Lucas May Petry
Maike de Paula Santos

Shelter Animal Outcomes

Data Mining

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

Florianópolis
Junho 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

Dados adicionais sobre as raças de cães e gatos foram coletados dos *websites Dogs Breeds List* (DOG... 2017) e *Cats Breeds List* (CAT... 2017). Os atributos extraídos são descritos na tabela a seguir.

Atributo	**Descrição**	**Valores Possíveis**
type	Tipo do animal	*Cat, Dog*
breed	Raça do animal	*Labrador Retriever, German Shepherd...*
other_names	Outros nomes da raça	*Labrador, Lab, Alsatian (UK)...*
size	Tamanho do animal	*Small, Medium, Large, Largest, Giant*
life_span_low	Tempo de vida mínimo estimado (em anos)	7, 8, 9, 10....
life_span_high	Tempo de vida máximo estimado (em anos)	10, 11, 12, 13...
price_low	Preço mínimo estimado (em dólares americanos)	80, 100, 200, 300...
price_high	Preço máximo estimado (em dólares americanos)	500, 600, 1000, 2000...

adaptability	Valor indicando o grau de adaptabilidade do animal	0, 1, 2, 3, 4, 5
child_friendly	Valor indicando quão amigável com crianças o animal é	0, 1, 2, 3, 4, 5
cat_dog_friendly	Valor indicando quão amigável com cachorros (para gatos) ou com gatos (para cachorros) o animal é	0, 1, 2, 3, 4, 5
grooming	Valor indicando a necessidade de cuidados de higiene, limpeza, etc	0, 1, 2, 3, 4, 5
health_issues	Valor indicando a propensão do animal a problemas de saúde	0, 1, 2, 3, 4, 5
intelligence	Valor indicando a inteligência do animal	0, 1, 2, 3, 4, 5
shedding_level	Valor indicando o quanto o animal solta pêlos.	0, 1, 2, 3, 4, 5

2. Pré-Processamento dos Dados

2.1. Descrição Geral da Coleta e Pré-Processamento

Para a realização do pré-processamento dos dados nós utilizamos a linguagem python. Do conjunto de dados originais extraímos e derivamos atributos, bem como padronizamos os valores dos atributos não padronizados. Os atributos são descritos na próxima seção.

A coleta dos dados dos websites teve uma certa complexidade. Primeiramente, elaboramos *crawlers* para a leitura das páginas web e extração dos dados relevantes para o problema. Apesar de os dados de cães e gatos terem sido extraídos de duas fontes diferentes, em sua maioria os dados mantinham as mesmas propriedades, o que tornou relativamente fácil a união dos dois conjuntos. Por fim, os dados foram padronizados aplicando-se transformação de unidades, derivação de atributos, entre outros.

A última etapa do pré-processamento compreendeu a complementação dos dados originais com os dados extraídos da *web*. Como a descrição das raças dos animais não seguia o mesmo padrão nos dois conjuntos, um *script* foi criado para fazer o mapeamento das raças dos conjuntos. Para comparação das raças utilizamos a medida de similaridade *Edit Distance*. Uma última análise manual foi necessária para verificação e inserção manual de cerca de 10% de todas as raças dos conjuntos. Após a realização deste, foi possível unir os dados adicionais ao conjunto de dados original.

2.2. 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...*
size	Qualitativo (Texto)	Tamanho do animal	Sim	*Small, Medium, Large, Largest, Giant*
life_span_low	Quantitativo (Inteiro)	Tempo de vida mínimo estimado (em anos)	Sim	7, 8, 9, 10....
life_span_high	Quantitativo (Inteiro)	Tempo de vida máximo estimado (em anos)	Sim	10, 11, 12, 13...
price_low	Quantitativo (Inteiro)	Preço mínimo estimado (em dólares americanos)	Sim	80, 100, 200, 300...
price_high	Quantitativo (Inteiro)	Preço máximo estimado (em dólares americanos)	Sim	500, 600, 1000, 2000...
adaptability	Qualitativo (Inteiro)	Valor indicando o grau de adaptabilidade do animal	Sim	0, 1, 2, 3, 4, 5
child_friendly	Qualitativo (Inteiro)	Valor indicando quão amigável com	Sim	0, 1, 2, 3, 4, 5

