Modelo de Aprendizado de Máquina para **predizer alocação** de pacientes com lombalgia crônica para unidades de **práticas integradas** em um sistema de cuidados de saúde baseados em valor.

## *Machine Learning model to predict back pain pacient allocation for integrated health practice units in a value based health care system*

Vitor Pereira Barbosaa\*, João Lucas Maehara Said dos Reisb*,* Vinicius Monteiro de Paula Guiradoc,

Natalia Neto Pereira Cerized

a Laboratório de Processos Químicos e Tecnologia de Partículas do Centro de Bionanomanufatura do Instituto de Pesquisas Tecnológicas do Estado de São Paulo

b, c Laboratório de Biotecnologia Industrial do Centro de Bionanomanufatura do Instituto de Pesquisas Tecnológicas do Estado de São Paulo

c Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo

\*e-mail: vitorpbarbosa@ipt.br

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## Resumo

O Resumo deve ser escrito em português, conter, no máximo, 250 PALAVRAS e estar disposto em um único parágrafo. Obedecendo ao seguinte estilo: Times New Roman 12, espaçamento 1,5, justificado, espaçamento depois de 1 linha. Deve expor, de forma sucinta, o problema científico e/ou o objetivo da pesquisa e resumir o conteúdo do trabalho. Deve ser conciso e indicar, de forma clara, os principais resultados e conclusões. O emprego de abreviações deve ser evitado. Não são admitidas referências no resumo.

Palavras-chave: corrosão; revestimento.

A lombalgia crônica é responsável por grande parte dos custos globais de saúde necessitando cuidados especializados por longo tempo. A imensa variabilidade de pacientes em seus itinerários terapêuticos entre hospitais e provedores torna os cuidados de saúde baseados em valor potencialmente menos viáveis. Para avançar mais efetivamente na gestão e nos esforços de prevenção secundária, terciária e quaternária, a adoção de ferramentas, métodos e técnicas de ciência de dados, como processamento de linguagem natural e aprendizado de máquina é necessária. Os modelos de aprendizado de máquina têm se mostrado confiáveis na previsão de resultados específicos em várias doenças neurológicas, mas, o desenvolvimento de estratégias de gerenciamento das unidades de prática integrada ainda não foram testados com auxílio da ciência de dados.

RESULTADO NO RESUMO

Palavras-chave: lombalgia crônica, aprendizado de máquina, QUAIS MAIS?

Abstract

*FAZER ABSTRACT*

## 1 Introdução

A complexidade de opções diagnósticas e terapêuticas disponíveis para área de cirurgia neurológica é evidente no início do século XXI. Da mesma forma, também é variado o espectro de resultados possíveis porque existem muitas dimensões de interpretação. Ainda influenciados pelo contexto social, pacientes e médicos estão sobrecarregados com informações da era digital e deste modo, a tomada de decisão hoje é cardinal e crítica. O recurso contemporâneo para essa demanda desafiante é a aplicação das tecnologias de sistemas de gestão da informação, como a inteligência artificial.

Simplificar a complexidade não é uma solução, portanto, a ciência aplicada aos dados, compreendida como conhecimento humano agregado às tecnologias digitais é a melhor alternativa disponível para tomada de decisão na área de cirurgia neurológica. Discutir opções baseadas em dados é uma decisão mais importante que uma incisão.

Avanços recentes em inteligência artificial (IA) estão criando novas oportunidades para personalizar intervenções de saúde baseadas em tecnologia para pacientes com dor crônica. Ferramentas presentes no campo de IA - ambientes de aprendizagem inteligente, geração de narratias interativa, modelagem de usuário e treinamento adaptativo - podem ser utilizadas para modelar a aprendizagem e o envolvimento de pacientes com dor crônica e fornecer suporte personalizado em tecnologias de saúde adaptativas. Muitas dessas tecnologias surgiram de aplicações centradas em atividades humanas para educação, treinamento e entretenimento. No entanto, sua aplicação na melhoria da saúde, até o momento, tem sido comparativamente limitada. Ilustramos as oportunidades fornecidas pelas tecnologias adaptativas impulsionadas pela IA para cuidados preventivos de saúde para pacientes com dor crônica, descrevendo uma visão de como as futuras intervenções preventivas de saúde para esse grande grupo de pacientes podem ser realizadas dentro e fora da clínica especializada.

