Lucas May Petry
Maike de Paula Santos

Shelter Animal Outcomes

Definição do Problema de Data Mining

Universidade Federal de Santa Catarina
Departamento de Informática e Estatística
Ciências da Computação

Florianópolis
Abril de 2017

1. Definição do Problema

1.1. Motivação

Todo ano, inúmeros animais domésticos são abandonados por seus donos, se perdem ou até mesmo são encontrados e retirados de situações de crueldade nos Estados Unidos. Aproximadamente 7.6 milhões destes animais acabam indo para abrigos, dentre os quais 2.7 milhões de gatos e cães são submetidos à eutanásia.

Através de um conjunto de dados provido pelo *Austin Animal Center*, acredita-se que seja possível descobrir padrões e tendências para o desfecho destes animais. O objetivo é dedicar mais esforços em animais específicos, visando maximizar o número de animais de estimação que encontram um novo lar (KAGGLE, 2016).

1.2. O Conjunto de Dados Original

O conjunto de dados foi provido pelo *Austin Animal Center*, coletado entre outubro de 2013 e março de 2016. Todos os animais recebem um identificador único ao entrarem no abrigo e o que acontece com o animal no abrigo também é registrado. Os possíveis resultados para um animal são adoção, morte, eutanásia, retorno ao dono e transferência para outro abrigo (KAGGLE, 2016). Os atributos do conjunto de dados original são descritos na tabela a seguir.

Atributo	**Descrição**	**Valores Possíveis**
AnimalID	Identificador do animal	A671945, A656520...
Name	Nome do animal	Hambone, Emily, Elsa...
DateTime	Data do registro	Valor no formato yyyy-MM-dd HH:mm:ss
OutcomeType	Tipo do evento que determinou a saída do animal do abrigo	*Adoption, Died, Euthanasia, Return to owner, Transfer*
OutcomeSubtype	Subtipo do evento de saída do animal do abrigo	*Suffering, Foster, Partner, Offsite*...
AnimalType	Tipo do animal (cão, gato)	*Cat, Dog*
SexuponOutcome	Sexo e informação sobre castração ou esterilização do animal	*Intact Female, Intact Male, Neutered Male, Spayed Female, Unknown*

AgeuponOutcome	Idade do animal ao deixar o abrigo	1 *year*, 2 *years*, 3 *weeks*, 5 *months*...
Breed	Raça do animal	Shetland Sheepdog Mix...
Color	Cor do animal	*Brown/White, Tan*...

1.3. O Problema de Data Mining

O problema de data mining consiste, principalmente, em responder a seguinte pergunta: **qual será o desfecho de um animal, levando em conta todas as informações que se tem disponível do mesmo?**

Com os conjuntos de treino e teste obtidos da plataforma *Kaggle*, deseja-se obter o melhor modelo a partir do conjunto de treino, que classifique precisamente as instâncias do conjunto de teste.

1.4. Conjuntos de Dados Adicionais

Dois conjuntos de dados sobre cães, os quais consideramos potencialmente relevantes para o problema, foram encontrados na *web*.

O primeiro conjunto consiste em dados de altura e peso de cães de acordo com a raça do mesmo (DATA.WORLD, 2017). Os atributos do conjunto são descritos na tabela a seguir:

Atributo	**Descrição**	**Valores Possíveis**
Breed	Raça do cão	Akita, Anatolian...
height_low_inches	Limiar inferior da altura em polegadas da raça	0, 1, 2...
height_high_inches	Limiar superior da altura em polegadas da raça	0, 1, 2...
weight_low_lbs	Limiar inferior do peso em libras da raça	0, 1, 2...
weight_high_lbs	Limiar superior do peso em libras da raça	0, 1, 2...

