UNIVERSIDADE DE SÃO
PAULO
FACULDADE DE FILOSOFIA, CIÊNCIAS E LETRAS DE RIBEIRÃO
featurePRETO PROGRAMA DE PÓS-GRADUAÇÃO EM FÍSICA
APLICADA À MEDICINA EBIOLOGIA

KAREN GONÇALVES TOZZI

# Predição da curva de calibração em dosimetria gel utilizando aprendizado de máquina

RibeirãoPreto – SP
2023

KAREN GONÇALVES TOZZI

# Predição da curva de calibração em dosimetria gel utilizando aprendizado de máquina

Dissertação apresentada à Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto da Universidade de São Paulo, para obtenção do título de Mestre em Ciências.

Área de Concentração: Física Aplicada à Medicina e Biologia

Orientador: Prof(a).Dr(a). Juliana Fernandes Pavoni

Ribeirão Preto -SP

2023

Catalogação na publicação
Departamento de Física
Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto da
Universidade de São Paulo

Tozzi, Karen Gonçalves

Predição da curva de calibração em dosimetria gel utilizando aprendizado de máquina. Ribeirão Preto, 2023.

106 p. : il. ; 30 cm

Dissertação de Mestrado, apresentada à Faculdade de Medicina de Ribeirão Preto/USP. Área de concentração: Física Aplicada à Medicina e Biologia.

Orientador: Pavoni, Juliana Fernandes.

1. Dosimetria gel. 2. Aprendizado de Maquina 3.Aprendizado Supervisionado 4. Radioterapia

# DEDICATÓRIA

*Para meus pais e minha irmã, sem vocês nada disso seria possível. Cada conquista tem vocês como base no incentivo, no apoio e no amor.*

# AGRADECIMENTOS

À minha parceira de laboratório Jéssica, que se tornou uma amiga, por todo conheciemento compartilhado, discussões, paciência nas inúmeras vídeochamadas, pelo companherismo nessa jornada que atravessou uma pandemia.

Aos colegas de laboratório que chegaram depois, mas não menos importantes. À Júlia pela troca experiencias, viagens em congressos e momentos que fortaleceram nossa amizade.

Aos meus amigos do "De volta aos bares" por toda descontração, parceria e convivência, o que torna a vida mais leve ao lado de vocês.

Aos meus pais, irmã, familiares e amigos de vida que sempre torceram pelas minhas conquistas e sempre estiveram presentes de alguma forma.

Ao Matheus, meu companheiro de vida, por todo suporte, apoio e incentivo, obrigada por toda compreensão, por sempre estar ao meu lado e por acreditar em mim.

À minha querida orientadora, Juliana Pavoni, por todos os ensinamentos, paciência, suporte em toda a caminhada até aqui.

Ao Departamento de Física e a todos os funcionários que contribui de alguma forma, para esclarescimentos de dúvidas, suporte informativo, organização de eventos, os responsáveis por semrpe nos proporcionar um ambiente limpo e aconchegante.

À CAPES e à FAPESP pelo apoio financeiro.

A todos que contribuiram, direta ou indiretamente, para a conclusão deste trabalho.

*“O que vale na vida não é o ponto de partida e sim a caminhada. Caminhando e semeando, no fim, terá o que colher”*

*(Cora Coralina)*

## RESUMO

TOZZI, Karen Gonçalves. **Predição da curva de calibração em dosimetria gel utilizando aprendizado de máquina**. 2023. 103p. Dissertação (Mestrado) – Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, 2023.

A dosimetria gel é uma técnica que permite a medida de dose em três dimensões e, portanto, tem grande potencial de aplicação na radioterapia moderna. A calibração desses dosímetros é um procedimento essencial para sua utilização e demanda um tempo significativo seja durante a irradiação dos lotes de géis ou durante sua leitura com alguma técnica de imagem que, na maioria das vezes é feita através de imagens de ressonância magnética (IRM). O aprendizado de máquina (AM) é uma técnica que vem ganhando espaço em todas as áreas do conhecimento, de maneira a ajudar a otimizar a solução de tarefas. Este trabalho tem como objetivo usar modelos de algoritmos de AM para desenvolver uma metodologia capaz de predizer a curva de calibração (coeficientes angulares e lineares) de um lote de dosímetro químico com base em extração de características radiômicas com e sem a aplicação de filtros *wavelets*, das imagens dos tubos de géis não irradiados. Dois modelos de regressão foram inicialmente propostos, *RandonForest* (RF) e *Categorical Boosting* (CB*)* combinados com três técnicas de seleção de características que mais influenciam na predição dos coeficientes: *Mean Decrease Impurity* (MDI), *Recursive Feature Elimination* (RFE) e *PowerShap* (PS). As IRM que compõe o conjunto de dados foram separadas em dois conjuntos de dados: *dataset* 1 utilizado para desenvolver os modelos e o *dataset* 2, formado com IRM de uma máquina diferente da utilizada para adquirir as amostras presentes no *dataset* 1. Este segundo *dataset* foi designado para estudar a aplicabilidade do modelo desenvolvido através do primeiro conjunto de dados. Para avaliar os modelos desenvolvidos, foram utilizados três métricas: Erro quadrático médio (Mean Squared Error - MSE), Erro absoluto médio (Mean Absolute Error - MAE) e Raiz quadrada do erro quadrático médio *(Root Mean Squared Error* - RMSE) , sendo que os modelos apresentaram melhor performance quando desenvolvidos com base no MSE. Assim, a combinação de técnicas que apresentou melhor acurácia para as predições utilizou o modelo de regressão RF selecionando as melhores características com a biblioteca PS para os dois coeficientes. O valor de MSE de 6,67 x $10^{-3}$ englobando 77% das predições dentro de um desvio de ±5% para o coeficiente angular, e 0,073 com 80% das predições dentro do mesmo desvio para o coeficiente linear. Para o *dataset* 2 os valores de MSE chegaram a 2,84 x $10^{-2}$ para o

coeficiente angular, diminuindo o desvio para ±2% em 94% das predições e 0,15 para o coeficiente linear, mantendo o desvio de incerteza de ±5% para 74% das predições. Também foram desenvolvidos três modelos de classificação para identificar as diferenças que as amostras de géis apresentam entre si, utilizando o modelo de classificação RF e selecionando as melhores características com o método PS. O modelo 1 com o objetivo de predizer o tipo de agente oxidante das amostras apresentou uma acurácia de 95% enquanto o modelo 2, desenvolvido para predizer o *bloom* das gelatinas utilizada na fabricação do gel, obteve 78% de acurácia. Já o terceiro modelo obteve uma acurácia de 80% para classificar se o lote de gel sofreu alguma alteração de luminosidade durante o processo de produção desses dosímetros.
Aplicando esses modelos no *dataset* 2, os três foram capazes de classificar todas as amostras às classes que elas pertencem.
O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001 e da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), processo nº 2021/02254-6.

Palavras-chave: dosimetria gel, aprendizado de máquina, dosimetria.

## ABSTRACT

TOZZI, Karen Gonçalves. **Calibration curve prediction in gel dosimetry using machine learning.** 2023. 103p. Dissertation (Master) – Faculty of Philosophy, Sciences and Literature of Ribeirão Preto, University of Sao Paulo, Ribeirão Preto, 2023.

Gel dosimetry is a technique that allows dose measurement in three dimensions and, therefore, has great potential for application in modern radiotherapy. The calibration of these dosimeters is a crucial procedure and demands considerable time in routine, either in the process of samples irradiation or during image acquisition, using most of the time, Magnetic Resonance Imaging (MRI). Machine Learning (ML) has been used in all knowledge fields, optimizing tasks and procedures. This study aims to use ML models to develop a methodology to predict the calibration curve (Coeficiente angular and Coeficiente linear) of gels batches based on radiomics features extraction of non-irradiated images, with and without wavelets filters. Two regression models were initially proposed: Random Forest (RF) and Categorical Boosting (CB) combined with three methods to feature selection: Mean Decrease Impurity (MDI), Recursive Feature Elimination (RFE), and PowerShap (PS). The data were composed of MRI, and two datasets were defined: dataset 1, used to build the models, and dataset 2, containing MRIs of a different machine than dataset 1, used to evaluate the model developed applicability. Three metrics were used to evaluate the models: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). The model showed better accuracy prediction when developed using MSE. The combination using model RF with the PS library showed better results for both models. The MSE value of 6,67 x $10^{-3}$ with 77% of its predictions within a deviation of ±5% for Coeficiente angular and 0.073 with 80% of its prediction within the same deviation for Coeficiente linear. For dataset 2, the values achieved were 2,84 x $10^{-2}$ for Coeficiente angular, reducing the deviation to ±2% for 94% of its prediction and 0.15 for Coeficiente linear maintaining de ±5% deviation for 74% of the predictions. Also, three classification models were developed to identify the differences between gel samples. RF classifier was the algorithm chosen, and for feature selection, PS was used. Model 1 aimed to

predict the oxidant agent type of samples and presented an accuracy of 95%. In comparison, model 2 was built to predict the gelatin bloom used in gel fabrication and got 78% accuracy. Model 3 obtained 80% accuracy in classifying if a gel batch suffered any luminous change during production.

Applying these models to dataset 2, the three could correctly predict all samples of this dataset to the classes they belong to.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001 and by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), process nº 2021/02254-6

Keywords: gel dosimetry, machine learning, dosimetry.

## LISTA DE FIGURAS

## LISTA DE TABELAS

**SUMÁRIO**

# Capítulo 1 – Introdução

## *1.1 Considerações iniciais*

A dosimetria é uma maneira de quantificar doses de radiação em um indivíduo ou em um meio. Para medir essa grandeza, existem vários instrumentos capazes de fornecer valores de dose pontual com alta precisão, exatidão, reprodutibilidade e/ou de maneira eficaz como a câmara de ionização, os detectores de cintilação, os dosímetros termoluminescentes, entre outros. Os filmes radiográficos ou radiocrômicos podem ser usados para medidas em duas dimensões, enquanto os dosímetros géis são os mais eficientes em medidas tridimensionais da distribuição da dose.

A dosimetria gel é uma técnica de dosimetria baseada em um dosímetro químico que sofre alteração em sua estrutura ou em suas propriedades físicas quando irradiado, sendo as mais comuns a polimerização ou a mudança se cor. A distribuição de dose medida em três dimensões pode ser obtida a partir da aquisição de imagens volumétricas do gel, sendo as mais comuns as imagens de ressonância magnética (IRM) para quantificação da polimerização sofrida pelo dosímetro ou a tomografia ótica para avaliação da alteração de cor. Trata-se de uma técnica ainda muito utilizada em pesquisa, devido a algumas limitações práticas para sua implementação, como os conhecimentos de química e das técnica de imagens necessárias(Baldock et al., 2010). O gel é tecido equivalente ao corpo humano, sendo possível assim, observar fenômenos e efeitos nesses géis de interesse no estudo em questão, buscando sempre uma evolução para a clínica e para os pacientes.

Por se tratar de um composto químico, esses lotes de géis fabricados estão propensos a vários fatores que podem influenciar sua sensibilidade e, logo, sua performance em uma medida. As principais influências para resposta dos dosímetros géis são variações na temperatura de preparo do dosímetro na adição dos químicos, de manuseio e de aquisição das imagens(Vandecasteele & De Deene, 2013). Por isso, há uma necessidade de calibração de cada lote de dosímetro preparado, para se considerar pequenas variações na sua sensibilidade.

O processo de calibração dos dosímetros químicos é feito através de uma irradiação de diferentes amostras do dosímetro do mesmo lote e armazenadas em tubos de calibração, com doses variadas (comumente utilizando doses entre 0 e 10 Gy) e, posteriormente adquirindo as imagens desse gel que identificam as alterações sofridas pelo dosímetro. Com essas imagens, é possível quantificar as alterações sofridas em função dos valores conhecidos de cada dose, construindo assim a curva de calibração característica desse lote através do ajuste da equação

de reta que melhor representa os dados. Costuma-se avaliar a sensibilidade do lote de dosímetro através da análise do coeficiente angular obtido. Esse processo todo demanda um tempo considerável para ser realizado, de forma que alguma ferramenta que otimize esta etapa pode ser de interesse clínico e facilitar o uso desse dosímetro no ambiente clínico(Gallo & Locarno, 2023).

O aprendizado de máquina (AM) é um ramo da inteligência artificial (IA) que vem ganhando espaço em todas as áreas do conhecimento, e hoje em dia é comum ver sua aplicação nos mais diferentes campos de atuação, assim como em análises de imagens médicas. Ensinar a máquina a realizar tarefas humanas é um desafio que vem com uma recompensa imensurável, uma vez que tarefas repetitivas e que seguem um padrão podem ser facilmente substituídas por algoritmos que, supervisionados por um profissional, otimiza a rotina e garante inovação de qualidade para o serviço.

Várias técnicas de extração de características de imagens são utilizadas para poder desenvolver modelos de algoritmos capazes de obter informações dos pixels e assim, trabalhar com elas para diferentes finalidades propostas, como é o caso do *Radiomics*(van Timmeren et al., 2020).

Através do *Radiomics* é possível extrair características de imagens radiológicas bidimensionais e tridimensionais usando algoritmos de caracterização de dados que encontrem padrões característicos dessas imagens, melhorando diagnósticos e prognósticos na medicina oncológica e sendo possível desenvolver métodos automatizados que melhorem a precisão de procedimentos realizados na clínica (Yan & Zhang, 2015), como o caso do trabalho aqui apresentado.

Assim, modelos podem ser desenvolvidos sendo essas características utilizadas como dados de entrada para o aprendizado de uma dada tarefa e retornar para o usuário a melhor resposta encontrada, baseada em vários treinamentos, otimizações, busca de melhor performance, correções e avaliações.

### *1.2 Objetivo*

O objetivo principal é estabelecer um protocolo baseado em aprendizado de máquina para a previsão dos coeficientes angulares e lineares associados aos lotes de dosímetros gel. Esse protocolo possibilitará a avaliação da sensibilidade desses dosímetros com base na análise de características radiômicas extraídas das imagens de ressonância magnética (IRM) dos géis não irradiados. Para tanto, serão implementados  dois modelos de regressão baseados no

*Random Forest Regressor* e no *Categorical Boosting*. Esses modelos serão combinados com três métodos de seleção de características: *Mean Decrease Impurity, Recursive Feature Elimination* e *PowerShap*. Além disso, utilizaremos três métricas de avaliação: Erro Quadrático Médio, Erro Absoluto Médio e Raiz Quadrada do Erro Quadrático Médio. A melhor combinação desses parâmetros será selecionada no final do estudo.

Adicionalmente, exploraremos a aplicação de três modelos de classificação com o propósito de identificar os componentes presentes na formulação dos dosímetros gel. Esses modelos irão classificar o agente oxidante (modelo 1), o grau de gelificação da gelatina (modelo 2) e determinar se houve interferência de luz durante o processo de fabricação do gel (modelo 3). Esses parâmetros podem impactar significativamente a resposta dos dosímetros químicos e, portanto, são de importância crítica para a análise.

# Capítulo 2 – Dosimetria Gel

Com o passar do tempo, os tratamentos radioterápicos evoluíram e foram aprimorados, principalmente devido aos avanços tecnológicos e ao desenvolvimento de software e hardware capazes de criar planos de tratamento mais eficazes e otimizados, especialmente em três dimensões. Uma das principais causas da evolução na capacidade de entrega de dose dos planejamentos foi a possibilidade de modular o feixe de radiação, aumentando a conformação do tratamento nos volumes alvo e diminuindo dose de radiação em tecidos sadios, dando origem a técnica de radioterapia com intensidade modulada (IMRT)(Schreiner, 2004,Cho, 2018). Assim como os tratamentos foram evoluindo, também é preciso garantir que o controle de qualidade de todo o sistema de tratamento até à sua entrega ao paciente seja feito rigorosamente(Kalaiselven D, 2012). Dessa forma, as técnicas de dosimetria também evoluíram passando da dosimetria pontual ou bidimensional usadas nos tratamentos convencionais, até a possibilidade da dosimetria tridimensional em tratamentos com IMRT.

Como já foi dito, uma forma de obter a distribuição de dose 3D é a utilização da dosimetria gel. Esses dosímetros possuem equivalência com o tecido, sendo possível registrar doses depositadas sem a dependência da direção de incidência do feixe, com alta resolução espacial em três dimensões (Baldock et al., 1998). A alta resolução espacial das medidas de dose permitem a detecção de pequenas variações no gradiente de dose característicos dos tratamentos de IMRT(Cho, 2018). Outra vantagem dessa técnica dosimétrica é a capacidade de confeccionar fantomas de diversos tamanhos e formatos, uma vez que o dosímetro é um composto gelatinoso capaz de se adequar ao formato do recipiente em que é armazenado.

Existem principalmente duas classes de dosímetros gel, os géis poliméricos e os géis radiocrômicos. O primeiro deles, como o próprio nome diz, responde à radiação sofrendo polimerização. Os polímeros produzidos pela radiação alteram a mobilidade das moléculas de água ao seu redor e podem ser quantificadas através da relaxometria em IRM. Já o segundo deles, apresentam uma mudança de cor após a irradiação e nesse caso, técnicas de imagem óticas podem ser empregadas para quantificação dos coeficientes de atenuação ótica, que se relacionam com a dose.

Neste estudo, utilizamos géis poliméricos, portanto, a seguir, detalharemos o seu mecanismo de resposta e leitura.

### *2.1 Géis poliméricos*

Esses dosímetros são compostos por monômeros em uma matriz aquosa de gel e respondem à irradiação com radiação ionizante pelo fenômeno de polimerização, sendo a dose absorvida no dosímetro proporcional ao grau de polimerização. Essa matriz gelatinosa (M) é composta por cerca 90% de água e quando é irradiada, são liberados íons e radicais reativos pelo processo de radiólise da água (equação 1). Este processo ocorre com uma taxa de dissociação proporcional à dose absorvida (*kd*) e os principais radicais produzidos ($R^{\bullet}$) são elétrons aquosos, hidroxila ($OH^-$) e o íon hidrônio ($H_3O^+$) que são os responsáveis por desencadear a polimerização ao reagirem com os monômeros suspensos na matriz do gel.

$$H_2O \xrightarrow{kd} R^{\bullet} \quad (1)$$

$$R^{\bullet} + M \xrightarrow{kl(n)} R\,M^{\bullet}_{n} \quad (2)$$

$$R\,M^{\bullet}_{n} + M_m \xrightarrow{kp(n,m)} RM^{\bullet}_{n+m} \quad (3)$$

O processo de polimerização se inicia após a formação do monômero com uma taxa de reação (*kl*) como mostra a equação 2, ocorrendo a ligação entre o radical e um elétron da dupla ligação presente no monômero. O valor de n inicia igual a um, pois no começo, não há polimerização, e esse valor vai aumentando à medida que o processo de polimerização vai acontecendo. A taxa de reação (*kp*) depende do número de monômeros na cadeia polimérica, assim como o kl, e essa reação ocorre pela ligação do monômero radical a outros monômeros (m=1) ou polímeros (m>1), como explica a equação 3.

