Universidade de São Paulo - USP

Instituto de Ciências Matemáticas e de Computação - ICMC
Departamento de Ciências de Computação - SCC

SCC5871 - Algoritmos de Aprendizado de Máquina

Professor: André Carlos Ponce de Leon Ferreira de Carvalho

Alunos: Jorge Valverde Tohalino
Rayner Harold Montes Condori

Relatório 1

1. **Análise dos dados**

   Tipos de variáveis

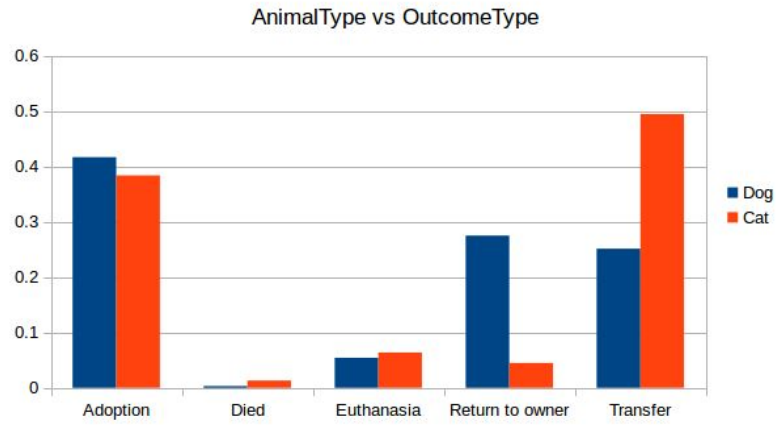
   a. **Numéricas:** DateTime, AgeUponOutcome
   b. **Categóricas:** Name, AnimalType, SexUponOutcome, Breed, Color, OutcomeType[Target]. Todas as variáveis são nominais.

2. **Análise da variáveis:**

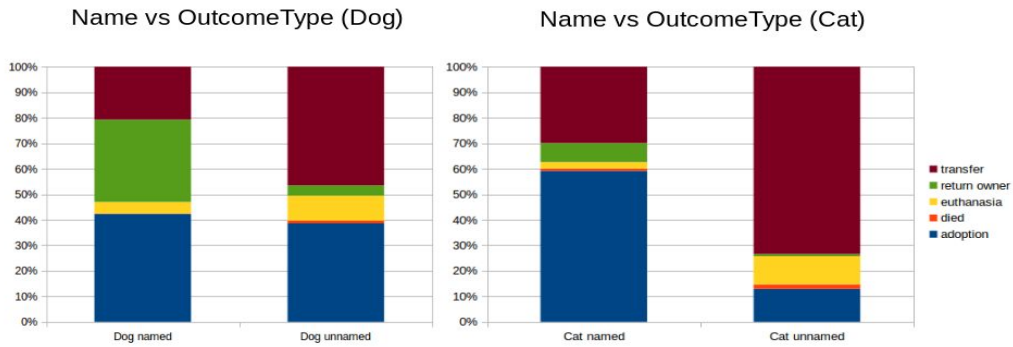
   a. OutcomeType (Target): A variável tem 5 categorias. Na seguinte tabela é mostrada a frequência e a porcentagem. As categorias Adoption e Transfer são as mais freqüentes.

OutcomeType	Frequência	Porcentagem
Adoption	10769	40.2896%
Died	197	0.7370%
Euthanasia	1555	5.8177%
Return_to_owner	4786	17.9056%
Transfer	9422	35.2501%

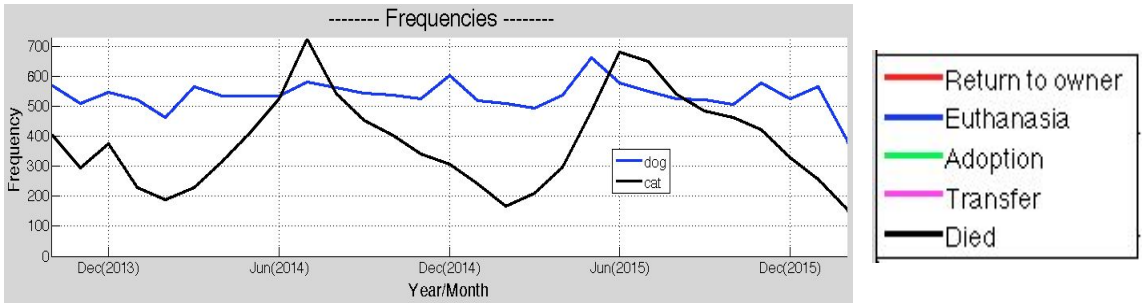
   b. AnimalType: Contém somente duas categorias: Cat (11134 amostras) e Dog (15595 amostras). Na figura a seguir pode ser visto que os cães e gatos são comumente adotados ou transferidos. Também, os cães são muito mais propensos a ser devolvidos aos seus donos do que os gatos.



