Título: Modelo de Machine Learning para **predizer alocação** de pacientes com lombalgia crônica para unidades de **práticas integradas** em um sistema de cuidados de saúde baseados em valor.

Independente do resultado mas seguir para publicação.

Metodologia (tratamento dos dados, seleção das principais variáveis utilizadas na predição, de acordo com a tabela que mostra as maiores correlações)

1. De todo o dataset, há 240 linhas, sendo que 138 responderam à questão se possuem dor na região lombar (D30 correspondente ao questionário de Inventário Breve de Dor), portanto, estas 138 linhas foram selecionadas para serem utilizadas na construção do modelo de predição
2. Pré-processamento dos dados, identificando quais variáveis apresentam respostas de caráter binário e quais de caráter ordinal para cada um dos 6 questionários
3. Variáveis que são de caráter binário receberam a transformação de One Hot Encoder
4. Dos 135 vetores colunas de variáveis preditoras, selecionou-se aquelas que possuem maior correlação com a variável se possuir dor na região lombar através das frequências apresentadas na tabela cruzada e do coeficiente de ponto bisserial (maior que 0.3 na tabela cruzada e maior que 0.5 no coeficiente de correlação de ponto bisserial. A validação estatística se deu com o teste de chi quadrado para a tabela cruzada e teste de t de student de Pearson para o coeficiente de correlação de ponto bisserial para o p-valor < 0.05.
5. Resultou-se portanto em 33 colunas, das quais filtrou-se para apenas 18 colunas de acordo com o critério de selecionar aqueles vetores coluna com menos de 10 % dos valores nulos
6. Substitiui-se os dados faltantes com a mediana correspondente a cada coluna
7. Entre os algoritmos testados, têm-se que o que apresentou o melhor resultado de predição foi o

Com os seguintes resultados:

Explicação do algoritmo:

Adaboost:

Stumps, os quais são compostos geralmente por uma série de árvores com apenas um nó e duas folhas são chamados de “*weak learners*”, ou seja, não possuem grande poder de predição, dado que individualmente suportam apenas uma variável por vez para servir de suporte à decisão.

O modelo Adaboost é um modelo de ensemble, ou seja, utiliza uma série de weak learners para realizar a predição final. Também é considera um modelo de boosting, ou seja, em seu treinamento, os próximos modelos ( no caso stumps) são construídos de modo a corrigir as classificações incorretas realizadas pelos weak learners anteriores.