O processo de polimerização cessa quando a combinação dos radicais gera um polímero estável. A presença de oxigênio na matriz do gel também pode acarretar o fim da reação de polimerização, uma vez que radicais peróxidos reagem com os outros radicais ali presentes, omitindo o crescimento das cadeias poliméricas e inibindo a polimerização. Para evitar que isso aconteça, uma solução utilizada na dosimetria gel polimérica é a confecção de géis em atmosferas inertes sem oxigênio, ou confecciona-se o dosímetro em atmosfera normal e utiliza-se um agente antioxidante na formulação do gel, que é responsável por capturar as moléculas de oxigênio presente no gel.

Existem diversas formulações de dosímetros químicos que foram desenvolvidos ao longo dos anos, podendo ser destacado dois tipos principais: géis baseados em acrilamida PAG (*Polymer Acrylamide Gelatine*) (Gore & Kang, 1984) e nPAG (*Normoxic Polymer Acrylamide*

*Gelatine*) (Olsson et al., 1992) e géis baseados em ácido metacrílico MAGIC (*Methacrylic and Ascorbic Acid in Gelatin Iniciated by Copper*).

O gel MAGIC inovou a dosimetria tridimensional pois foi o primeiro gel fabricado com adição do agente antioxidante (J. P. Fernandes et al., 2008). Neste estudo usamos o dosímetro MAGIC-f, derivado do gel MAGIC.

### *2.2 Gel MAGIC-f*

O gel *MAGIC-f* é uma versão aprimorada do gel MAGIC a partir da adição de formaldeído a sua composição, o que provoca o aumento do seu ponto de fusão (J. P. Fernandes et al., 2008) e facilita o seu manuseio em temperatura ambiente. Sua composição e concentração em massa de cada reagente é apresentada na tabela 1.

*Tabela 1 - Composição do gel MAGIC-f*

| Reagentes | Concentração em massa (%) |
|---|---|
| Água Mili-Q | 82,70 |
| Gelatina – 250bloom (Gelita) | 8,25 |
| Ácido metacrílico (Sigma-Aldrich) | 6,00 |
| Formaldeído (Sigma-Aldrich) | 3,00 |
| Ácido ascórbico (Sigma-Aldrich) | 0,03 |
| Sulfato de cobre | 0,02 |

A Gelatina juntamente com a água Mili-Q constitui a base da matriz onde estão dissolvidos os monômeros (ácido metacrílico). A dureza da gelatina é quantificada pelo valor do seu *bloom*, usualmente utiliza-se o valor de 250 *bloom*. Os agentes antioxidantes são o sulfato de cobre e o ácido ascórbico, que atuam pela formação de um complexo organometálico que capturam o oxigênio (Baldock et al., 2010b). Por fim, o formaldeído é responsável por aumentar o ponto de fusão do gel.

### *2.3 Imagens por Ressonância Magnética*

Uma IRM é adquirida quando prótons de nucleares são colocados em um campo magnético externo e seus momentos de dipolo magnético começam seu movimento de

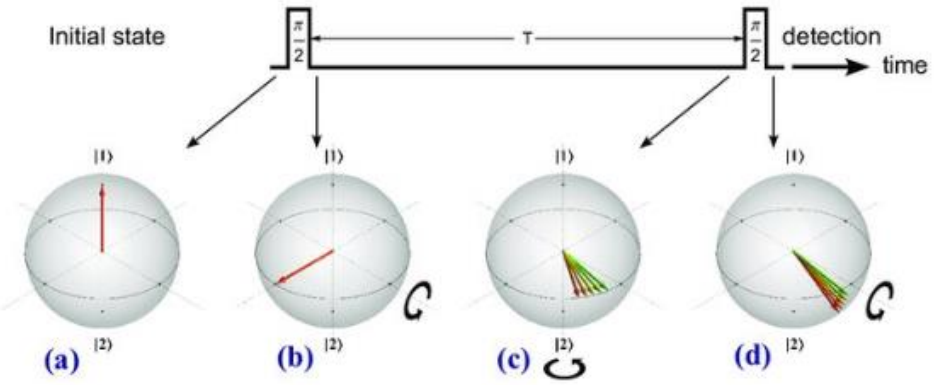
precessão (spin) ao redor desse campo e se alinham com ele, gerando um vetor de magnetização (Bradley, 1997). Gradientes de campo magnético em sequências são aplicados para localizar espacialmente os sinais a serem adquiridos. Assim, pulsos de excitação são aplicados e os núcleos absorvem energia, posteriormente ocorre o fenômeno de relaxação e os núcleos passam a induzir o sinal de RM nas bobinas receptoras, onde o sinal é adquirido e processado por meio de funções estatísticas, uma vez que a imagem é formada ponto a ponto em uma matriz.

Os tempos de relaxação T1 (longitudinal) e T2 (transversal) são os principais parâmetros alterados na IRM dos géis poliméricos quando há polimerização, alterando, consequentemente a taxa de spin-spin (R2 = 1/T2) e spin-rede (R1 = 1/T1). Dessa forma, a determinação dos mapas de relaxometria é a maneira pela qual conseguimos determinar os mapas de dose absorvida pelo dosímetro. A taxa de relaxação spin-spin ou R2 possui uma maior sensibilidade que a taxa de relaxação R1 para quantificação da dose e, portanto, é a usada em dosimetria gel polimérica (Maryanski et al., 1993) (Berg et al., 2001).

Os mapas de R2 podem ser calculados a partir de imagens ponderadas em T2 adquiridas com sequencias de *spin echo* (SE). Preferencialmente, utilizando sequencias *multi spin echo* (Multi SE) pelo fato de várias imagens serem adquiridas dentro do mesmo tempo de medição e assim apresentar um aumento da relação sinal-ruido (SNR) nas imagens de R2 calculadas (Baldock et al., 2010b). Quanto mais alta for a SNR, menor é o efeito do ruido de fundo na medição do sinal.

A figura 1 mostra o comportamento de magnetização da amostra a ser imageada com uma sequência SE. Inicialmente há apenas a magnetização longitudinal ($M_z$) presente. Em seguida, um do pulso de radiofrequência (RF) de 90° é aplicado e a magnetização é transferida para o eixo x, ocasionando em uma defasagem da magnetização transversal ($M_{xy}$). O pulso de 180° faz com que a magnetização gire ao longo do plano x, reorientando e realinhando os spins, produzindo assim, o sinal (eco). Esse eco acontece após um tempo chamado tempo de eco (TE), que é computado a partir da aplicação do pulso inicial de 90°. Para uma sequência Multi SE, a sequência de ecos é obtida por meio da aplicação do pulso de RF de 90° seguido de múltiplos pulsos de 180°, após cada pulso de 180° o sinal de eco é capturado, sendo TE o tempo entre os ecos adquiridos.

*Figura 1 - Descrição de um spin eco no formalismo da esfera de Bloch*




Fonte: Valentin Ivannikov (2017)

O tempo de repetição é outro parâmetro importante na aquisição das IRM e se refere a tempo decorrido para a repetição de sucessivas sequencias de pulsos aplicadas ao mesmo corte. O TR juntamente com o TE são parâmetros que definem o contraste nas IRM e podem ser manipulados de acordo o objetivo da aquisição.

O sinal de eco na sequência SE é modelado pela equação 4. Sendo o $S_0$ a função densidade de prótons e $S_{BG}$ o sinal de fundo.

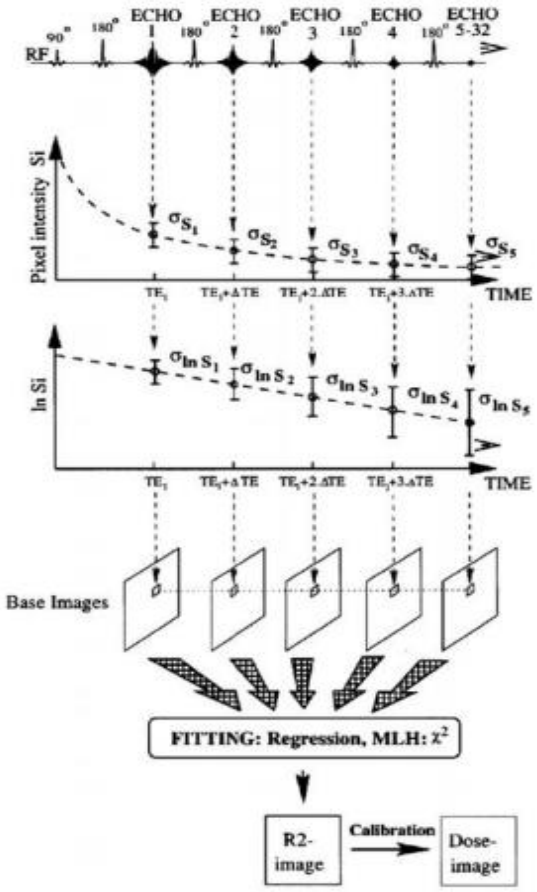
$$SE = S_0 \left( 1 - e^{-\frac{TR}{T1}} \right) . \; e^{-\frac{TE}{T2}} + S_{BG} \qquad (4)$$

Como, para medir T2 (e consequentemente R2) em imagens spin eco devemos utilizar um TR muito maior em comparação ao TE, a equação 4 pode ser simplificada no que é apresentado na equação 5:

$$SE = S_0 . e^{-\frac{TE}{T2}} \qquad (5)$$

A equação 5 possui duas incógnitas ($S_0$ e T2), por isso são necessárias duas ou mais imagens com diferentes TEs para que os valores de T2 ou R2 sejam determinados. Assim, o mapa de R2 é determinado através do ajuste da curva de decaimento exponencial correspondente a cada imagem adquirida, como é mostrado na Figura 2.

*Figura 2- Reconstrução do mapa de R2 e dose, através da imagem de ressonância magnética, utilizando sequência Multi SE*




Fonte: De Deene et al. Mathematical analysis and experimental investigation of noise in quantitative magnetic resonance imaging applied in polymer gel dosimetry. 1998, Signal Process 70:2:85–101

# Capítulo 3 – Aprendizado de Máquina

O AM é considerado uma área dentro da IA, com características multidisciplinares e grande predominância de probabilidade e estatística (Carbonell et al., 1983). Segundo Carbonell, é uma ferramenta utilizada para ensinar ao computador processos de tarefas humanas com base em experiência, melhorando sua performance com o aumento da experiência. O principal foco do AM é automatizar o conhecimento existente para que seja acessível e até melhorado, tirando viés de possíveis erros humanos. Quanto mais simples e discriminado o processor for, maior a probabilidade de um AM bem sucedido.

Os algoritmos de AM vieram ganhando destaque em diversas áreas do conhecimento, automatizando e otimizando atividades, inovando o meio de trabalho. A mineração de dados é um exemplo de como a AM lida bem com grande quantidade de dados que são analisados automaticamente em busca de algum objetivo pré definido que seja útil para cada caso em questão.

## *3.1 Tipos de aprendizados*

Segundo Bishop (Bishop, 2006), os três principais tipos de aprendizados, resumidamente são: supervisionado, não supervisionado e o aprendizado por reforço, como ilustra a figura 3.

O aprendizado supervisionado ocorre quando o modelo aprende a partir de resultados pré definidos. O modelo utiliza conjunto de dados já rotulados, ou seja, que já possuem uma resposta correta, para aprimorar seu desempenho e alcançar os resultados esperados (Bishop, 2006). Árvores de decisão e k-vizinhos são algumas das técnicas mais utilizadas desse tipo de aprendizado.

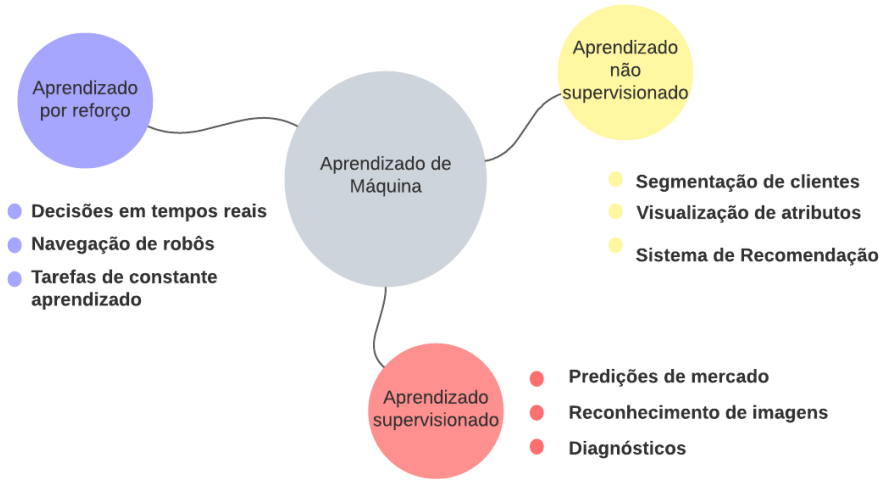
Quando se fala em aprendizado não supervisionado, o modelo trabalhará com dados não rotulados. Ele encontra critérios de funcionamento para identificar um padrão de aprendizado sem resultados pré definidos . É muito interessante para trabalhar com grande quantidade de dados, onde uma segmentação seja prioridade em relação a tarefas mais minuciosas (Bishop, 2006). Redução de dimensionalidade e clusterização são exemplos de técnicas abordadas no aprendizado não supervisionado

Por fim, o aprendizado por reforço. Nesse método, a máquina tenta aprender qual é a melhor decisão a ser tomada levando em consideração as circunstâncias na qual a decisão será

executada. Esse tipo de aprendizado não ocorre apenas uma vez, e sim N vezes até que ache a melhor resposta para o problema estudado, diminuindo erros a cada interação (Murphy, 2012)

Cada tipo de pesquisa dentro do AM terá aquele tipo de aprendizado que melhor se encaixe no seu conjunto de dados a ser analisado. Todos têm igual importância assim como suas indicações e contraindicações, que serão estudadas e definidas de acordo com o problema a ser solucionado.

*Figura 3 - Representação visual de algumas aplicações dos três tipos de aprendizado de máquina: supervisionado, não supervisionado e por reforço*




Fonte: O autor (2023)

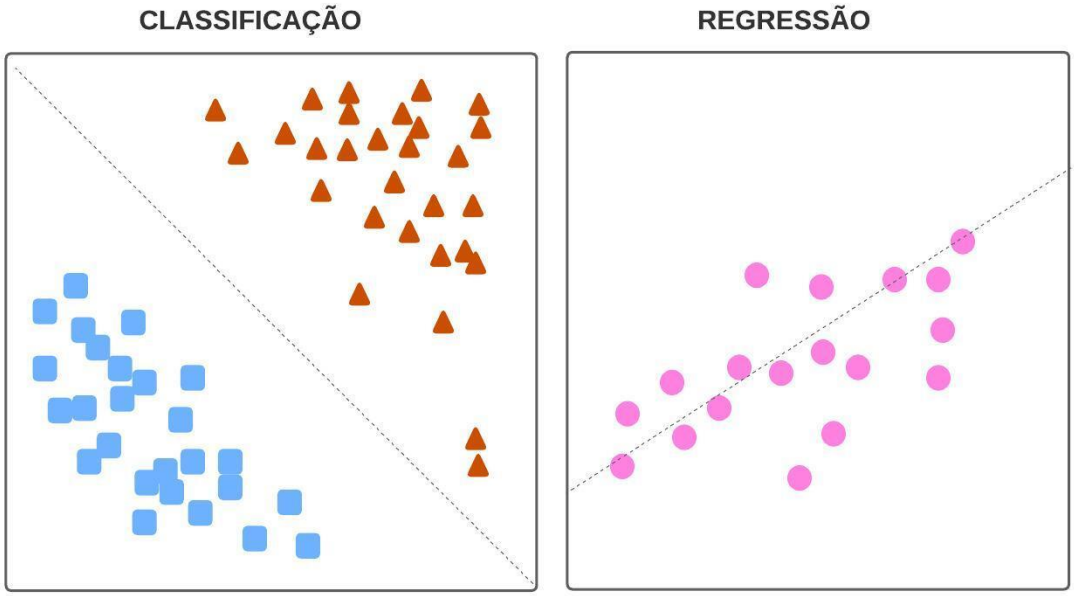
### *3.2 Tipos de problemas no aprendizado supervisionado*

Dentre todos os tipos de AM apresentados anteriormente, este tópico focará no aprendizado supervisionado, que é utilizada neste trabalho.

Os problemas a serem destrinchados dentro desta área se dividem em duas possíveis categorias (Figura 4). A primeira inclui os problemas que necessitam identificar duas ou mais classes, que são definidos como problemas de classificação. Neste caso, o modelo de AM desenvolvido procura encontrar padrões que gerem um classificador qualitativo de um conjunto de dados nunca vistos pelo algoritmo com bases em dados de entrada característicos das classes a serem preditas. Já, a segunda categoria, inclui problemas de regressão, que utilizam os dados

de entrada (X) para obter um uma resposta em valor numérico específico (Y), e não apenas uma distinção de classes. Essas implementações, calculam uma função estatística que melhor aproxime Y dos seus valores reais com base nas descrições do preditor X.

*Figura 4 – Representação gráfica dos problemas de Classificação (direita) e Regressão (esquerda). Na imagem à direita, amostras de duas classes diferentes são representadas pelos símbolos azuis e rosas e a linha pontilhada representa a separação das duas classes, onde o modelo aprende a classificar as amostras. Na figura à esquerda, amostras representando valores numéricos são retratadas pelo símbolo rosa e a linha pontilhada representa uma função gerada pelo aprendizado do modelo que melhor se adequa no conjunto de dados, para predizer os valores das amostras.*




Fonte: O autor(2023)

Neste trabalho foram utilizados modelos de classificação e também de regressão.

### *3.3 Qualidade dos dados*

Para um modelo aprender de forma eficaz, é necessário que os dados trabalhados sejam livres de viés que possam atrapalhar o aprendizado, como dados nulos no conjunto de dados, características que não agregam na construção do modelo (colunas com valores iguais ou próximos de zero por exemplo, irão ocupar espaço e tempo de processamento e não tem valor significativo para a construção do algoritmo), desbalanceamento na quantidade de valores entre as classes, *outliers* no conjunto de dado utilizado, entre outros fatores.