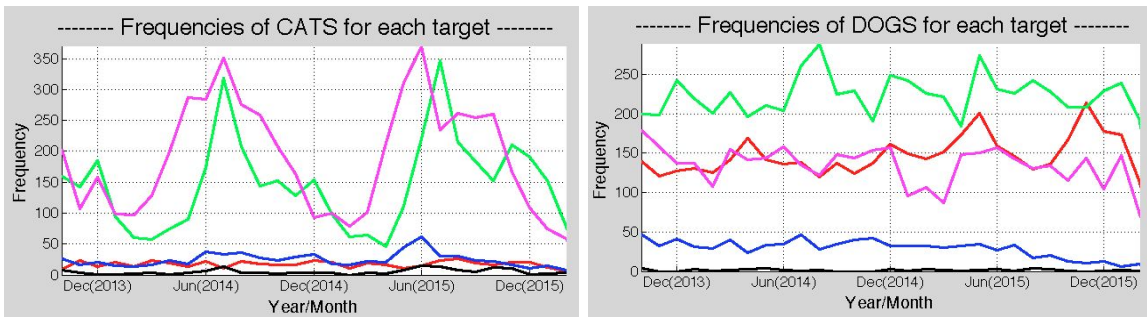
   c. Name: De acordo com a figura a seguir: cães com nomes são mais propensos a voltar ao seu proprietário e gatos são 4 vezes mais propensos a ser adotados se eles são nomeados. Aparentemente a variável Name tem importância.



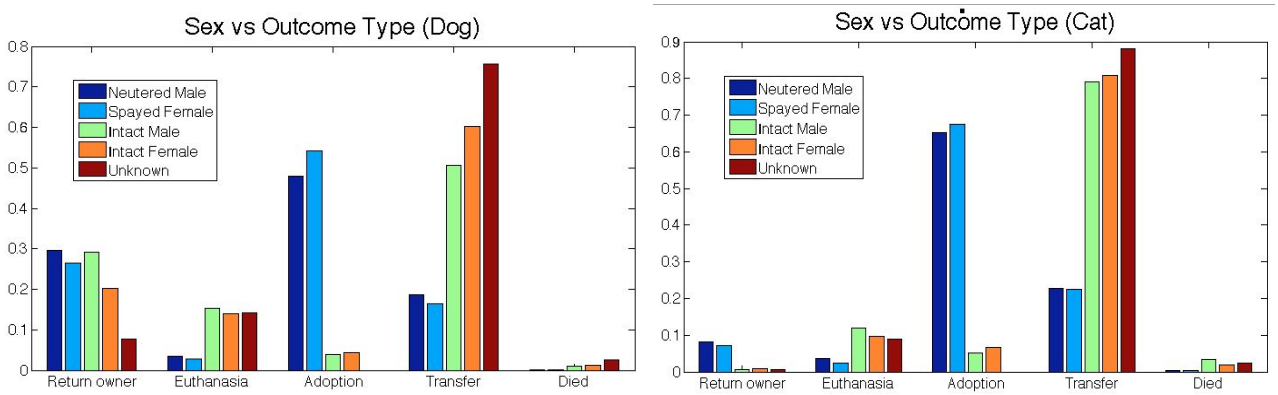
d. Datetime: Contém datas de Outubro/2013 até Fevereiro/2016. Pelo visto, indicam a data de entrada do animal ao abrigo. Na seguinte figura mostramos quantos gatos e cachorros entraram no abrigo cada mês. Podemos notar que a entrada de cachorros ao abrigo é relativamente constante. Porém, no caso dos gatos, tem meses onde a frequência de casos é muito alta (junho, julho) e outros onde é muito baixa (março, abril).