		crianças o animal é		
cat_dog_friendly	Qualitativo (Inteiro)	Valor indicando quão amigável com cachorros (para gatos) ou com gatos (para cachorros) o animal é	Sim	0, 1, 2, 3, 4, 5
grooming	Qualitativo (Inteiro)	Valor indicando a necessidade de cuidados de higiene, limpeza, etc	Sim	0, 1, 2, 3, 4, 5
health_issues	Qualitativo (Inteiro)	Valor indicando a propensão do animal a problemas de saúde	Sim	0, 1, 2, 3, 4, 5
intelligence	Qualitativo (Inteiro)	Valor indicando a inteligência do animal	Sim	0, 1, 2, 3, 4, 5
shedding_level	Qualitativo (Inteiro)	Valor indicando o quanto o animal solta pêlos.	Sim	0, 1, 2, 3, 4, 5
outcome (classe)	Qualitativo (Texto)	Desfecho do animal no abrigo.	Não	*Adoption, Died, Euthanasia, Return to owner, Transfer*

2.2.1. 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.

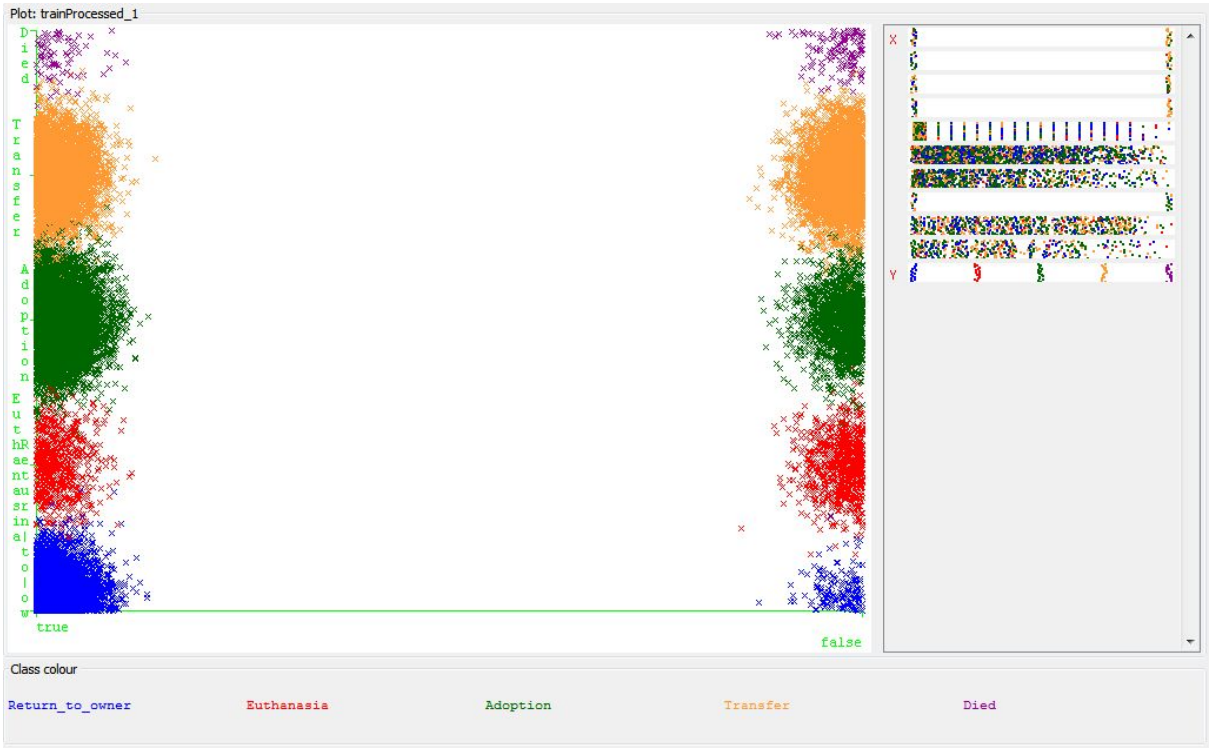

Figura 2.1: Gráfico *outcome* vs. *hasName*.

2.2.2. 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).