Como ilustrar isso?

Revisão bibliográfica?

Artificial intelligence to improve back pain outcomes

and lessons learnt from clinical classification approaches:

three systematic reviews (TAGLIAFERRI et al., 2020)

## 2 Procedimento metodológico

### Coleta dos dados

Os dados disponíveis são resultados da avaliação clínica em uma unidade de cuidados integrados multidisciplinares especializados em dor crônica de fevereiro a dezembro de 2019. Na amostra disponível foram incluídos 240 pacientes submetidos a avaliação clínica. Os dados oriundos dos questionários são de natureza sócio-demográfica e clínica, os quais estão enumerados a seguir:

1 - Questionário básico de sintomas da coluna vertebral (Protocolo de Gotemburgo),

2 - Inventário Breve de Dor (BPI),

3 - Índice Oswestry 2.0 de Incapacidade,

4 - Questionário de Incapacidade de Roland Morris (RMDQ),

5 - Questionário para avaliação da qualidade de vida 12-Item Short Form Health Survey (SF-12)

6 - Questionário para Diagnóstico de Dor Neuropática 4 (DN-4).

### Pré processamento dos dados

As questões presentes nos questionários foram divididas em duas categorias, quais sejam, questões com respostas de caráter binário e questões com respostas de caráter ordinal. Esta abordagem permite que a observação da correlação entre as variáveis binárias seja realizada através de tabela cruzada (também denominada de tabela de contingência) e que a correlação entre variáveis binárias e variáveis de caráter ordinal seja realizada através do coeficiente de correlação de ponto bisserial.

Sobre as variáveis que apresentam respostas de caráter binário, utilizou-se a técnica de *One Hot Encoder* para gerar novas colunas.

A base de dados original gerada, resultante da junção dos 6 questionários previamente citados, é composta por variáveis, ou seja, 118 vetores colunas.

As seguintes variáveis foram removidas:

Gotemburgo - Profissão

Gotemburgo - Estado\_nascimento

Gotemburgo – Se de outro estado, há quanto tempo está em São Paulo

Gotemburgo – Em qual hospital realizou cirurgia

Gotemburgo - Faz uso de quais medicamentos

Gotemburgo – Quais outras doenças possui

Inventário Breve de Dor - Quais medicamentos você está recebendo para dor

Desta forma, resultou-se em 111 variáveis, das quais após a aplicação da técnica de One Hot Encoder para variáveis binárias, obteve-se 146 vetores coluna, dos quais, 101 de caráter binária e 45 de caráter ordinal.

Para a aplicação do modelo de aprendizado de máquina para classificação do paciente quanto à presença ou não de dor lombar, removeu-se as os pacientes que não preencheram esta opção, correspondente a dor de número 30 no questionário Inventário Breve de Dor. Como resultado, têm-se 138 pacientes na base de dados final.

Sabe-se que com esta quantidade de vetores coluna (146), e apenas 138 pacientes, têm-se um caso de dados esparsos, nos quais está presente o fenômeno de *Curse of Dimensionality,* introduzido por Bellman, (1957). A esparsidade dos dados se torna um problema para obter resultados com significância estatística em um modelo de aprendizado de máquina. Parar obter resultados com relevância estatística, faz-se necessário que o número de observações cresça exponencialmente com a dimensão (VENKAT, 2010).

Para reduzir o número de variáveis preditoras, utilizou-se os resultados derivados das análises de correlação entre variáveis binárias e entre variáveis binárias e variáveis de caráter ordinal, de modo a selecionar aquelas que apresentam os maiores valores.

Adicionalmente, utilizou-se a análise de PCA com 2 componentes para visualização das observações pertencentes ao grupo de pessoas que possuem dor na região lombar e as que assinalaram a opção de não apresentarem esta dor.

