Atividade Avaliativa

Disciplina: Coleta e gestão de dados

Aluno: Rennê Ruan Alves Oliveira

Resumo

A partir de uma base de dados disponível, utilizando a ferramenta Orange, devemos realizar uma análise exploratória dos dados, submetê-los aos tratamentos necessários e aplicar modelos de machine learning, resgatando os resultados dos modelos aplicados comparando as diversas abordagens de como os dados podem ser tratados.
Foram utilizados dados de um abrigo de animais de Austin (EUA), o mesmo possui predominantemente dados categóricos, consequentemente foram utilizados modelos de aprendizados que suportassem entradas e saídas categóricas, como Naive Bayes, Regressão Logística e Random Forest. Os resultados foram satisfatórios frente aos atributos utilizados e notou-se melhora na acurácia dos modelos na medida que os dados foram tratados.

Coleta do conjunto de dados

Os dados utilizados foram coletados a partir da plataforma Kaggle. Eles se referem a dados de saída dos animais do abrigo de Austin (Austin Animal Center) , localizado nos Estados Unidos.

O dataset pode ser acessado pelo link: https://www.kaggle.com/datasets/aaronschlegel/austin-animal-center-shelter-outcomes-and

O mesmo se encontra desatualizado tendo sua última atualização datada há 7 anos atrás, possuindo 147 upvotes e 7741 downloads. Cabe ressaltar que o site do abrigo de Austin disponibiliza uma API para resgate dos dados mais recentes, além de disponibilizá-los em formato CSV. Porém para a proposta deste trabalho, um conjunto de dados pequenos e mais direcionados cumpriam o propósito, sendo os disponibilizados pelo Kaggle mais fáceis de serem carregados a ferramenta Orange.

Além do recorte inicial de data, utilizamos o dataset com as saídas apenas de gatos do abrigo (aac_shelter_cat_outcome_eng.csv), com isso poderiamos realizar uma análise e predição mais direcionada para uma amostra populacional específica.

O conjunto de dados conta com 37 colunas: age_upon_outcome, animal_id, animal_type, breed, color, date_of_birth, datetime, monthyear, name, outcome_subtype, outcome_type, sex_upon_outcome, count, sex, Spay/Neuter, Periods, Period Range, outcome_age_(days), outcome_age_(years), Cat/Kitten (outcome), sex_age_outcome, age_group, dob_year, dob_month, dob_monthyear, outcome_month, outcome_year, outcome_weekday, outcome_hour, breed1, breed2, cfa_breed, domestic_breed, coat_pattern, color1, color2, coat.

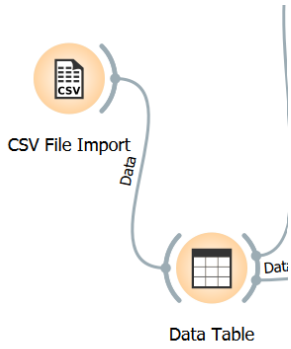
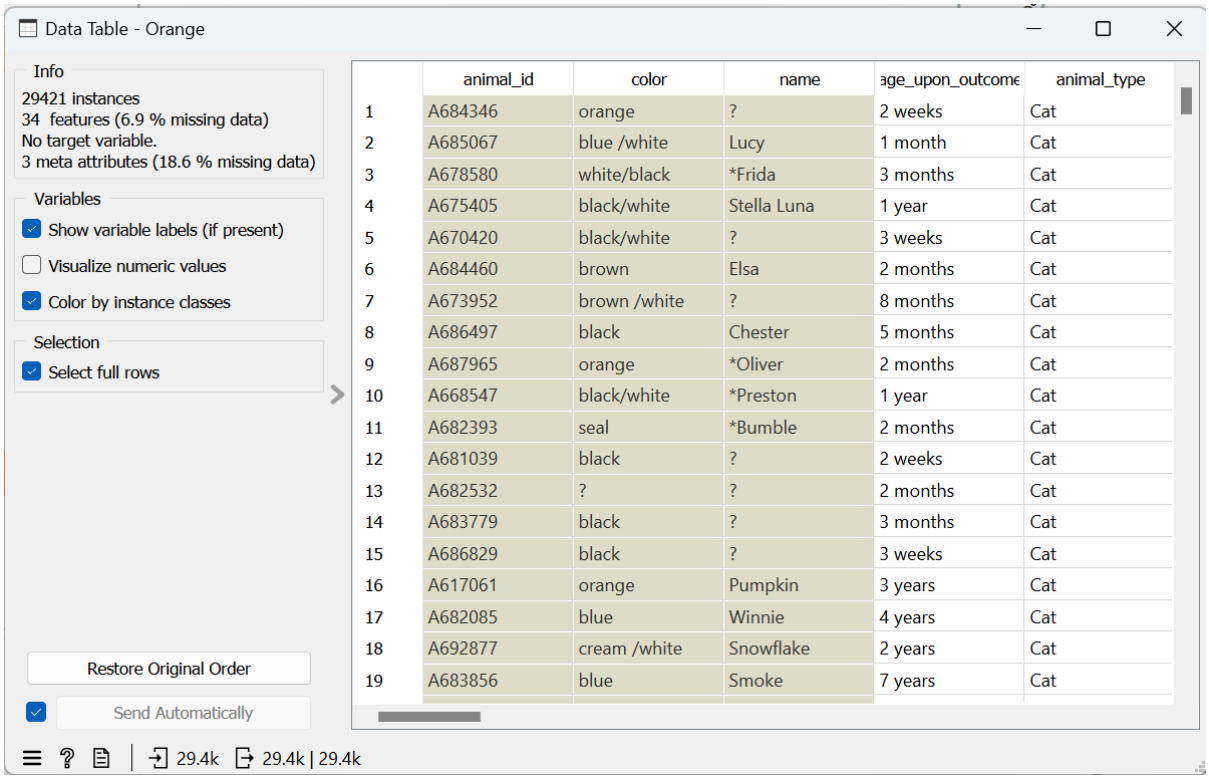


Figura 1: Widgets para carregamento dos dados.