Quando algum fator acima citado não é devidamente tratado, pode ocorrer influência direta na predição do modelo, por isso é necessário fazer um pré-processamento e uma análise descritiva e exploratória para deixar no conjunto de dados apenas dados influentes na predição, sendo esta considerada a etapa mais importante na construção de um modelo. Excluindo pontos fora da curva e dados sem significância para a predição e trabalhando no balanceamento de classes, a qualidade do conjunto de dados que se inicia o processo já aumenta significantemente.

O desbalanceamento de classes pode também ocasionar *overffiting* no aprendizado do modelo, ou seja, se uma classe possui mais amostras que a outra, o modelo encontrará mais exemplos desta classe majoritária e pode ser enviesado a predizer um valor como pertencente a essa categoria sem necessariamente ser o correto. Então uma etapa muito importante na preparação de dados, é lidar com essa diferença numérica entre as classes(Haibo et al., 2013).

Há várias maneiras de resolver esse impasse de desbalanceamento, sendo aqui destacadas três delas, que são mais utilizadas em modelos de classificação: *undersampling*, *oversasmpling* e SMOTE ( *Synthetic Minority Oversampling Technique*) (Shelke et al., 2017).

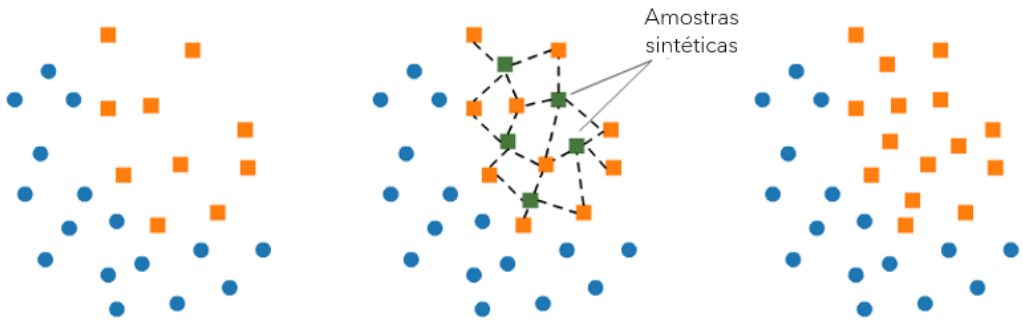
No *ovesampling*, o modelo cria cópias das amostras da classe minoritária para que o número das duas classes fique iguais. Já no *undersampling*, o modelo exclui amostras da classe majoritária para equilibrar o número de amostras entre as classes (figura 5). Ambos os métodos apresentam significativas desvantagens, pois ao excluir amostras de um conjunto de dados, também se perde informações que podem interferir na qualidade do modelo e, ao criar cópias de dados já existentes pode-se enviesar o modelo, acarretando uma predição equivocada (Shelke et al., 2017).

*Figura 5 - Técnicas de balanceamento de dados Oversampling e Undersampling*


Fonte: https://vivekrai1011.medium.com/undersampling-and-oversampling-imbalanced-data-bf7e9405fcad

A técnica SMOTE (Figura 6) possui a vantagem de, ao invés de criar cópias das amostras, ele cria amostras sintéticas com bases na combinação de amostras vizinhas, o que não é uma cópia idêntica, tendo menor probabilidade de o modelo sofrer *overffiting* (Dablain et al., 2022).

*Figura 6 - Técnica SMOTE de balanceamento de dados*




Fonte: traduzido de https://medium.com/@asheshdas.ds/oversampling-to-remove-class-imbalance-using-smote-94d5648e7d35

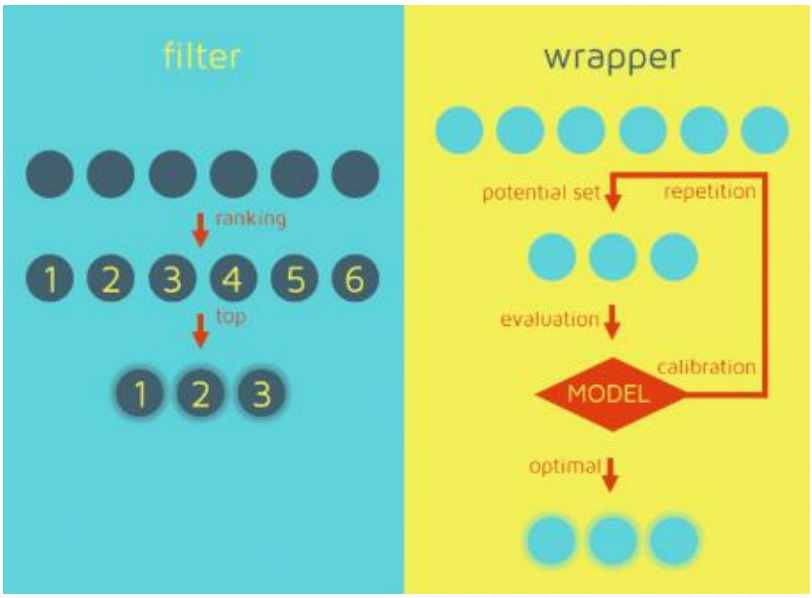
Já na regressão, onde se trabalha com funções estatísticas, uma técnica para melhorar o aprendizado do modelo é estudar a distribuição de dados, removendo valores atípicos, que podem atrapalhar este processo, e realizando uma análise exploratória na etapa de pré processamento(Tan et al., 2007).

Outra abordagem para melhorar a qualidade dos dados trabalhados é selecionar os dados de entradas que mais tem relação com a resposta de saída do modelo. Para um modelo de AM conseguir aprender e dar uma resposta a um problema, é necessário fornecer dados de entrada que caracterizam à predição final. Muitas vezes os dados que entram para o ensinamento do modelo são grandes em quantidades, mas não fornecem uma influência significativa para o ele aprender, então selecionar apenas aquelas características que têm maior peso na predição, diminui o tempo de processamento e otimiza o processo de aprendizagem do algoritmo(H. Liu & Motoda, 2013)

As técnicas de seleção de características mais utilizadas podem ser destacadas em dois tipos: métodos de filtros e métodos Wrapper (Figura 7). Na maioria das vezes, os métodos Wrapper resultam em melhores performances, mas podem precisar de grande poder

computacional e levar um tempo significante para entregar um resultado, principalmente quando se trata de dados com grandes dimensões(H. Liu & Motoda, 2013). Já os métodos de filtros são mais rápidos, mas podem sofrer de desvantagens como a necessidade de estabelecer intervalos de treshold para agrupar as características de acordo com seus valores de importâncias para a predição. A maioria dos métodos de filtro não levam em consideração as correlações das características e ignoram a interação delas com o modelo.

*Figura 7 - Métodos de seleção de características. À direita, representação do método filtro e a figura à esquerda ilustra o método Wrapper. Amos métodos discutidos neste capítulo.*




Fonte: https://quantdare.com/what-is-the-difference-between-feature-extraction-and-feature-selection/

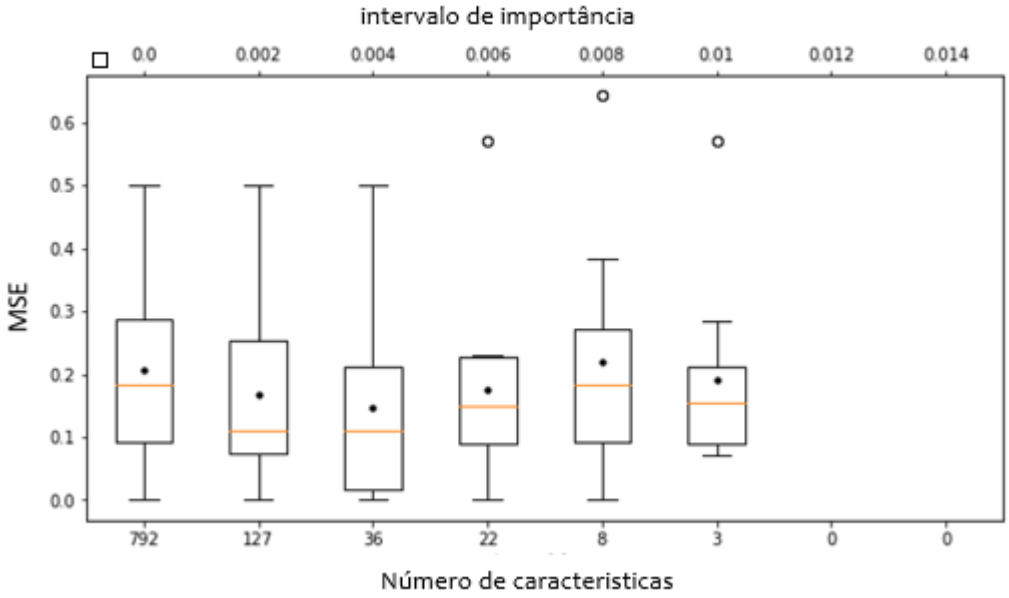
Método Wrapper avalia diversos conjuntos de características utilizando procedimentos que adicionam ou removem caraterísticas com a finalidade de encontrar uma combinação ótima que maximize a performance do modelo (Figura 7) (H. Liu & Motoda, 2013).

Neste trabalho foram usadas três maneiras diferentes de fazer a seleção das melhores características para o modelo: Mean Decreased Impurity (MDI), PowerShap (PS) e Recursive Feature Selection (RFE), todas aplicadas para os dois modelos, RF e CB.

O primeiro método utilizado, MDI, é um método de filtro de selecionar as características e opera com impuridade Gini. Um intervalo é considerado importante se nele contém um decaimento de impuridade considerável (Wehenkel et al., 2018). As regras de divisão de uma RF maximizam a redução de impurezas. Essa impuridade está relacionada ao quanto a divisão de nós de uma árvore de decisão foi bem feita a ponto de que não prejudique suas predições.

Essa redução na impuridade é a diferença entre a impureza de um nó e a soma ponderada das impurezas dos nós existentes na mesma divisão a que pertencem (Rigatti, 2017). Cada feature recebe um valor de importância e é ranqueada, desde a que não possui nenhuma influência na predição até aquela que possui maior valor de impacto. Diferentes grupos de características são selecionados aplicando intervalos de *treshold*, de mesmo tamanho, pré-estabelecidos pelo usuário. O grupo de característica escolhido, será aquele que o modelo utilizar e devolver o maior score alcançado, aqui, no caso, o menor valor de Mean Squared Error (MSE), já que quanto mais próximo de zero, melhor avaliado é o modelo por essa métrica, como é ilustrado na figura 8. Observa-se que este método, como o próprio nome diz, filtra as características a serem selecionadas de acordo com o que é pedido, em uma forma de ranqueamento. É uma abordagem mais interativa com o usuário, ao estabelecer os treshold e selecionar os grupos de características, tem um viés humano atrelado a ele.

*Figura 8 - Método de seleção de características MDI. Boxplot plotados levando em consideração grupos de características pertencentes a intervalos de importância pré estabelecidos (eixo x) e que alcançam determinados valores de uma métrica (mse – eixo y)*

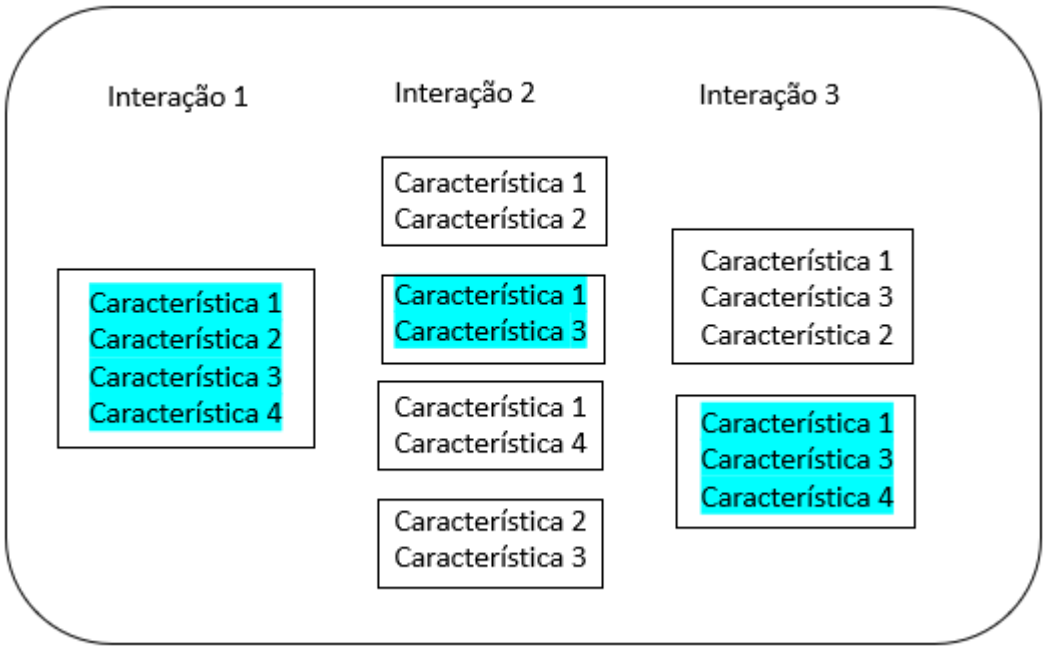



Fonte: O autor (2023)

O RFE, que se encaixa nos métodos Wrapper, trabalha selecionando as características considerando recursivamente conjuntos pequenos de características. O estimador é treinado no conjunto original de características (todas as características) e a significância de cada ponto é

atribuída através de cada atributo específico ou previamente denominado e então o próximo conjunto de características é comparado com o anterior (Figura 9). Esse processo é repetido recursivamente até que o número de características requerido seja alcançado(H. Liu & Motoda, 2013). Nesse modelo, é possível escolher o número de características a serem selecionadas, sendo o modo default metade das características existentes no dataset.

*Figura 9 - Método RFE de seleção de características. A melhor combinação das características escolhida na próxima interação e é representada pelo quadrado azul*




Fonte: O autor (2023)

Ao contrário dos dois métodos apresentados, que possuem um esquema de interação com o usuário, o PS possui um mecanismo mais direto de devolver um conjunto de características consideradas mais eficiente para o aprendizado do modelo.

Ele seleciona as características através de hipóteses de testes estatísticos e poder de cálculo nos valores de *shapely*, permitindo uma seleção de características rápida e intuitiva baseada nos métodos wrapped (Verhaeghe et al., 2023). Valores de *shapely* calculam a importância de um recurso comparando o que o modelo prevê com e sem o recurso (Marden & Shamma, 2018).

PS traz a ideia de que uma característica aleatória conhecia deveria ter, em média, impacto menor na predição do que uma característica informativa. Uma característica é

considerada informativa quando sua importância no dataset original é superior à sua importância no dataset após aplicar a função *shuffle*, ou seja, reordenar seus registros aleatoriamente (Marden & Shamma, 2018). Para realizar a seleção das características, o algoritmo do PS consiste em dois componentes: o componente *Explain* e o componente *core powershap.*

No primeiro componente, *Explain*, múltiplos modelos são treinados usando diferentes random seeds em diferentes subconjuntos dos dados. Cada subconjunto é composto por todas as características originais juntas com uma random feature selecionada através do *Random Uniform*. Uma vez que os modelos são treinados, uma média de impacto das características é explicada utilizando valores de *shapely* em um conjunto de dados fora da amostra dos treinos, para evitar algum possível enviesamento na avaliação. No fim desta etapa, é definido o valor absoluto de todos os valores de *shapley*, assim, sendo possível obter uma média (μ) total do impacto de cada feature. Esse processo é repetido por I interações e em cada uma o modelo é retreinado com uma random feature e subconjuntos de dados diferentes. Com isso é possível quantificar os valores de s*hapely* e adquirir uma distribuição empírica das médias dos impactos que, mais tarde, será usada para fins de comparações estatísticas.

Com a média dos impactos de cada feature para cada interação, no componente core *PowerShap*, os impactos das características originais são estatisticamente comparados com a random feature. Essa comparação é quantificada usando a fórmula do percentil (Equação 7), onde s indica uma matriz de valores *Shapely* médios para uma única feature com mesmo tamanho do número de interações, x representa um valor único e $\mathbb{I}$ a função indicadora. Esse valor único x pode ser interpretado como *p-value*, pois essa fórmula calcula a fração de interação em que x for maior que shap -value (o impacto das variáveis levando em consideração outras variáveis).

$$\text{Percentil (s,x)} = \sum_i^n \frac{\mathbb{I}(x > si)}{n} \tag{7}$$

A hipótese PS afirma que o impacto da random feature deve ser menor do que qualquer feature informativa, então todos os impactos das random features são recalculados, resultando em um único valor que pode ser usado na função do percentil, derivando em um *p-value* para cada feature original. Esse valor de p representa a fração de casos em que uma feature é menos importante, em média, do que a *random feature*. Portanto, tendo a hipótese e os valores de p calculados, chega-se a uma implementação de teste estatístico t-Student, onde a hipótese nula

afirma que a random feature (H1 – distribuição) não é mais importante do que a característica testada (H0 - distribuição)(Hahs-Vaughn & Lomax, 2020). Essa implementação não assume um impacto de pontuação na distribuição nas características testadas, como em um teste estatístico padrão *t-Student*, onde se assume um padrão de distribuição Gaussiana. Então, dado um intervalo de *treshold* α é possível encontrar um conjunto de dados de características informativas.

Neste trabalho, o algoritmo PS foi rodado no modo automático. Para isso, é necessário setar dois hiper parâmetros: α (o *treshold* do *p-value*) e I (o número de interações). A escolha desses hiper parâmetros é feita também de maneira automática e otimiza e determina I usando poder de cálculo para α, aqui se encontra a origem do nome *Powershap*.