Correlações e validação da estatística

### Correlação entre variáveis binárias

Para avaliar a correlação entre as variáveis que se apresentam com caráter binário, fez-se uso da ferramenta de tabela cruzada (também denominada de tabela de contingência)

Tabela 1 – Exemplo de tabela cruzada

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Variável B | |
|  |  | 0 (não) | 1 (sim) |
| Variável A | 0 (não) | a | b |
| 1 (sim) | c | d |

Na Tabela 1 têm-se um exemplo de tabela cruzada, na qual a, b, c e d são números inteiros. Avaliou-se neste estudo se a mudança de proporção não/sim da variável B possui correlação com a variável B.

Para avaliar a relevância estatística dos resultados, utilizou-se o teste de hipótese de chi-quadrado:

H0: Independência – A não depende de B

H1: Dependência – A depende de B

O cálculo do chi-quadrado se dá através da seguinte expressão:

Em que:

, , e , são as contagens de observações presentes na Tabela 1

:

:

:

:

Pode-se interpretar este teste de hipótese como a diferença na distribuição de frequência da variável A devido a inserção da variável B. Deste modo, quando o valor do chi-quadrado calculado é maior ou igual ao chi-quadrado tabelado, ao adotar nível de significância de 5 %, pode-se rejeitar a hipótese nula da independência e considerar a hipótese alternativa de dependência da variável A em relação à variável B.

### Correlação entre questões binárias e questões com resposta de caráter ordinal

### A correlação entre questões que se apresentam dicotômicas e questões que apresentam em escala ordinal, foi realizada através do coeficiente de ponto bisserial.

Considera-se a divisão de um grupo em (0 – não) e (1 – sim) e, portanto, pode-se calcular o coeficiente de ponto bisserial de acordo com a seguinte expressão:

Em que:

: Desvio padrão considerando todos os dados da população

e : Média da variável que apresenta caráter ordinal para os pacientes que responderam a questão de forma negativa (0 – não) e de forma positiva (1 - sim), respectivamente

e : Número de pessoas pertencentes a cada um dos grupos (0 – não) e (1 – sim)

A avaliação da significância estatística deste coeficiente é realizada pelo teste de hipótese do coeficiente de correlação de Pearson, dado que a correlação de ponto bisserial é um caso específico daquela para uma das variáveis sendo dicotômica. Neste teste de hipótese têm-se:

H0: = 0

H1: ≠ 0

O teste utiliza a distribuição de *student*, de modo que o valor de t calculado se dá por:

Em que:

: Coeficiente de correlação de ponto bisserial

Número de observações

De acordo com o critério estatístico adotado, se o valor de calculado for superior ao valor de tabelado ao adotar nível de significância de 5 %, rejeita-se a hipótese nula que aponta o coeficiente como nulo. Deste modo, neste documento foram apresentados apenas as correlações que passaram por este teste.

### Modelo de aprendizado de máquina supervisionado

Foram avaliadas as performances dos seguintes modelos de aprendizado de máquina supervisionado para classificação:

• Regressão Logística com regularização do tipo l1 e l2 (elasticnet), com fator de regularização l1 de no valor de 0.8. O algoritmo de otimização utilizado para obtenção dos coeficientes foi o lbfgs (*Limited-memory Broyden–Fletcher–Goldfarb–Shanno*).

• Rede neural com 1 camada oculta com 5 neurônios com a função de ativação *ReLu* (unidade linear retificada) e com a função de ativação *sigmoid* na camada de saída. A rede foi treinada através de tamanhos discretos de amostras no valor de 5 observações por 200 épocas.

• *Random Forest:* Fez-se uso de 50 árvores de decisão, com profundidade média de 8 níveis. O critério escolhido para realização do split foi o da impureza de gini.

• SVM (*Suppor Vector Machine*): Fez-se uso do kernel de Função de Base Radial, a qual permite, através das ferramentas de produto escalar e da expansão da série de Taylor, obter uma relação entre as observações em uma dimensão infinita. As probabilidades geradas por este modelo foram calibradas através do método do dimensionamento de Platt, o qual aplica a função regressão logística sobre as probabilidades originais geradas, conforme detalhado em (NICULESCU-MIZIL; CARUANA, 2005).

