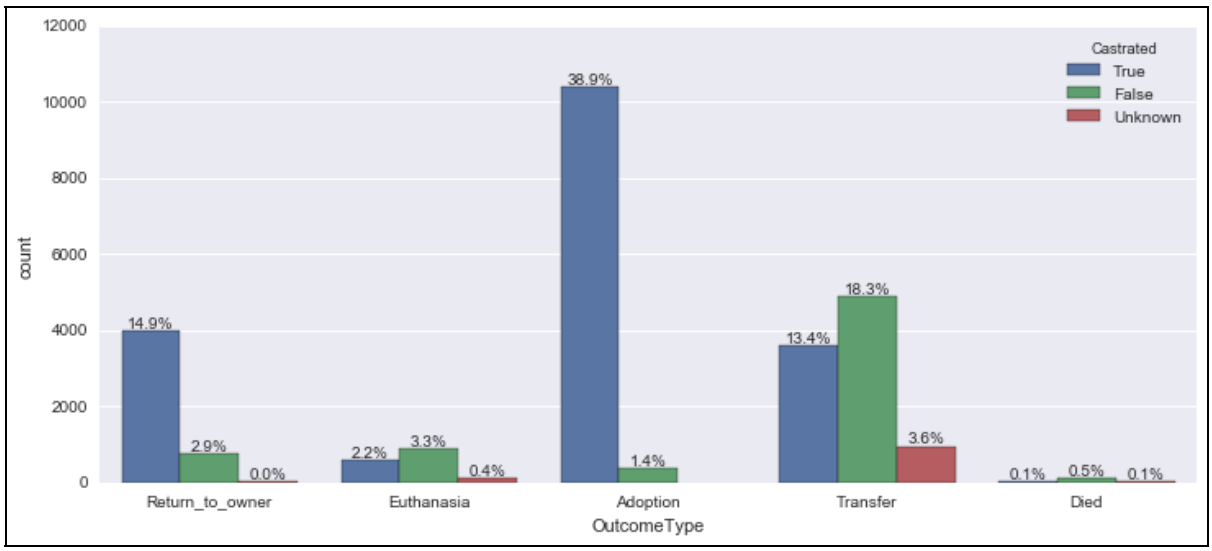

Figura 6: Relação da castração com o desfecho do animal.

d. Breed

Este atributo foi provavelmente o mais desafiador de ser trabalhado, pois a sua alta diversidade de valores possíveis, aliada ao fato de que as informações para gatos eram extremamente genéricas, descrevendo mais o tipo de pelagem do indivíduo do que sua raça, fizeram com que processos de generalização e transformação de atributos categóricos para numéricos fossem aplicados de maneiras diferentes para cães e gatos. Dessa maneira discutiremos cada processo de maneira separada.

i. Cães

A partir de uma fonte de raças de cães foi compilada uma lista de com as raças e suas principais características (ESTADOS UNIDOS, 2016), para que dessa maneira seja possível agrupar as diversas raças em algumas categorias básicas, que gerou uma nova coluna chamada *SimplifiedBreed*.

Os valores para a agregação de raças são: *Herding, Pit Bull, Non-Sporting, Terrier, Working, Sporting, Hound, Toy, Exotic* e *Domestic*. Estes 10 valores agregados generalizam as mais de 100 diferentes raças indicadas inicialmente. Pode-se ver a distribuição de cada valor abaixo no gráfico da figura 6.