Um segundo conjunto de dados sobre a inteligência de cães também foi obtido. Os dados classificam raças de cães de acordo com a probabilidade de que ele obedeça ao primeiro comando e o número de repetições necessárias para aprender um comando. Os dados provém de uma pesquisa do professor Stanley Coren da *University of British Columbia* (DATA.WORLD, 2017). Os atributos são descritos a seguir:

Atributo	**Descrição**	**Valores Possíveis**
Breed	Raça do cão	Border Collie, Poodle...
Classification	Classificação do animal de acordo com a probabilidade de que ele obedeça ao primeiro comando	*Brightest Dogs, Excellent Working Dogs, Above Average Working Dogs, Average Working/Obedience Intelligence, Fair Working/Obedience Intelligence, Lowest Degree of Working/Obedience Intelligence*
obey	Limiar inferior da probabilidade de que o animal obedeça ao primeiro comando	0 - 100%
reps_lower	Limite inferior de repetições para aprender novos comandos	0, 1, 2...
reps_upper	Limite superior de repetições para aprender novos comandos	0, 1, 2...

Adicionalmente à estes dados, encontramos três *websites* que contém as mesmas informações, entre outras, os quais podem ser relevantes para o problema:

- *Dogs Breeds List* (http://www.dogbreedslist.info);
- *Cats Breeds List* (http://www.catbreedslist.com);
- Portal do Dog (http://portaldodog.com.br/cachorros/racas-de-cachorros).

Porém, devido à dificuldade de adaptar os dados ao problema proposto, foi decidido não utilizá-los na primeira iteração do processo de KDD.

2. Pré-Processamento dos Dados

2.1. Atributos Selecionados

Os atributos do conjunto original de dados foram pré-processados e são descritos na tabela abaixo. Em seguida, descrevemos, para cada atributo, as transformações realizadas no mesmo, motivação para utilização do atributo e análise do mesmo.

Atributo	**Tipo**	**Descrição**	**Vazio?**	**Valores Possíveis**
hasName	Qualitativo (Booleano)	Valor indicando se o animal possui nome	Não	*true, false*
animalType	Qualitativo (Texto)	Valor indicando o tipo do animal	Não	*Cat, Dog*
sex	Qualitativo (Caractere)	Sexo do animal	Sim	F, M
isIntact	Qualitativo (Booleano)	Valor indicando se o animal não foi castrado/esterilizado	Sim	*true, false*
monthsOld	Quantitativo (Inteiro)	Idade do animal em meses.	Sim	0, 1, 2...
breed1	Qualitativo (Texto)	Raça do animal	Não	Cairn Terrier, Pit Bull Mix...
breed2	Qualitativo (Texto)	Raça secundária do animal	Sim	Labrador Retriever...
isMix	Qualitativo (Booleano)	Valor indicando se a raça é um mix	Não	*true, false*
color1	Qualitativo (Texto)	Cor do animal	Não	*Red, Black, White...*
color2	Qualitativo (Texto)	Cor secundária do animal	Sim	*White, Tan, Black...*
outcome (classe)	Qualitativo (Texto)	Desfecho do animal no abrigo.	Não	*Adoption, Died, Euthanasia, Return to owner, Transfer*

2.1.1. Atributo *hasName*

Este atributo indica se um animal possui um nome ou não. Acreditamos que a existência ou não de um nome para um animal tenha influência no desfecho do mesmo. Por exemplo, uma pessoa provavelmente gostaria de adotar um animal que já possui um nome, de modo que ela não precise ensinar a ele um novo nome. Segundo os dados do *Austin Animal Center*, Aproximadamente 71% dos animais possui nome.

O gráfico abaixo relaciona o desfecho do animal em relação à existência de um nome ou não. Note que, apesar de a proporção de animais que possuem nome ser bem maior, o número absoluto de animais classificados como *Died* ou *Euthanasia* parecem ser bem similares, o que nos sugere que talvez animais sem nome estejam mais propensos a Eutanásia ou morte por outras causas.

