Trabalho Final – Data Mining

Aluno: Filipe Faria Rodrigues

Matrícula: 211.101.222

Professora: Manoela Kohler

1. Introdução

Foi sugerido o problema de classificação da base "`horse.csv`", que possui diversos atributos que descrevem o estado de saúde de cavalos.

A base de dados fornecida possui três possíveis rótulos de saída, indicando se o cavalo sobreviveu, morreu, ou se foi submetido a eutanásia.

O estudo foi realizado em Python, e o código se encontra no arquivo `Analise-Horse.ipynb`.

2. Objetivo

O trabalho tem por objetivo prever se um cavalo pode, ou não, sobreviver, baseado nas condições médicas informadas.

3. Apresentação dos dados

Para a realização do trabalho, foram fornecidas duas bases de dados: `horse.csv` e `horseTest.csv`, que correspondem, respectivamente, às bases de dados de treino e teste.

Os dados apresentados possuem originalmente 27 atributos, com um total de 388 entradas, sendo elas 299 da base de treino, e 89 da base de teste. Dentre os atributos apresentados, 11 são numéricos e 16 categóricos.

O detalhamento de cada um dos atributos se encontra no arquivo `datadict.pdf`.

É possível observar que alguns atributos possuem valores faltantes, que será objeto de tratamentos posterior.

4. Análise dos dados

Verificando os rótulos de saída das bases apresentadas, vemos que a saída *lived* está em quantidade bem acima das demais, porém considerando que as saídas *died* e *euthanized* indicam que o cavalo não sobreviveu, temos um bom balanceamento entre as saídas das bases apresentadas, conforme podemos observar nas tabelas abaixo.

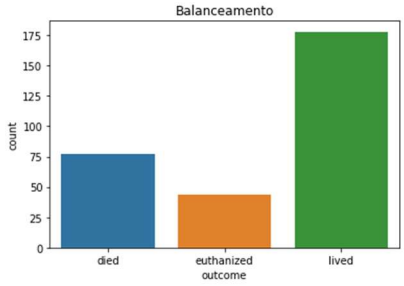

Figura 1. Balanceamento da base Treino

Saída	Quantidade	Percentual
lived	178	59,53%
died	77	25,75%
euthanizaed	44	14,72%

Figura 2. Distribuição de saídas da base Treino

Saída	Quantidade	Percentual
lived	53	59,55%
died	23	25,84%
euthanizaed	13	14,61%

Figura 3. Distribuição de saídas da base Teste

Conforme podemos observar nas tabelas acimas, o balanceamento entre as saídas indica que temos cerca 60% das entradas indicando que o cavalo sobreviveu, e cerca de 40% das entradas indicando que o cavalo não sobreviveu.

5. Análise de atributos desnecessários e valores faltantes

Inicialmente, foram analisados possíveis atributos não significantes para o estudo. Conforme apresentado no arquivo `datadict.pdf`, o atributo "`cp_data`" pode ser desconsiderado para o estudo em questão. Além disso, a atributo "`hospital number`"

também foi desconsiderado, por representar apenas um `id` de identificação do cavalo, e não corresponder a uma condição médica.

Pode-se observar também a partir dos histogramas gerados para a base treino, que os atributos "`lesion_2`" e "`lesion_3`" também não representam significância para a avaliação em questão, sendo também desconsiderados.

