



TRABALHO DATA MINING

FREDY OSORIO

Ing.fredyosorio@gmail.com

BASE DE DADOS

PREVIA DOS DADOS E
PREPARAÇÃO PARA O
ANALISE

ANALISE EXPLORATÓRIA

REVISÃO DOS ATRIBUTOS

01

02

CONTEÚDO

03

04

PRÉ-PROCESSAMENTO

PREPARAÇÃO DOS DADOS

MODELOS DE CLASSIFICAÇÃO

APLICAÇÃO DE DIFERENTES
MODELOS E AVALIAÇÃO

01

BASE DE DADOS



PROBLEMA

Classificação de chances de sobrevivência de Cavalos

O trabalho aqui descrito, usa a base de dados de Horse Colic. A BD oferece 27 diferentes atributos relacionados com a saúde de cavalos, e três classes, que indicam se o animal conseguiu sobreviver, morreu ou teve que ser aplicada a eutanásia como efeito de seu estado de saúde.



O alvo é de reforçar os conhecimentos sobre Data Mining e Machine Learning, e criar modelos que tentem prognosticar se um cavalo pode sobreviver de acordo ao seu estado de saúde atual

BASE DE DADOS

- A base de dados foi entregue em dois arquivos, horse.csv e horsetest.csv.
- Juntaram-se num Dataframe só, pois isso facilitava o análise e pré-processamento.
- A base dados inicialmente contém 388 casos, 27 atributos e indicadores médicos e 3 classes.
- Entre as 27 dimensões se incluem dados categóricos (16) , numéricos(7) e inclusive codificados(3) .



BASE DE DADOS

- Na revisão dos valores nulos foram achados 3 Atributos que estavam com falta de más da metade dos valores. As colunas foram eliminadas da base de dados.
- Na folia de explicação de atributos foi indicado que as colunas 'hospital_number', 'respiratory_rate' e 'cp_data', não forneciam informação relevante ou duvidosa, por tanto foram eliminadas da base de dados também
- Assim mesmo, eliminados os casos com 50% ou mais de atributos faltantes.
- A classe 'Euthanazed' foi eliminada e os dados foram trocados por 'died'.

ATRIBUTO	Valores Nulos
nasogastric_reflux_ph	321
abdomo_appearance	209
abdomo_protein	258





02

ANALISE EXPLORATÓRIA

ANALISE EXPLORATÓRIA

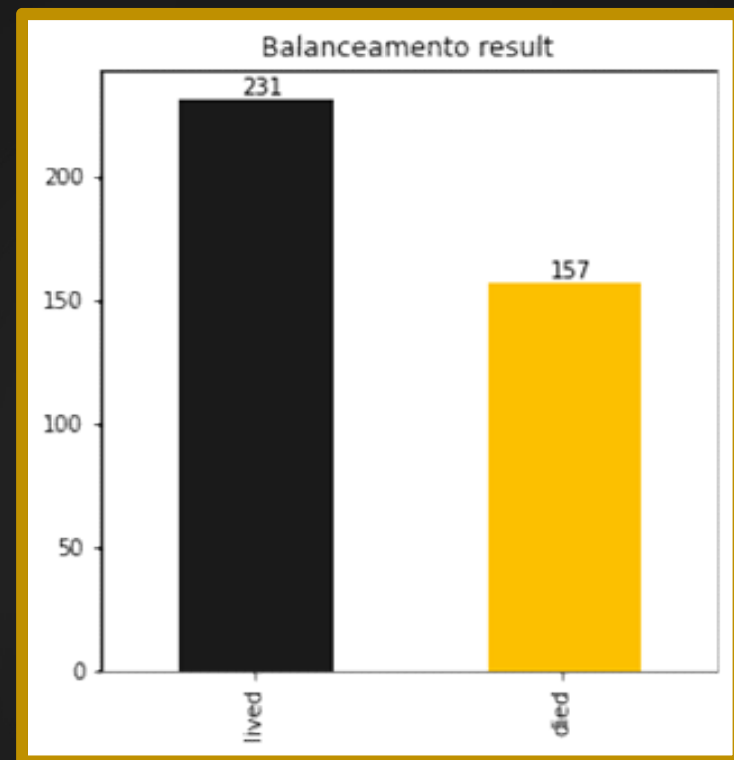
Balanceamento

Após tirar as filas e colunas com muitos dados nulos temos uma Base de dados com o seguinte balance