O poder estatístico de um teste é dado por 1 – β, onde β é a probabilidade de ocorrer falsos negativos. Um valor de falso negativo indica que uma característica que não é informativa é dita como informativa. Se um teste estatístico de uma amostra testada tem como saída uma *p-value* α, significa que a chance da amostra testada pode ser sinalizada erroneamente como significativa para os dados. Esse conceito é calculado pela equação 8.

$$\alpha\,(x) = F_{H0}\,(x) \tag{8}$$

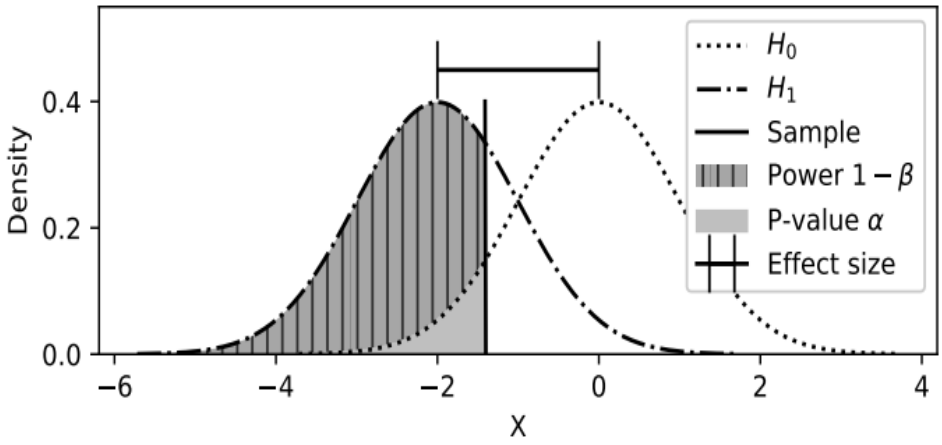
Para um dado α o poder associado deve ser o mais próximo de 1 possível, para evitar falos negativos. Então, o poder do teste estatístico pode ser calculado usando uma função de distribuição cumulativa F da distribuição testada e implícita H1 usando a equação (9).

$$\text{Power}\,(\alpha) = F_{H1}\,(F_{H0}^{-1}(\alpha)) \tag{9}$$

A figura 10 resume todo esse conceito de maneira visual. De maneira sucinta, H0 representa o impacto da distribuição da *random feature* e H1 o impacto da distribuição da *feature* testada.

O poder de cálculo requer a distribuição cumulativa F, mas a distribuição implícita do impacto calculado das *características* é desconhecida. O PS lida com esse empasse mapeando as distribuições implícitas para dois padrões de distribuição *t-student* (Figura 17). Primeiramente, o desvio padrão agrupado (*pooled standard deviation*) através da equação 10, utilizando o desvio padrão σ das duas distribuições.

*Figura 10 - Visualização dos parâmetros calculados pelo PS*




Fonte: https://doi.org/10.48550/arXiv.2206.08394

$$pooled\_std\ (s_1, s_2) = \frac{\sqrt{\sigma^2(s1) + \sigma^2(s2)}}{2} \quad (10)$$

Após isso, a distância d, também chamada de tamanho de efeito (*effect size*) entre as distribuições é calculado, em termos do desvio padrão pooled, usando o tamanho de efeito de *Cohen* (Equação 11) .

$$effectSize\ (s_1, s_2) = \frac{\mu(s2) - \mu}{pooled\_std(s1,s2)} \quad (11)$$

Através desses conceitos estatísticos e suas derivações que são muito bem detalhadas e explicadas no trabalho de *Jarne Verhaeghe* (Verhaeghe et al., 2023), PS é capaz de realizar as a escolha dos melhores hiperparâmetros α e I e executar o comando de seleção de carcteristicas no modo automático. Ao final deste processo, cada caracteristica terá um *p-value*. No modo automático, somente as características que possuírem valores de $p > 0.01$, mostrado importância maior que a *random feature*, serão selecionadas. A implementação do algoritmo do PS é feita em Python, de maneira open-source(Verhaeghe et al., 2023).

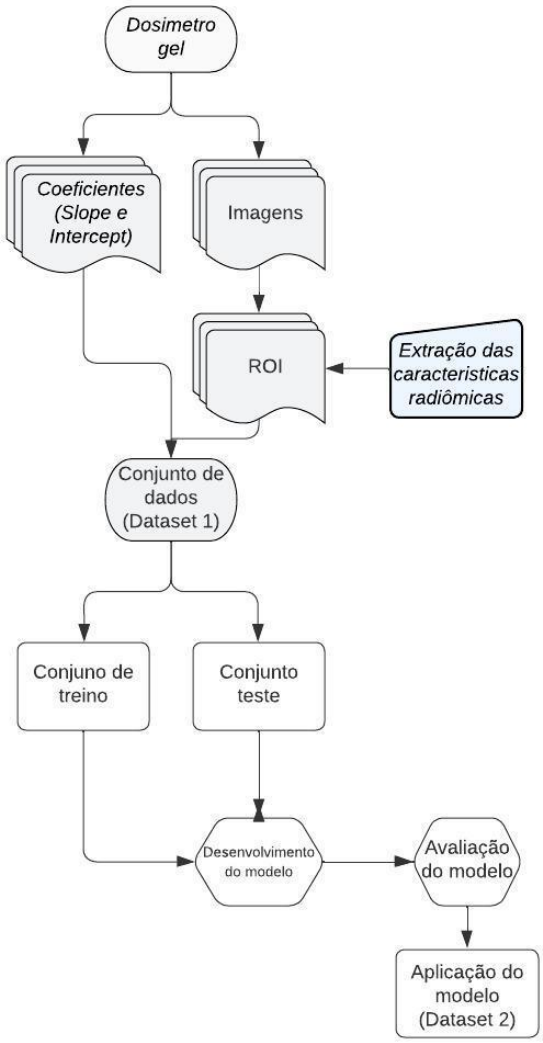
## Capítulo 4 – Metodologia

A metodologia deste trabalho apresenta uma contribuição pioneira ao usar uma ferramenta de aprendizado de máquina para a calibração de dosímetros químicos, sendo o primeiro trabalho a explorar a sensibilidades dos géis por meio de um algoritmo desenvolvido para esta tarefa.

Através das de IRM dos dosímetros gel MAGIC-f é possível obter a curva de calibração relacionando dose e R2 em cada IRM. As características radiômicas do pacote *Pyradiomics* são extraídas das regiões de interesse (ROI - *region of interest*) dos tubos de géis não irradiados e, juntamente com dados característicos da amostra de gel e valores de R2 das amostras, formaram o conjunto de dados de entradas na construção dos modelos preditores dos coeficientes angulares e lineares das curvas de calibração.

Assim, com o conjunto de dados formado, os modelos propostos puderam ser desenvolvidos, avaliados e, ao final do processo ter sua aplicabilidade analisada em um segundo conjunto de dados (*dataset* 2). Esses processos podem ser visualizados na figura 9.

*Figura 11- Representação da metodologia empregada neste estudo*

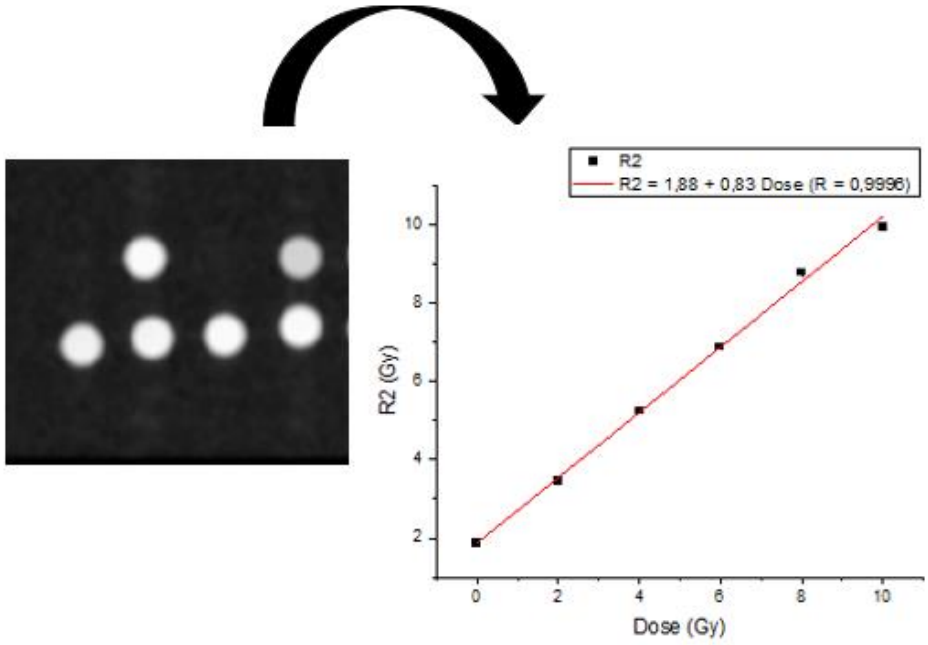



Fonte: O autor (2023)

Foram usadas 145 IRM de calibração de diferentes lotes do dosímetro gel *Magic-f* pertencentes ao acervo do laboratório DART-3D do Departamento de Física da Faculdade de Filosofia, Ciências e Letras – FFCLRP da Universidade de São Paulo, adquiridas por membros que já passaram por este laboratório e fizeram estudos com esse dosímetro para diversas outras finalidades (J. P. Fernandes et al., 2008)(Lizar et al., 2021) (Pavoni & Baffa, 2012) (Pavoni et al., 2017) (Acurio et al., 2021). Essas imagens correspondem a aquisição simultânea no plano axial de tubos de calibração (12 mm de diâmetro e 75 mm de comprimento), contendo a mesma quantidade de gel do mesmo lote de fabricação, que foram irradiados com doses entre 0 e 10

Gy, e sempre um deles permaneceu não irradiado (Figura 10 - esquerda). Como cada tubo foi irradiado com uma dose conhecida e seu valor de R2 foi calculado utilizando um *software*, desenvolvido pelo grupo de pesquisa em MATLAB (Mathworks Inc) (J. Fernandes et al., 2003), é possível obter a curva de calibração para cada imagens (Figura 10 - direita).

*Figura 12 – IRM dos tubos de géis irradiados com doses de 0 a 10 Gy (esquerda) e o gráfico R2 x Dose gerado a partir desta irradiação (direita)*




Fonte: O autor (2023)

Os coeficientes lineares e angulares foram adquiridos através da curva de calibração formada pelos valores de R2 e dose (Gy) e adicionado ao conjunto de dados, já que estes serão os valores que os modelos a serem desenvolvidos irão predizer.

As IRM foram adquiridas com dois aparelhos de Ressonância Magnética diferentes, ambos do Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto – Universidade de São Paulo (HCFMRP – USP): *Siemens Magneton Vision* de 1,5T e *Philips Achieva* de 3T. As aquisições foram feitas a partir de sequências SE (TEs de 22, 60 e 120 ms, com TR de 3000 ms; e TEs de 40, 60, 80 e 100 ms, com TR de 3000 ms) e multi SE (32 ecos múltiplos de 12.5 ms com TR de 2000 ms; 16 ecos múltiplos de 20ms, com TR de 4000 ms; 8 ecos múltiplos de 35 ms, com TR de 1000 ms; e 16 ecos múltiplos de 22.5 ms, com TR de 1000 ms).

O fato do conjunto de imagens utilizado neste trabalho ter sido composto por várias

imagens de géis feitas por pessoas diferentes com objetivos diferentes, resultou em um agrupamento de uma ampla amostra com características heterogêneas, com alguns géis apresentando particularidades em sua confecção como por exemplo o *bloom* da gelatina utilizada (250, 270 ou 300), presença da hidroquinona em algumas preparações, alguns lotes de géis foram privados de luz e também em alguns casos, variou-se a orientação de gelificação de alguns tubos de calibração. Todas essas informações são representadas na tabela 2 e foram levadas em consideração durante o desenvolvimento do modelo de AM, incluindo os valores de R2 de cada imagem, pois a completa caracterização da amostra favorece o aprendizado. Esse conjunto de dados compõe o *dataset* 1 deste estudo.

*Tabela 2 - Diferenças encontradas nas amostras de géis que compõem o Dataset 1.*

| | **Número de IRM (total 145)** |
|---|---|
| Agente oxidante (hidroquinona/sulfato de cobre e ácido ascórbico) | 38/107 |
| Ausência de luz(sim/não) | 112/33 |
| Poder de gelificação (250 / 270 / 300) | 79/12/54 |
| Força de campo RM (1.5T / 3.0T) | 96/49 |
| Aquisição RM (SE / Multi SE) | 58/87 |

Um segundo conjunto de dados (*dataset* 2) foi utilizado como forma de avaliar a aplicação do modelo em dados nunca vistos ao final deste trabalho. Ele é composto 19 IRM adquiridas em um terceiro aparelho de ressonância magnética, *Philips Achieva* de 3T do Hospital Sírio Libanês em São Paulo, Brasil. Esse dado também já estava disponível no acervo do grupo, oriundo de outro trabalho já realizado (Pavoni et al., 2017). O protocolo de aquisição desse conjunto de dados foi uma sequência de aquisição Multi SE semelhante à uma das sequências usadas para treinar o modelo (sequência Multi SE com 8 TEs múltiplos de 35ms e TR de 1000 ms). Esse lote de gel foi produzido com a formulação padrão do gel *MAGIC-f* (250 bloom de poder de gelificação, sulfato de cobre e ácido ascórbico como agentes oxidantes e ambiente típico de exposição à luz – tabela 1).

### *4.1 Extração das características*

Para cada imagem foi feita uma máscara correspondente selecionando uma ROI do tubo não irradiado. Com isso, foi possível usar a biblioteca *Pyradiomics* (van Timmeren et al., 2020) do software Python 3.6 para extrair as características radiômicas dessas máscaras. O processo que envolve esta extração inclui processamento de imagem, segmentação e, por fim, extração das características que podem ser relacionadas à textura, morfologia e às análises

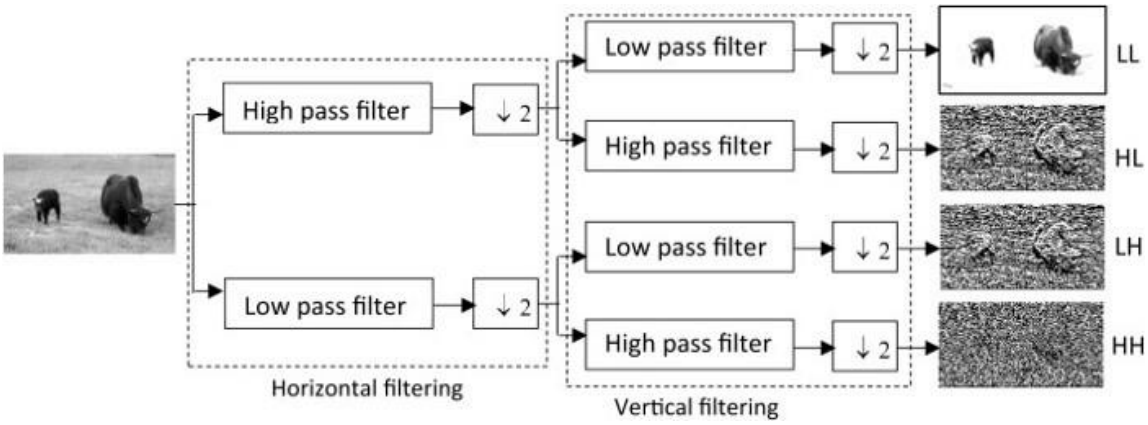
estatísticas, ou de intensidade, referentes aos pixels dessas imagens(Larue et al., 2017).

As características morfológicas estão relacionadas a propriedades geométricas da ROI trabalhada, como área, largura, diâmetro, todas puramente quantitativa, mas, para este trabalho elas não serão utilizadas, uma vez que não há variações morfológicas para estes dados. Já as de intensidade, são características denominadas de primeira ordem que descrevem a distribuição das intensidades de sinais (IS) dentro da ROI, como valores mínimos e máximos, média, mediana, desvio padrão, percentis das intensidades (normalmente 10 e 90), entre outras. Quando se fala nas características de textura, entende-se como características de segunda ordem que descrevem a complexidade espacial e a relação das IS entre os pixels vizinhos, como por exemplo contraste, correlação, entropia ou uniformidade do nível de cinza. Podendo destacar cálculos importantes como co-occurrence matrix (GLCM) descrita por Haralick (Haralick et al., 1973), run-length matrix (GLRLM) descrita por Galloway (Galloway, 1975), ray-level size-zone matrix (GLSZM) (Thibault et al., 2014), gray-level distance-zone matrix (GLDZM) (Thibault et al., 2014), neighborhood gray-tone difference matrix (NGTDM) (Amadasun & King, 1989), and neighborhood gray-level dependence matrix (NGLDM) (Sun & Wee, 1983).

Outra técnica utilizada para extração de novas características foi a aplicação de filtro wavelet na IRM original com objetivo de enfatizar algumas propriedades da imagem, como escala de cinza e delimitações, sensibilizando as intensidades dos valores das características radiômicas, positiva ou negativamente, podendo comprimir ou reduzir o ruído das imagens. Filtros passa alta (H) e passa baixa (L) foram empregados e com eles, a IRM foi decomposta e a informação de frequência de sinal foi separada em componentes de alta e baixa frequência (Mallat, 1989). Assim a imagem é filtrada na direção horizontal e vertical (linhas e colunas da matriz de pixels), originando imagens filtradas a partir da combinação desses dois filtros (figura 11).

Em seguida, novas características radiômicas foram extraídas dessas IRM de maneira semelhante à descrita anteriormente.

*Figura 13 - Decomposição de uma imagem 2D com a transformada wavelets*




Fonte: https://medium.com/@koushikc2000/2d-discrete-wavelet-transformation-and-its-applications-in-digital-image-processing-using-matlab-1f5c68672de3

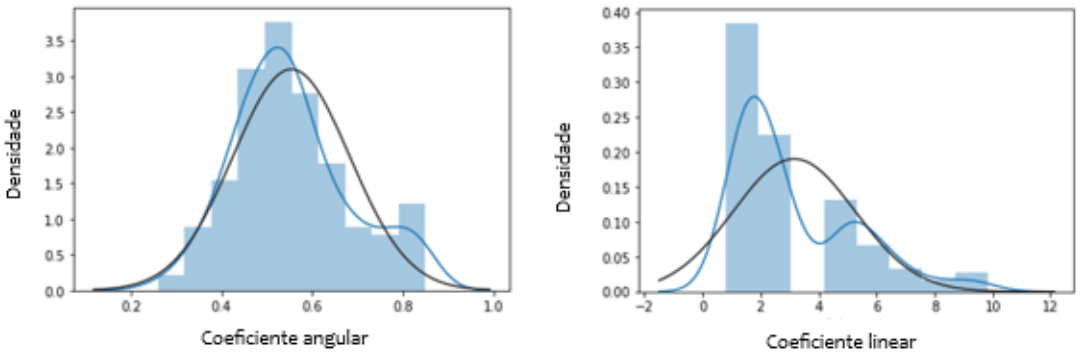
Com isso, o conjunto de dados utilizado neste trabalho no *Dataset* 1 e *Dataset* 2 são matrizes de 145 x 941 e 19 x 941, onde o número de linhas indica a quantidade de imagens utilizada e o número de colunas as características dessas imagens, tanto características radiômicas extraídas com e sem filtro wavelets, quanto o valor de R2 das imagens e características de preparo dos géis, já mencionadas anteriormente (tabela 2).