Foram utilizados os widgets (Figura 1) CSV file import para importação dos dados e estes foram visualizados utilizando o widget Data Table (Figura 2).



	animal_id	color	name	age_upon_outcome	animal_type
1	A684346	orange	?	2 weeks	Cat
2	A685067	blue /white	Lucy	1 month	Cat
3	A678580	white/black	*Frida	3 months	Cat
4	A675405	black/white	Stella Luna	1 year	Cat
5	A670420	black/white	?	3 weeks	Cat
6	A684460	brown	Elsa	2 months	Cat
7	A673952	brown /white	?	8 months	Cat
8	A686497	black	Chester	5 months	Cat
9	A687965	orange	*Oliver	2 months	Cat
10	A668547	black/white	*Preston	1 year	Cat
11	A682393	seal	*Bumble	2 months	Cat
12	A681039	black	?	2 weeks	Cat
13	A682532	?	?	2 months	Cat
14	A683779	black	?	3 months	Cat
15	A686829	black	?	3 weeks	Cat
16	A617061	orange	Pumpkin	3 years	Cat
17	A682085	blue	Winnie	4 years	Cat
18	A692877	cream /white	Snowflake	2 years	Cat
19	A683856	blue	Smoke	7 years	Cat

Figura 2: Visualização inicial dos dados

Como informado pelo Widget Data table, o dataset possui 29421 registros, 37 colunas, sendo 3 delas de meta atributo e 34 de atributos, possuindo 6,9% de dados faltantes.

Mais sobre as colunas e o que cada uma representa será discutido na próxima seção de tratamento dos dados Seção 2.

Limpeza e preparação dos Dados

Numa análise exploratória inicial podemos perceber que a base de dados contém muitas variáveis. Ao analisar inicialmente, nota-se que muitas delas possuem valores repetidos, porém com um outro formato, enquanto outras quase não possuem dados preenchidos.

Vamos repassar o significado de cada coluna, para verificar quais colunas podem ser utilizadas de alvo e de atributo.

As colunas de meta-dados são:

- animal_id: Id único do animal no abrigo
- name: Nome do animal
- color: Cor do animal, concatenando as diferentes pelagens

As demais colunas são:

- age_upon_outcome: Idade do animal no momento de saída do abrigo, campo de texto sem unidade definida
- animal_type: Tipo do animal analisado, para esta base de dados terá um único valor, pois há apenas gatos
- breed: Raça do gato
- date_of_birth: Data de nascimento do animal
- datetime: Data da saída do animal
- monthyer: Possui os mesmos valore de datetime
- outcome_type: Tipo da saída do animal do abrigo
- outcome_subtype: Sub-tipo de saída do animal
- sex_upon_outcome: Estado reprodutivo do animal ao sair do abrigo
- count: Coluna sem descrição, com apenas valores 1
- sex: Sexo do animal
- Spay/Neuter: Estado de castração do animal
- Periods: Sem descrição formal, assume-se que é a quantidade de cios que o animal possuiu
- Period Range: Sem descrição formal, provavelmente é o período médio de cio do animal
- outcome_age(days): Tempo para saída do animal em dias
- outcome_age(years): Tempo para saída do animal em anos
- Cat/Kitten(outcome): Se animal classificado como gato adulto ou filhote
- age_group: Sem descrição formal, intervalos de idade em que o animal se encaixa
- dob_year: Sem descrição formal
- dob_month: Sem descrição formal
- dob_monthyear: Sem descrição formal
- outcome_month: Mês da saída
- outcome_year: Ano da saída
- outcome_weekday: Dia da semana da saída
- outcome_hour: Hora de saída
- breed1: Mesmos dados de raça da coluna breed
- breed2: Possui 100% dos valores como nulo
- cfa_breed: Raça possui certificação/pedigree
- domestic_breed: Se é uma raça doméstica
- coat_pattern: Tipo de padrão de pelagem
- color1: Tipo de cor primária, semelhante a coluna color, porém possui apenas a cor predominante
- color2: Tipo da cor secundária, em conjunto com a coluna color1 gera a coluna color, 65% dos valores nulos
- coat: Possui os mesmos valores que color1

Os dados apresentam uma grande necessidade de limpeza, em boa parte das colunas que podem ser utilizadas. Temos que o melhor candidato para uma coluna alvo visando realizar determinada predição ou estudo de análise seja a coluna `outcome_type` que significa o tipo de saída do animal do abrigo. Com ela

podemos predizer de acordo com as características do animal, qual será o destino dele, se será adotado, transferido ou no pior dos casos eutanasiado.

As medidas de limpeza tomadas foram:

- Descarte de colunas sem especificação definida.
- Descarte de todas as colunas de tempo e data referente a saída do animal.
- Utilização de apenas uma coluna para a idade do animal no momento de saída, foi utilizada a coluna `outcome_age(years)`.
- Descarte das colunas com muitos dados faltantes como `breed2` e `color2`
- Descarte das colunas referentes aos períodos de cio.
- Descarte da coluna de nascimento e nome do animal.

Ao final temos as colunas utilizadas para criação do modelo: `breed, sex_upon_outcome, sex, Spay Neuter, outcome_age(years), Cat/Kitten (outcome), sex_age_outcome, breed1, cfa_breed, domestic_breed, coat_pattern, color1, coat`.

Temos que as colunas selecionadas trazem informações, principalmente referente a pelagem, sexo e idade do animal.