lived: 222 (59.7%)
died: 150 (40.3%)
TOTAL 372 Casos

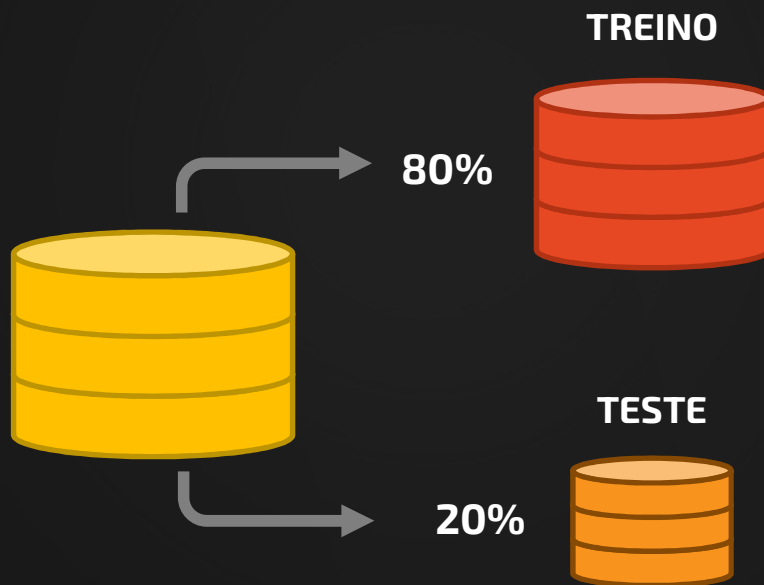
22 Atributos, 1 coluna de 2 classes

Não há necessidade de resampling pois a base esta suficientemente balanceada.



ANALISE EXPLORATÓRIA

Separação de Bases



A base de Dados é então dividida para fazer Treinamento com o 80% dos dados, o 20% restante será usado como Prova do funcionamento dos modelos de **Machine Learning**



ANALISE EXPLORATÓRIA

Valores Nulos ou Faltantes

MISSING VALUES POR COLUNA

A abordagem dos dados nulos foi de usar a **moda** nos atributos categóricos e a **média** nos atributos numéricos.

Tirando as referências da **Base de Treino**. Isto com o fim de evitar adicionar uma tendência que interfira com a avaliação do modelo mais na frente

surgery	0
age	0
rectal_temp	66
pulse	19
respiratory_rate	63
temp_of_extremities	53
peripheral_pulse	72
mucous_membrane	41
capillary_refill_time	21
pain	51
peristalsis	38
abdominal_distention	52
nasogastric_tube	111
nasogastric_reflux	118
rectal_exam_feces	120
abdomen	137
packed_cell_volume	24
total_protein	31
surgical_lesion	0
lesion_1	0
lesion_2	0
lesion_3	0
result	0

ANALISE EXPLORATÓRIA

Lesões

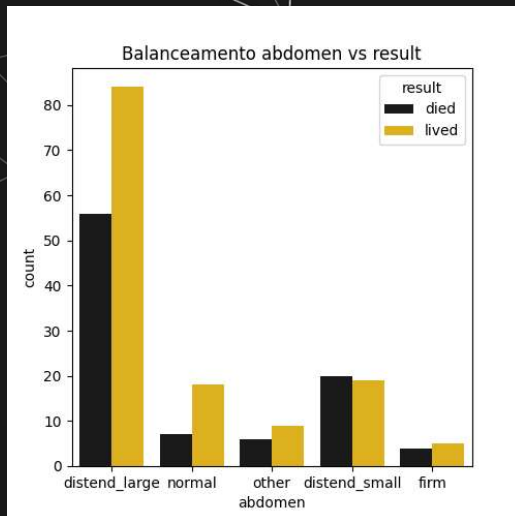
6	1	1	2
---	---	---	---

No caso das colunas referentes às lesões (1,2 e 3), a informação estava 'codificada', cada número representa um local, tipo, subtipo e código específico, respectivamente.