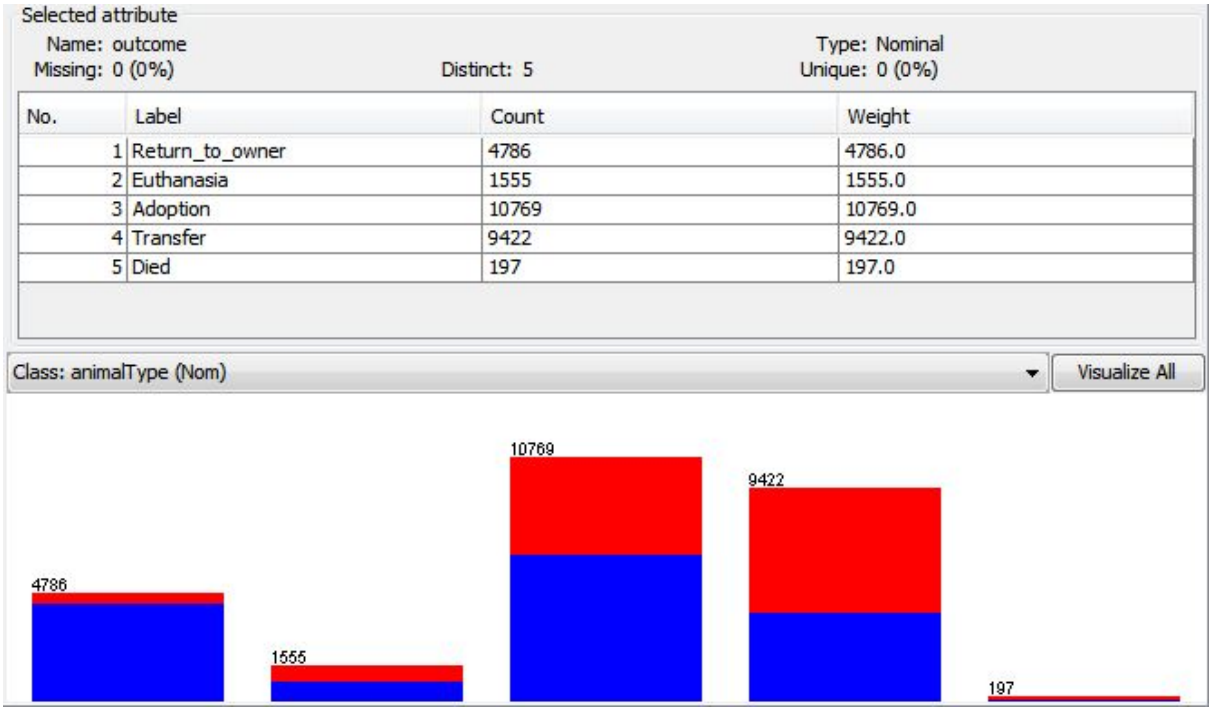

Figura 2.2: Gráfico *outcome* vs. *animalType*.

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*.

2.2.3. 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*.

2.2.4. 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.

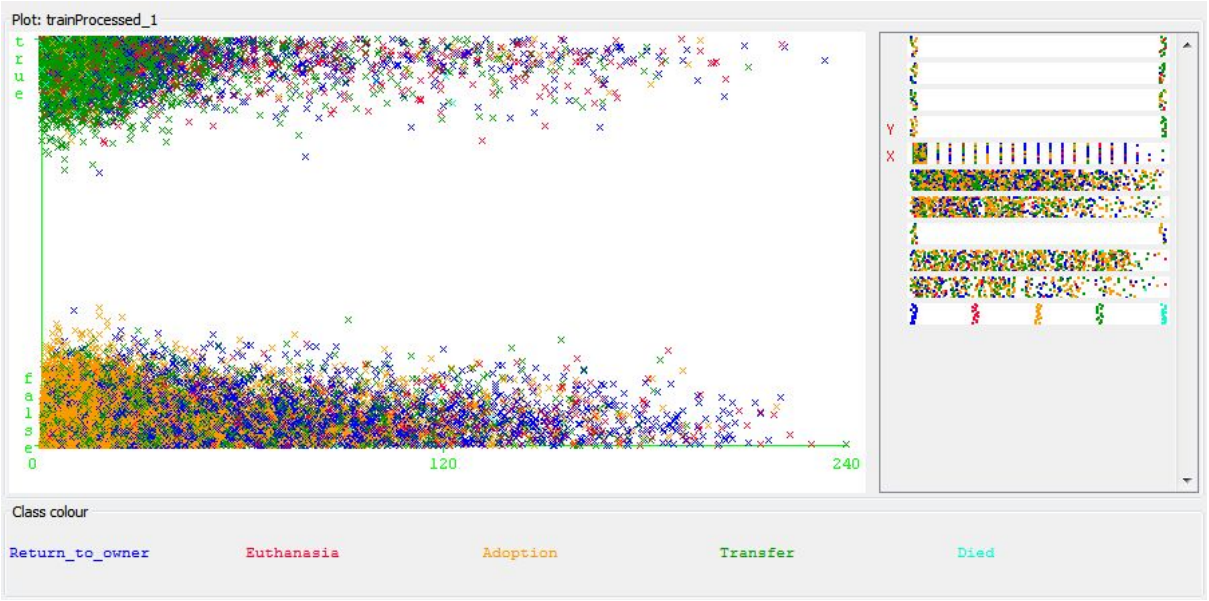

Figura 2.5: Gráfico *isIntact* vs. *monthsOld*.