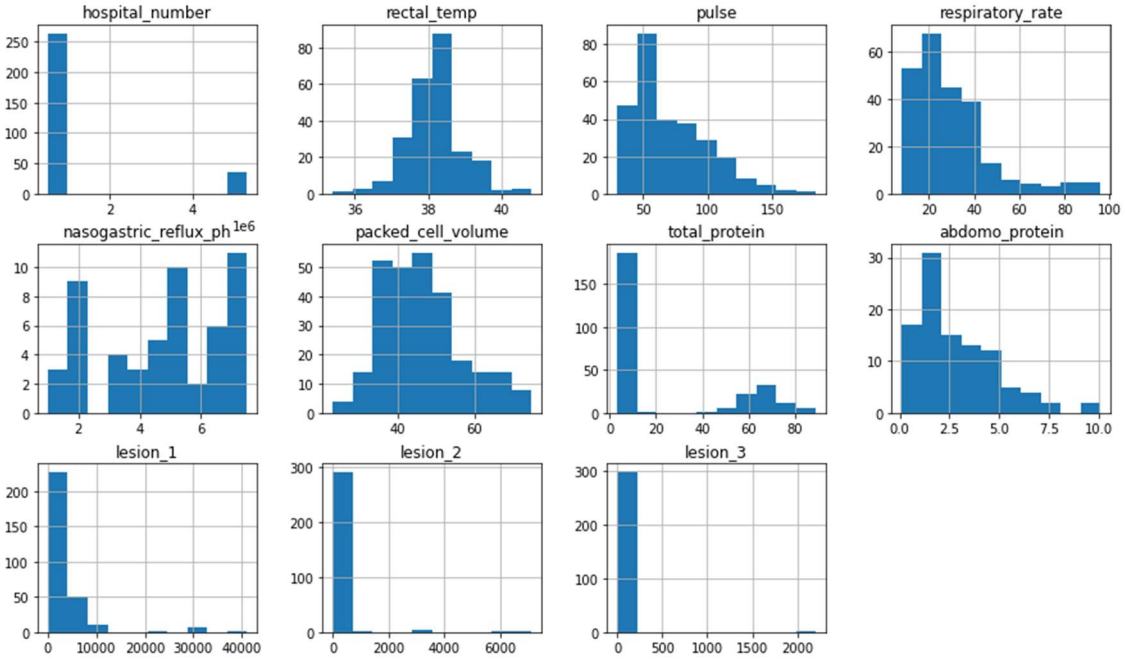

Figura 4. Histogramas da base Treino

Em seguida, foi realizada uma avaliação dos valores faltantes (*missing values*) nos atributos de treino, onde pode-se constatar que os atributos "`nasogastric_reflux_ph`" e "`abdomo_appearance`" possuem elevados percentuais de *missing values* (82% e 66%, respectivamente), sendo, portanto, também considerados não significativos para o estudo e descartados. Tal avaliação também foi realizada para a base de teste.

Atributo	*% Missing*
nasogastric_reflux_ph	82,27%
abdomo_protein	66,22%
abdomo_appearance	55,18%
abdomen	39,46%
nasogastric_reflux	35,45%

Figura 5. Percentual valores faltantes da base Treino

Atributo	% *Missing*
nasogastric_reflux_ph	84,27%
abdomo_protein	67,42%
abdomo_appearance	49,34%
abdomen	39,33%
nasogastric_reflux	38,20%

Figura 6. Percentual valores faltantes da base Teste

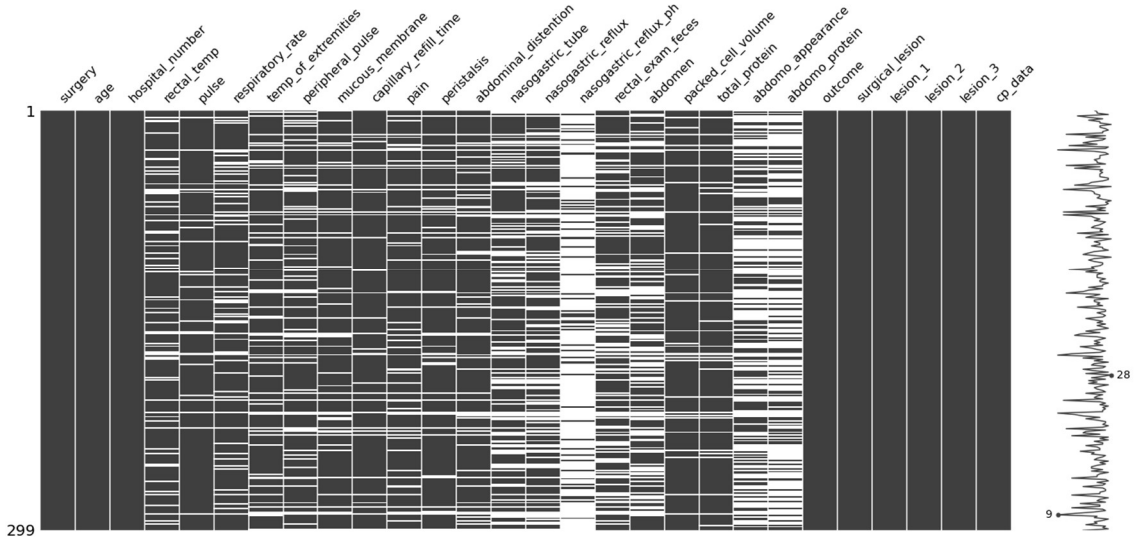

Figura 7. Matriz de distrubuição de valores faltantes pré-tratamento

6. Tratamento dos dados

Inicialmente, foi realizada a exclusão dos atributos desnecessários identificados anteriormente.

Em seguida, foi realizada a substituição do rótulo *"euthanized"* por *"died"*, uma vez que ambos indicam que o cavalo não sobreviveu, de forma que ao final dessa etapa restassem dois rótulos: *lived* e *died*.

Na sequência, foi realizado o preenchimento dos valores faltantes, sendo considerada a média para atributos numéricos, e a moda para atributos categóricos.