### *4.2 Modelos desenvolvidos*

Foram desenvolvidos dois modelos de regressão para a predição da curva de calibração baseados nas características radiômicas dos géis não irradiados, um para predizer o coeficiente angular e outro para predizer o coeficiente linear. Além disso, também foram desenvolvidos três modelos de classificação para predizer se os géis se encontram dentro ou fora do padrão de preparo, pois isso pode impactar em sua sensibilidade (ou coeficiente angular). Esses modelos de classificação separam géis em classes que possuem ou não hidroquinona em sua confecção, qual o tipo de gelatina usada e se foram ou não privados de luz em algum momento do procedimento.

Primeiramente, uma análise exploratória dos dados foi feita, checando o padrão de distribuição de dados, através da função *Stats* da biblioteca *Spicy* (Varoquaux et al., 2015) e removendo os valores atípicos por meio do intervalo interquartil. Após a limpeza dos dados resultante da análise exploratória, um perfil de uma distribuição mais próxima de uma distribuição normal foi alcançado para os coeficientes.

*Figura 14 - Distribuição dos dados dos coeficientes angulares (esquerda) e lineares (direita). A linha preta representa a distribuição Gaussiana e a linha azul a distribuição de ajuste gaussiano dos dados destes coeficientes*

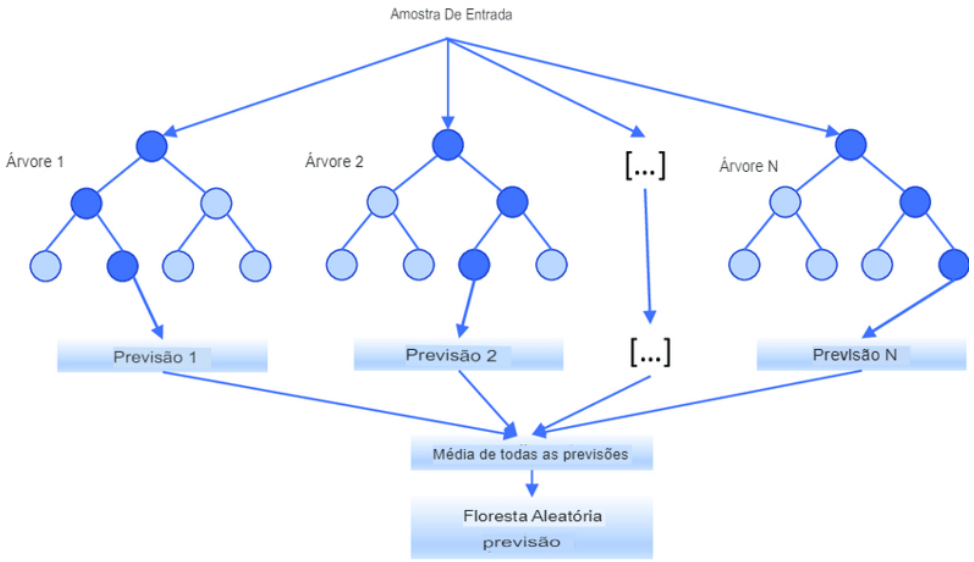



Fonte: O autor (2023)

Foram utilizados dois modelos neste estudo para fins comparativos: *Randon Forest (RF)* e *Categorical Boosting (CB).*

A RF é um algoritmo de *ensemble* de árvores de decisão para problemas de classificação ou regressão. Cada árvore presente na floresta representa um fluxograma de tomadas de decisões que são decompostas recursivamente no espaço das características em regiões contendo observações com valores preditos semelhantes (Strobl et al., 2009). A predição final da RF é uma média das predições de cada árvore, melhorando a predição e evitando *overffiting*.

*Figura 15- Representação de uma RandomForest*




*Fonte:* (Segura et al., 2022)

O CB é um algoritmo que usa um gradiente de reforço nas árvores de decisões. Um conjunto de árvores de decisão é construído consecutivamente durante o treino. Cada arvore sucessiva é construída com uma redução de perda do aprendizado da árvore anterior, tendo o objetivo de melhorar a porcentagem de acerto do modelo e com isso, o aprendizado. Mesmo tendo esse nome, o modelo suporta tanto variáveis categóricas quanto estruturas numéricas e textuais (Jabeur et al., 2021).

O modelo de RF foi escolhido pelo fato dele cobrir limitações de *overffiting* encontradas em árvores de decisões que tem sido amplamente utilizada na área médica e porque esse modelo apresenta performance superior de aprendizado em vários estudos ao longo dos anos (Strobl et al., 2009)(Su et al., 2020). CB, por sua vez, foi escolhido com propósitos comparativos, levando em consideração que também é um modelo de *ensemble* com o adicional da ideia da técnica de *boosting*, que pode enriquecer o aprendizado (Jabeur et al., 2021).

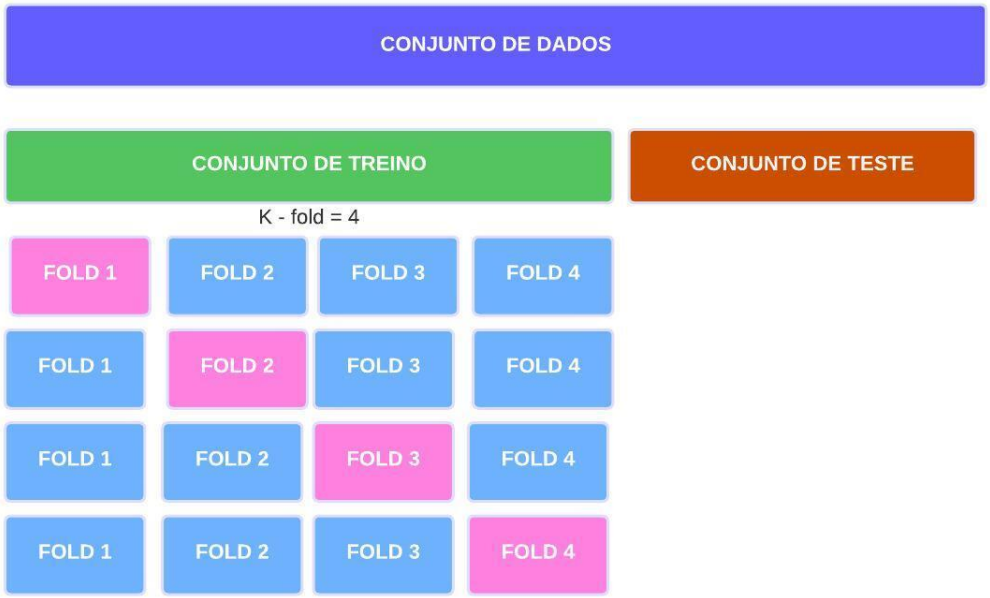
Para o desenvolvimento dos modelos de regressão e classificação, foi utilizada a biblioteca *Sklearn* (Pedregosa et al., 2011).

### *4.3 Afinamento do modelo por meio da escola dos melhores hiperparâmetros*

Com o intuito de melhorar o aprendizado do modelo, foi feita a escolha dos melhores hiperparâmetros através da ferramenta *Randomized Search CV*, encontrada também no *Sklearn*. Através dessa ferramenta é possível achar a melhor combinação existente dos hiperparâmetros de um modelo que sejam capazes de predizer o *output* com o maior *score* e menor incerteza. A otimização desses termos é feita através de uma validação cruzada *K-fold*.

A validação cruzada é uma técnica utilizada para avaliar modelos de AM por meio de treinamento de vários subconjuntos (de mesmo tamanho) do conjunto de dados, de forma randômica (C. H. B. Liu et al., 2017). Aqui, o modelo treina em um subconjunto e avalia a sua performance no subconjunto complementar, utilizando no treinamento final do modelo, aquela combinação que melhor se performar. O número de subconjunto a ser dividido os dados é chamado de K e, normalmente, o valor mais utilizado na literatura é 10 por alguns estudos trazerem esse valor como sendo o que traz melhor desempenho (Borra & Di Ciaccio, 2010) (Kim, 2009).

*Figura 16 - Representação de uma validação cruzada. Após dividir o conjunto de dados em subconjuntos de treino e teste, a porção de treino é novamente subdividida k vezes e em cada interação o modelo treina em uma parte dos novo subconjunto (representados pela cor azul) e testa em outra (representado pela cor rosa).*




Fonte: O autor (2023)

Os hiperparâmetros utilizados para a RF (Varoquaux et al., 2015), tanto na regressão quanto na classificação, foram:

- ***n_estimator:*** 200 - número de árvores pertencentes à RF
- ***max_deph:*** 20 -profundidade máxima da floresta, o número máximo de camadas que as árvores serão divididas
- ***max_feature:*** modo default("sqrt" – raiz quadrada do número de características preente no conjunto de dados)número de características que serão levadas em consideração para encontrar o melhor intervalo de divisão
- ***min_samples_split:*** 10 - número mínimo de amostras necessário para dividir um nó interno **(**os nós internos dão origem a outros nós na próxima camada**)**
- ***min_samples_leaf:*** 4 - número mínimo de amostras necessárias para estar em um nó terminal (que não terá mais nós derivados)

Já para o CB (Gulin Andrey & et al), os hiperparâmetros utilizados, também para os dois tipos de problemas, foram:

- ***interations:*** 100 - número de interações com o melhor valor perda no conjunto de dados de validação, o que ajuda a diminuir a probabilidade de ocorrer *overffiting* no aprendizado
- ***learning_rate***: 0.1 - reduz a etapa do gradiente. Se tiver uma taxa de aprendizado muito pequena, serrão necessárias mais interações e maior será o tempo do treinamento
- ***deph:*** 2 - ***s***imilar a RF, definirá a quantidade de camadas de arvores no modelo
- ***l2_leaf_reg:*** 0.5 - *L2 regularization* é uma técnica utilizada para dataset com grandes números de características para evitar *overffiting* e/ou criar modelos complexos que agem com cautela desnecessária, pelo grande número de entrada de dados e isso cause um viés no aprendizado. Para isso, um coeficiente *(l2_leaf_reg)* será atribuído à função de perda para achar a melhor função que represente o conjunto de dados

### *4.4 Treinamento, teste e avaliação dos modelos*

Os dois modelos (RF e CB) desenvolvidos foram construídos utilizando o grupo de características selecionada pelo melhor dos três métodos propostos (MDI, RFE e PS), sendo a escolha baseada na melhor porcentagem de acerto alcançada por eles, sendo assim seis modelos inicialmente foram desenvolvidos.

Para compor os conjuntos de treino e teste, o *dataset* 1 foi dividido de maneira randômica em 70% e 30%, respectivamente, sendo a validação cruzada aplicada no subconjunto de treino. O mesmo *random_state*, ou seja, o ponto de divisão dos dados nos subconjuntos treino e teste, usado na construção do modelo de regressão, também foi utilizado no desenvolvimento dos modelos de classificação das *características*.

Para avaliar as performances dos modelos, neste trabalho foram utilizadas 3 métricas para escolher a que apresenta o melhor resultado nos modelos desenvolvidos. A MSE é a primeira delas, onde n é o número de pontos preditos, y o valor real e $y_i$ o valor predito pelo modelo (equação 12).

$$\text{MSE} = \frac{1}{n}\sum(y - y_i)^2 \tag{12}$$

A segunda, *Root Mean Squared* (RMSE) é calculada a partir da raiz quadrada do MSE (equação 13). Essa métrica traz o resultado um pouco mais próximo do real quando interpretado, pois, retira do resultado a unidade elevada ao quadrado. Essas duas métricas penalizam valores muitos distantes dos reais, conseguindo ser sensível a outliers que podem estar presentes no modelo (Chai & Draxler, 2014)

$$\text{RMSE} = \sqrt{MSE} \tag{13}$$

Por fim, *Mean Absolute Error* (MAE). Como mostra a equação 14, essa métrica calcula a diferença entre os valores reais (y) e preditos (yi), de modo que, por ser a diferença absoluta, não penaliza valores distantes do esperado como nas outras duas métricas citadas acima.

$$\text{MAE} = \frac{1}{n}\sum |y - yi| \tag{14}$$

Além disso, uma análise em relação à porcentagem total de acerto (acurácia) de cada coeficiente foi analisada considerando um desvio de incerteza para as predições.

Os modelos de classificação foram avaliados utilizando o *Classification report*, também da biblioteca *Sklearn*. Nele, consta a matriz confusão dos dados relacionando os valores reais com os preditos pelo modelo e essa relação é avaliada por algumas métricas (Figura 18). Compondo essa matriz, tem-se os valores de falso positivos (FP), falso negativo (FN), verdadeiros positivos (TP) e verdadeiros negativos (TN) . Esses valores citados, são utilizados para calcular alguns parâmetros como acurácia, precisão, revocação e *f1-score*.

*Figura 17 - Matriz de confusão*

| | | Valores Reais | |
|---|---|---|---|
| | | POSITIVOS | NEGATIVOS |
| Valores Preditos | POSITIVOS | VP | FP |
| | NEGATIVOS | FN | VN |

Fonte: O autor (2023)

A acurácia verifica a quantidade de acerto que o modelo teve em relação ao conjunto de dados completo (equação 13). A precisão, analisa a capacidade do modelo de predizer corretamente amostras verdadeiras dentre todas aquelas ditas como verdadeiras (equação 14). A revocação analisa as predições verdadeiras em relação àquelas que realmente são (TP) com as que são verdadeiras, mas são preditas como falsas (FN) (equação 15). Por fim, o *f1-score* (equação 16) é a média harmônica entre os parâmetros precisão e revocação e seu valor varia entre 0 e 1 para baixa e alta performance do modelo, respectivamente. Ele representa a performance do modelo para cada classe especificamente.

$$\text{Acurácia} = \frac{VP+VN}{VP+VN+FP+FN} \tag{13}$$

$$\text{Precisão} = \frac{VP}{VP+FP} \tag{14}$$

$$\text{Revocação} = \frac{VP}{VP+FN} \tag{15}$$

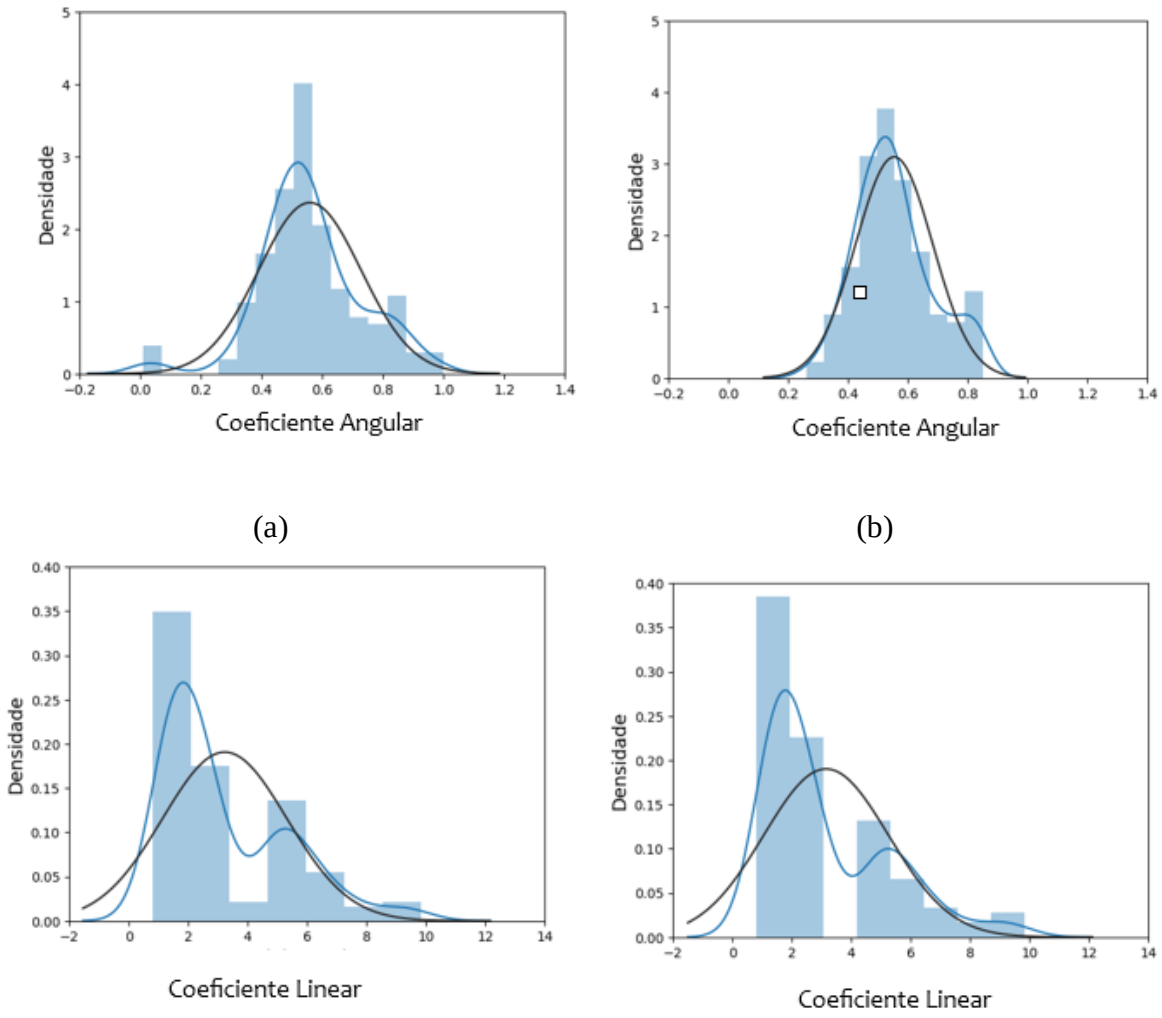
$$F1 - score = 2 \text{ x } \frac{Precisão\ x\ revocação}{Precisão+revocação} = \frac{2VP}{2VP+FP+FN} \tag{15}$$

## Capítulo 5 – Resultados e Discussão

### *5.1 Análise inicial dos dados*

A figura 18 apresenta a distribuição dos dados antes (Figura 18 - a e c) e após (Figura 18 - b e d) o processamento para o *dataset* 1. Após o pré-processamento dos dados, foi possível obter uma distribuição mais perto do normal para os coeficientes, o que é mais notável para o coeficiente angular (Figura 18 a e b).

