**Relatório de Data Mining**

Horse Colic Data Set



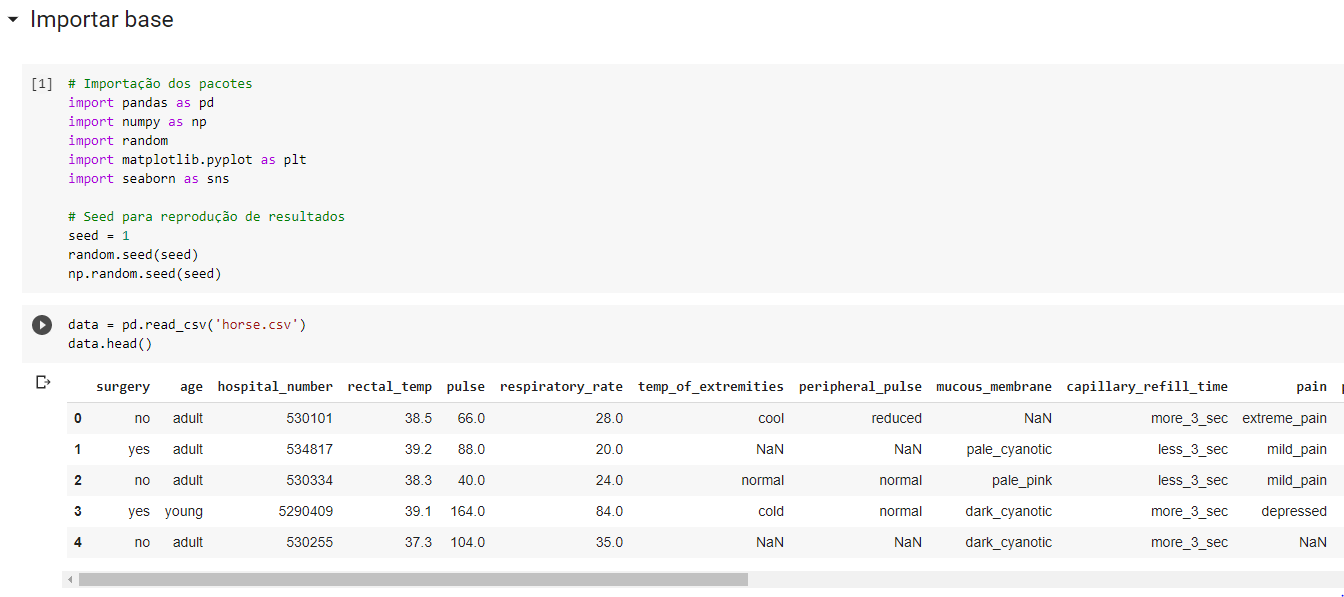
Felipe L. Meira Alves

RELATÓRIO DE DATA MINING 1

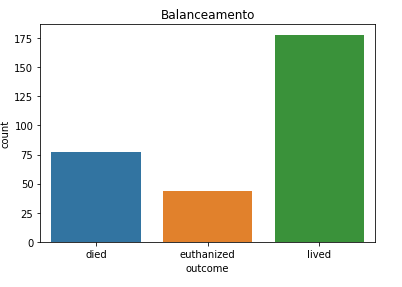
**Análise exploratória, missing values e atributos desnecessários**

1ª Etapa: Consiste na análise da documentação e carga do arquivo “horse.csv” para uma visão geral da base. A primeira análise foi realizada através da utilização do comando .head() para conhecer melhor a base e seus respectivos campos.

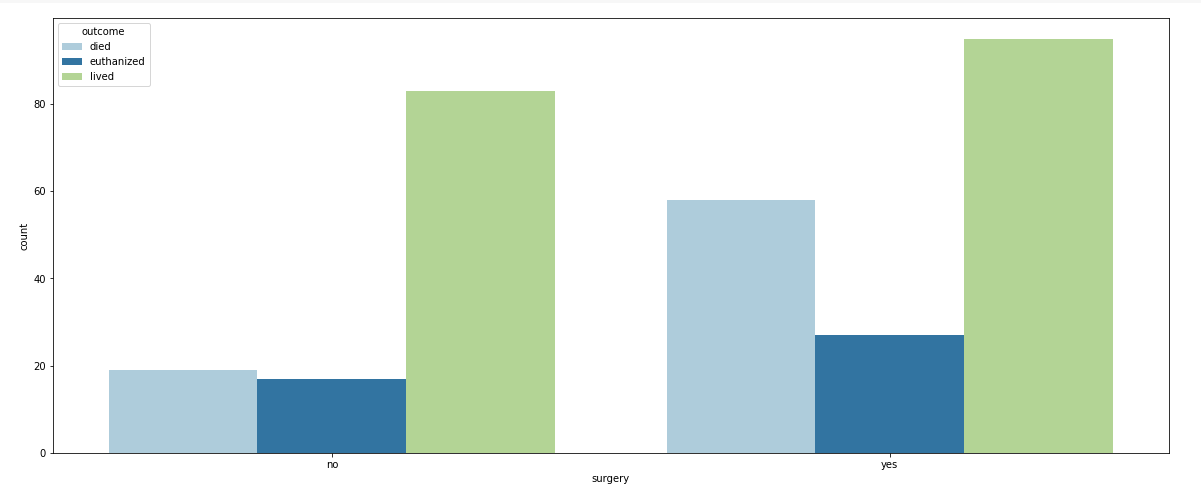
Para uma análise exploratória inicial, foram gerados alguns gráficos para maior aprofundamento do assunto.