Outra etapa realizada foi a separação entre entradas e saídas (X_treino, y_treino, X_teste, y_teste).

Por fim, foi realizado o *encoding* dos atributos categóricos, além de ser processar o *encoding* também das saídas de treino e teste, para utilização em alguns modelos.

Ao final do processo, foi gerada nova matriz para checagem dos valores faltantes, verificando-se a eficácia dos tratamentos realizados.

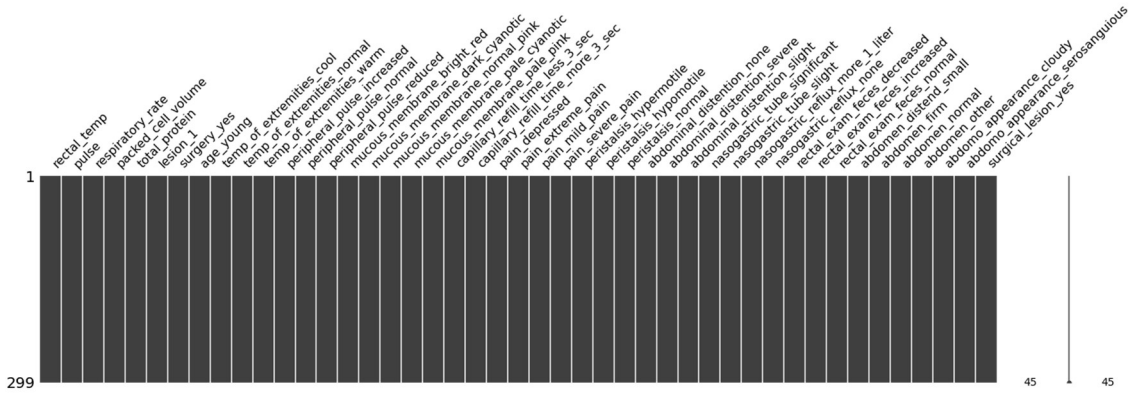

Figura 8. Matriz de distribuição de valores faltantes pós-tratamento

7. Avaliação dos modelos

Realizadas as etapas anteriores, foi realizado o treinamento e avaliação dos seguintes modelos: Árvore de Decisão, SVM, Random Forest, KNN, Grid Search e Logistic Regression. Para melhor avaliação dos modelos, foram realizadas as predições com a base de teste, mas também foi realizada a avaliação com a base de treino. As tabelas abaixo apresentam os resultados obtidos.

Base Teste

Modelo	Acurácia	Kappa	F1
Árvore de Decisão	0,8427	0,6705	0,8704
SVM	0,9326	0,8601	0,9434
Random Forest	0,8764	0,7400	0,8990
KNN	0,8202	0,6268	0,8202
Grid Search	0,8090	0,5981	0,8075
Logistic Regression	0,8315	0,6454	0,8301

Base Treino

Modelo	Acurácia	Kappa	F1
Árvore de Decisão	0,8829	0,7561	0,9025
SVM	0,8896	0,7669	0,9106
Random Forest	0,8796	0,7468	0,9016
KNN	0,7960	0,5726	0,7949
Grid Search	0,7960	0,5726	0,7949
Logistic Regression	0,8027	0,5866	0,8016

Seguem a matriz de confusão obtidas para cada um dos modelos.

Figura 9. Matriz de confusão - Árvore de Decisão

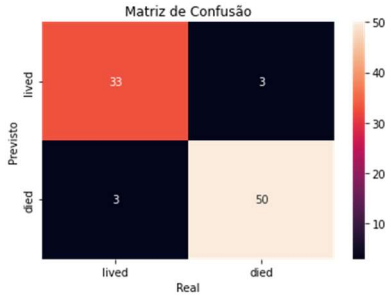

Figura 10. Matriz de confusão – SVM

Figura 11. Matriz de confusão - Random Forest

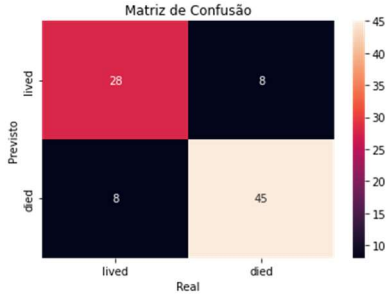

Figura 12. Matriz de confusão – KNN

Figura 13. Matriz de confusão - Grid Search

Figura 14. Matriz de confusão - Logistic Regression

8. Conclusão

Baseado nos resultados obtidos, pode-se concluir que o modelo *SVM* se mostrou mais eficaz para a previsão se sobrevivência, ou não, de um cavalo baseado em suas condições médicas.