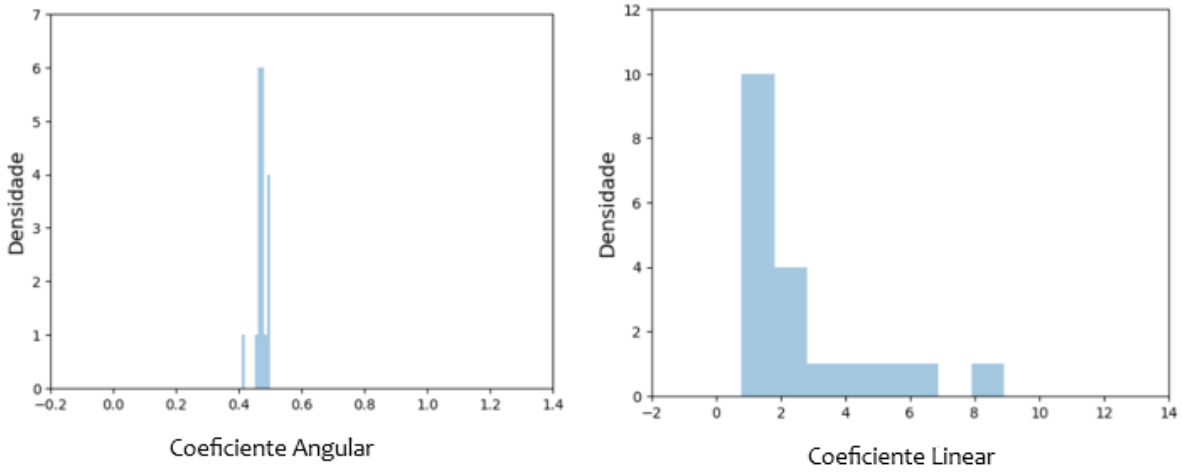
*Figura 18 - Distribuição dos dados para os coeficientes angulares e lineares antes (a e c) e após (b e d) o pré processamento dos dados. A linha preta representa a distribuição gaussiana e a linha azul a distribuição de ajuste gaussiano dos coeficientes*




Fonte: O autor (2023)

As distribuição dos dados do *dataset* 2 se encontram na Figura 19. É possível observar que, para este conjunto de dados, os valores do coeficiente angular (figura 19 - a) se concentram na região central da distribuição de dados deste mesmo coeficiente no *dataset* 1.

*Figura 19 - Distribuição dos dados dos coeficientes angular (esquerda) e linear (direita) para o dataset 2*




### *5.2 Modelos de regressão*

A metodologia descrita no capítulo 4 foi seguida para o desenvolvimento de todos os modelos. A seguir será apresentado todos os passos para a construção dos modelos de regressão que foram avaliados com as métricas MSE, RMSE e MAE e acurácia(porcentagem de acerto das predições do modelo em relação a todos as predições), respectivamente.

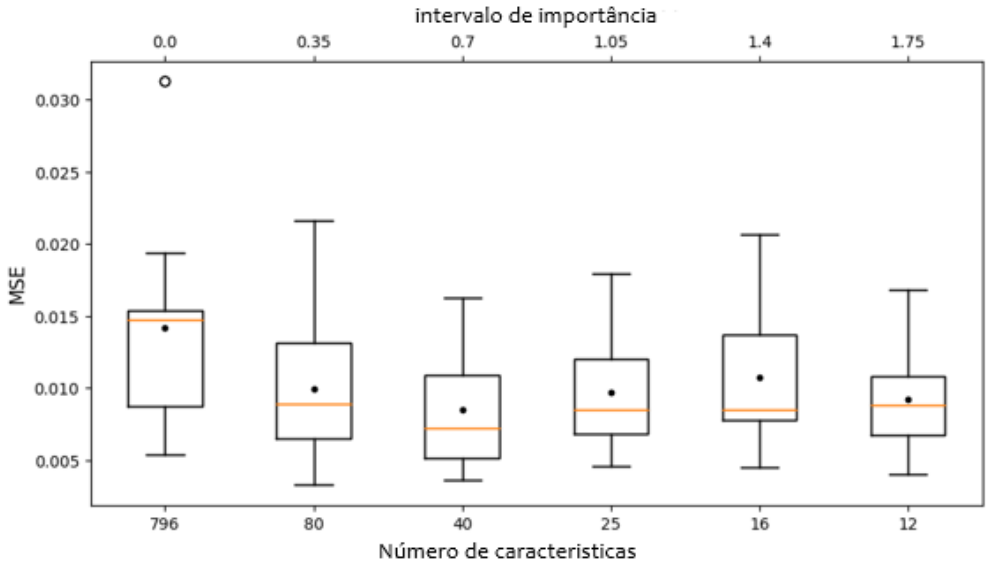
#### 5.2.1 Seleção das características utilizando a métrica MSE

Os grupos de características foram selecionados através dos 3 métodos descritos no capítulo 4 (MDI, RFE e PS) para os dois modelos desenvolvidos, RF e CB.

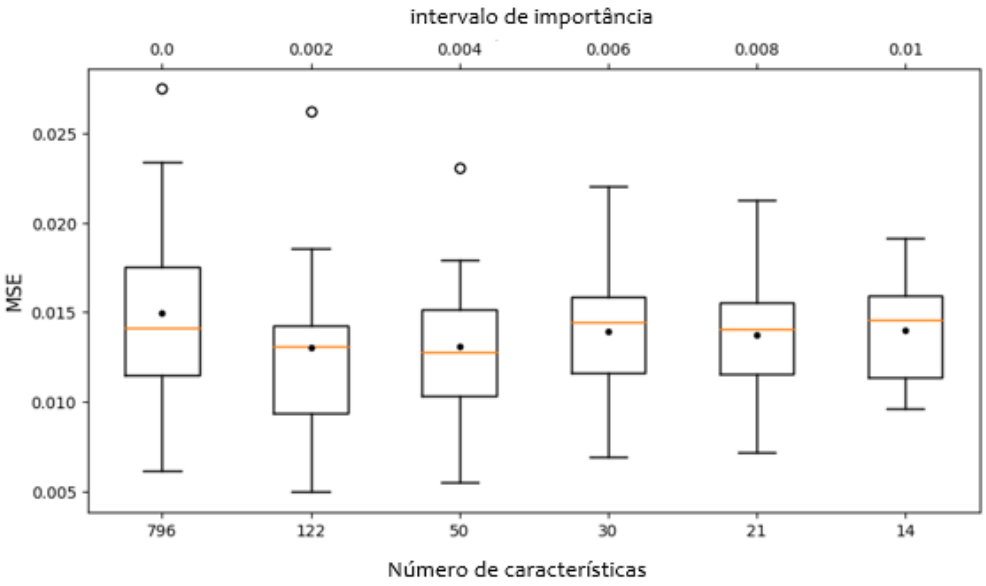
O grupo de característica selecionado neste estudo é aquele que apresenta a maior quantidade dos seguintes critérios: menor valor de MSE máximo, menor MSE médio, menor dispersão (diferença entre o primeiro e terceiro quartil) e valores de média e mediana próximos. Com base nesses critérios, 12 características foram selecionadas para o coeficiente angular com

o modelo CB (Figura 20 - a) e 40 para o modelo RF (Figura 20 - c); já para o coeficiente linear, foram selecionadas 12 características com a CB (Figura 20 - b) e 42 com o RF (Figura 20 - c).

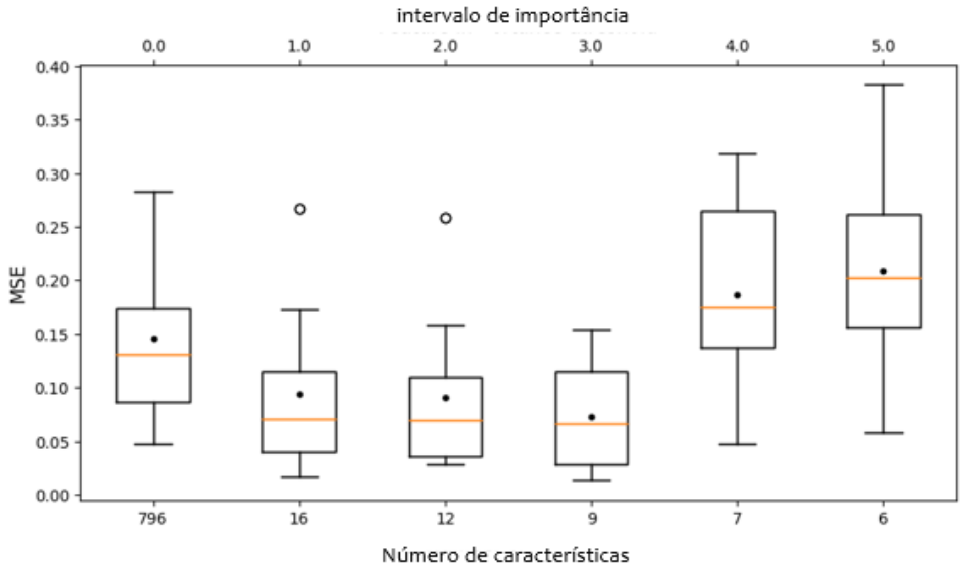
*Figura 20 - MDI Boxplot (a) Boxplot para o coeficiente angular com o modelo CB e intervalos de treshold de 0.35 (b) Boxplot para o coeficiente linear com o modelo CB e intervalos de treshold de 1.0 (c) Boxplot para o coeficiente angular com o modelo RF e intervalo de treshold de 0.002 (d) Boxplot para o coeficiente linear com o modelo RF e intervalo de treshold de 1.0*



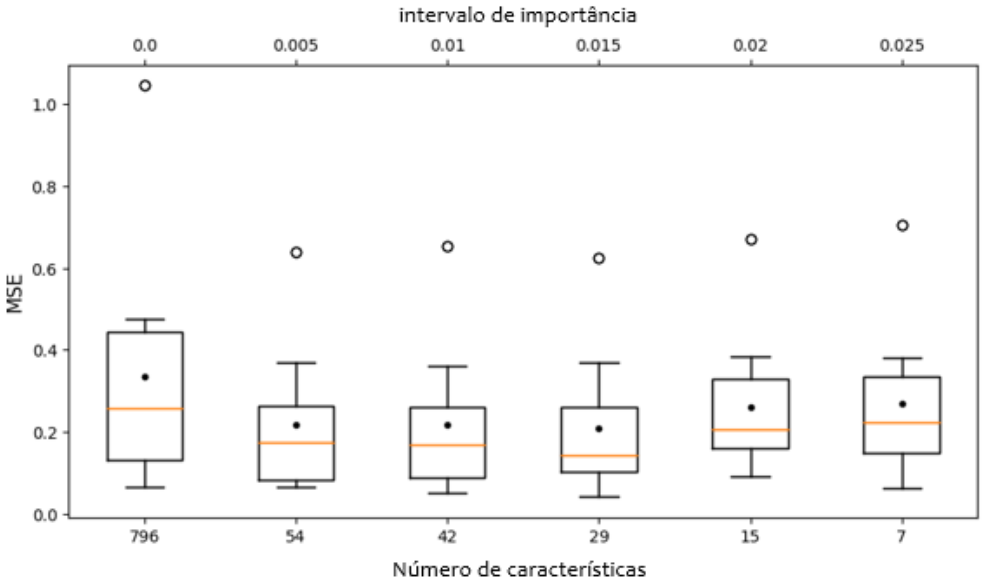

(a)




(b)




(c)




(d)

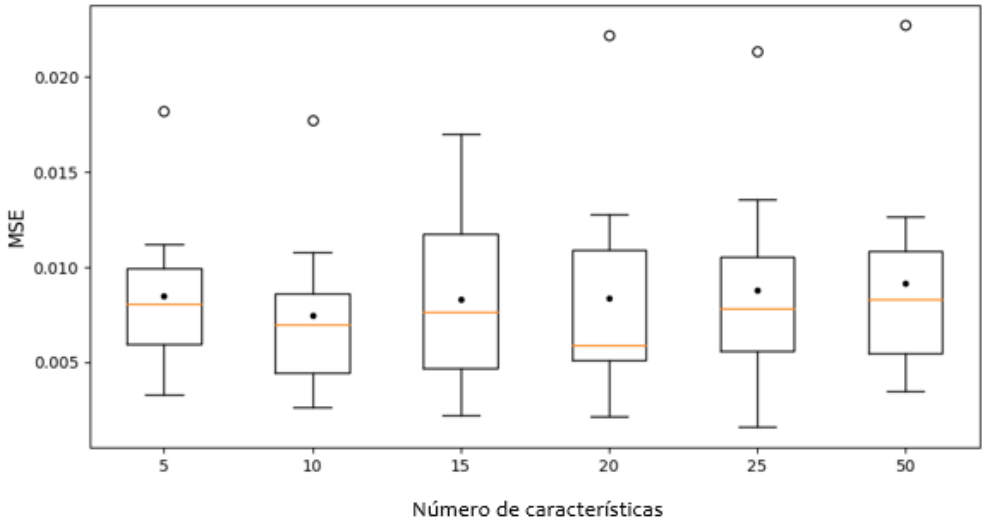
Fonte: O autor (2023)

As tabelas A1 a A4 contém o grupo de características escolhido em cada gráfico e se encontram no Anexo A.

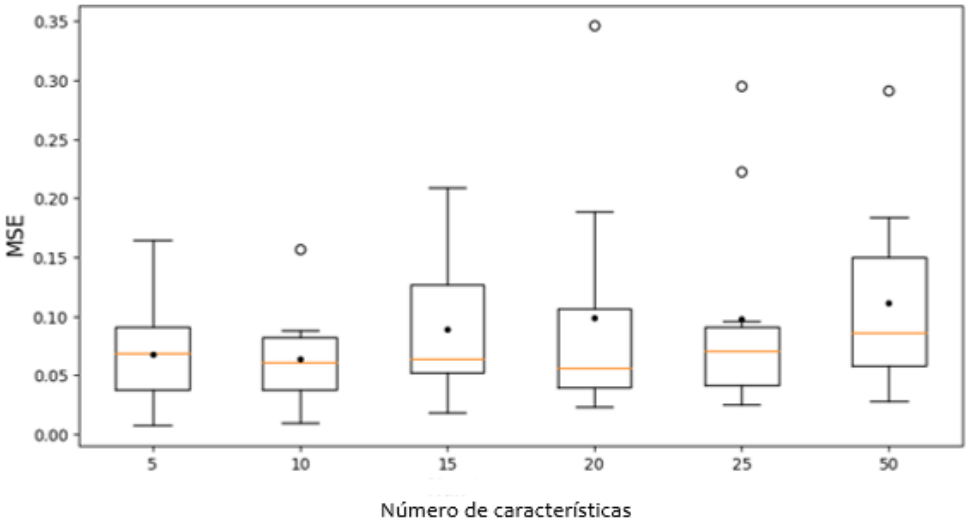
Nesta análise foram escolhidos grupos de 5, 10, 15, 20, 25 e 50 características pois, grupos contendo mais que 50 características geraram modelos de baixa performance. A Figura 21 mostra o Boxplot dos valores de MSE para os grupos de *características* selecionados com o

RFE para os coeficientes angular e linear. Nesta etapa também foi feita uma validação cruzada com 10 *folds* com o objetivo de um resultado sem probabilidade de viés e utilizado o mesmo critério para a escolha do boxplot usado para o método MDI. Para o coeficiente angular, com os modelos CB e RF foi escolhido o grupo contendo 10 características (Figura 21 - a e c, respectivamente) e para o coeficiente linear, foi escolhido o grupo contendo 10 características com o modelo CB (figura 21 - b) e o grupo contendo 5 para a RF (figura 21 - d). As tabelas de A5 a A8 do anexo A apresentam as características escolhidas em cada caso.