Foi considerada informação importante e se fez uma função que separasse a informação corretamente em **quatro colunas novas para cada lesão.**

As colunas originais foram eliminadas

ANALISE EXPLORATÓRIA



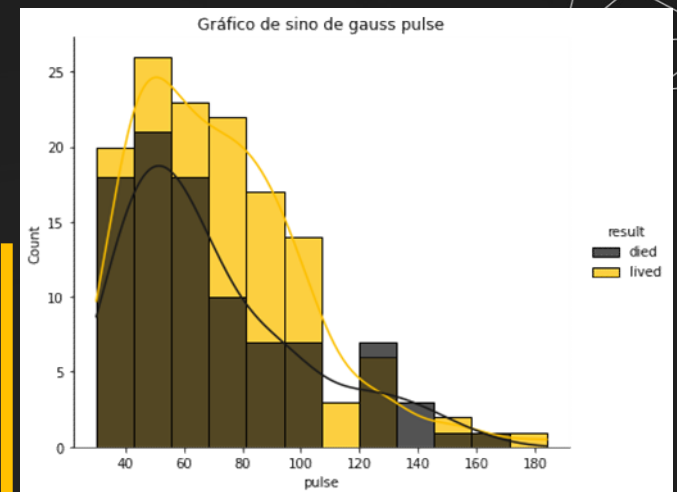
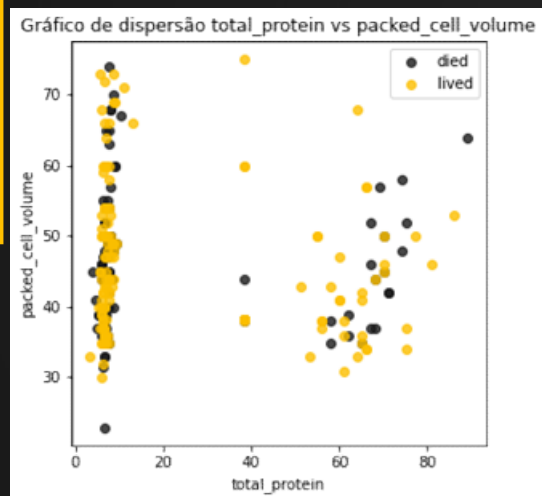
BARRAS

Gráfico de barras com atributos categóricos.

No atributo abdome foi claro ver que ter o abdômen distendido tem uma proporção maior de casos de morte.

Os diagramas de dispersão, comparando dois atributos numéricos não deixaram ver algum tipo de separação óbvia entre classes

DISPERSÃO



HISTOGRAMA

Se fizeram histogramas com os atributos numéricos

O mais relevante na inspeção visual foi o pulso: Os dados apontam que a proporção de cavalos com pulso alto e morreram é maior do que ao contrário.