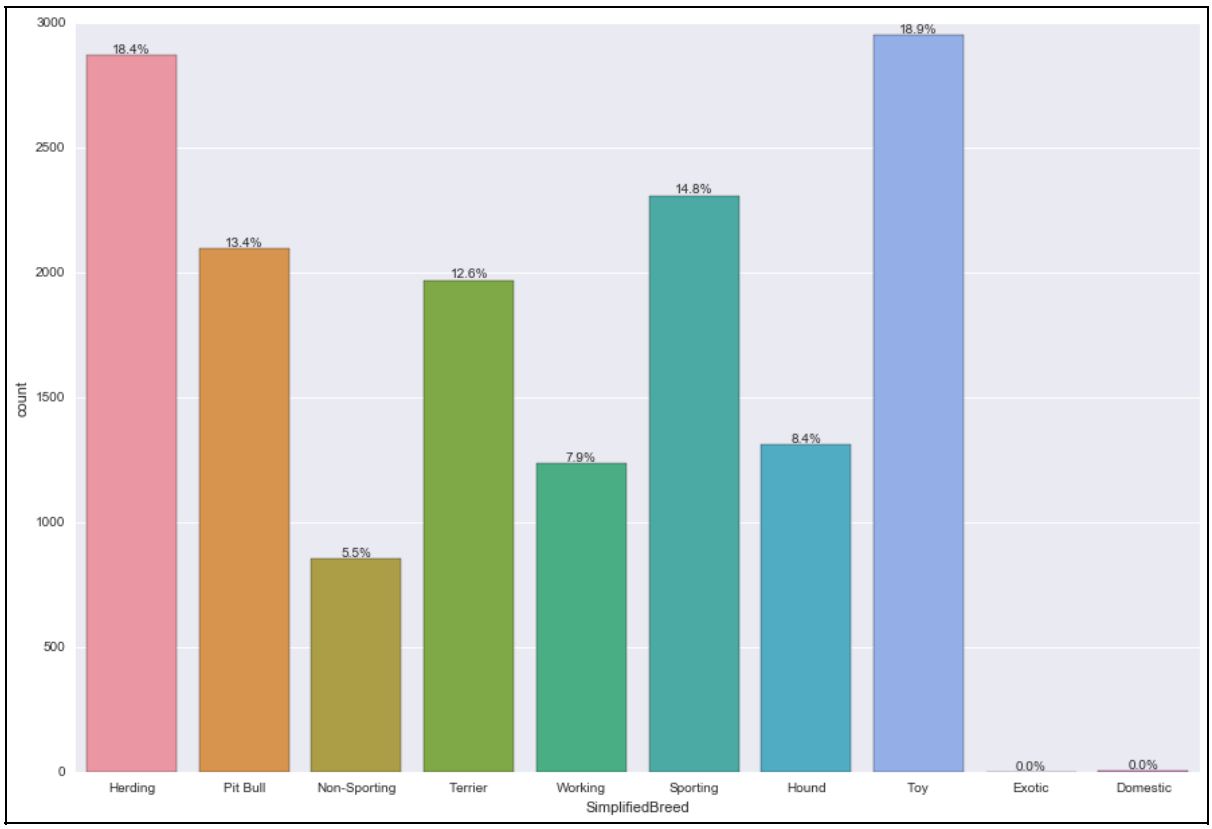

Figura 7: Distribuição das agregações de raças de cães.

A partir das informações de raça foi possível derivar novos atributos que indicam se o animal é um mestiço (duas colunas foram criadas, devido ao modo como misturas de raça são representadas na tabela original), se o tamanho do animal pode ser considerado "miniatura" e se a raça do animal é agressiva não. Dessa maneira é esperado que tenha-se retirado o maior número de informações relevantes das informações apresentadas por essa coluna, no caso dos cães.

ii. Gatos

Para os felinos parece existir uma dificuldade maior em se diferenciar entre raças, pois a maioria dos valores presentes era uma generalização baseada no tipo de pelagem (Ex.: Pelo Curto Doméstico). Dessa maneira se faz impossível realizar o mesmo processo que foi feito para cães.

Sendo assim, foi decidido agrupar as pelagens por 5 categorias: *Short Hair, Medium Hair, Long Hair, Wirehair* e *Hairless* (existindo o valor N/A que é aplicado para todos os cães). Essas categorias nominais foram mapeadas, respectivamente, para os valores numéricos de 1 a 5 - o número 0 representando N/A - para facilidade de processamento por parte de algoritmos de classificação.