• XBGClassifier

• Adaboost

### Métricas de avaliação do modelo

Para avaliar a performance dos modelos quanto à classificação correta dos pacientes quanto à presença de lombalgia crônica, utilizou-se as métricas Acurácia, *Precision*, *Recall* e *F1-Score*. Todas estas métricas podem ser calculadas a partir da matriz de confusão, a qual é apresentada na Tabela 2.

Tabela 2 - Exemplo de matriz de confusão

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Real | |
|  |  | 0 (não) | 1 (sim) |
| Predição | 0 (não) | TP | FN |
| 1 (sim) | FP | TP |

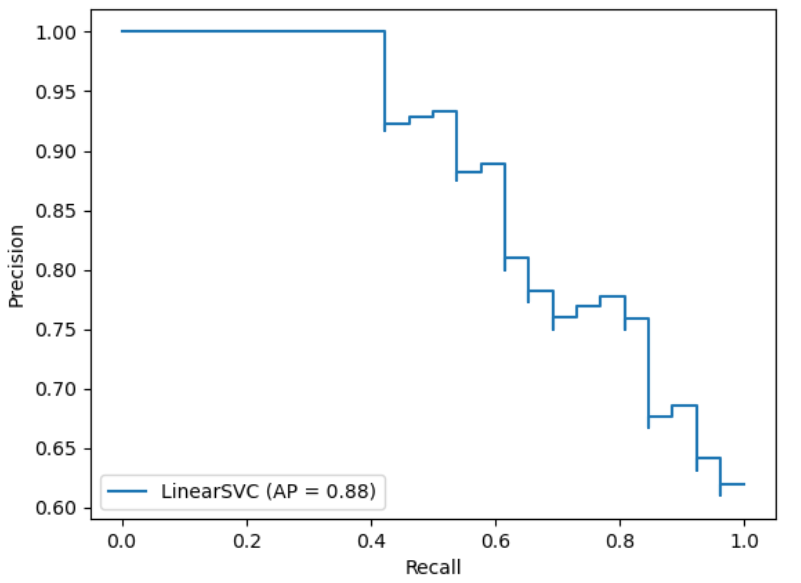
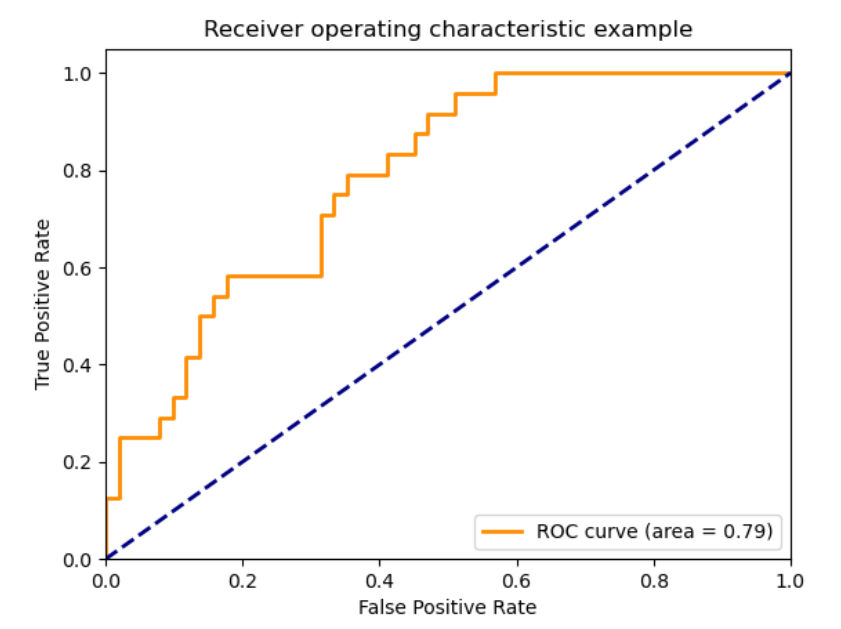
A partir do exposto na Tabela 2, define-se portanto as seguintes métricas:

A métrica *Precision* permite que obter a razão entre o número de positivos preditos corretamente em relação a todos positivos preditos. Entretanto, a métrica *Recall* é particularmente ais importante para o foco de análise deste trabalho*,* dado que o cálculo é realizado a partir da razão entre o valor total de Verdadeiros Positivos em relação a soma de Verdadeiros Positivos e Falsos Negativos. REFERÊNCIA

A razão da utilização da mesma se encontra no fato de que um modelo de aprendizado de máquina para previsão da ocorrência ou não de lombalgia crônica em pacientes deve apresentar um baixo valor de Falsos Negativos e consequentemente um valor alto de *Recall*. REFERÊNCIA

A métrica F1-Score permite obter uma média harmônica entre *Precision* e *Recall*. REFERÊNCIA

A avaliação do modelo também foi realizada a partir das curvas ROC (Figura X a), área sobre a curva ROC (Area Under Curve – AUC) e *Precision-Recal* (Figura X b)*.*



A Curva ROC permite observar como ocorre a distribuição entre a Razão de Verdadeiros Positivos (TPR) e a Razão de Falsos Positivos (FPR), para diferentes *thresholds*, o qual é definido entre 0 e 1.

O cálculo de ambos é definido por:

O mesmo princípio está presente na curva de *Precision* x *Recall*. A primeira

## 3 Resultados e discussão

## Estatísticas iniciais presentes no próprio resumo do Lefort

## Correlações

## FAZ SENTIDO deixar um trecho inteiro para dor neuropática?

# Principais correlações das variáveis do Questionário para Diagnóstico de Dor Neuropática 4 (DN4)

Na Figura 1 são apresentadas as principais correlações com as variáveis presentes no questionário para diagnóstico de dor neuropática 4

Figura 1 – Principais correlações com as variáveis presentes no Questionário para Diagnóstico de Dor Neuropática 4 (DN4)



Observa-se da Figura 1 que 79,19 % dos pacientes que assinalaram que apresentem o sintoma de adormecimento na região em que sentem dor, também marcaram que a dor apresenta a característica de queimação. Destaca-se também que 68,6 % das pessoas que disseram apresentar o sintoma de queimação na região em que sentem dor, marcaram a opção de sentirem dor nas costas.

Figura 2 – Principais correlações entre a questão do questionário DN4, sobre apresentar ou não a característica de alfinetada e agulhada na região da dor



A Figura 2 apresenta quais foram as questões de caráter ordinal que apresentam maior correlação com a questão de caráter binária se o paciente apresenta ou não o sintoma de alfinetada e agulhada na região de sua dor. Através da Figura 3 têm-se um exemplo do porquê a correlação no valor de 64,42 para a questão BPI – Relacionamento com outras pessoas.

Figura 3 – Gráfico de *boxplot* para visualização da distribuição dos dados da questão 2-b do questionário DN4 com a interferência da dor no relacionamento com outras pessoas



Na Figura 3 constata-se que os pacientes que assinalaram apresentarem o sintoma de alfinetada e agulhada na região da dor, também responderam valores maiores para a interferência desta dor para o relacionamento com outras pessoas, como é observado na maior densidade de pontos, e no *boxplot* de cor azul, correspondente à opção 1 da questão binária.

Ressalta-se que apenas esta questão (2 – b – Alfinetada e agulhada) do Questionário para diagnóstico de Dor Neuropática apresentou um número superior a 30 pessoas que responderam sim ou não a questão. As demais questões do questionário DN4 não apresentaram significância estatística no coeficiente de correlação ponto bisserial ou apresentaram dados desbalanceados entre sim ou não, ou seja, uma proporção muito maior de pessoas responderam a opção sim ou não em relação à outra alternativa.

# Principais correlações entre as variáveis binárias dor nas costas, lombalgia, dor nas pernas e estar ou não desempregado.

Figura 4 - Principais correlações das variáveis binárias com lombalgia, dor nas costas, dor nas pernas e estar ou não desempregado



Na Figura 4, é possível visualizar quais as questões mais correlacionadas com as variáveis lombalgia, dor nas pernas, dor nas costas e estar ou não desempregado. Nota-se que 65,57 % das pessoas que marcaram a opção de apresentar dor na região lombar (lombalgia) também marcaram que possuem dor nas pernas.