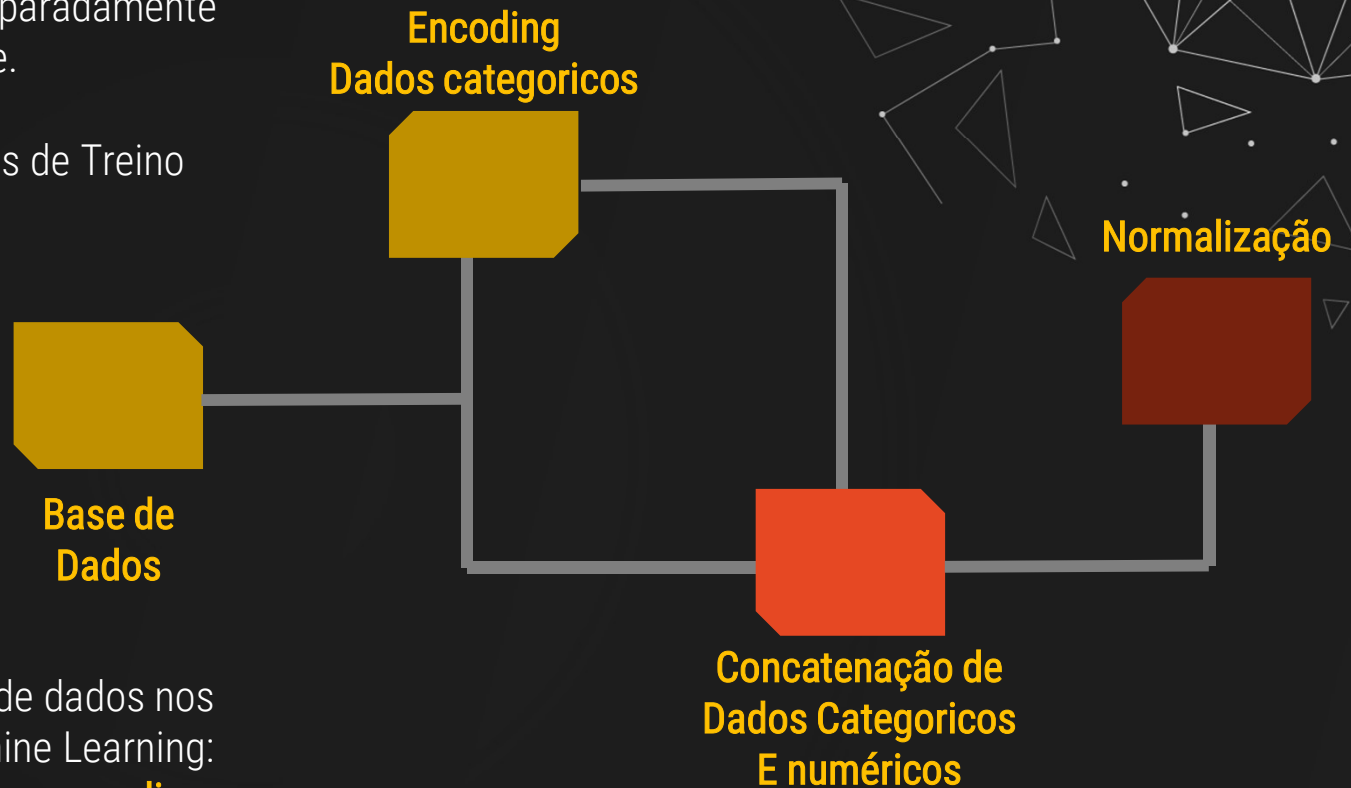
03

PRÉ-PROCESSAMENTO

PRÉ-PROCESSAMENTO

O pré-processamento é feito separadamente entre as bases de Treino e teste.

Usando sempre a Base de dados de Treino como referência para o Teste



Usaram-se duas bases de dados nos modelos de Machine Learning:

Uma normalizada e outra sem normalizar



04

MACHINE LEARNING

MACHINE LEARNING

Cada modelo foi Treinado e validado 4 vezes:

1. Base de dados
2. Base de dados e modelo otimizado com GridSearch
3. Base de dados normalizada (Scaled - Sc)
4. Base de dados normalizada (Scaled - Sc) e modelo otimizado com GridSearch

Foram usados 3 métricas de Avaliação:

1. Accuracy
2. Kappa
3. F1

A Otimização foi feita tendo em conta a métrica de F1 (binária) para TODOS os modelos



MACHINE LEARNING

Arvore de decisão (AD)

Random Forest(RF)

