Problema de Classificação de Drogas

gabrielstella28@gmail.com
Gabriel Felipe Dalla Stella

1 Introdução ao Problema

O conjunto de dados Drug Classification foi extraído do site Kaggle (Encontrado Aqui).

Este conjunto de dados possui os informações médicas obtidas de pacientes que utilizaram alguma dentre as drogas denominadas A, B, C, X, e Y, nomes concedidos pelo organizador do conjunto de dados.

O objetivo deste trabalho foi utilizar as variáveis preditoras para descobrir qual a droga mais provavelmente utilizada pelos pacientes e a partir deste modelo tentar responder as seguintes perguntas:

- O que cada droga faz nas pessoas?
- Há diferença entre homens e mulheres no efeito das drogas?
- A idade tem algum efeito relevante na interação com as drogas?
- Supondo que uma pessoa com 20 anos, pressão alta, colesterol baixo e a razão de sódio e potássio de $13,093$, qual é a droga mais provável que essa pessoa deve ter usado?

As variáveis preditoras utilizadas do *dataset* são:

- Age: Idade do paciente
- Sex: Gênero do paciente (M, F)
- BP: Pressão sanguínea do paciente (Low, Normal, High)
- Cholesterol: Concentração de colesterol (Normal, High)
- Na_to_K: Razão entre a concentração de sódio e potássio do paciente.

Para realizar essa análise, a linguagem de programação utilizada foi Python, utilizando a IDE Jupyter Notebook.

2 Pré-processamento

Sobre as variáveis foram utilizadas as seguintes transformações:

- Age: `MinMaxScaler()`
- Sex: `OneHotEncoding(drop=‘‘first")` (Classe positiva: M)
- BP: `OneHotEncoding(drop=‘‘first")` (Classes utilizadas: Low, Normal)
- Cholesterol: `OneHotEncoding(drop=‘‘first")` (Classe positiva: Normal)
- Na_to_K: `MinMaxScaler()`

A opção drop="first"é utilizada para evitar problemas de colinearidade entre as variáveis, para melhorar o desempenho dos modelos treinados.

Também utilizamos as técnicas de resampling para melhorar o precision e o recall dos modelos, visto que os dados estão desbalanceados.

3 Análise Gráfica

Segue o pairplot do dataset, gerado antes do balanceamento dos dados:

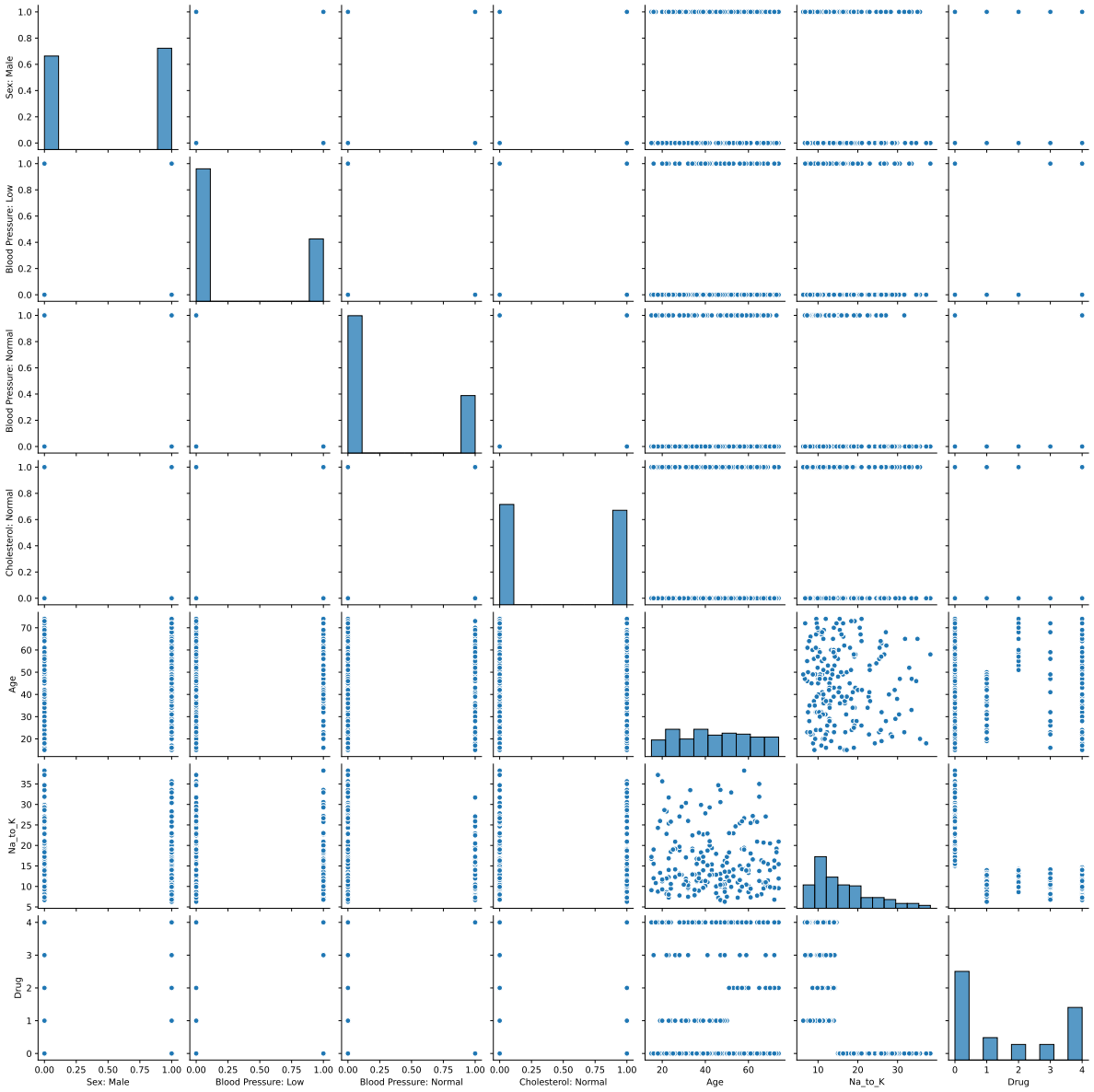

Figura 1: Pairplot

Com a análise deste gráficos podemos gerar as seguintes hipóteses:

- Vemos que a concentração a razão Na_to_K mais alta que a mediana aparentemente indica que a droga mais provável é a que corresponde à droga Y, que corresponde ao valor 0.
- Em geral, a droga B (valor 2) é utilizada pela população acima dos 50 anos, enquanto que a droga A (valor 1) é utilizada pela população abaixo dos 50. As drogas restantes são bem distribuídas para cada uma das idades.

- Aparentemente a droga C (valor 3) causa um aumento no colesterol, se os pacientes, cujas informações foram coletadas, forem usuários frequentes dessa droga. Lembrando a frase clássica "Correlação não implica causalidade"
- Quanto à pressão, as drogas A e B (valores 1 e 2) parecem ter correlação positiva com a pressão alta. A droga C (valor 3) parece estar correlacionada com a pressão baixa.

4 Treino e Teste de modelos

4.1 Rede Neural

O primeiro modelo treinado é uma rede neural profunda com as seguintes configurações:

- A rede é do tipo feedforward com 6 camadas escondidas.
- Cada camada é composta por 7 neurônios, Batch Normalization e a função de ativação GELU.
- Função de saída LogSoftmax(dim=1).
- A função de perda (*Loss Function*) é dada por `NLLLoss()` (*Negative Log-Likelihood Loss*)
- O otimizador utilizado foi o Adam com *learning rate* (taxa de aprendizado) de 8×10^{-5} e *weight decay* (Regularização ℓ_2) de 1×10^{-5}.
- O treinamento foi realizado com *batch size* de 64, número de épocas 10000 e EarlyStopping com patience 6.
- A divisão utilizada foi de 50%, 25% e 25%, para os dados de treino, validação e teste, respectivamente.