2.2.5. 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.

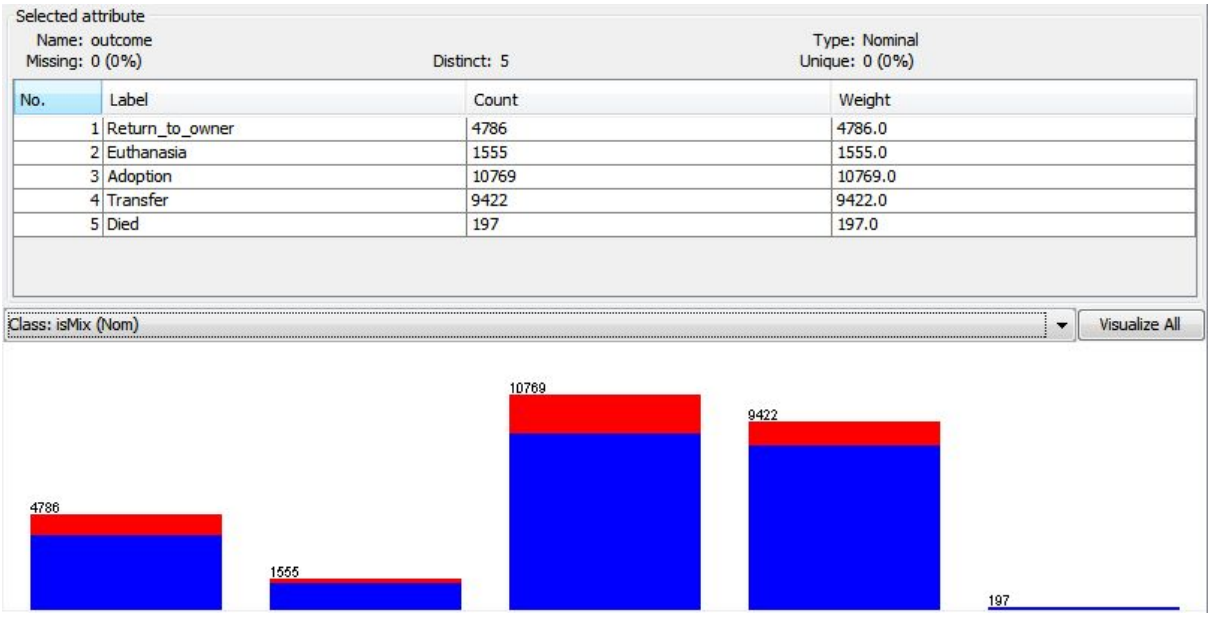

Figura 2.6: Gráfico *outcome* vs. *isMix*.

2.2.6. 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.

2.2.7. Atributos *life_span_low* e *life_span_high*

Presentes no conjunto de dados adicional, nós acreditamos que a expectativa de vida de um animal seja um fator decisivo em alguns casos para o seu desfecho, principalmente quando relacionada à idade do animal quando o desfecho ocorreu. Por exemplo, para animais em que a idade do mesmo no desfecho seja muito próxima a expectativa de vida máxima dele, muito provavelmente seu desfecho será *Died*.

2.2.8. Atributos *price_low* e *price_high*

Inicialmente nós acreditamos que o preço estimado do animal seria importante para a classificação do desfecho do mesmo. Pensamos que animais com um preço elevado muito provavelmente teriam donos e, portanto, retornariam aos mesmos. Porém, após visualizar e analisar os dados, não foi possível verificar este padrão, tendo em vista que existem diversos outros fatores que podem influenciar no desfecho do animal.

2.2.9. Atributos *adaptability*, *child_friendly*, *cat_dog_friendly*, *grooming*, *health_issues*, *intelligence* e *shedding_level*

Atributos como adaptabilidade, inteligência, cuidados necessários com saúde e higiene, por exemplo, podem ser relevantes para a classificação do animal. A título de exemplo, cães mais inteligentes e menos propensos a desenvolver problemas de saúde podem ser mais preferidos para adoção.