Importante mencionar também que 63,93 % dos pacientes que apresentam dor nas pernas também marcaram a opção no Questionário DN4 que a sua dor apresenta a característica de queimação (DN – queimação). Nota-se também a alta porcentagem de pacientes que fazem uso de analgésico, sendo 84,38 % para aqueles que possuem dor nas pernas e 90 % para quem possui dor nas costas.

Ressalta-se também que não houve nenhuma correlação maior que 60 e com nível de significância de 5% para a variável que considera se o paciente já realizou alguma cirurgia na coluna

A partir das correlações apresentadas, definiu-se o critério de utilizar os seguintes critérios para seleção de variáveis preditoras:

- De acordo com a tabela cruzada, aquelas variáveis que apresentam porcentagem maior que 60 % de correlação

- Da correlação entre a variável binária Lombalgia (se possui ou não), selecionar apenas as que apresentam correlação superior a 0,5 ou inferior a -0,5.

Deste modo, as variáveis selecionadas foram:

- Gotemburgo – Se faz ou não uso de Analgésicos

- Gotemburgo – Se possui dor nas costas

- Gotemburgo – Se possui dor nas pernas

- Gotemburgo – Qual sexo (Masculino ou feminino)

- BPI – Se possui dor na região 29 da questão 2 do questionário de Inventário Breve de Dor

- Gotemburgo – Há quanto tempo está com dor nas pernas

- BPI – Intensidade da dor no momento

- Short Form Health Survey – Realizou menos do que queria nas atividades diárias devido ao problema de dor

## Análise de Componentes Principais

A Figura 5 traz a Análise de Componentes Principais aplicada sobre as 9 variáveis de modo que podemos observar que aproximadamente 59 %, 70 % e 89 % da variância dos dados pode ser explicada por 2, 3 e 5 componentes principais, respectivamente.

Figura 5 -Variância explicada acumulada por componente principal



|  |  |
| --- | --- |
| Componente Principal | Variável mais importante para o componente |
| PC1 | Intensidade da dor nas pernas |
| PC2 | Se possui ou não dor nas pernas |
| PC3 | Qual sexo (masculino ou feminino) |
| PC4 | Se possui dor na região 29 do questionário BPI |
| PC5 | Se possui dor na região 29 do questionário BPI |
| PC6 | Se possui ou não dor nas costas |
| PC7 | Realizou menos do que queria nas atividades diárias devido ao problema de dor |
| PC8 | Se possui ou não dor nas pernas |

Na Figura 6 pode-se observar como ocorre a distribuição dos pacientes com e sem lombalgia crônica a partir da representação em 2 eixos, com a utilização dos 2 primeiros componentes principais, os quais explicam até 59 % da variabilidade dos dados.