O gráfico da função de perda em cada iteração é dado por:

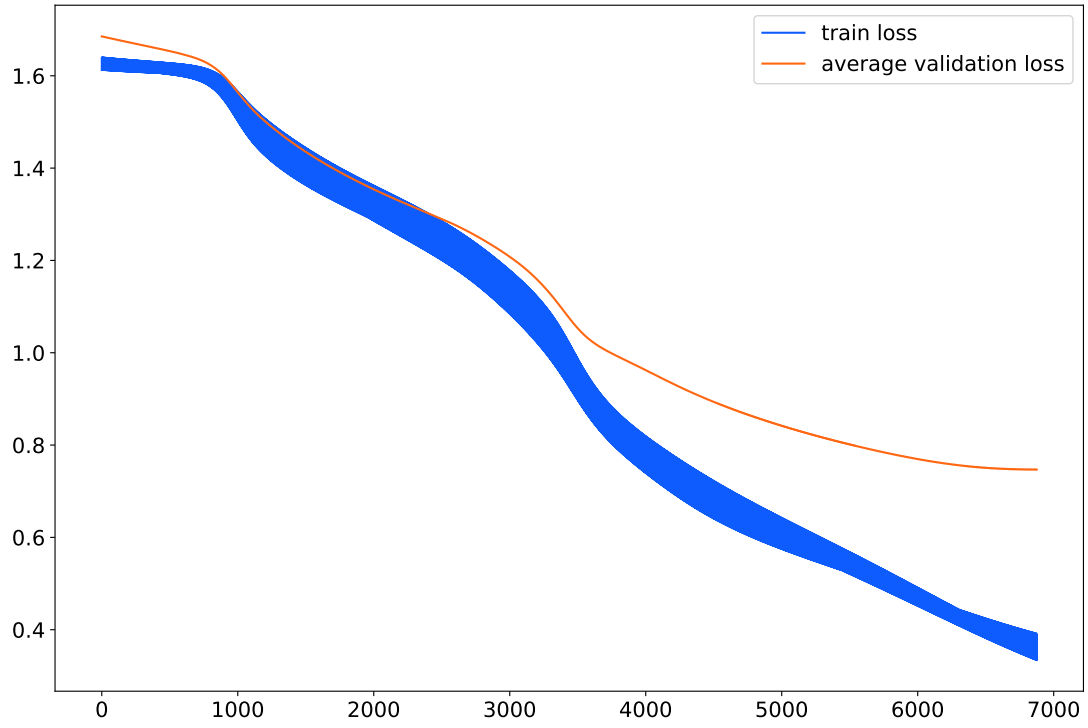

Figura 2: Função perda por iteração

Nos dados de treino e validação, o desempenho foi de 88% de precisão.

Verificando o desempenho do melhor classificador nos dados teste, que corresponde à um total de 50 dados, obtemos as seguintes tabelas utilizando a função classification_report:

	precision	recall	f1-score	support
0	1.000000	0.925926	0.961538	27
1	1.000000	1.000000	1.000000	2
2	0.666667	1.000000	0.800000	2
3	0.833333	1.000000	0.909091	5
4	1.000000	1.000000	1.000000	14
accuracy			0.960000	50
macro avg	0.900000	0.985185	0.934126	50
weighted avg	0.970000	0.960000	0.962140	50

Esse classificador possui um desempenho interessante nos dados de teste. Pode ser que o número pequeno de amostras influencie a ter um alto desempenho.

4.2 Random Forest

O segundo modelo testado foi o Random Forest, com o método de RandomSearchCV com 20 iterações e o hiperparâmetro estimado foi o max_depth entre 1 e 40.