2.3. Atributos Desconsiderados

Dentre os atributos selecionados, foram desconsiderados os valores dos atributos ID, *Name* e *DateTime* do conjunto de dados original.

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*.

Do conjunto de dados adicional, foram desconsiderados os atributos *type*, *breed* e *other_names*, pois os mesmos eram atributos repetidos do conjunto original e apenas foram utilizados para fazer a união dos conjuntos de dados.

3. Técnica de Mineração 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.

4. Resultados da Mineração

Como mencionado na seção anterior, o algoritmo C4.5 seria utilizado pelo fato de superar as limitações mencionadas, presentes nos outros algoritmos. Ainda assim, testamos a mineração com outros algoritmos como o *RandomTree* e o *REPTree*. Para cada algoritmo, computamos a acurácia do mesmo sobre o conjunto de treino e, em seguida, computamos a classificação sobre o conjunto de teste fornecido.

A tabela abaixo resume os algoritmos e parâmetros utilizados, o *score* obtido na plataforma *Kaggle* e a acurácia no conjunto de treino. É importante destacar que as configurações com a melhor acurácia sobre o conjunto de teste resultaram em *scores* ruins na plataforma *Kaggle*. Uma das possíveis razões para esta questão é que nestas configurações o modelo ficou "viciado" no conjunto de treino e, portanto, não obteve um bom desempenho para conjuntos de dados inéditos.

Algoritmos e Resultados			
Algoritmo Utilizado	**Parâmetros**	**Score Kaggle**	**Acurácia no Conjunto de Treino**
C4.5 (J48)	-C 0.25 -M 3	0.93655	64.2037%
C4.5 (J48)	-C 0.25 -M 2	0.96796	64.3122%
REPTree	-M 2 -V 0.001 -N 3 -S 1 -L 1 -I 0.0	1.80906	66.7627%
C4.5 (J48)	-C 0.7 -M 2	1.86157	71.5328%
RandomTree	-K 0 -M 1.0 -V 0.001 -S 1	6.14302	83.9687%

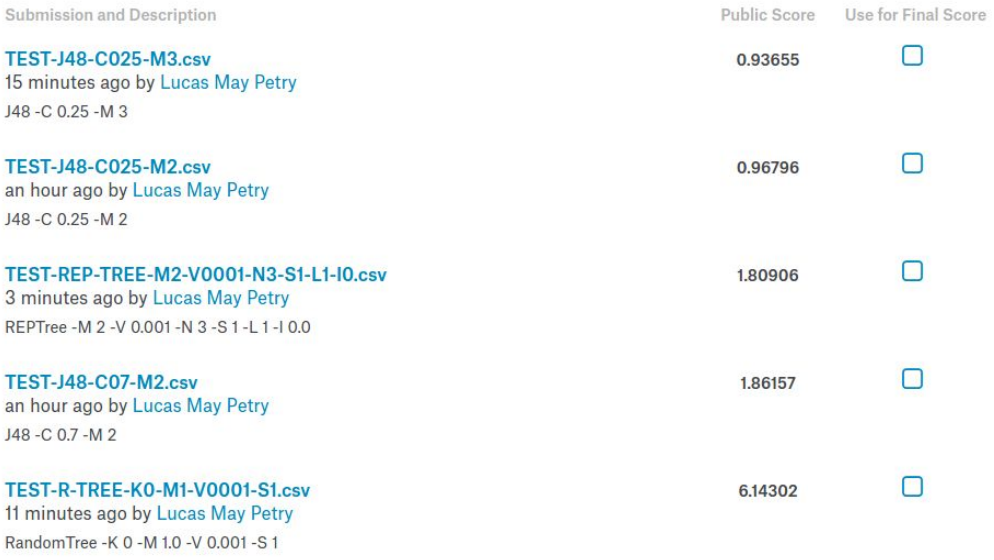

Figura 4.1: Resumo de submissões na plataforma *Kaggle*.

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.
CAT Breeds List: Search 60+ cat breeds with pictures. Search 60+ cat breeds with pictures. Disponível em: <http://www.catbreedslist.com/>. Acesso em: 12 maio 2017.
DOG Breeds List: Search 300+ dog breeds info and pictures. Search 300+ dog breeds info and pictures. Disponível em: <http://www.dogbreedslist.info/>. Acesso em: 12 maio 2017.