Após a seleção inicial, verificamos as problemáticas de cada coluna selecionada. Verificamos havia diversos valores nulos na coluna `coat_pattern`. Porém a informação de nulidade também agrega informação ao animal, ou seja podemos classificar o padrão de pelagem como outro ou não definido. Para sanar isto, foi utilizado o widget Impute, com as configurações presentes na figura 3. Para os demais dados, foi utilizado o método padrão de popular dados faltantes com a Média/Moda, apenas para a coluna `coat_pattern` atribuímos a um novo valor com a configuração `As a distinct value` onde os valores nulos foram substituídos por `N/A`, possível visualizar o resultado na figura 4

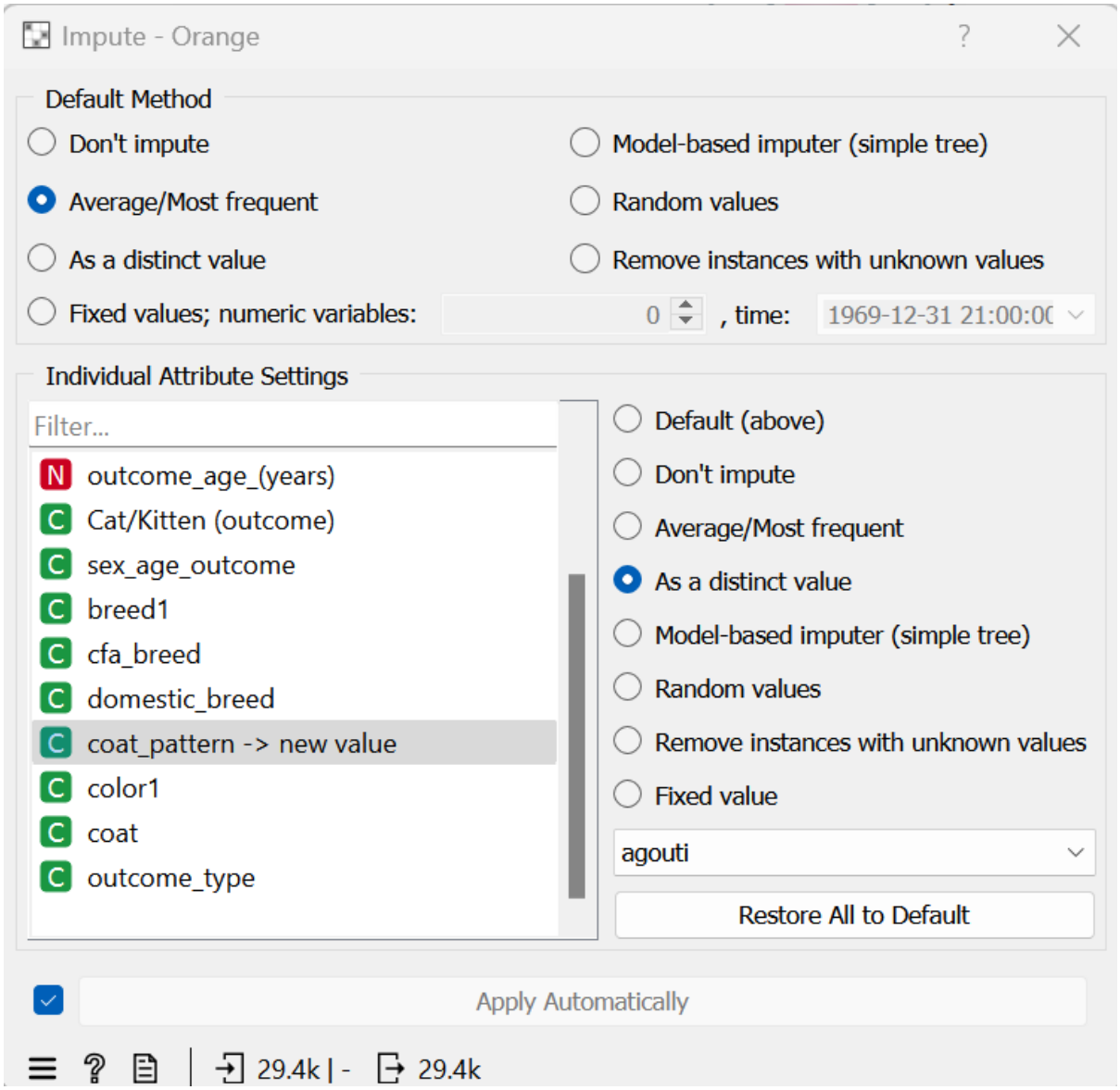


Figura 3: Configurações Impute

cfa_breed	domestic_breed	coat_pattern	color1	coat
False	True	tabby	orange	orange
False	True	tabby	blue	blue
False	True	N/A	white	white
False	True	N/A	black	black
False	True	N/A	black	black
False	True	tabby	brown	brown
False	True	tabby	brown	brown
False	True	tabby	black	black
False	True	tabby	orange	orange

Figura 4: Dados nulos substituídos

Em seguida devemos tratar o formato das variáveis, como temos uma saída categórica e das variáveis selecionadas, todas são categóricas, exceto a idade de saída `outcome_age(years)`, seria prudente apenas

converter a variável de idade numérica para valores categóricos. Foi realizada a discretização dos valores com o widget Preprocess. As configurações utilizadas podem ser visualizadas na figura 5