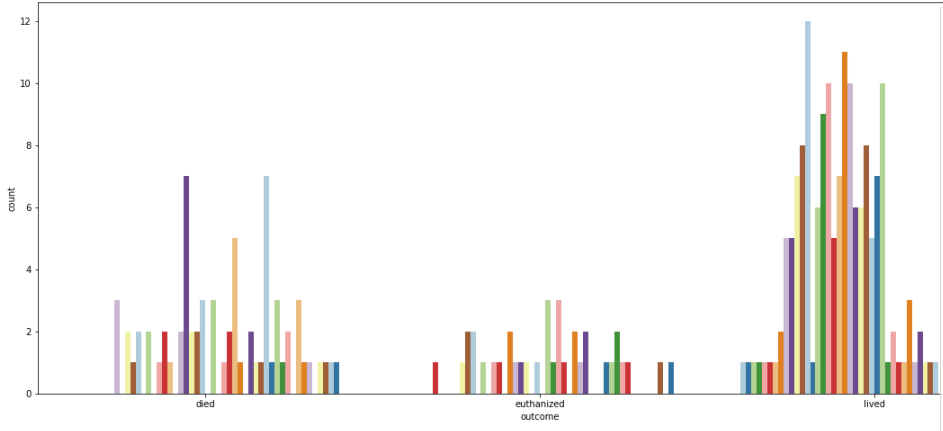
Análise Exploratória:



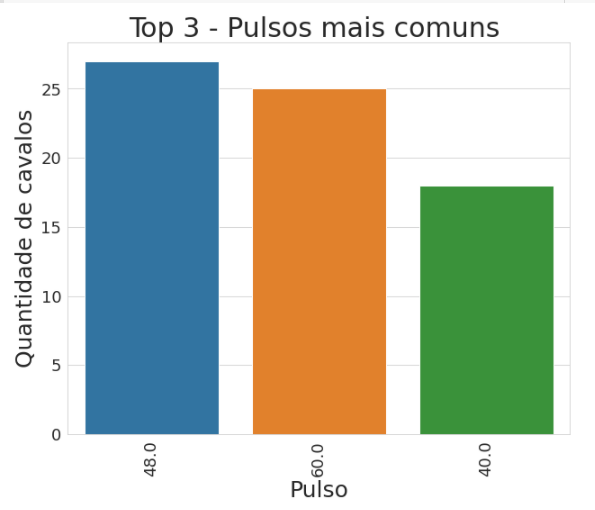
O gráfico apresenta uma visão geral de balanceamento entre as classes de outcome, nesse momento eram divididas entre “Died”, “Euthanized” e “Lived”.



O gráfico apresenta uma visão dos casos de “died”, “Euthanized” e “Lived” informando se os cavalos passaram por cirurgia ou não. O que chama atenção é o numero alto de cavalos que morreram e que já haviam passado por cirurgia.



O gráfico apresenta uma visão de nível de hematócrito no sangue, os cavalos que morreram possuíam um nível bem menor do que os vivos.