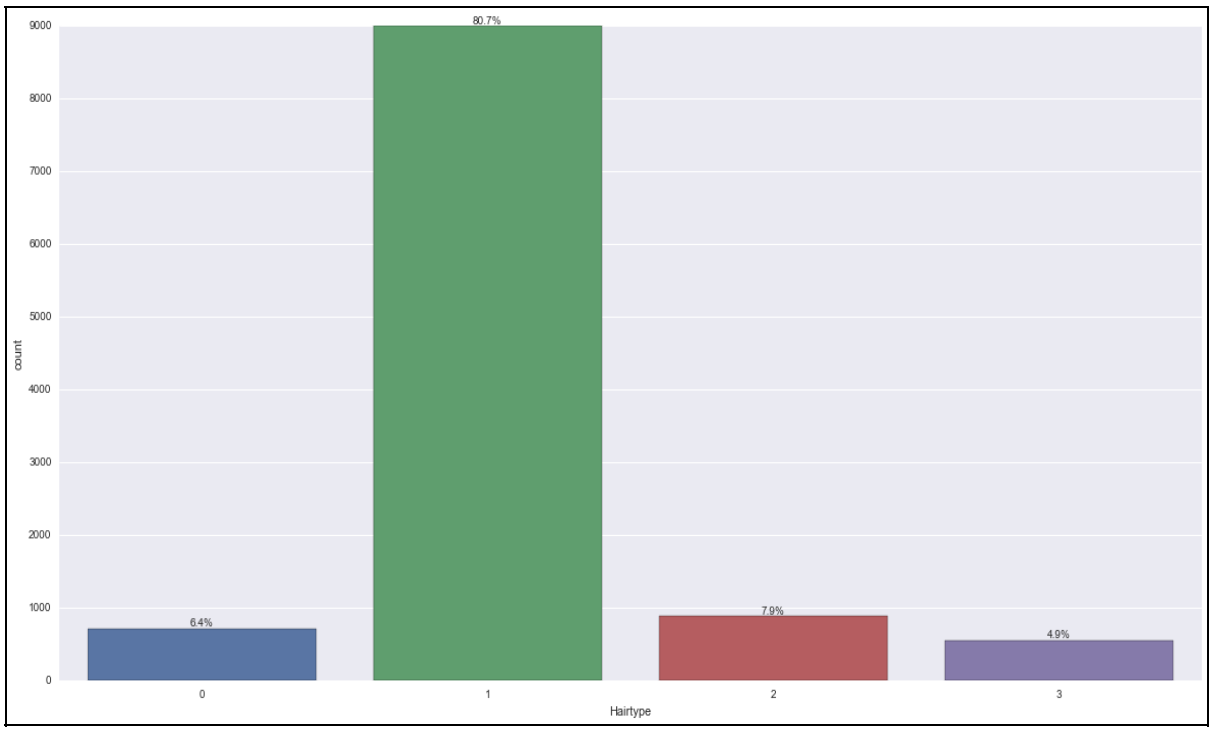

Figura 8: Distribuição dos tipos de pelagem de gatos.

e. Color

O processamento deste atributo ocorreu de forma similar ao das raças de cães do atributo breed. As cores foram divididas nos seguintes grupos: *Light* (cores claras). *Dark* (cores escuras), *Medium* (cores de intensidade média), combinações desses valores para animais com duas cores (*Medium/Dark*, *Dark/Light*) e Tricolor no caso de animais com três cores.

Com estes 10 grupos, foi possível representar os mais de 80 valores definidos para o atributo *Color*. O gráfico da figura 9 representa a distribuição dos novos valores do atributo.