Este modelo foi treinado com os mesmos dados de treino e validação da rede neural, visto que com o EarlyStopping o conjunto de validação também faz parte do treino, na rede neural. O conjunto de teste também se mantém inalterado.

A melhor performance é dada por max_depth de 34, com desempenho de 99% de precisão.

Verificando o desempenho do melhor classificador nos dados teste, que corresponde à um total de 50 dados, obtemos as seguintes tabelas utilizando a função classification_report:

	precision	recall	F_1-score	support
0.0	1.0	1.0	1.0	27
1.0	1.0	1.0	1.0	2
2.0	1.0	1.0	1.0	2
3.0	1.0	1.0	1.0	5
4.0	1.0	1.0	1.0	14
accuracy			1.0	50
macro avg	1.0	1.0	1.0	50
weighted avg	1.0	1.0	1.0	50

Vemos um melhor desempenho nesse modelo, tanto em termos de validação, quanto em termos de dados de teste.

5 Análise de impactos

Para analisar os impactos de cada variável sobre as probabilidades de cada classe utilizamos a biblioteca shap.

Agora, utilizando o modelo Random Forest, que obteve um melhor desempenho, podemos realizar a inferência sobre os dados. Nesse caso podemos responder as seguintes perguntas:

- O que cada droga faz nas pessoas?
- Há diferença entre homens e mulheres no efeito das drogas?
- A idade tem algum efeito relevante na interação com as drogas?
- Supondo que uma pessoa com 20 anos, pressão alta, colesterol baixo e a razão de sódio e potássio de $13,093$, qual é a droga mais provável que essa pessoa deve ter usado?

5.1 Droga Y

A seguir temos a figura contendo as informações dos Shapley Values gerada pela biblioteca shap:

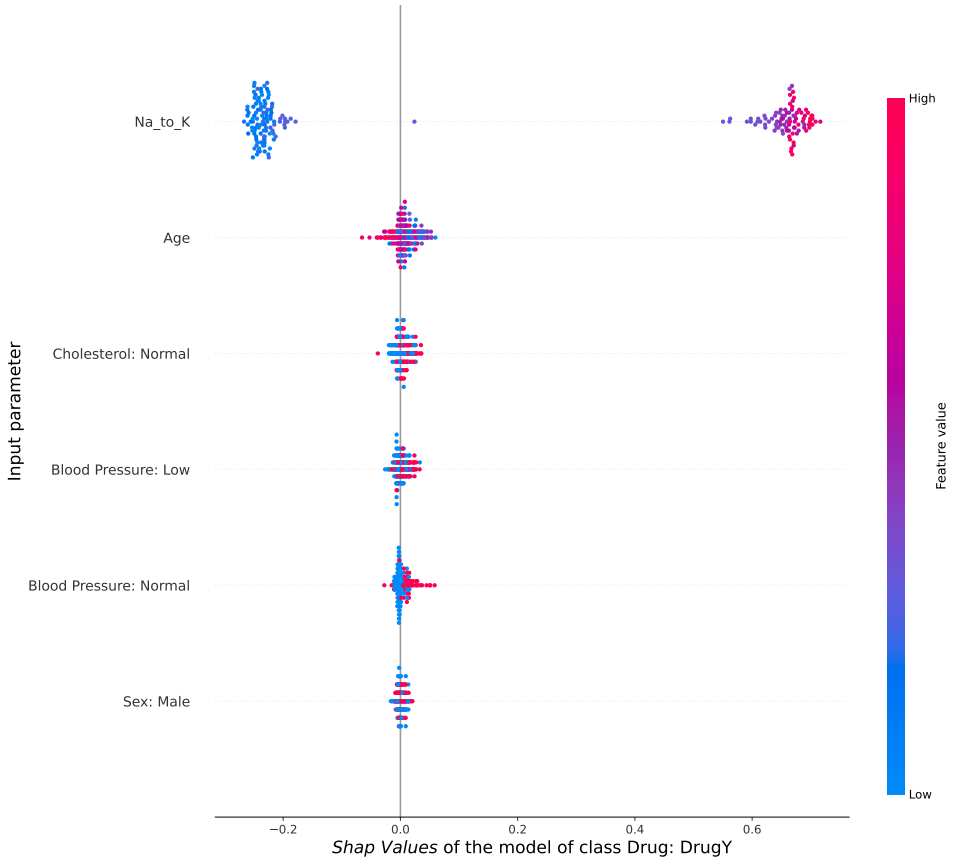

Figura 3: Shapley Values referentes à droga Y

Para a droga Y temos as seguintes conclusões:

- Podemos ver que o maior impacto dessa droga é na razão entre as concentrações de sódio e potássio. Mais precisamente, quando a razão é menor, a tendência é que não seja a droga Y e para valores maiores da razão, é esperado que a droga predita seja Y.

  Lembrando que essa afirmação está de acordo com a análise gráfica, onde comentamos que valores mais altos de razão Na_to_K tende a ser a droga Y.

- As outras variáveis tem pouca influência sobre a predição.

- O gênero é pouco relevante sobre a probabilidade de ser a droga Y.
- A idade possui pouca influência sobre a predição dessa droga.

5.2 Droga A

A seguir temos a figura contendo as informações dos Shapley Values gerada pela biblioteca shap:

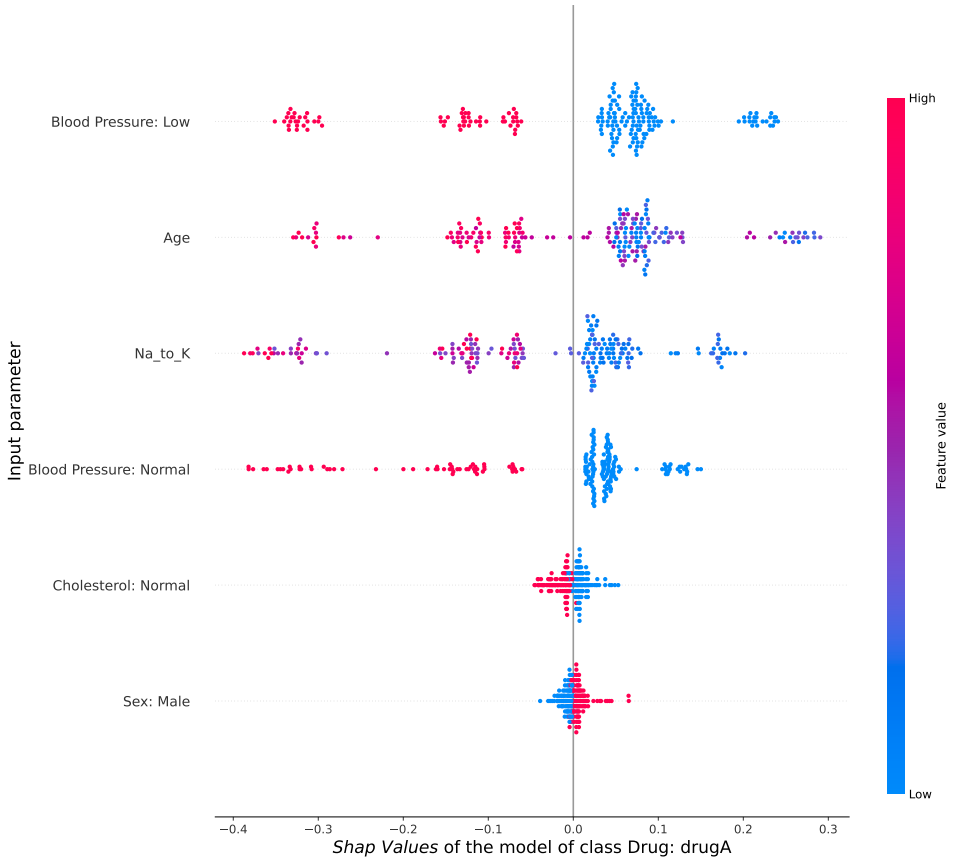

Figura 4: Shapley Values referentes à droga A

Para a droga A temos as seguintes conclusões:

- Cruzando as informações de Blood Pressure: Low e Blood Pressure: Normal, vemos que pressão alta tende a aumentar a probabilidade da de ser a droga A.