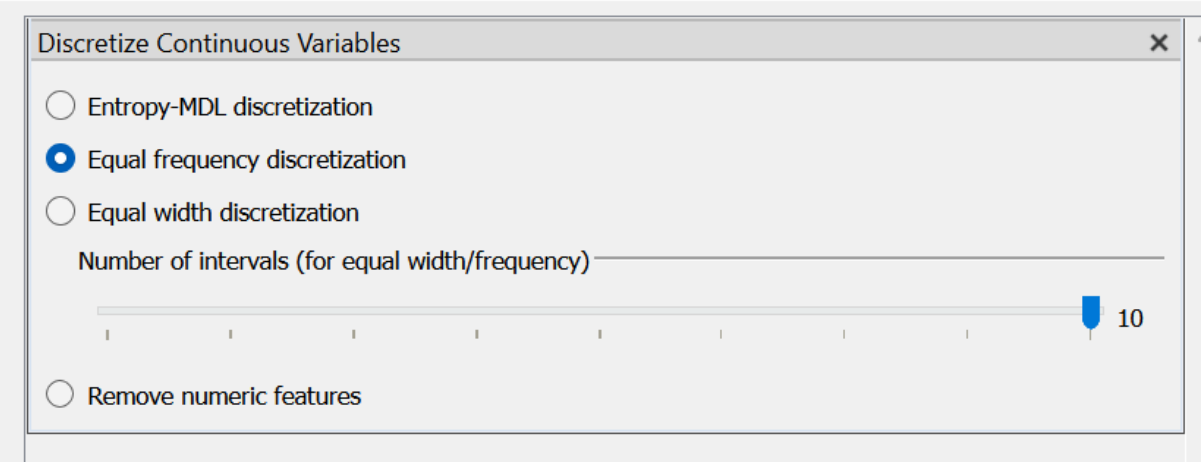


Figura 5: Configurações para discretização

Por último, foi identificado utilizando o widget Pivot que os valores de coloração apresentavam diferentes escritas para um mesmo tipo de valor. ou seja `black` também estava escrito como `black ` , em que este segundo há um espaço em branco ao final da string. Para realizar o tratamento dessas strings, foi utilizado o framework Edit Domain, em que os valores podem ser alterados manualmente, as configurações podem ser vistas na figura 6.

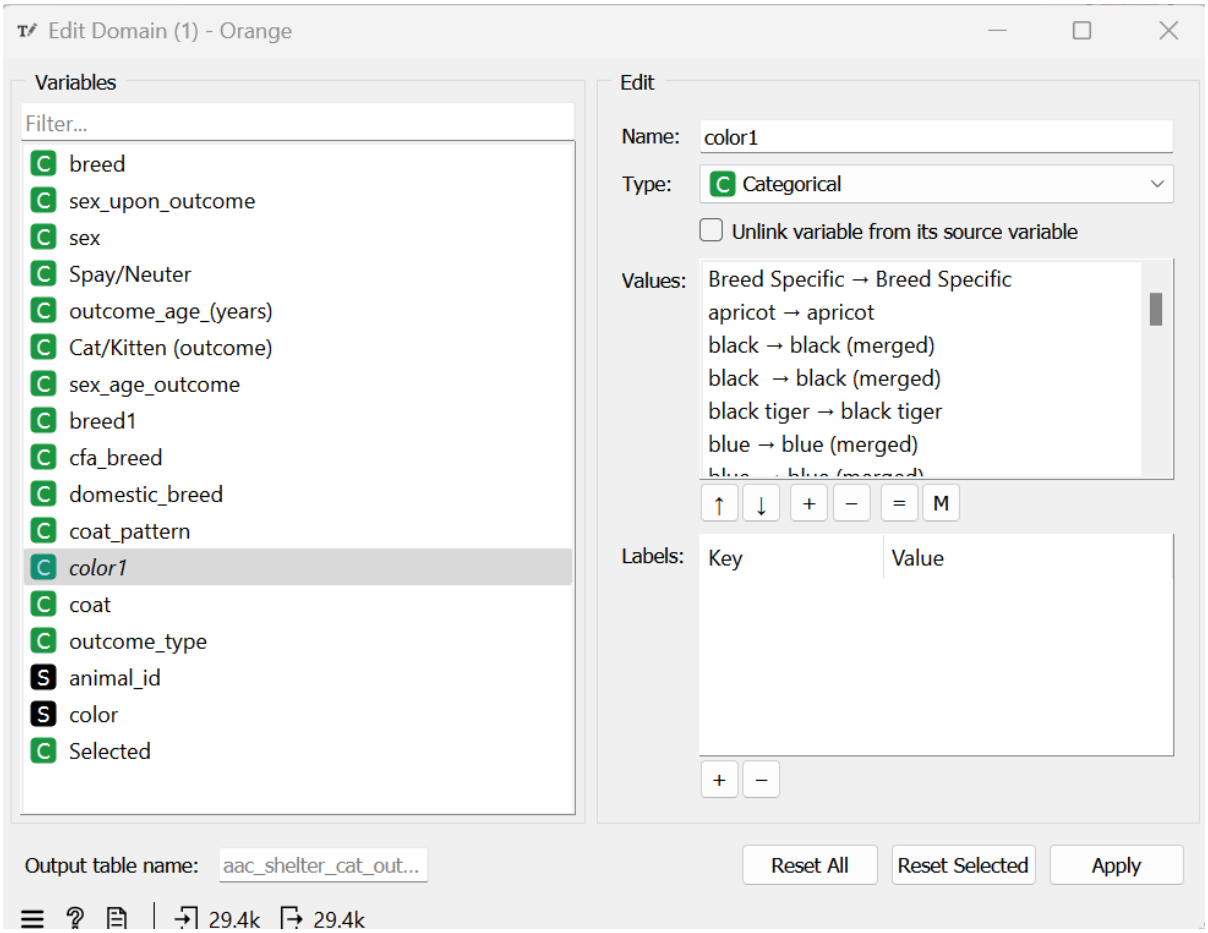


Figura 6: Edição de valores com diferentes escritas

Análise exploratória dos dados

Em seguida realizamos a análise exploratória dos dados. Temos predominantemente variáveis categóricas, ou seja, seria difícil aplicar gráficos como de dispersão, sendo o mais aplicável, gráficos de barra ou distribuição. Foi utilizado o framework Distributions, com ele é possível verificar graficamente como as variáveis se relacionam com a saída.