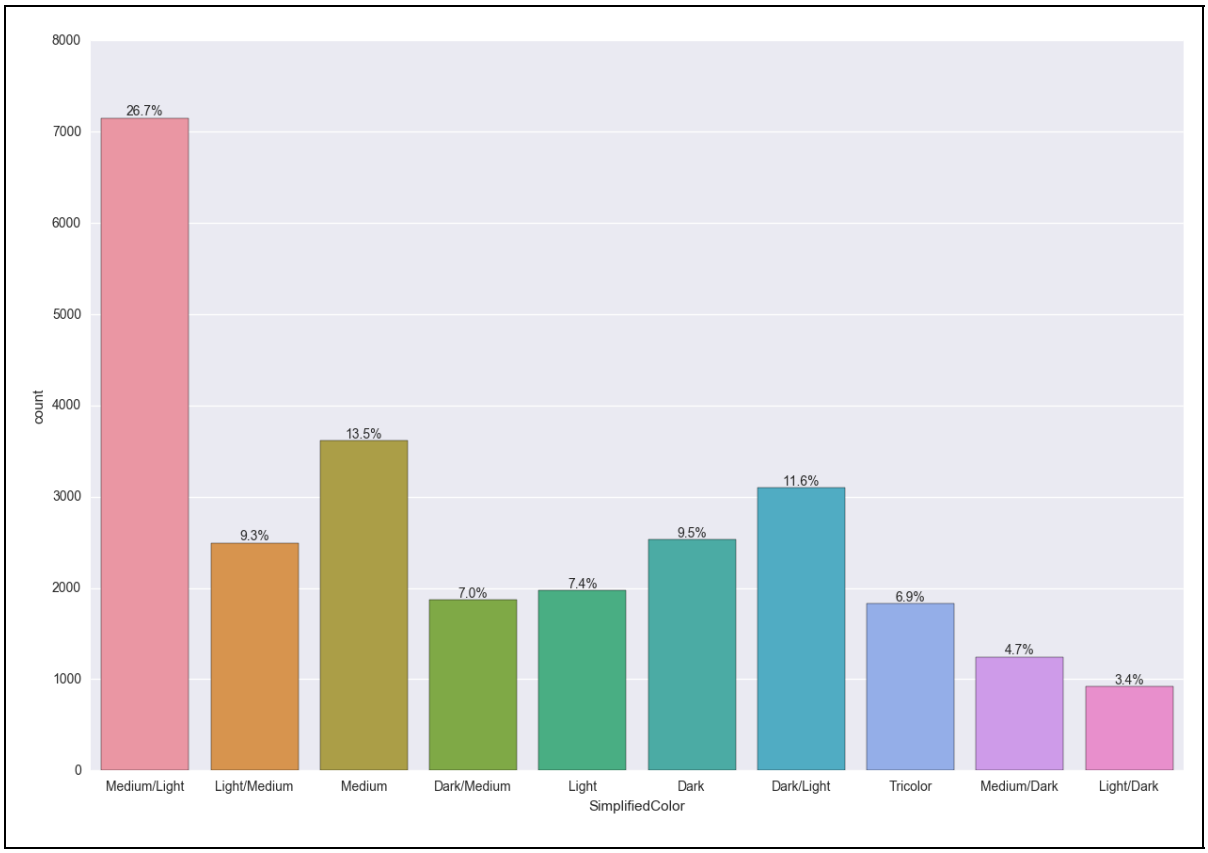

Figura 9: Distribuição das agregações de cores.

3. Mineração preliminar

Afim de se saber quais atributos são os mais relevantes, e se as operações e transformações aplicadas a cada atributo foram interessantes, uma mineração preliminar utilizando o algoritmo *Random Forest* (BREIMAN, 2001) foi realizada. É importante ressaltar que o atributo de coloração de pelagem foi ignorado para esta análise preliminar.

Na figura 10 podemos observar quais atributos o algoritmo classificou como mais relevantes para a predição da classe *OutcomeType*. Notamos que a idade do animal, e se ele é castrado ou não, são os fatores de maior peso na hora de decidir o destino de cada indivíduo. Podemos observar, também, que o agrupamento de raça e se o animal possui um nome são bons indicadores.

