Relatório Final - Horse Colic Data Set

Disciplina: **Data Mining**
Aluno: **Antonioni Barros Campos**
Contato: **(81) 99104-7437**
Email: antonioni.campos@gmail.com
Professora: **Manoela Kohler**

Introdução

O projeto sugerido foi o de classificação da sobrevivência ou não de cavalos baseados em suas condições médicas. Originalmente, os dados são representados por 27 atributos que são explicados no arquivo `datadict.pdf` e possuem 368 instâncias no total.

Para o projeto, foram fornecidos dois grupos de dados: 1 data set de treinamento com 299 instâncias (representado em `horse.csv`) e 1 data set de teste com 89 instâncias (representado em `horseTest.csv`).

O estudo do data set foi feito em Python e o código está no arquivo `analise.ipynb`.

Repositório do Projeto: https://github.com/antonionicampos/horse-dataset

Análise e tratamento dos dados

Inicialmente, foram carregados os arquivos .csv (todas as análises dos dados nos arquivos foram feitas utilizando a biblioteca pandas).

Em seguida, foi feita uma exploração dos dados. Dos 27 atributos, foi constatado que 7 atributos eram valores numéricos enquanto o restante eram dados categóricos.

A primeira operação feita no data set foi de retirada do atributo `cp_data` devido ao fato de que essa variável não tem significância devido aos dados de patologia não são incluídas e coletadas nesses casos.

Depois, foi feito um filtro de valores faltantes (NA) e constatou-se que existiam muitos atributos com valores em falta. O atributo com mais alto número de NA foi o atributo `nasogastric_reflux_ph` com 246 de 299 valores faltantes. Nesse caso, foi decidido retirar esse atributo do data set por não ter condições de representar sua contribuição na modelagem.

Após, foi decidido retirar também o atributo `hospital_number` devido a sua alta cardinalidade e não trazer nenhuma contribuição nos seus dados.

Finalmente, foram retirados os atributos `lesion_2` e `lesion_3` por terem variabilidade quase inexistente (enquanto `lesion_2` possuía apenas 7 eventos diferentes de zero, `lesion_3` possuía apenas 1).

O atributo que identifica se o cavalo morreu ou sobreviveu é o `outcome`. Como no data set original, existiam três classes (lived, died e euthanized). Como o objetivo da análise é detectar se o cavalo sobreviveu, foram agrupados em uma única, as classes died e euthanized.

Para o tratamento dos valores faltantes, foram criado abordagens para dados numéricos e para os dados categóricos. Para os dados numéricos, os dados faltantes foram preenchidos pela **média** dos dados existentes

para o atributo específico. Para os dados categóricos, os dados faltantes foram preenchidos pela **moda** dos dados existentes para o atributo específico.

Para finalizar, utilizou-se a técnica de One-Hot Encoding para os dados categóricos devido ao fato de que esses dados não são ordenados.

Treinando modelos

Antes de qualquer treinamento, foi realizado uma normalização dos dados para não inserir pesos para atributos em relação a outros. A normalização escolhida foi a Min-Max devido aos dados categóricos estarem compreendidos entre 0 e 1 (devido ao One-Hot Encoding).

A seguir, foi realizado treinamento usando uma série de algoritmos de classificação. Foram realizados utilizando **validação cruzada K-fold** (k = 30). Abaixo, uma tabela com a média da acurácia para cada algoritmo no conjunto de dados de treino e de teste.

Algoritmo	**Treino**	**Teste**
Logistic Regression	0.764	0.785
Decision Tree	1.000	0.805
Random Forest	1.000	0.985
Gradient Boosting	1.000	0.992
Gaussian Naive Bayes	0.757	0.757
Support Vector Machine	0.750	0.654
K-Nearest Neighbors	0.866	0.696

Conclusão

Como pode-se notar na tabela acima, os algoritmos Random Forest e Gradient Boosting foram os que mostraram melhores resultados no data set de testes, chegando a ter classificadores com 100% de acurácia.

Devido ao baixo número de dados e quantidade de valores faltantes, o processo de treinamento tornou-se bastante difícil. Com o One-Hot Encoding, houve um grande aumento no número de atributos tornando ainda mais complexo o treinamento e modelo.