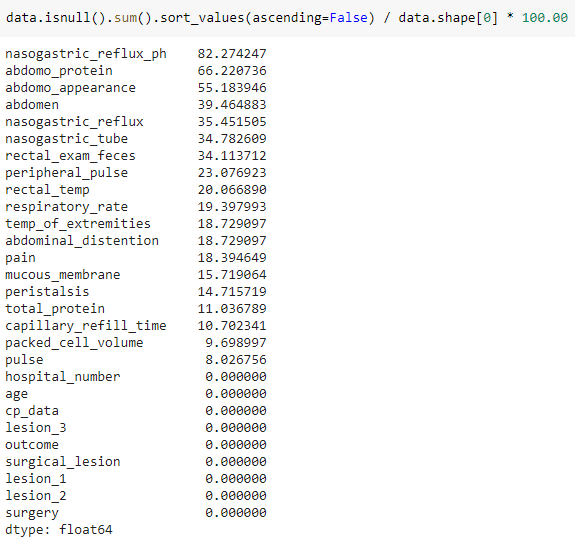


O gráfico acima apresenta os pulsos mais comuns encontrados nos cavalos, entre os 3 pulsos mais encontrados podemos perceber que estes têm uma variação de 40 até 60.

2º Etapa – Missing values e limpeza de atributos desnecessários.

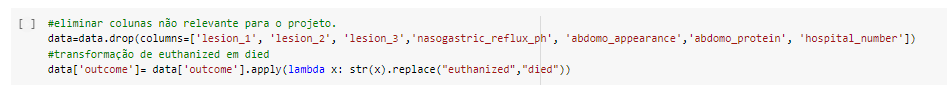
Essa etapa consiste na limpeza de alguns atributos que se mostraram desnecessários e os casos de missing values.

Existem 3 colunas com % de missing values acima de 40%, considerando esse critério para uma melhor qualidade da informação da base deverão ser retiradas as colunas ‘nasogastric\_reflux\_ph', 'abdomo\_protein', 'abdomo\_appearance'.

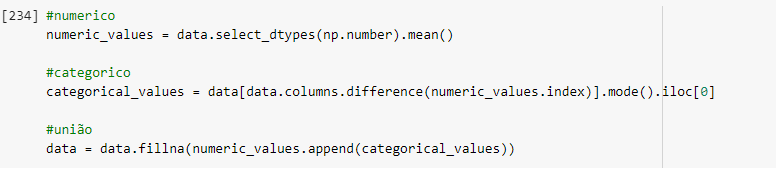


Eliminar colunas como lesion\_1, lesion\_2 , lesion\_3, segundo a documentação só será possível ter esse resultado depois da cirurgia ou realização da autopsia. O ‘hospital\_number’ não acrescenta informação relevante para o objetivo do trabalho.

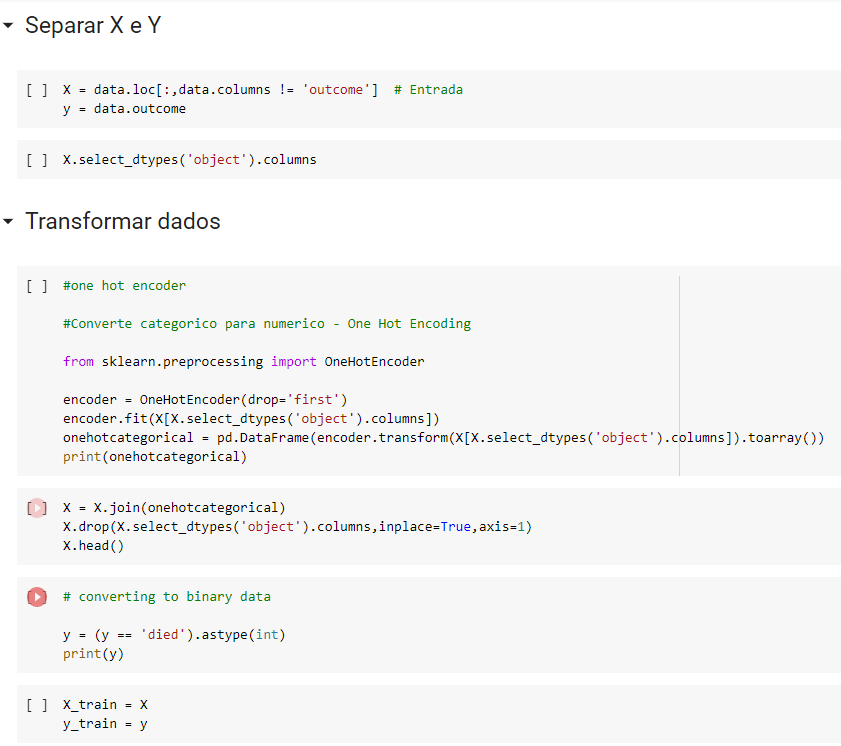
A coluna outcome está com 3 itens possíveis de retorno, sendo que Euthanized e Died refletem o mesmo cenário. Dessa forma, os casos de Euthanized foram transformados em Died.



Tratamento de NaN – As colunas com dados numéricas que possuem “NaN” serão substituídas por média e as categóricas pela moda.