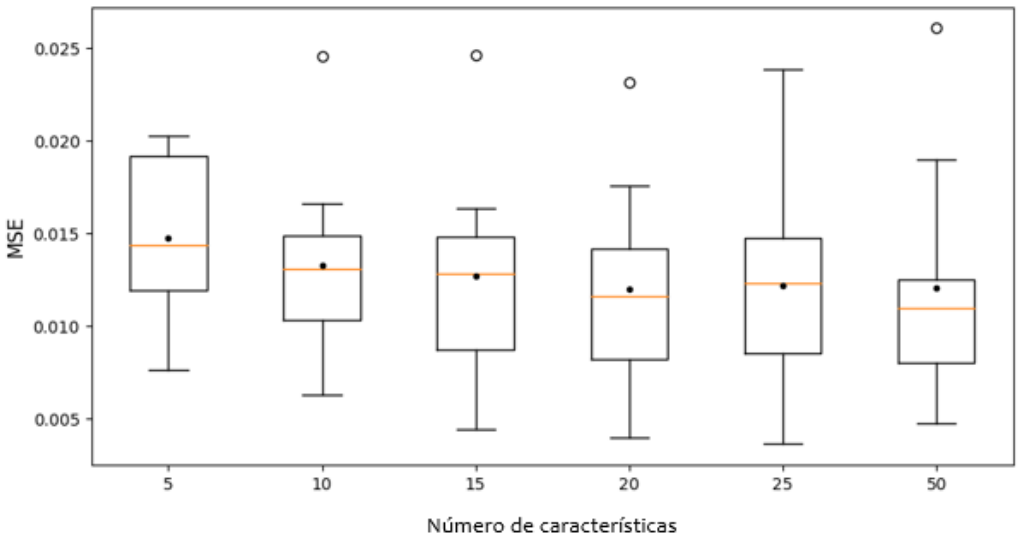
*Figura 21 - RFE Boxplot (a) para o Coeficiente angular com o modelo CB (b) para o Coeficiente linear com o modelo CB (c) para o Coeficiente angular com o modelo RF (d) para o Coeficiente linear com o modelo RF*



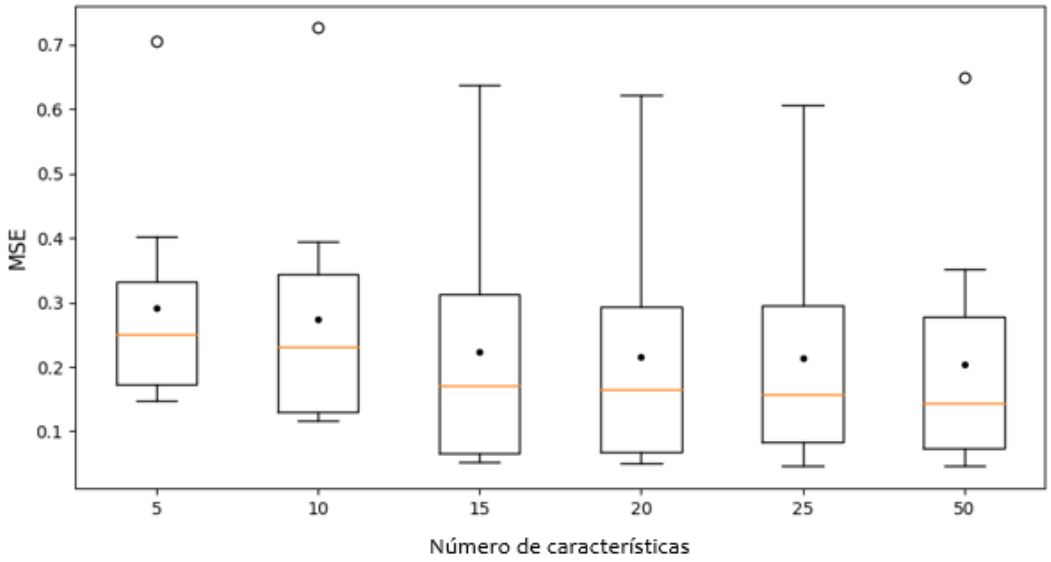

(a)




(b)




(c)




(d)

Fonte: O autor (2023)

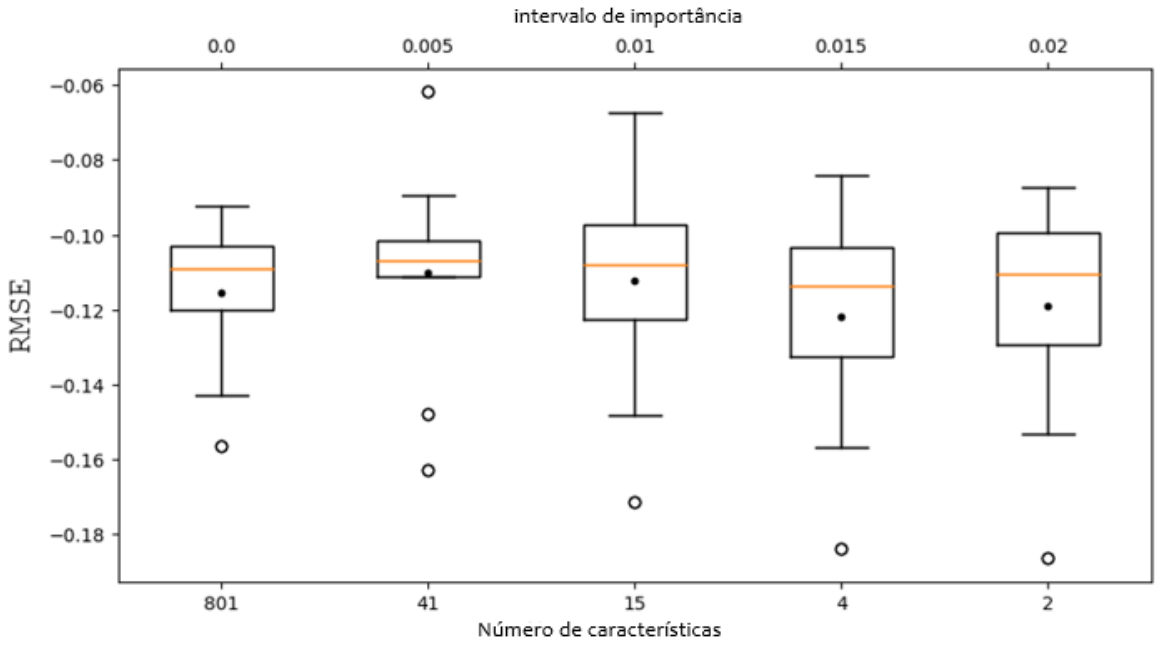
As características selecionadas pelo método PS foram 6,28 e 30 para o coeficiente angular com o modelo RF, coeficiente linear com o modelo RF e coeficiente linear com o modelo CB, respectivamente, e se encontram nas tabelas A9, A10 e A11 do anexo A . Por se tratar de um método mais autônomo comparado aos outros dois, o PS não necessita de interação por parte do usuário, devolvendo o grupo de características de forma direta. Para o modelo de regressão CB, a biblioteca PS não obteve êxito para selecionar um conjunto de características para predizer o coeficiente angular, não sendo possível obter um resultado com essa biblioteca usando este modelo de regressão.

### 5.2.2 Seleção das características utilizando a métrica RMSE

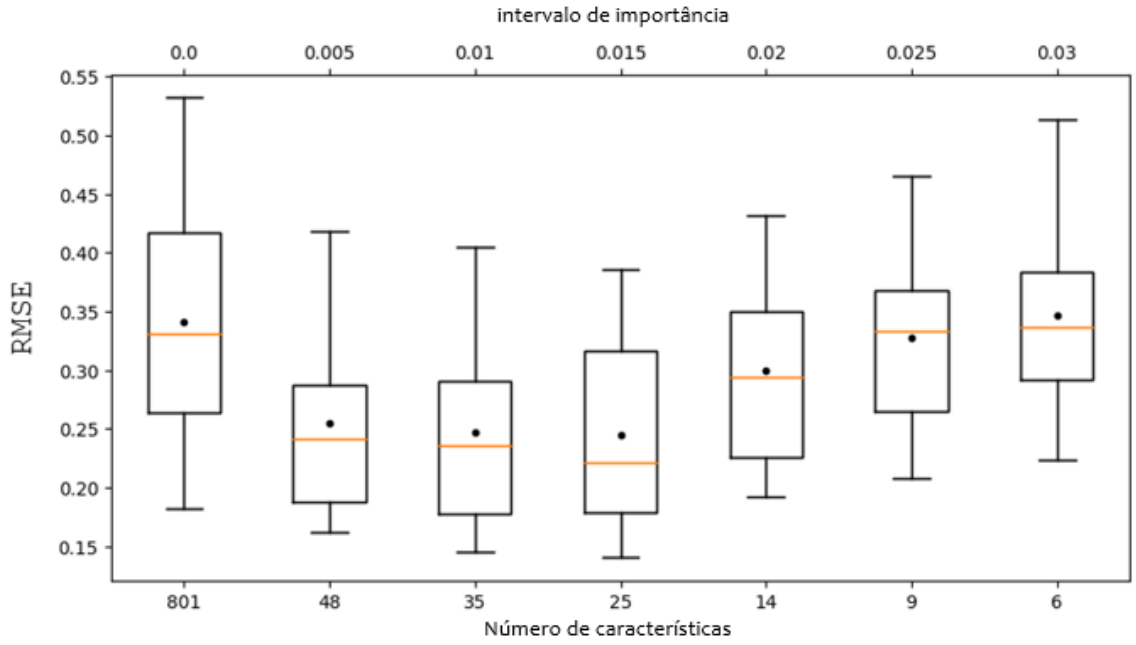
Os três métodos de seleção de características agora foram utilizados tendo como base nos cálculos internos a métrica RMSE, sendo a mesma metodologia descrita na seção anterior utilizada nesta análise.

A figura 22 mostra os gráficos obtidos através do método MDI para os dois coeficientes utilizando os modelos RF e CB. Com base nos critérios já mencionados com base, nesta seção nos valores de RMSE, 41 características foram selecionadas para o coeficiente angular com a RF (Figura 22 - a), e 18 com o modelo CB (Figura 22 - c) para este mesmo coeficiente. Para o coeficiente linear, 48 características foram selecionadas com o modelo RF (Figura 22 - b) e 8 com o CB (Figura 22 - d). A descrição dos nomes de *características* selecionadas em cada grupo, encontra-se no Anexo B (Tabelas B1, B2, B3 e B4).

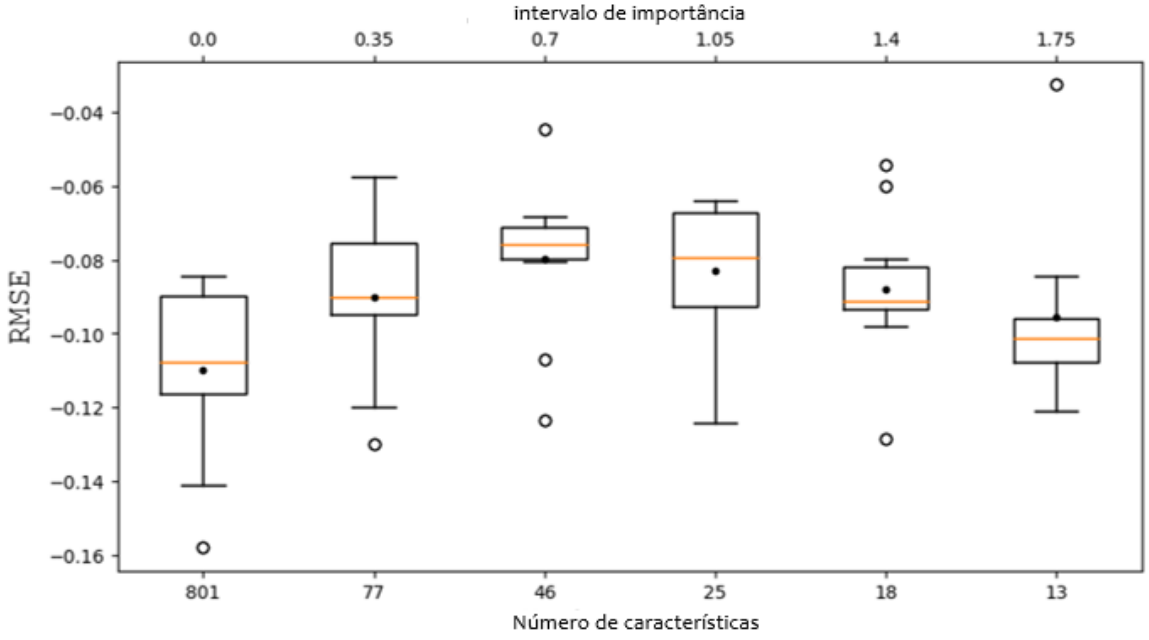
*Figura 22 - MDI Boxplot (a) Boxplot para o Coeficiente angular com o modelo RF e intervalos de treshold de 0.005 (b) Boxplot para o Coeficiente linear com o modelo RF e intervalos de treshold de 0.005 (c) Boxplot para o Coeficiente angular com o modelo CB e intervalo de treshold de 0.35 (d) Boxplot*



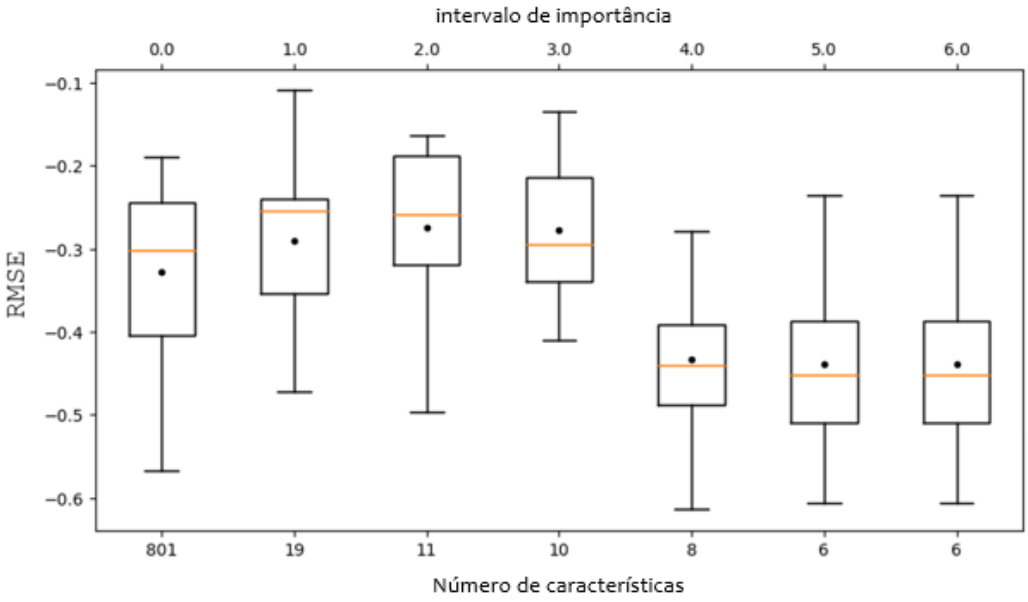

(a)


intervalo de importância
0.0
0.005
0.01
0.015
0.02
0.025
0.03
RMSE
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
801
48
35
25
14
9
6
Número de características

(b)


intervalo de importância
0.0
0.35
0.7
1.05
1.4
1.75
RMSE
−0.04
−0.06
−0.08
−0.10
−0.12
−0.14
−0.16
801
77
46
25
18
13
Número de características

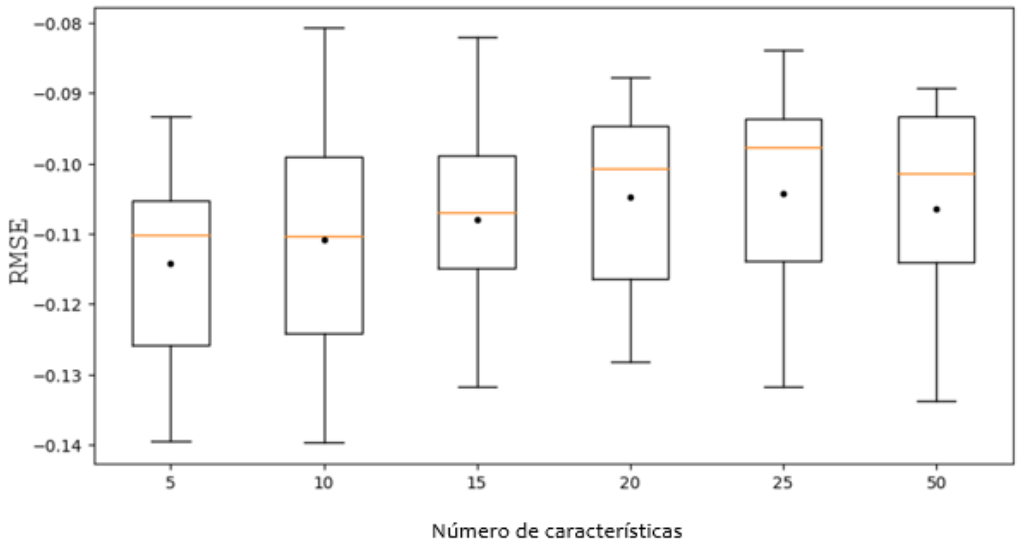
(c)



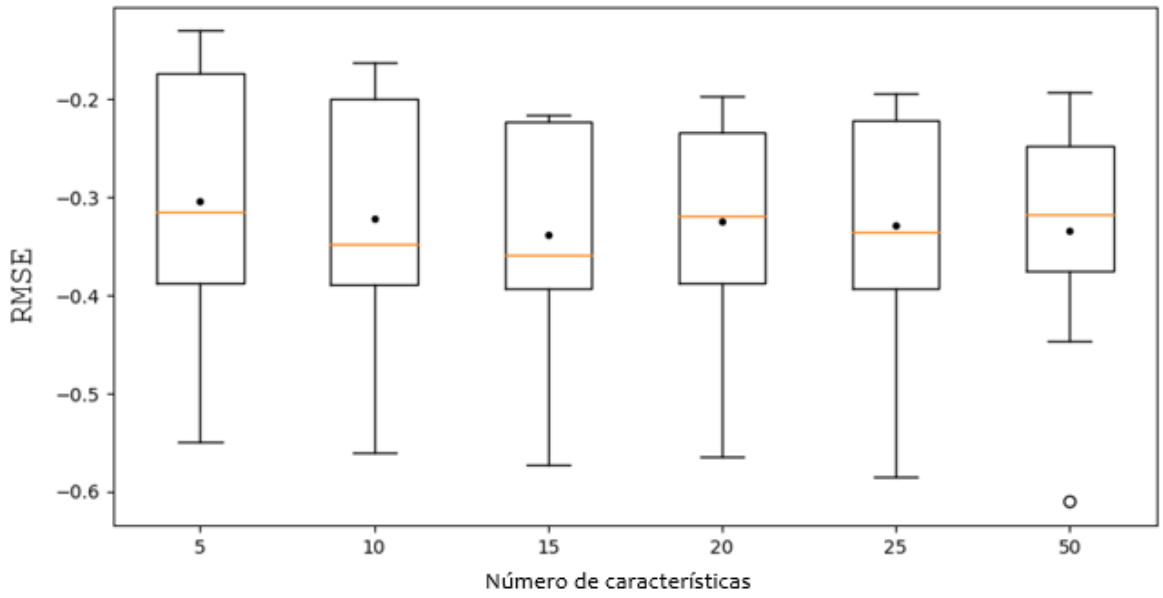

(d)

Os gráficos gerados a partir das análises feitas com o método RFE se encontram na Figura 23. Para este método, 5 características foram selecionadas com o modelo RF para o coeficiente angular (Figura 23 - a ) e 50 com o modelo CB (Figura 23 - c) para este mesmo coeficiente. 20 características selecionadas para o coeficiente linear com os modelos RF e CB (Figura 23 – b e d). Esses grupos selecionados para cada modelo estão apresentados nas tabelas de B5 a B8 no anexo B.