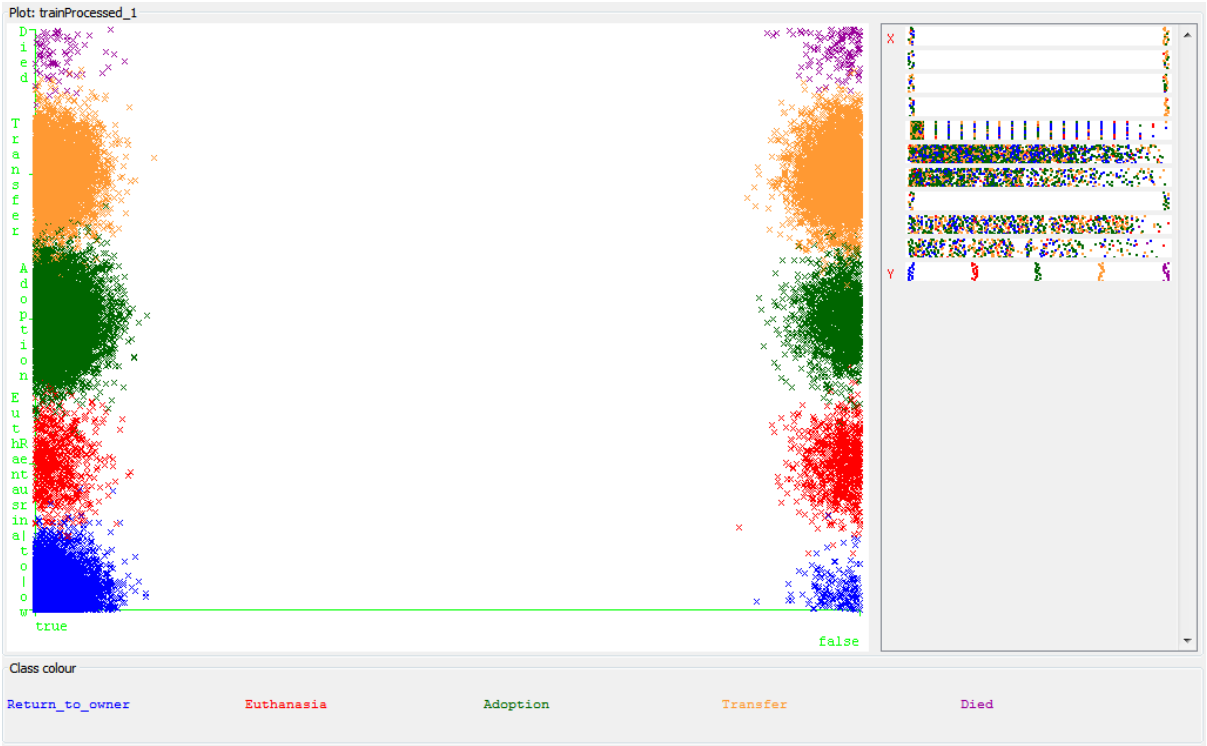

Figura 2.1: Gráfico *outcome* vs. *hasName*.

2.1.2. Atributo *animalType*

O atributo *animalType* indica se o registro é de um cão ou de um gato. Ele é importante, pois há divergência entre pessoas quanto à preferência de gatos ou cães. No conjunto de dados provido, cerca de 58% dos animais são cães. No gráfico da figura 2.2, podemos visualizar o tipo do animal (vermelho para gato e azul para cão) para cada desfecho possível. Podemos destacar, por exemplo, que gatos tendem a sofrer transferência com mais frequência do que cães, visto que o número

absoluto de gatos transferidos é maior do que o de cães (e proporcionalmente maior).