3ª Etapa – Consiste em separar os dados em X e Y. Em seguida, transformar os dados de X categóricos em numéricos através do One Hot Encoding e separar os dados de y.



**Base de teste**

Foram realizados os mesmos procedimentos da base treino para a base teste.





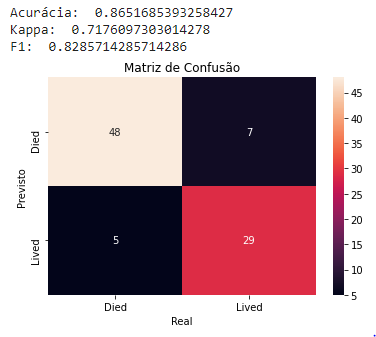
**Testando diferentes modelos**

O estudo proposto em aprendizado supervisionado busca avaliar a qualidade dos modelos de classificação, foram realizados experimentos com estes modelos para medir sua capacidade de rotular se o cavalo irá morrer ou viver dado os parâmetros de entrada fornecidos.

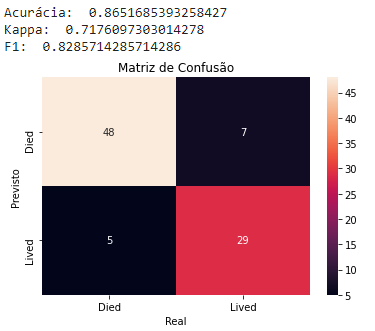
Como premissa para a avaliação dos modelos, todos foram executados e passaram por normalização e Grid Search para uma verificação se seria possível melhorar a sua qualidade.