  Esta informação é compatível com o que foi visto na análise gráfica.

- Valores médios a altos da razão Na_to_K tendem a diminuir o a probabilidade de ser a droga A.
- Vemos que a tendência é que os mais jovens utilizem a droga A, como visto na análise gráfica.
- A hipótese sobre o gênero é que os homens usam levemente mais essa droga que as mulheres. Porém isso deve ser melhor investigado para obter conclusões mais sólidas.
- Outra hipótese é que o uso ou não da droga A tem leve relação positiva com o colesterol alto.

5.3 Droga B

A seguir temos a figura contendo as informações dos Shapley Values gerada pela biblioteca shap:

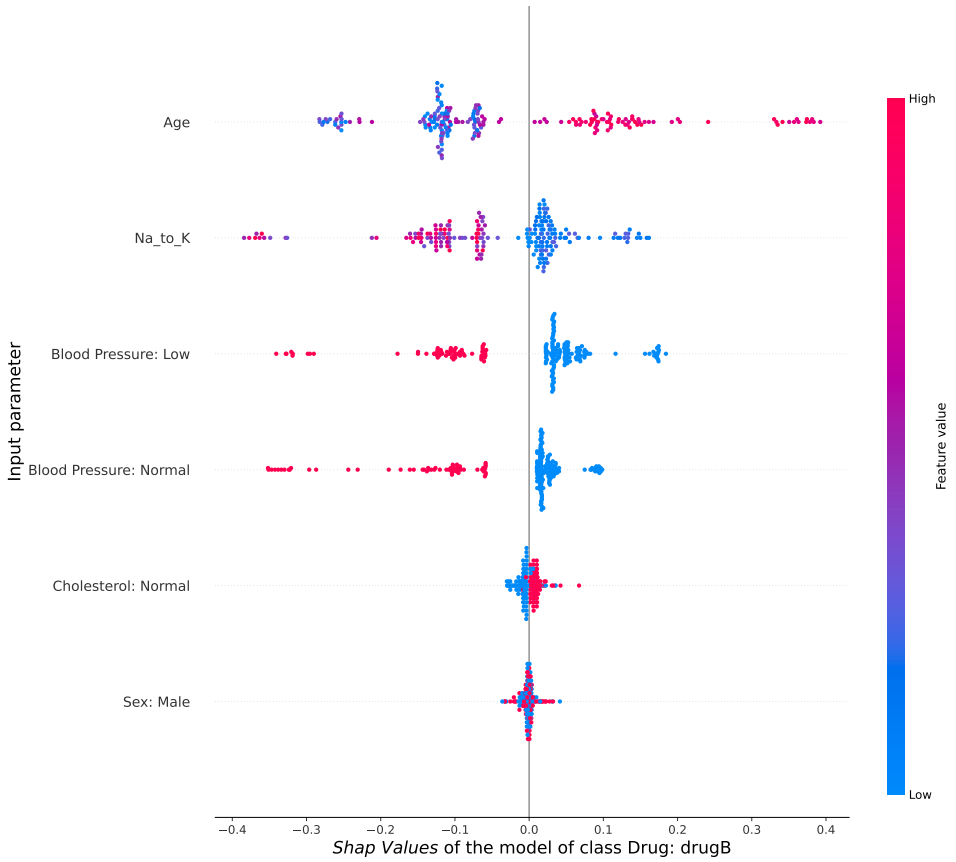

Figura 5: Shapley Values referentes à droga B

Para a droga B temos as seguintes conclusões:

- Cruzando as informações de Blood Pressure: Low e Blood Pressure: Normal, vemos que pressão alta tende a aumentar a probabilidade da de ser a droga B.