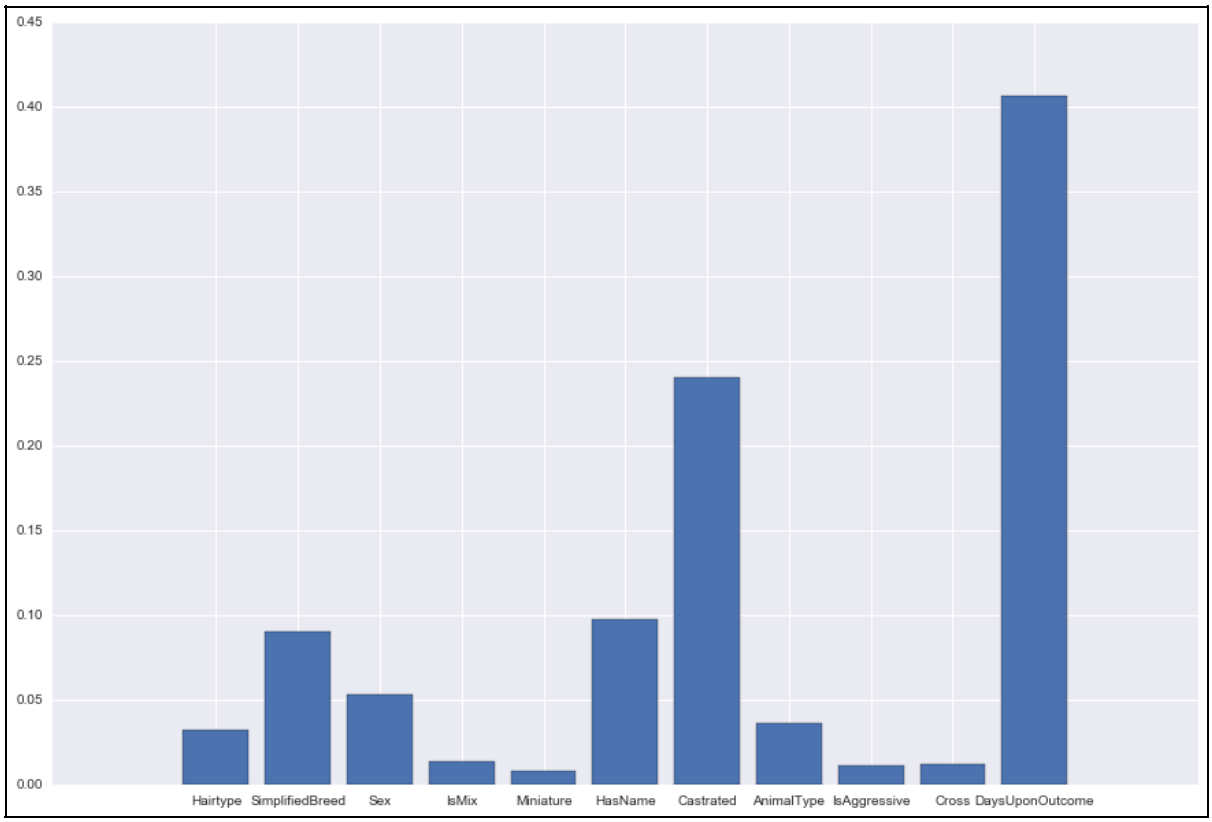

Figura 10: Gráfico de relevância dos atributos segundo algoritmo *Random Forest*.

Um ponto estranho neste gráfico é a importância da coluna *HasName*, que a primeira vista pensamos em se tratar de um atributo que iria contribuir pouco para a predição. A variável foi estudada com maior atenção, para se ter uma melhor idéia de sua importância, e relevância, para o resultado que se espera atingir com esta mineração.