Figura 6 – Divisão dos pacientes a partir da característica Lombalgia



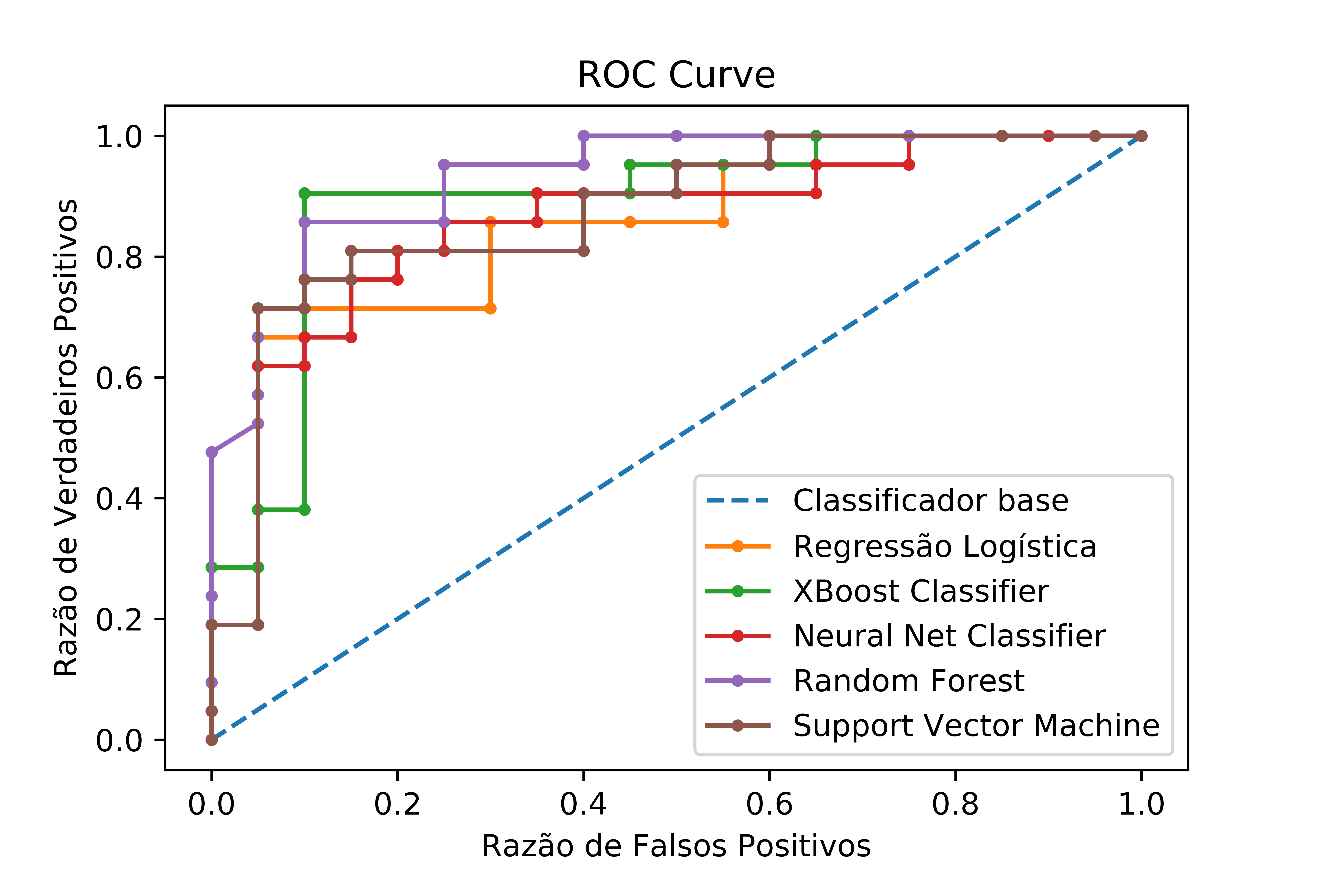
Observa-se da Figura 6 que há um grupo de pontos na cor vermelha (não apresenta lombalgia) na região esquerda superior e outro pequeno grupo de pontos azuis (apresenta lombalgia) na região direita superior. Regiões como estas em que não ocorre sobreposição dos pontos são amostras que podem ser previstas com maior facilidade por um modelo de aprendizagem estatística.