  Esta informação é compatível com o que foi visto na análise gráfica.

- Valores médios a altos da razão Na_to_K tendem a diminuir o a probabilidade de ser a droga B.

- Vemos que a tendência é que os mais velhos utilizem a droga B, como visto na análise gráfica.

- Não há um gênero de preferência para o uso da droga B, mais precisamente, o gênero não nos dá informação relevante para a previsão.

- Outra hipótese é que o uso ou não da droga A tem leve relação positiva com o colesterol normal.

5.4 Droga C

A seguir temos a figura contendo as informações dos Shapley Values gerada pela biblioteca shap:

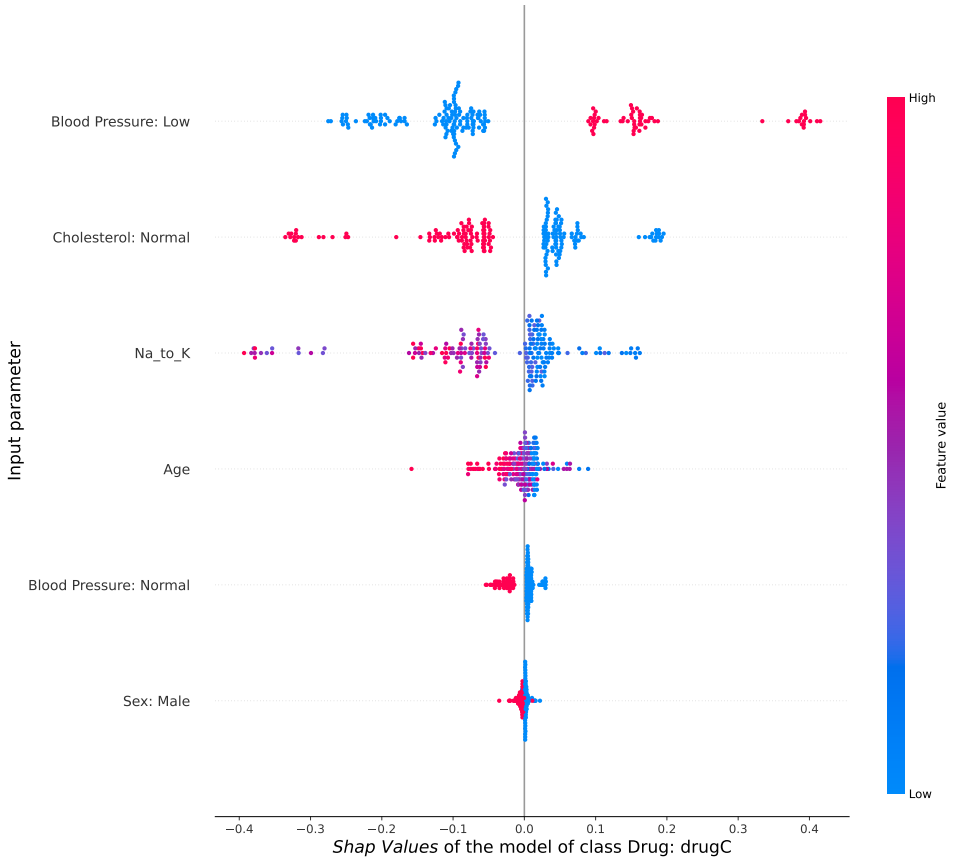

Figura 6: Shapley Values referentes à droga C

Para a droga C temos as seguintes conclusões:

- Cruzando as informações de Blood Pressure: Low e Blood Pressure: Normal, vemos que pressão baixa tende a aumentar a probabilidade da de ser a droga C. Essa variável tem uma influência muito alta sobre a probabilidade.