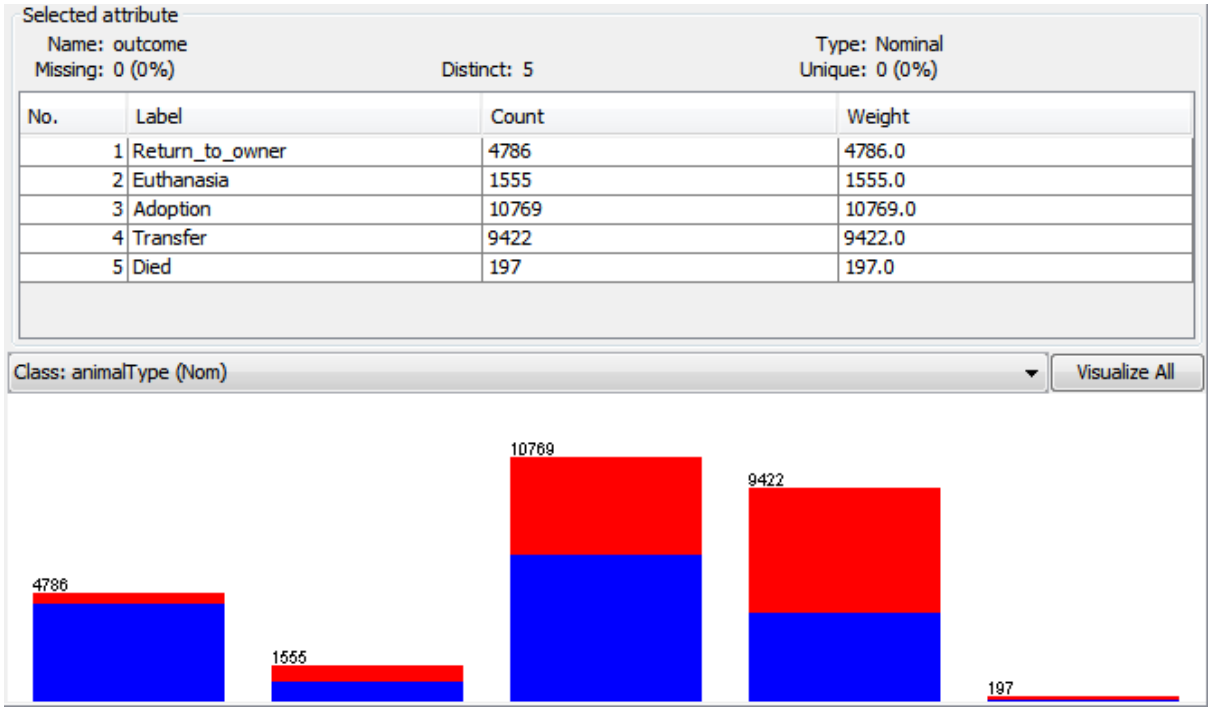

Figura 2.2: Gráfico *outcome* vs. *animalType*.

O gráfico da figura 2.3 relaciona o tipo do animal com o atributo *hasName*. Em uma análise do gráfico, percebemos que há uma maior quantidade de transferências para gatos que não tenham nome.