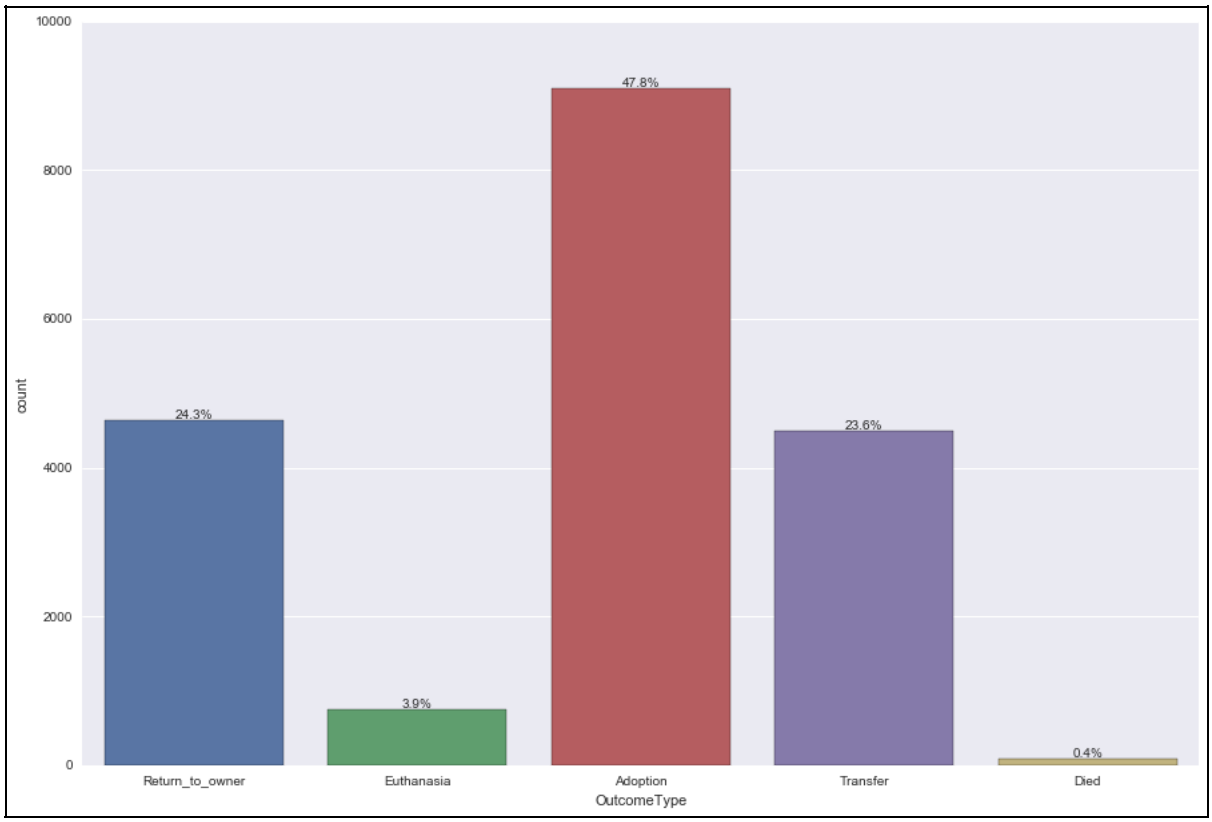

Figura 11: Gráfico da situação final para animais com nome.

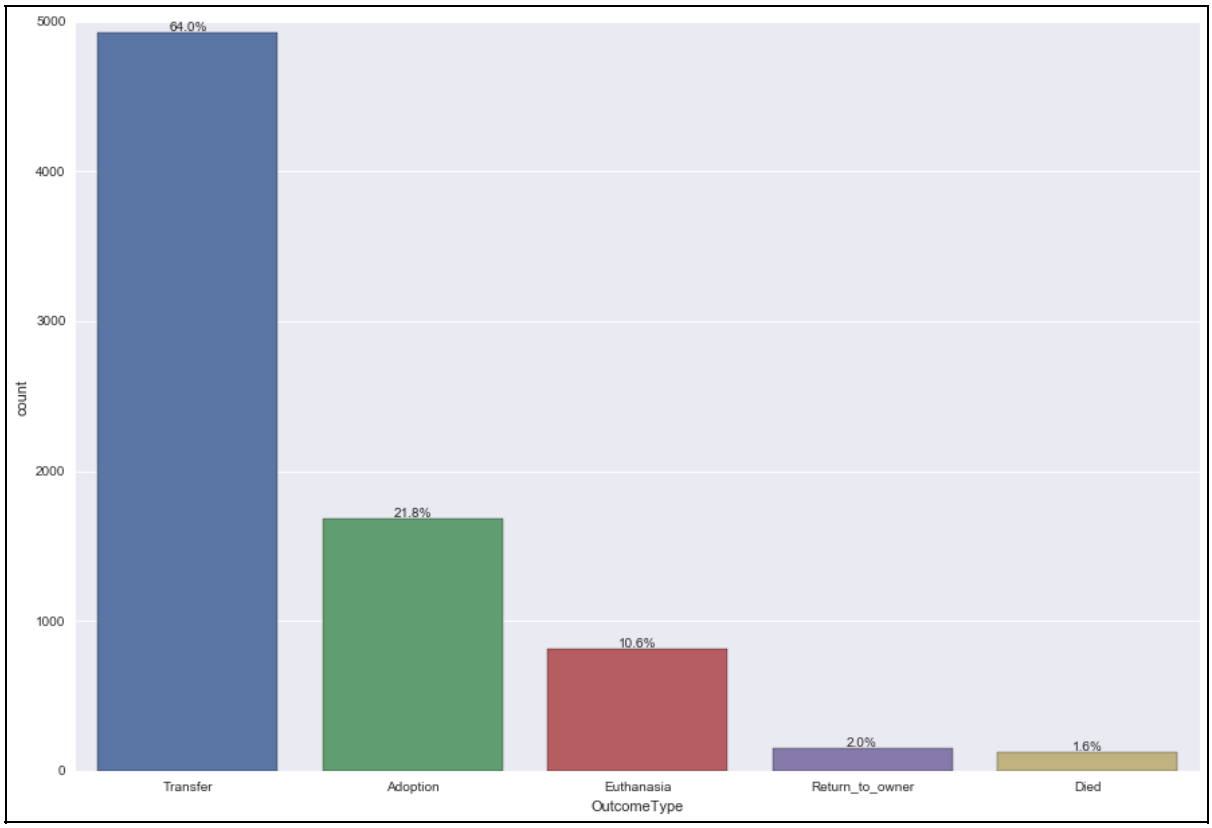

Figura 12: Gráfico da situação final para animais sem nome.