Percebemos que a cor de maior predominância entre os animais do abrigo, são os gatos de cor preta, em seguida os de coloração marrom. Por haver muitas classes nas variáveis, encontramos dificuldades na visualização do eixo x. (Figura 7)

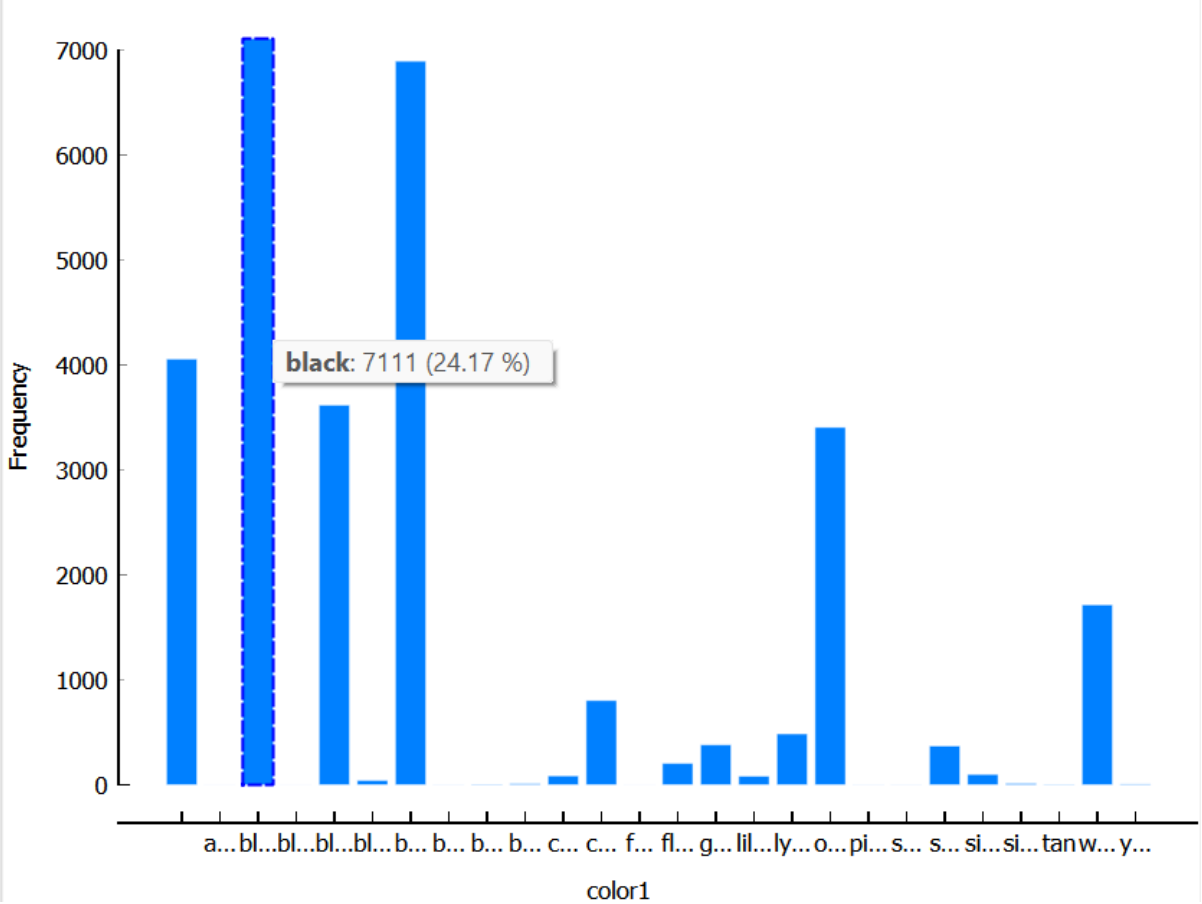


Figura 7: Gráfico de coloração

Referente as datas de saída dos animais em ano (figura 8) temos a maior porcentagem centrada no intervalo de 0,130 (46 dias) a 0,205 (75 dias). Gatos mais jovens tem maior frequência nos dados analisados, isso pode ser visualizado com melhor exatidão antes da etapa de discretização.