**Modelo Random Forest**

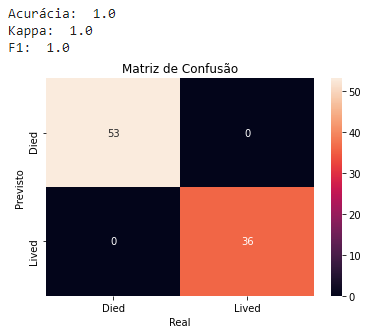
**Modelo:**



**Normalizar:**

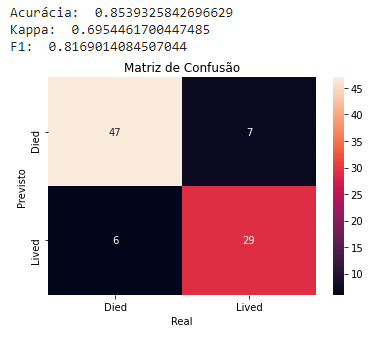


**Grid Search:**

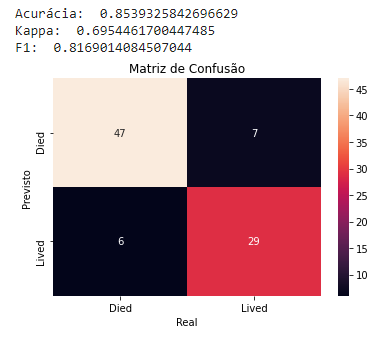


**Modelo Árvore de decisão**

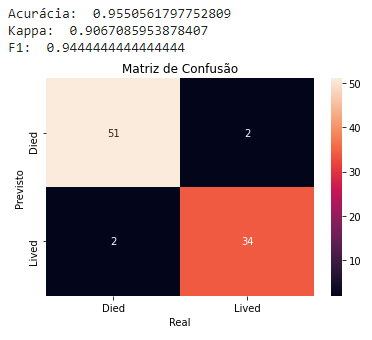
**Modelo:**



**Normalizar:**

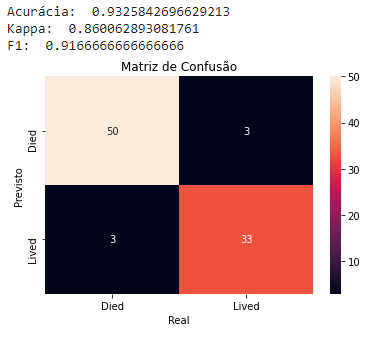


**Grid Search:**

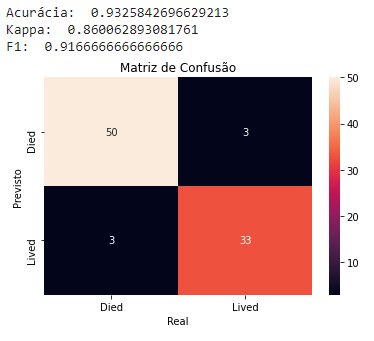


**Modelo SVM**

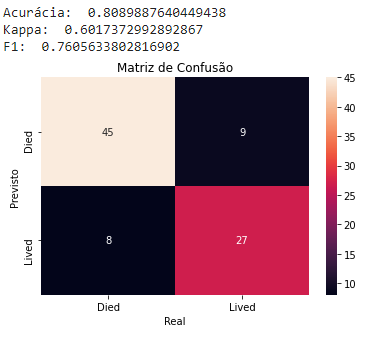
**Modelo:**



**Normalizar:**

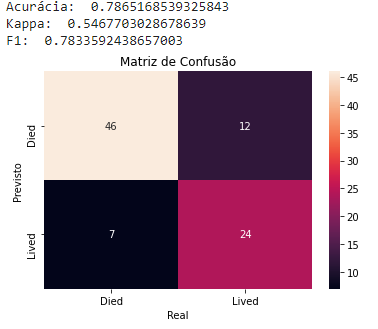


**Grid Search:**

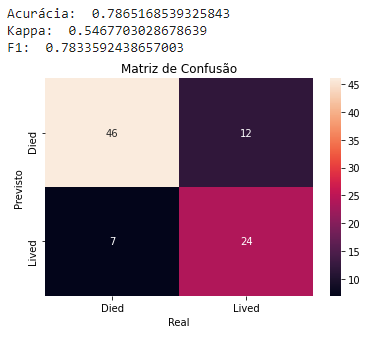


**Modelo KNN**

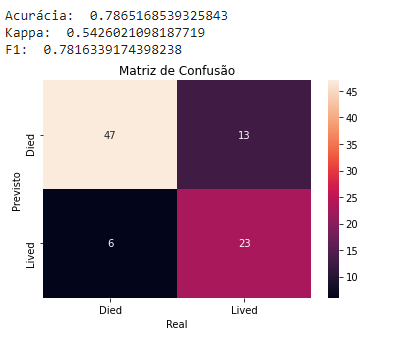
**Modelo:**



**Normalizar:**

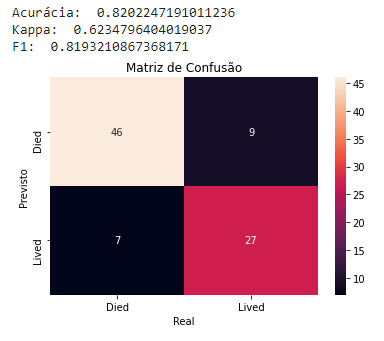


**Grid Search:**

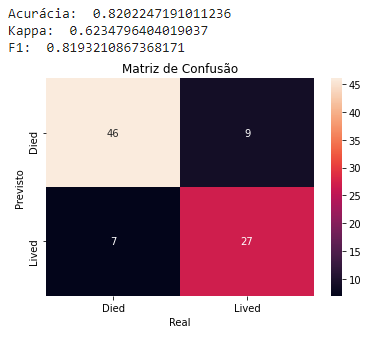


**Modelo Regressão Logística**

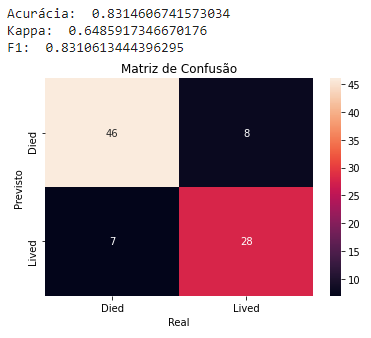
**Modelo:**



**Normalizar:**



**Grid Search:**



**Conclusão:**

No quadro abaixo, podemos ter uma visão geral dos modelos classificatórios em suas versões mais otimizadas em termos de capacidade de fazer uma predição. Em uma análise destes indicadores, o que obteve o melhor resultado foi o modelo Random Forest com a otimização de seus parâmetros através do Grid Search, permitiu o modelo alcançar 1.0 para acurácia, Kappa e F1. O modelo que obteve o pior resultado foi o KNN, apresentando uma capacidade de predição baixa com uma acurácia de 0.78, Kappa 0.54 e F1 de 0.78 mesmo tentando normalizar ou utilizar o grid Search o modelo não apresentou melhora.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Indicadores** | **Random Forest** | **Árvore de decisão** | **SVM** | **Regressão Logística** | **KNN** |
| Acurácia | 100% | 95.5% | 93.2% | 83.1% | 78.7% |
| Kappa | 100% | 90.6% | 86.0% | 64.8% | 54.6% |
| F1 | 100% | 94.4% | 91.6% | 83.1% | 78.3% |