Ao analisarmos estes gráficos podemos notar a enorme discrepância nas taxas de adoção e transferência. Contudo, o que mais chama a atenção são as diferenças nas taxas de eutanásia e de devolução aos donos, que refletem, respectivamente, um aumento de 3 vezes e uma diminuição de 12 vezes em sua probabilidade de ocorrência. Podemos especular que isso se deve ao fato de que animais que possuem, ao chegar no abrigo, identificações com seus nomes estejam sendo procurados por seus donos, ou sejam mais dóceis (por já terem vivido algum tempo em um lar), do que animais chegando sem algum tipo de identificação. Dessa maneira decidimos que o atributo deve ser mantido, apesar de em primeiro momento acharmos que ele atrapalharia o processo de classificação.

De posse do conhecimento trazido à tona por essas análises e gráficos, podemos dizer que em momento futuro a decisão de remover alguns dos atributos menos relevantes pode vir a beneficiar o desempenho do algoritmo, visto que algumas dessas colunas podem estar servindo apenas como barulho. Mas é necessária uma análise mais profunda de atributos que apresentam valores de importância discrepantes, pois (como ilustrado pela presença, ou não, de um nome) pode-se descartar elementos importantes por se pensar que eles apenas dificultariam o bom funcionamento dos algoritmos a serem implementados.

4. Seleção dos possíveis algoritmos de classificação

Durante a análise preliminar dos dados o algoritmo *Random Forests* foi aplicado, porém acreditamos ser necessária a geração de diversos modelos preditivos criados através de outros algoritmos. Após certa pesquisa foi encontrado o algoritmo criado por Chen e Guestrin (2016) *XGBoost*, o qual se encaixa perfeitamente para resolver a tarefa proposta neste trabalho.

De conhecimento de ambos estes algoritmos foi decidido que seria mais efetivo selecionar o algoritmo que gerasse o melhor resultado na geração de apenas um modelo, sendo o *XGBoost* o vencedor e o escolhido para a execução da classificação.

A princípio o algoritmo será configurado para gerar 1000 iterações e particionar os dados em 5 partes distintas. Estas configurações iniciais podem vir a ser alteradas em data posterior, caso se constate que em se fazer isso a eficácia seja elevada.

5. Conclusões finais

Nesta etapa do trabalho foi realizado todo o processo de ETL dos dados, além de um estudo do que cada característica tem a oferecer para a avaliação. Também foram testados e selecionados algoritmos de classificação.

Como discutido em seções anteriores, é necessário aplicar os algoritmos e estudar os resultados, para que se seja possível decidir se é necessário alterar o processo realizado até o momento - ou adicionar mais informações de fontes externas - ou as configurações iniciais aplicadas ao algoritmo *XGBoost*.

Na próxima etapa do projeto espera-se que seja possível responder a esses questionamentos e, caso seja necessário, aplicar as devidas correções, para que seja obtido o melhor modelo preditivo o possível.

6. Referências

ANDRAS ZSOM (Estados Unidos). **Solution of team 'Kaggle for the paws', no outcome datetime features!** 2016. Disponível em: <https://www.kaggle.com/c/shelter-animal-outcomes/forums/t/22538/solution-of-team-kaggle-for-the-paws-no-outcome-datetime-features/130815#post130815>. Acesso em: 25 set. 2016.

ESTADOS UNIDOS. AMERICAN KENNEL CLUB. . **Dog Breeds.** 2016. Disponível em: <http://www.akc.org/dog-breeds/>. Acesso em: 25 set. 2016.

CHEN, Tianqi; GUESTRIN, Carlos. XGBoost. **Proceedings Of The 22nd Acm Sigkdd International Conference On Knowledge Discovery And Data Mining - Kdd '16,** [s.l.], p.1-13, 2016. Association for Computing Machinery (ACM). http://dx.doi.org/10.1145/2939672.2939785.

BREIMAN, Leo. Random Forests. **Machine Learning,** [s.l.], v. 45, n. 1, p.5-32, 2001. Springer Nature. http://dx.doi.org/10.1023/a:1010933404324.