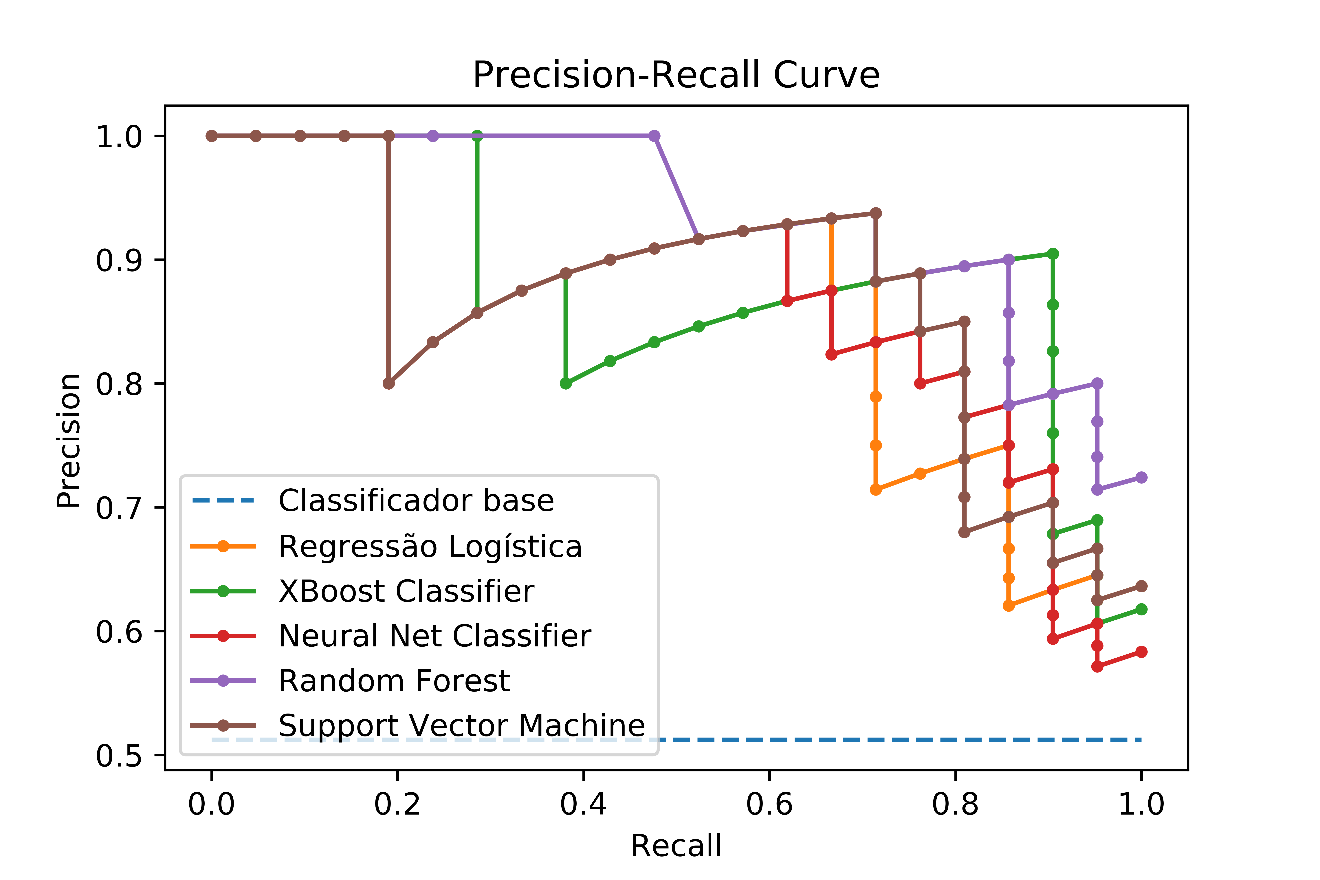
Entretanto, pode-se observar regiões em que há sobreposição entre os pontos e, portanto, o algoritmo de classificação apresentará maiores dificuldades para encontrar um padrão que explique a relação entre as variáveis. Ressalta-se, contudo, que a representação da Figura 6 é apenas uma representação em 2 eixos.

## Modelo de aprendizado de máquina

Resultados de cada um deles em uma tabela?

*Curvas ROC e Precision-Recall para todos de acordo com diferentes thresholds*





|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Acurácia | *Precision* | *Recall* | *F1-Score* |
| Regressão Logística | 0.780256 | 0.815357 | 0.72619 | 0.732547 |
| Rede Neural | 0.707317 | 0.827153 | 0.809699 | 0.818333 |
| Random Forest | 0.778755 | 0.767641 | 0.740476 | 0.735991 |
| Support Vector Machine | 0.785348 | 0.875595 | 0.690476 | 0.721166 |
| XBoost Classifier | 0.808635 | 0.808065 | 0.804701 | 0.789293 |

## Avaliou-se qual o melhor Threshold na curva ROC, e o melhor foi para o XBoost Classifier, o threshold de

## 4 Conclusões

## 5 Agradecimentos

## 6 Referências

# ANEXOS (utilize caso necessário)

APÊNDICES (utilize caso necessário)

# ANEXO 1 – Unidades

# Anexo 2 - Abreviaturas