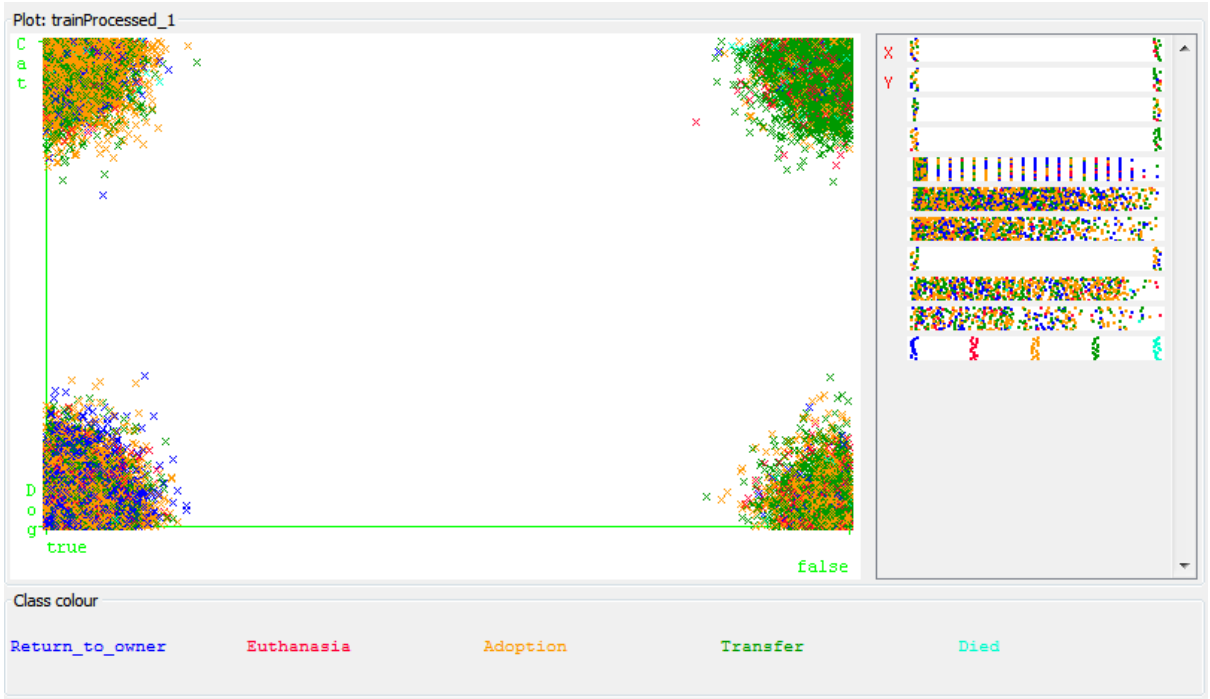

Figura 2.3: Gráfico *animalType* vs. *hasName*.

2.1.3. Atributos *sex* e *isIntact*

Os atributos *sex* e *isIntact* foram derivados do atributo original *SexuponOutcome*. Pensamos que seria mais apropriado considerar o sexo e o uma castração/esterilização separadamente. Assim, se o animal está intacto, *isIntact* recebe o valor *true*. Avaliamos a figura 2.4 e podemos visualizar, por exemplo, que há muitas transferências para animais que estão intactos. Adoções se concentram, em sua maioria, para animais que não estejam intactos, o que talvez se deve ao fato de existirem muito mais animais castrados/esterilizados do que intactos.

Figura 2.4: Gráfico *isIntact* vs. *sex*.

2.1.4. Atributo *monthsOld*

O atributo original *AgeuponOutcome* foi transformado e padronizado de forma a se construir o atributo *monthsOld*, que representa a idade do animal em meses. Esta transformação foi fundamental, pois a idade agora é um valor contínuo e mais facilmente visualizável. Animais com menos de um mês de vida foram representados com o valor zero.

A figura 2.5 relaciona a idade do animal de acordo com a situação do mesmo quanto à castração/esterilização. Pode-se notar que muitos animais intactos e com poucos anos de vida são transferidos. Já para os animais não intactos, a quantidade de adoções se concentra também para animais jovens.