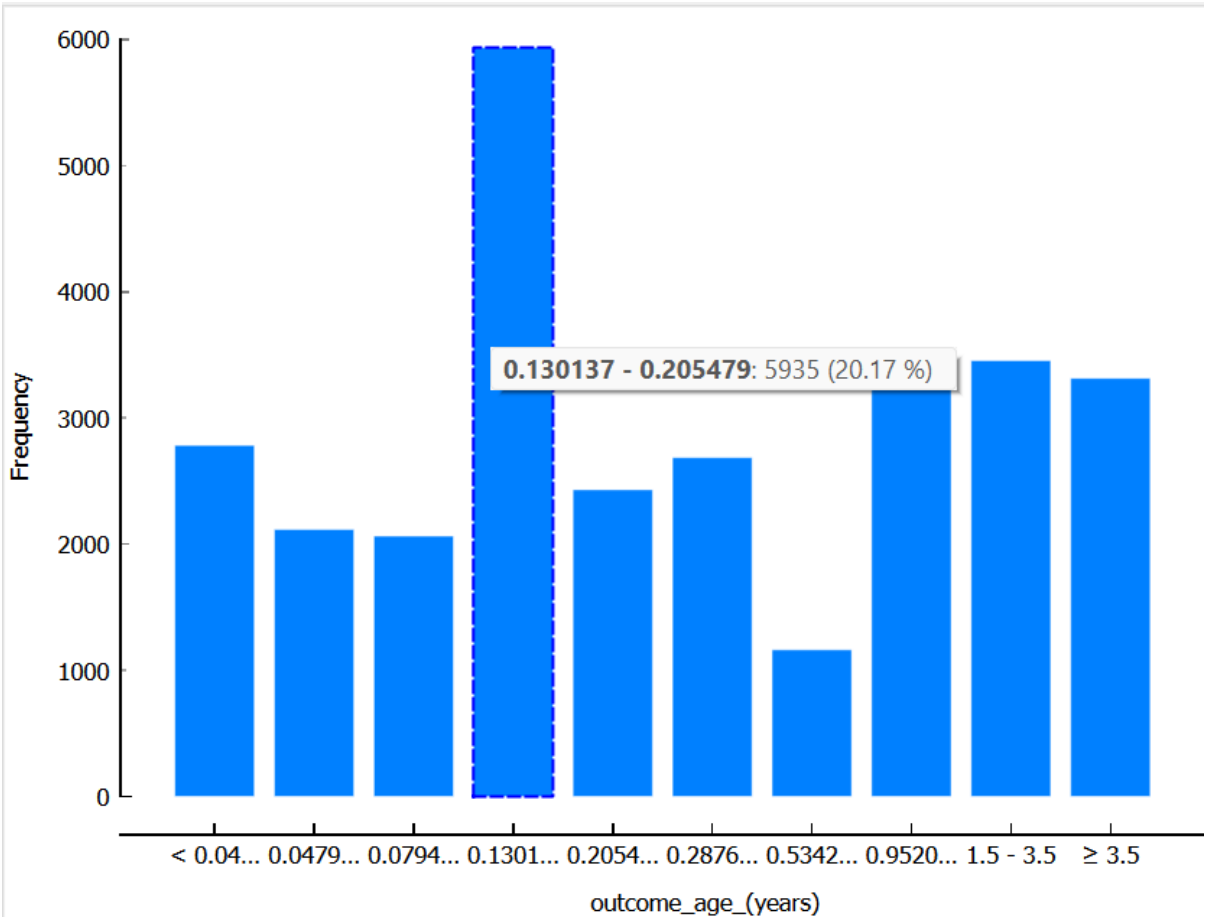


Figura 8: Histograma com o tempo de saída em anos

As saídas possuem 2 classes com maior predominância, caracterizando um desbalanceamento nas demais classificações, isso pode influenciar no resultado das predições e modelos aplicados. As classes de maior predominância são, Adoption e Transfer, o gráfico referente pode ser visualizado na figura 9.

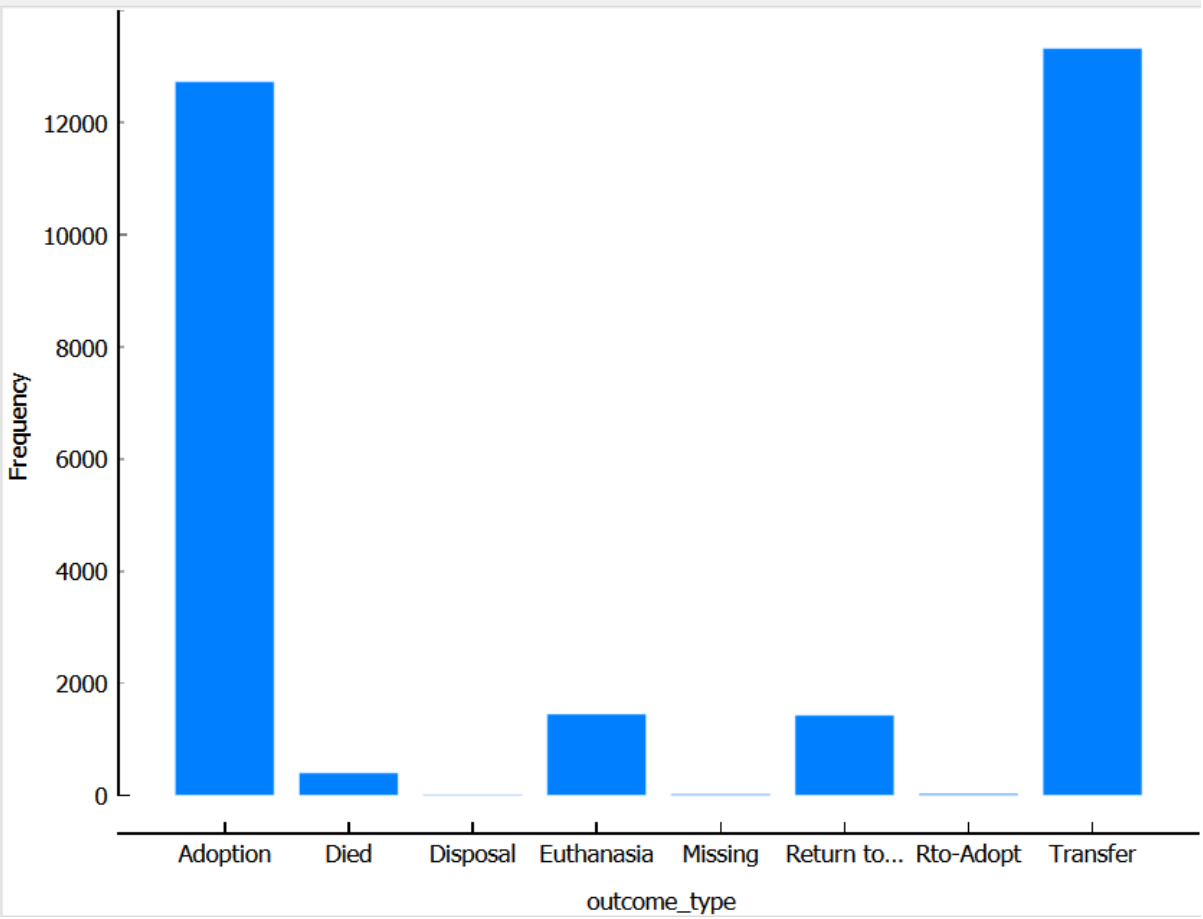


Figura 9: Gráfico com tipos de saída

Mineração dos Dados

Foram aplicados alguns algoritmos possível para dados categóricos com saídas categóricas, sendo eles: Naive Bayes, Logistic Regression, Random Forest e Gradiente Boosting. A aplicação de Árvores de Decisão não se tornou possível pois o Orange não permite que sejam aplicadas variáveis com mais de 16 classes diferentes.

O algoritmo com melhor desempenho foi o de regressão logística (Logistic Regression) com AUC de 0.881, mais informações acerca das acurácias obtidas serão discutidas na próxima seção, avaliação dos resultado.

Aplicado ao modelo previamente citado, utilizamos o widget de Predictions para verificar se foram realizadas predições condizentes com o esperado e como as saídas estavam se comportando.