SVM

KNN

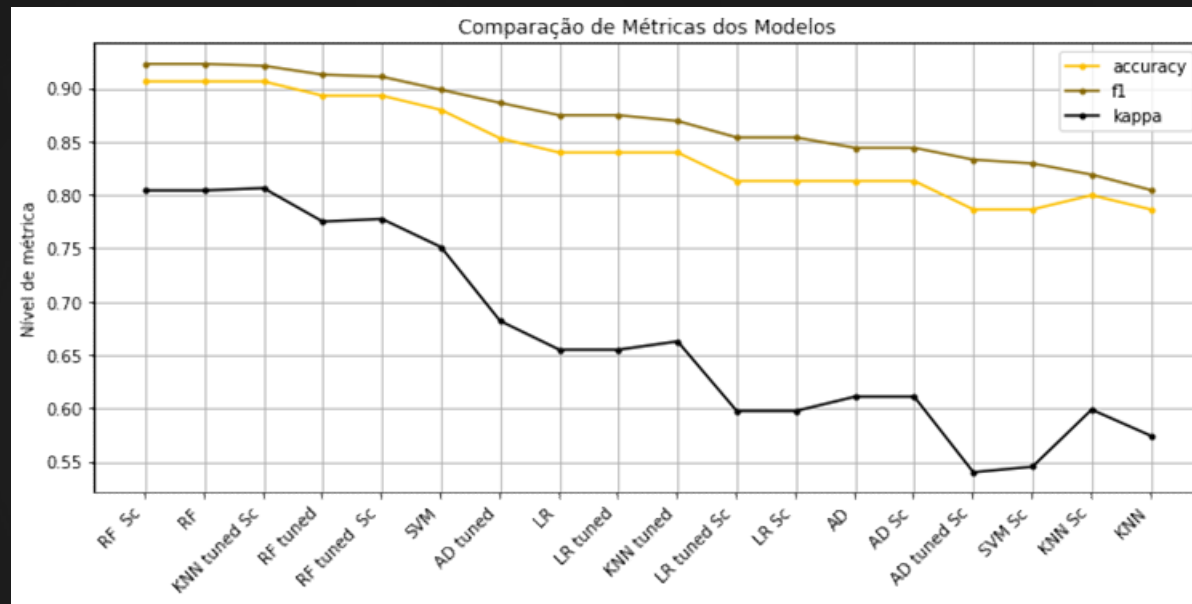
Regressão Logística (LR)

Modelos	Accuracy	Kappa	F1
AD	0.813	0.611	0.844
AD Sc	0.813	0.611	0.844
AD tuned	0.853	0.682	0.887
AD tuned Sc	0.787	0.54	0.833
KNN	0.787	0.574	0.805
KNN Sc	0.8	0.599	0.819
KNN tuned	0.84	0.663	0.87
KNN tuned Sc	0.907	0.807	0.921
LR	0.84	0.655	0.875
LR Sc	0.813	0.598	0.854
LR tuned	0.84	0.655	0.875
LR tuned Sc	0.813	0.598	0.854
RF	0.907	0.804	0.923
RF Sc	0.907	0.804	0.923
RF tuned	0.893	0.775	0.913
RF tuned Sc	0.893	0.778	0.911
SVM	0.88	0.751	0.899
SVM Sc	0.787	0.545	0.83

Todos os modelos apresentaram
F1 maior de **0.8**

Random Forest demonstrou ser o
modelo mais eficiente em todas
as avaliações

MACHINE LEARNING




Neste gráfico podemos enxergar de uma melhor maneira a avaliação dos modelos e suas métricas



CONCLUSÕES

A limpeza dos dados é o passo que mais tomou tempo, porém, é **indispensável** para obter melhores resultados.

Fazer o pré-processamento com as bases de dados de treino e teste separadas aumenta a fiabilidade da avaliação.






CONCLUSÕES

O **KNN Otimizado com Base de Dados normalizada** obteve a terceira melhor avaliação, além disso, sem diferenças significativas com os primeiros dois. Na minha opinião pessoal, este seria o modelo a escolher pois gera menor custo computacional do que RF.

Todos os modelos de **Random Forest** conseguiram melhores avaliações, estranhamente teve piores resultados quando foi otimizado. Possivelmente o parâmetro mais relevante é o número mínimo de folhas, o qual não foi incluso na otimização



CONCLUSÕES

Conseguiu-se que os modelos tivessem uma efectividade superior ao 80%. Conseguiu-se o objetivo de criar um modelo que consiga prognosticar se um cavalo pode sobreviver de acordo ao seu estado de saúde atual com sucesso.

Testaram-se diferentes modelos de **Machine Learning** vistos em aula, e o trabalho outorgou um grande desafio que enriqueceu os conhecimentos da aula e de programação.



OBRIGADO