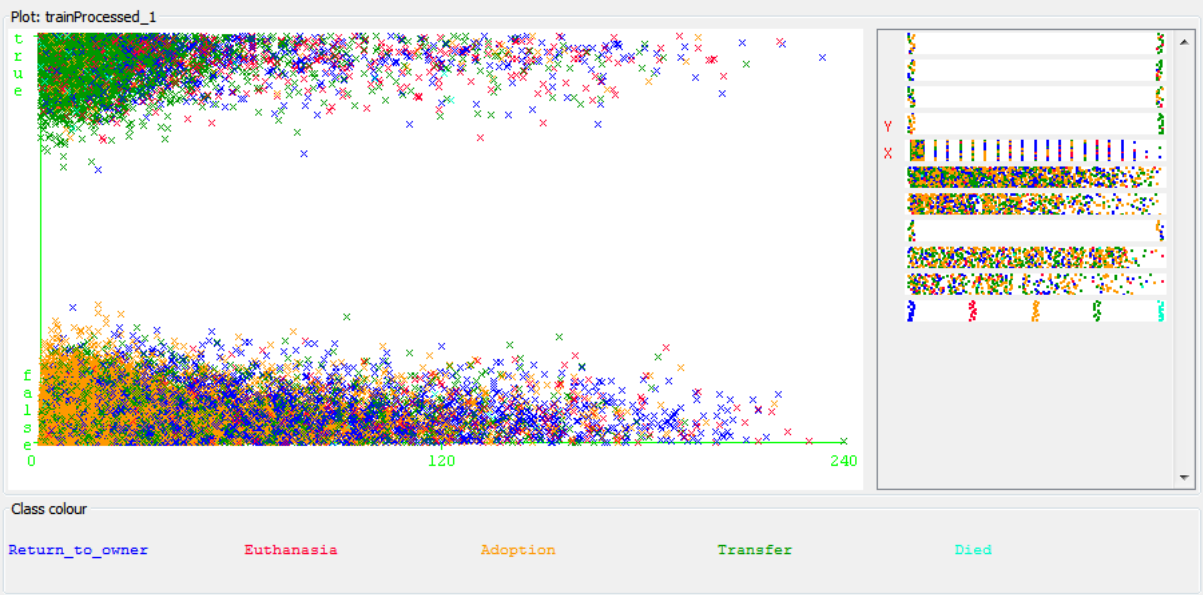

Figura 2.5: Gráfico *isIntact* vs. *monthsOld*.

2.1.5. Atributos *breed1*, *breed2* e *isMix*

O atributo *Breed* foi transformado nos atributos *breed1*, *breed2* e *isMix*. *breed1* representa a raça primária do animal, enquanto que *breed2* representa a raça secundária. Para as raças que eram uma mistura (continham a palavra *Mix*), criamos o atributo *isMix* para indicar tal fato. Com essas transformações, o número de valores distintos para a raça caiu consideravelmente. Com isso, os algoritmos de mineração podem encontrar padrões para raças puras, por exemplo, que não acontecem para misturas de raças.

O gráfico da figura 2.6 relaciona os possíveis desfechos com o atributo *isMix*, onde azul representa verdadeiro e vermelho, falso.