	Logistic Regression (1) (1)	error	outcome_type
1	0.00 : 0.02 : 0.00 : 0.01 : 0.00 : 0.00 : 0.00 : 0.96 → Transfer	0.040	Transfer
2	0.24 : 0.06 : 0.00 : 0.09 : 0.00 : 0.01 : 0.00 : 0.61 → Transfer	0.763	Adoption
3	0.89 : 0.00 : 0.00 : 0.00 : 0.00 : 0.01 : 0.00 : 0.09 → Adoption	0.107	Adoption
4	0.44 : 0.01 : 0.00 : 0.02 : 0.00 : 0.06 : 0.00 : 0.47 → Transfer	0.939	Return to Owner
5	0.00 : 0.02 : 0.00 : 0.04 : 0.00 : 0.00 : 0.00 : 0.93 → Transfer	0.073	Transfer
6	0.95 : 0.00 : 0.00 : 0.00 : 0.00 : 0.00 : 0.00 : 0.04 → Adoption	0.049	Adoption
7	0.00 : 0.02 : 0.01 : 0.07 : 0.00 : 0.03 : 0.00 : 0.87 → Transfer	0.132	Transfer
8	0.79 : 0.00 : 0.00 : 0.00 : 0.00 : 0.03 : 0.00 : 0.17 → Adoption	0.212	Adoption
9	0.93 : 0.00 : 0.00 : 0.00 : 0.00 : 0.01 : 0.00 : 0.06 → Adoption	0.069	Adoption
10	0.38 : 0.01 : 0.00 : 0.03 : 0.00 : 0.08 : 0.00 : 0.50 → Transfer	0.495	Transfer
11	0.38 : 0.03 : 0.00 : 0.07 : 0.01 : 0.01 : 0.00 : 0.50 → Transfer	0.499	Transfer
12	0.00 : 0.02 : 0.00 : 0.03 : 0.00 : 0.00 : 0.00 : 0.94 → Transfer	0.059	Transfer
13	0.96 : 0.00 : 0.00 : 0.00 : 0.00 : 0.00 : 0.00 : 0.04 → Adoption	0.041	Adoption
14	0.88 : 0.00 : 0.00 : 0.00 : 0.00 : 0.01 : 0.00 : 0.10 → Adoption	0.115	Adoption
15	0.00 : 0.02 : 0.00 : 0.04 : 0.00 : 0.00 : 0.00 : 0.93 → Transfer	0.073	Transfer
16	0.33 : 0.00 : 0.00 : 0.04 : 0.00 : 0.18 : 0.01 : 0.44 → Transfer	0.995	Died
17	0.48 : 0.00 : 0.00 : 0.01 : 0.00 : 0.34 : 0.00 : 0.17 → Adoption	0.524	Adoption
18	0.41 : 0.01 : 0.00 : 0.03 : 0.00 : 0.13 : 0.01 : 0.42 → Transfer	0.582	Transfer
19	0.38 : 0.00 : 0.00 : 0.02 : 0.00 : 0.44 : 0.00 : 0.16 → Return to Owner	0.623	Adoption
20	0.22 : 0.04 : 0.00 : 0.12 : 0.00 : 0.00 : 0.00 : 0.61 → Transfer	0.392	Transfer
21	0.02 : 0.01 : 0.00 : 0.08 : 0.00 : 0.02 : 0.00 : 0.87 → Transfer	0.130	Transfer
22	0.02 : 0.01 : 0.00 : 0.13 : 0.00 : 0.04 : 0.00 : 0.80 → Transfer	0.197	Transfer

Figura 10: Predições realizadas pelo algorimto de regressão logística

Na figura 10 temos exemplos de predições realizadas, realizando uma análise superficional neste pedaço de amostrado coletado, percebe-se que o algoritmo te dificuldade de predizer casos além dos predominantes, como Adoption ou Transfer, tal problema levante a necessidade de desbalancear os dados, ou seja, prover dados com quantidades semelhantes de animais com outros tipos de saída.

As associações obtidas pelo widget Frequent Itemsets e Association não foram de fácil interpretação, e não proveram informações acerca dos tipos de saída, porém serviram como uma visão geral das porcentagens dos tipos de dados. Podemos visualizar o resultado do widget na figura 11

Itemsets	Support	%
domestic_breed=True	27720	94.22
coat_pattern=tabby	13361	45.41
coat_pattern=N/A	9953	33.83
cfa_breed=False	27678	94.08
domestic_breed=True	27678	94.08
coat_pattern=tabby	13349	45.37
coat_pattern=N/A	9935	33.77
coat_pattern=tabby	13349	45.37
coat_pattern=N/A	9935	33.77
breed1=domestic shorthair	23728	80.65
cfa_breed=False	23722	80.63
domestic_breed=True	23722	80.63
coat_pattern=tabby	11750	39.94
coat_pattern=tabby	11750	39.94
domestic_breed=True	23728	80.65
coat_pattern=tabby	11753	39.95
coat_pattern=tabby	11753	39.95
breed=domestic shorthair	23720	80.62
breed1=domestic shorthair	23720	80.62
cfa_breed=False	23720	80.62
domestic_breed=True	23720	80.62
coat_pattern=tabby	11750	39.94
coat_pattern=tabby	11750	39.94

Figura 11: Regras de associação

Temos que as informações e valores predominantes são de atributos com pouca variedade de classificações, geralmente Sim/Não. Não ficou claro se as regras de associação utilizam da coluna target para sua saída, uma vez que não houve nenhuma associação que remetesse a `outcome_type`.