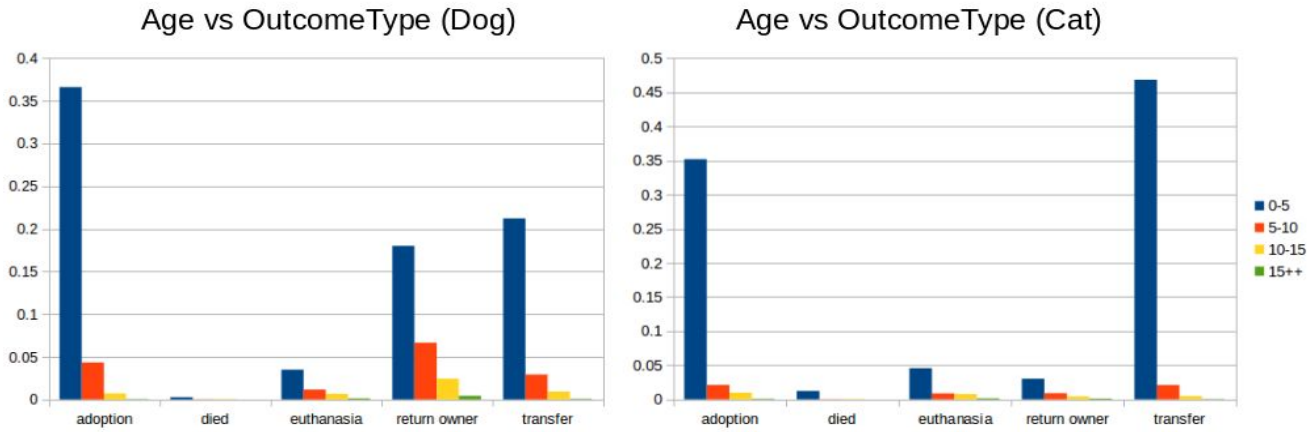
Ao fazer os gráficos de frequência por cada variável alvo, observamos a mesma tendência nas categorias “Transfer” e “Adoption”.



e. SexUponOutcome: Esta variável tem 5 categorias, nos gráficos seguintes são mostrados, em proporções, o destino dos gatos e cachorros em relação com esta variável. Note que os cachorros e gatos com categorias “Neutered Male” e “Spayed Female”, tem grandes probabilidades de serem adotados e é improvável de eles morrer por eutanásia. No caso das categorias Intact Male e Intact Female, eles tem maior probabilidade de serem transferidos.



f. **AgeUponOutcome:** Na figura a seguir pode ser visto que os animais jovens são mais propensos a ser adotados do que animais adultos. Eles também são mais propensos a ser transferidos. Também, cães jovens são propensos a ser devolvidos aos seus proprietários.



g. **Breed:** Esta variável tem 1380 categorias. Portanto, um passo aconselhável seria reduzir esse conjunto. Na tabela seguinte é mostrada a distribuição de frequências para esta variável.

AnimalType	Mix	No Mix	Total
Dogs	182	1138	1320
Cats	31	29	60

h. **Color:** Existem 366 tipos diferentes de cores. Onde 312 são cores compostos e 54 são cores únicos. Esta variável tem muitos níveis, portanto, devemos reduzir esses níveis. As segundas palavras de cada cor pode ser usada (Brindle, Tabby, Cream, etc.). Estas palavras poderiam descrever padrões importantes. A tabela a seguir mostra a distribuição de frequência.

Padrão	Frequência	Porcentagem
None	19094	71.4355%
Tabby	4904	18.3471%

Brindle	1018	3.8086%
Cream	608	2.2747%
Point	586	2.1924%
Merle	320	1.1972%
Tick	108	0.4041%
Smoke	75	0.2806%
Tiger	16	0.0599%

3. **Trabalhos Futuros:** Serão utilizados duas abordagens.
    a. ***Todas as variáveis categóricas:*** As variáveis *AgeUponOutcome* e *DataTime* têm que ser mudadas para valor categórico. *AgeUponOutcome* serão amostrado para os intervalos [0-5][5-10][10-15][15+]. No caso de *DateTime*, as categorias serão por anho ou mês. Finalmente, o algoritmo Naive Bayes será utilizado.
    b. ***Todas as variáveis numéricas:*** Nesta abordagem, todas as variáveis exceptuando *AgeUponOutcome* e *DataTime* devem ser transformadas para uma representação numérica. Devido a que as variáveis: *SexUponOutcome* e *AgeUponOutcome* apresentam poucas categorias, elas serão transformadas com One Hot encoding. A variável *AnimalType* é uma variável binária, portanto somente é necessário utilizar uma substituição com 0 e 1, de forma semelhante a variável *Name* será transformada para uma variável binaria com as categorias: “tem nome (0)” e “não tem nome (1)”. As variáveis *Breed* and *Color* contém muitas categorias, portanto não é factível usar One Hot encoding. Uma solução consiste em reduzir o número de categorias da variável. No caso da variável *Color*, o número de cores podem ser reduzidas, utilizando um mapeamento dessas cores para as suas cores primarias mais próximas, ou definir somente dois categorias: “Claro” e “Escuro”. No caso da variável Breed, nos fóruns de Kaggle propuseram mapear as raças dos cachorros, nas seguintes novas categorias: Toy, Hound, Sporting, Non-Sporting, Herding, Terrier, e Working. Finalmente, algoritmos de aprendizado de máquina como Redes Neurais e Random Forests serão aplicadas.
    c. Uma possível ideia seria utilizar ensembles, possivelmente (Redes Neurais, Naive Bayes, Random Forest).