  Esta informação é compatível com o que foi visto na análise gráfica.

- Valores médios a altos da razão Na_to_K tendem a diminuir o a probabilidade de ser a droga C.

- Vemos que uma leve tendência é que os mais jovens utilizem a droga C.

- O gênero tem influência praticamente nula sobre a probabilidade de ser a droga C.

- Outra hipótese é que o uso ou não da droga C tem forte relação positiva com o colesterol alta.

5.5 Droga X

A seguir temos a figura contendo as informações dos Shapley Values gerada pela biblioteca shap:

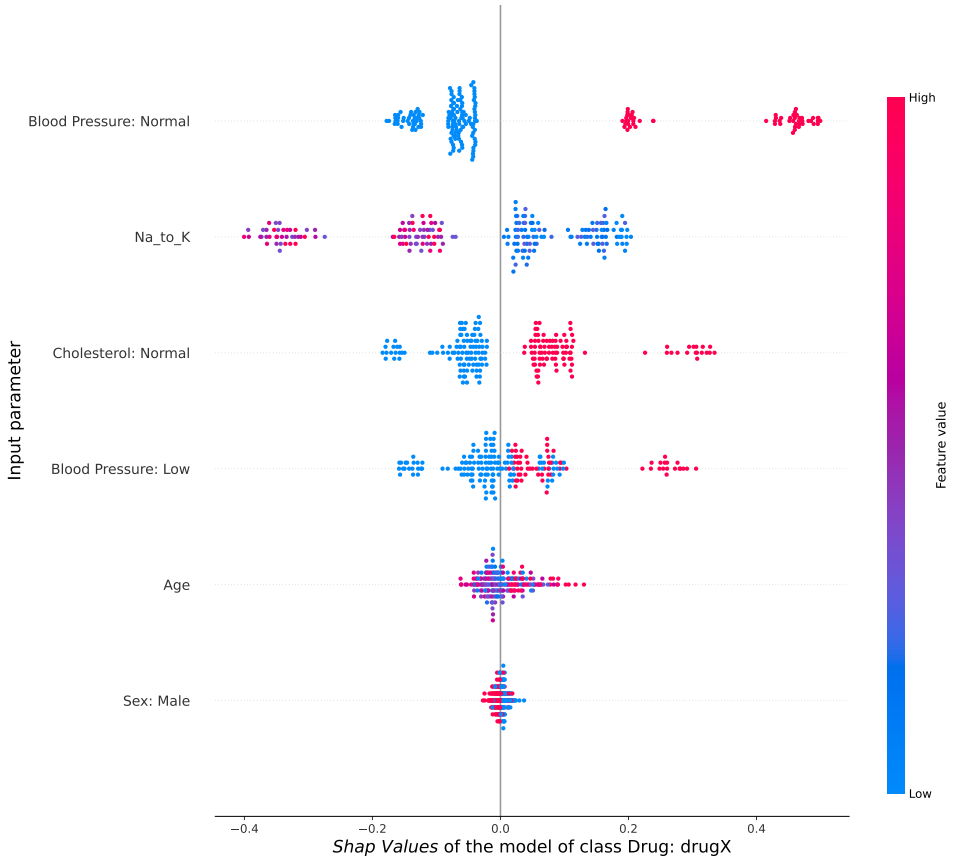

Figura 7: Shapley Values referentes à droga X

Para a droga X temos as seguintes conclusões:

- Cruzando as informações de Blood Pressure: Low e Blood Pressure: Normal, vemos que os usuários da droga X tendem, em geral, a apresentar pressão baixa ou normal na coleta dos dados.
- Valores médios a altos da razão Na_to_K tendem a diminuir o a probabilidade de ser a droga X.
- A idade tem relação pouco clara com o uso da droga X.
- O gênero tem influência praticamente nula sobre a probabilidade de ser a droga X.
- O uso ou não da droga X tem forte relação positiva com o colesterol normal.