Avaliação dos Resultados

Inicialmente foi utilizado o algoritmo Naive Bayes para verificar como o tratamento dos dados influenciavam no refinamento da acurácia do modelo e no AUC obtido. Submetemos o Naive Bayes a diferentes formatos dos dados, (1) sem nenhum tratamento, (2) colunas selecionadas e valores nulos sanados, (3) adicional 2 com junção de classes de texto semelhantes mas escritas diferentes, (4) adicional ao 2 com discretização dos valores numéricos, (5) todos os processamentos anteriores realizados. 12

Tipo de Pré-processamento	AUC	Precision	Recall
(1)	0.854	0.742	0.640
(2)	0.854	0.743	0.638

Tipo de Pré-processamento	AUC	Precision	Recall
(3)	0.854	0.743	0.640
(4)	0.863	0.753	0.640
(5)	0.864	0.753	0.641

Como resultado, visualizamos que realizar os pré-processamentos trouxe uma melhoria no resultado do modelo, não foram mudanças expressivas nos valores, porém corroboramos a necessidade do tratamento de dados inicial. Deste ponto agora podemos comparar diferentes modelos para verificar qual melhor se aplica aos dados utilizados. Utilizamos: Naive Bayes, Logistic Regression, Random Forest e Gradiente Boosting, podemos visualizar o resultado dos diferentes modelos na figura 12

Model	Train	Test	AUC	CA	F1	Prec	Recall	MCC
Naive Bayes (1) (1) (1)	0.071	0.008	0.864	0.641	0.678	0.753	0.641	0.473
Logistic Regression (1) (1)	100.819	1.503	0.881	0.733	0.694	0.687	0.733	0.532
Random Forest	7.682	1.329	0.863	0.726	0.693	0.681	0.726	0.523
Gradient Boosting	532.721	3.239	0.881	0.734	0.694	0.686	0.734	0.534

Figura 12: Resultados de modelos aos dados tratados e processados.

Temos que o modelo de Logistic Regression e o Gradient Boosting tiveram resultados semelhantes, sendo o de Logistic Regression 5 vezes mais rápido.

Conclusão

Temos que o algoritmo de regressão logística obteve o melhor resultado na predição de saídas dos animais do abrigo com AUC de 0.881. Melhores insights e refinamento do modelo podem ser obtidos ao sanar o desbalanceamento das diferentes classes de saída. A maior porcentagem de gatos no abrigo são os que possuem características mais propensas para gatos de rua, como pelagem malhada, coloração preta ou marrom e não possuem raça definida.

O fluxo inteiro pode ser visualizado na figura 13

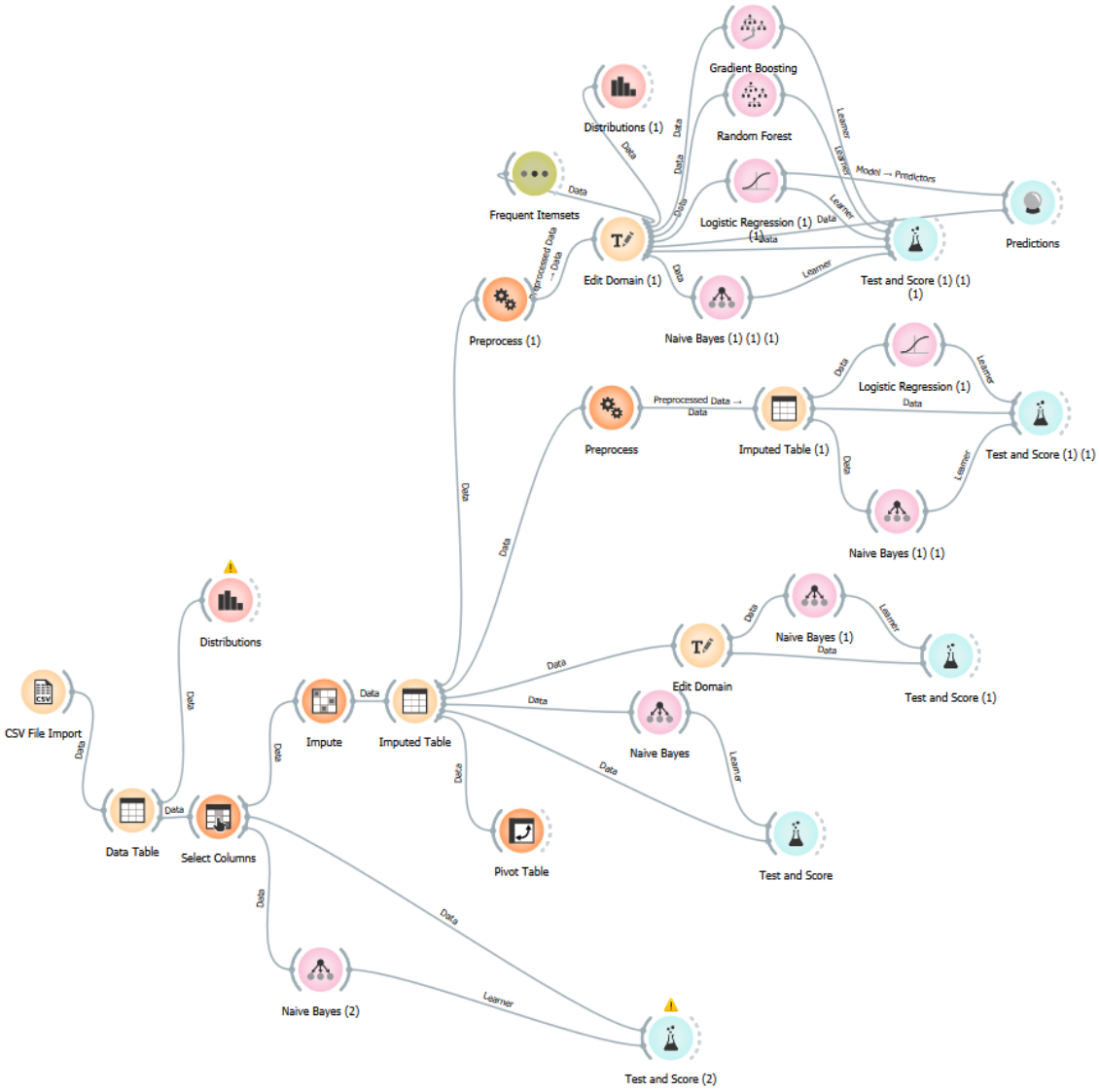


Figura 13: Fluxo de widgets do Orange.