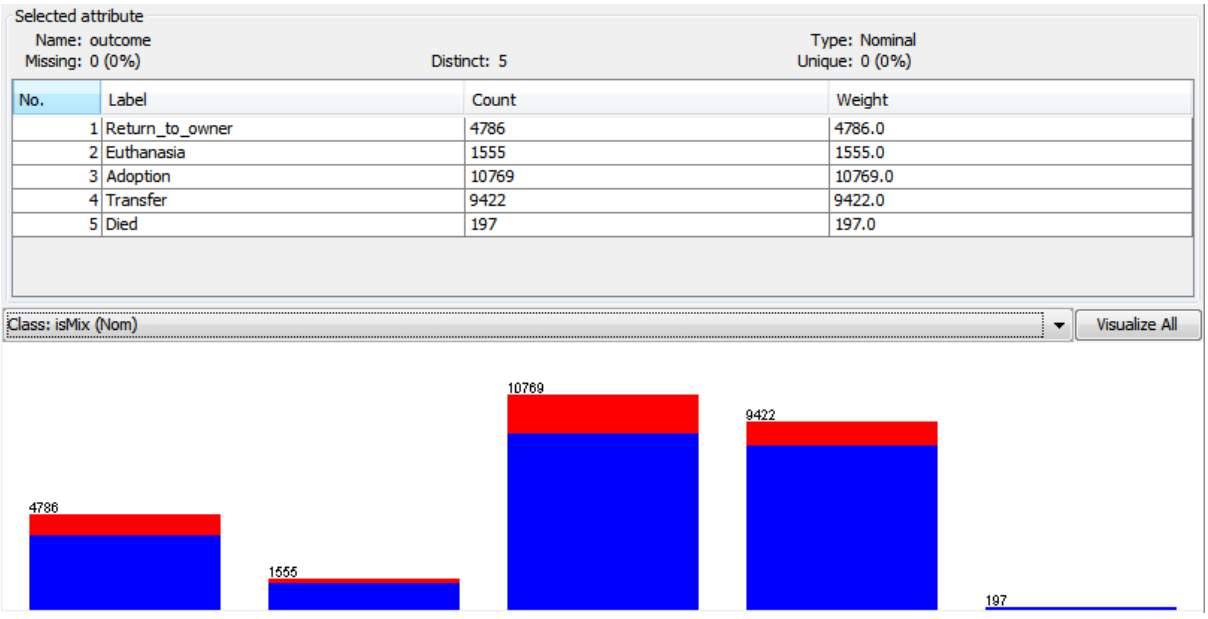

Figura 2.6: Gráfico *outcome* vs. *isMix*.

2.1.6. Atributos *color1* e *color2*

O atributo *Color* originalmente possuía valores do tipo “Cor/Cor”, significando que o animal possui duas cores distintas. Então, para facilitar o entendimento do atributo foram criados dois outros atributos, *color1* e *color2*, utilizando o caractere “/” como separador. Em casos onde o animal possui apenas uma cor o atributo *color2* fica vazio. 48% dos animais possuem apenas uma cor.

2.2. Atributos Desconsiderados

Dentre os atributos selecionados, foram desconsiderados os valores dos atributos ID, *Name* e *DateTime*.

O atributo ID não é interessante para a mineração, uma vez que cada animal possui um ID único e novos animais também receberão identificadores diferentes.

Do atributo *Name* nós apenas derivamos o atributo *hasName*, explicado anteriormente. Não seria interessante considerar os nomes dos animais, pois nome é um atributo baseado em preferências pessoais.

Finalmente, não consideramos o atributo *DateTime* por não termos informações suficientes a respeito dele. O autor dos dados não menciona se o atributo representa a data de ocorrência do desfecho, a data de registro no sistema, a data de entrada do animal no abrigo ou a data de saída do animal do abrigo. Por essa razão, não seria pertinente considerar padrões que levem em conta o *DateTime*.

3. Técnica de Mineração que Será Utilizada

Dentre os algoritmos vistos em aula, temos como possibilidade utilizar o k-NN, *k-Nearest Neighbor*s, que utiliza uma métrica pré-definida específica para o problema que diz o quão parecido um dado novo é em relação aos outros já existentes. Este não é um bom algoritmo para resolver o problema deste trabalho, pois é muito difícil definir uma métrica de similaridade para os dados, além de que seu desempenho pode ser afetada pela quantidade de dados existentes.

Além deste, temos o SVM, *Support Vector Machine*, que possui um desempenho excelente para casos onde os dados não possuem ruídos e é necessário apenas classificar estes em duas classes, o que não é o caso.

Outra opção é utilizar o algoritmo de *Naïve Bayes*, que é uma abordagem puramente estatística ao problema, porém este considera que não existe dependência entre os atributos.

O algoritmo ID3 seria um ótimo candidato para resolver o problema de mineração dos dados, por ser relativamente simples e eficiente. Contudo, ele é utilizado somente em conjuntos de dados onde os valores das variáveis são discretos e não existem campos com valores vazios.

Por fim, o algoritmo C4.5 foi desenvolvido a partir do ID3, resolvendo as limitações previamente apontadas. Logo, concluímos que este seria o algoritmo mais adequado para abordar o problema em questão.

Referências

KAGGLE. **Shelter Animal Outcomes:** Help improve outcomes for shelter animals. 2016. Disponível em: <https://www.kaggle.com/c/shelter-animal-outcomes>. Acesso em: 17 abr. 2017.
DATA.WORLD (Comp.). **Dog/Canine Breed Size (AKC).** Disponível em: <https://data.world/len/dog-canine-breed-size-akc>. Acesso em: 17 jul. 2017.
DATA.WORLD (Comp.). **Dog size/intelligence linked?** Disponível em: <https://data.world/len/dog-size-intelligence-linked>. Acesso em: 17 jul. 2017.