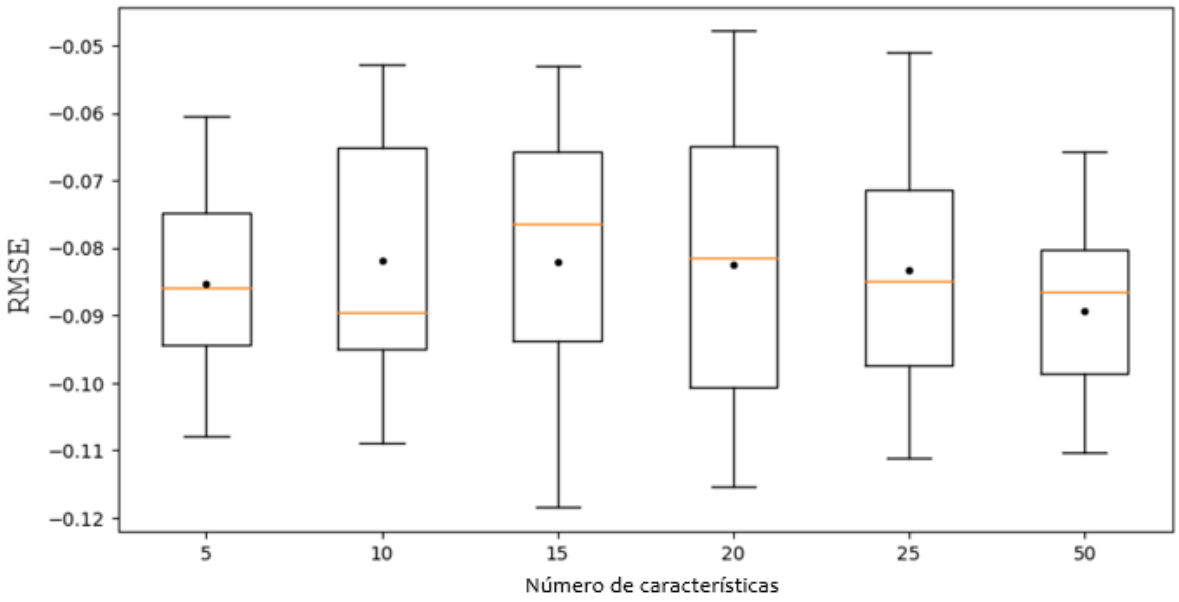
Figura 23 - RFE Boxplot (a) Boxplot para o Coeficiente angular com o modelo RF (b) Boxplot para o Coeficiente linear com o modelo RF (c) Boxplot para o Coeficiente angular com o modelo CB (d) Boxplot para o Coeficiente linear com o modelo CB



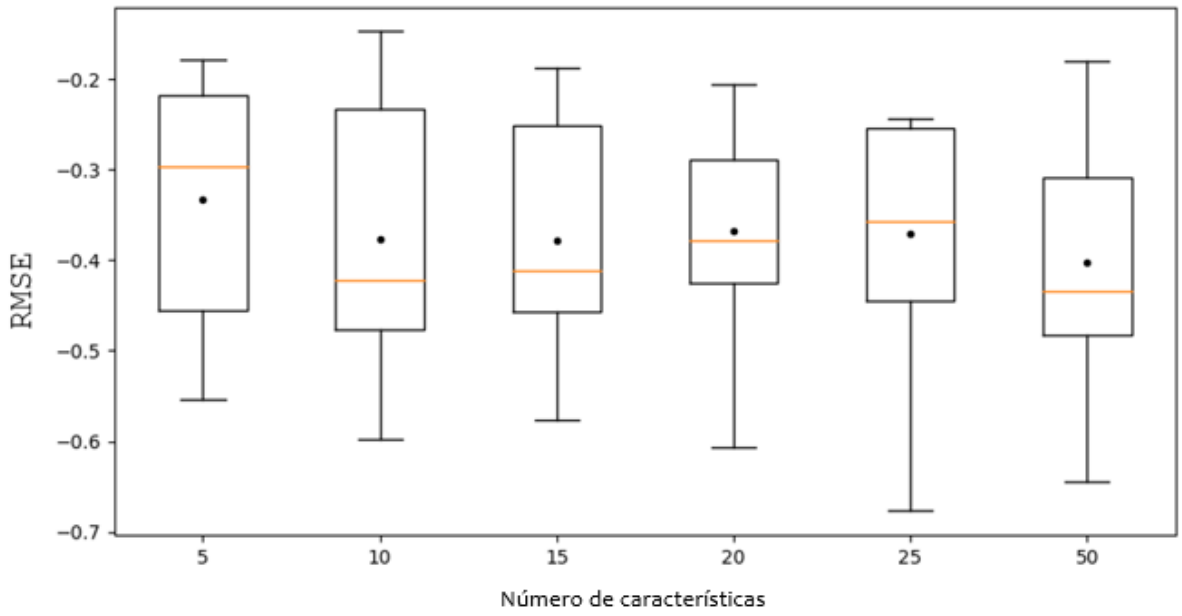

(a)


RMSE
−0.2
−0.3
−0.4
−0.5
−0.6
5
10
15
20
25
50
Número de características

(b)


RMSE
−0.05
−0.06
−0.07
−0.08
−0.09
−0.10
−0.11
−0.12
5
10
15
20
25
50
Número de características

(c)




(d)

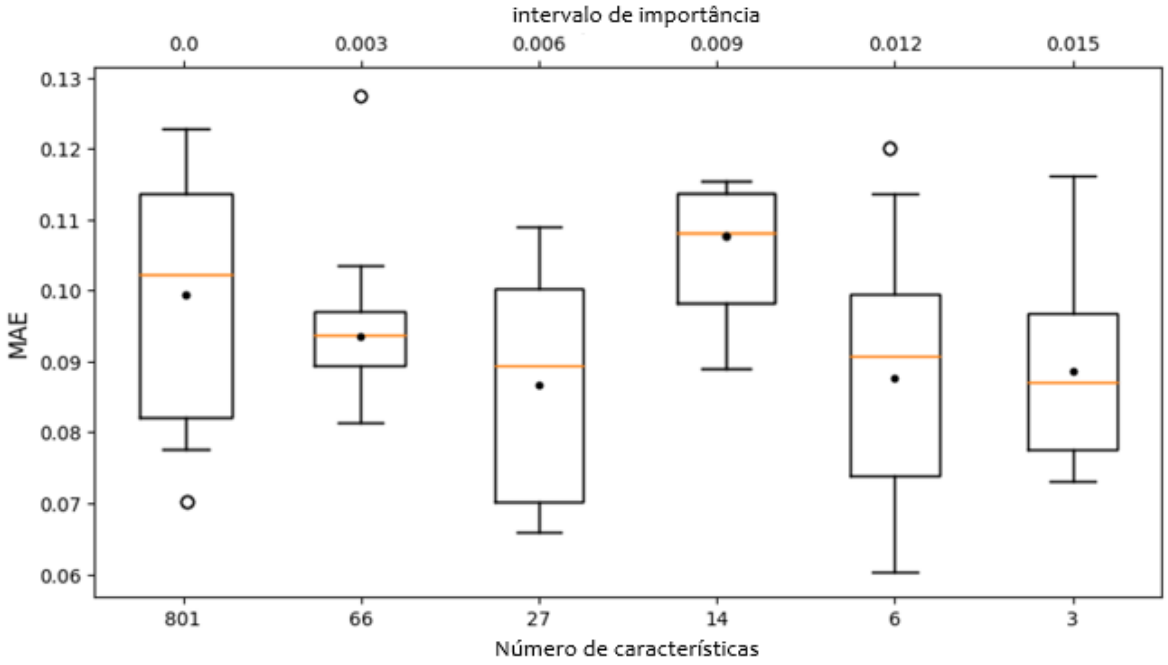
Os grupos de características selecionados com a biblioteca PS estão também presentes nas tabelas B9, B10 e B11 no anexo B. Para o coeficiente angular, essa método selecionou 17 características com o modelo RF e para o coeficiente linear, 16 características foram selecionadas com o modelo RF e 5 com o modelo CB.

### 5.2.3 Seleção das características utilizando a métrica MAE

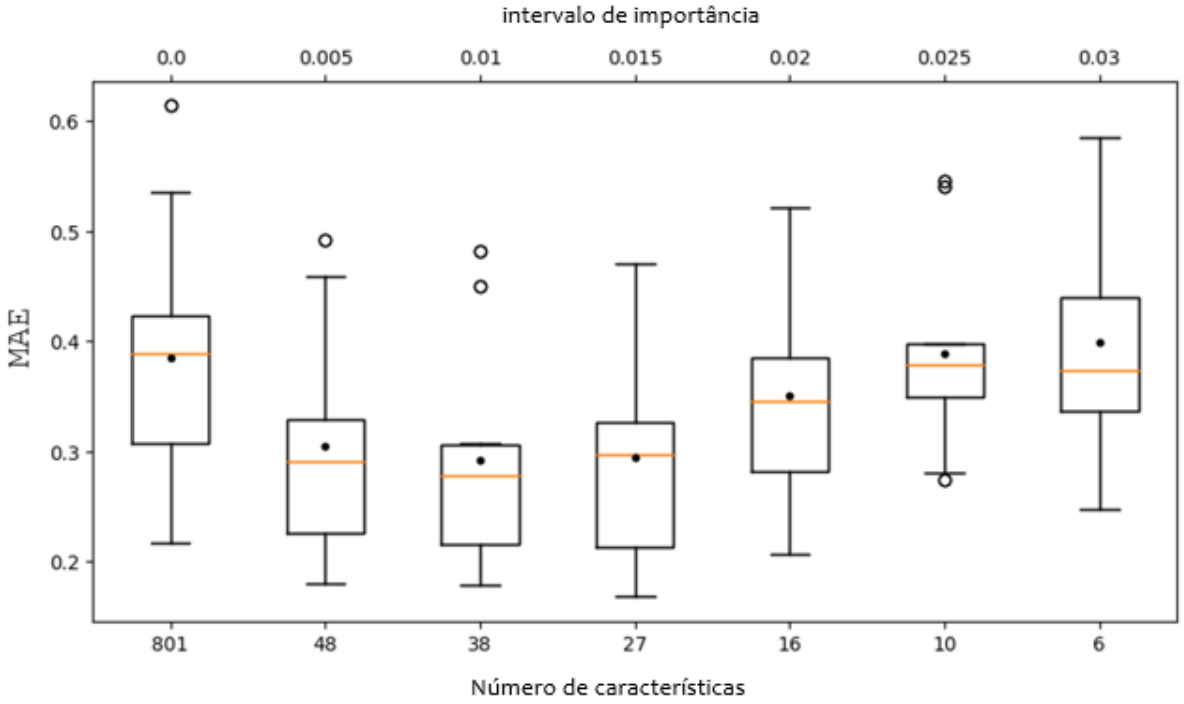
A figura 24 mostra os resultados obtidos pelo modelos (RF e CB) quando a métrica MAE foi utilizada para a seleção ds caracteristicas. Os critérios de escolha são os mesmos utilizados para as outras duas métricas (MSE e RMSE) e que ja foram mencionados. Sendo assim, o modelo RF mostrou uma melhor configuração usando como dados de entradas 66 caracteristicas (Figura 24 - a ) e o modelo CB utilizando 10 para o mesmo coeficiente (Figura 24 - c ). 38 características foram selecionada com o modelo RF (Figura 24 - b) e 10 com o modelo CB (Figura 24 - c ) para o coefieciente linear. Esses grupos estão expostos no anexo C, nas tabelas de C1 a C4.

*Figura 24 - MDI Boxplot (a) Boxplot para o Coeficiente angular com o modelo RF e intervalos de treshold de 0.003 (b) Boxplot para o Coeficiente linear com o modelo RF e intervalos de treshold de 0.005 (c) Boxplot para o Coeficiente angular com o modelo CB e*

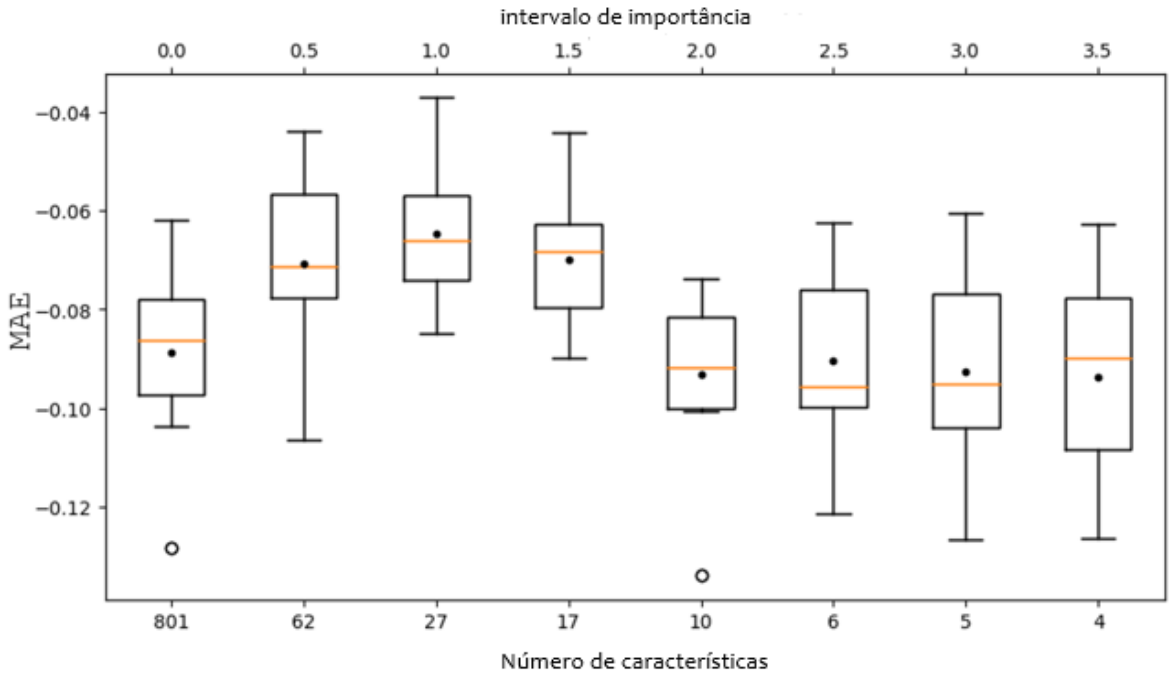
*intervalo de treshold de 0.5 (d) Boxplot para o Coeficiente linear com o modelo RF e intervalos de treshold de 1.0*



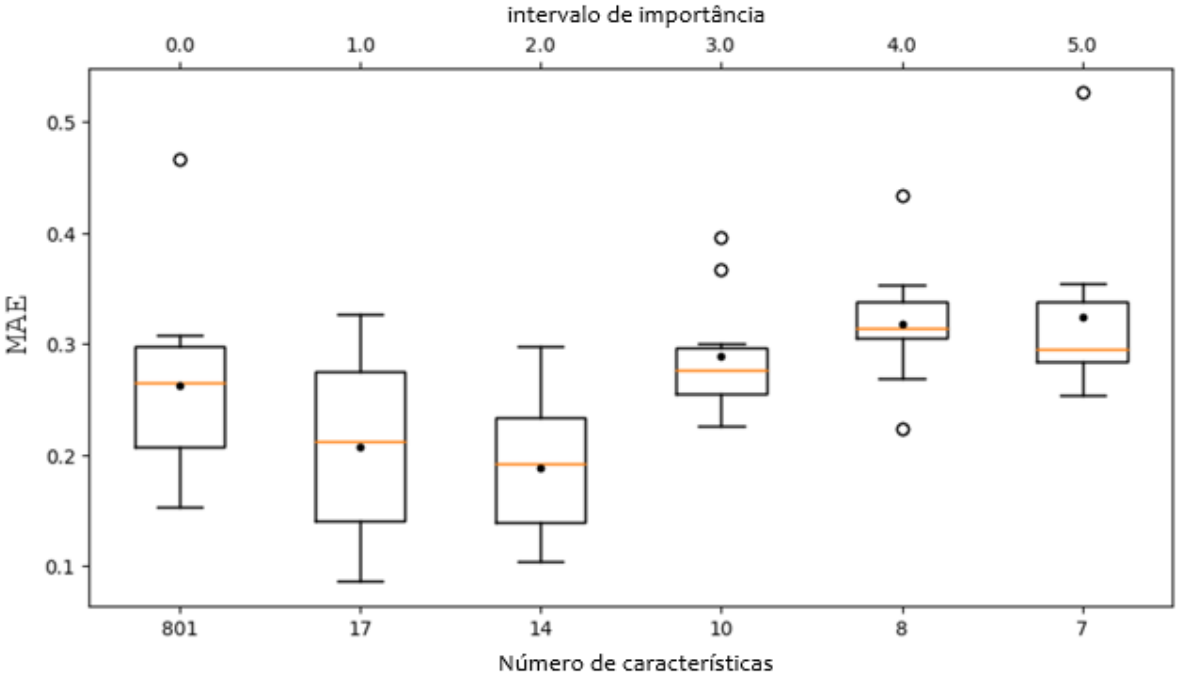

(a)




(b)




(c)

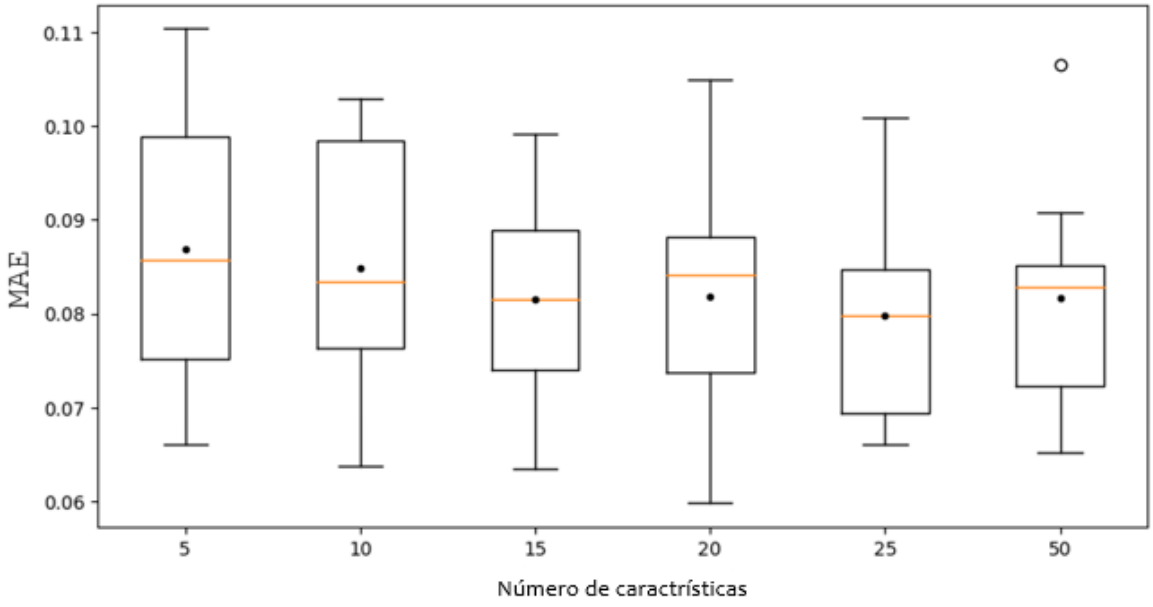



(d)

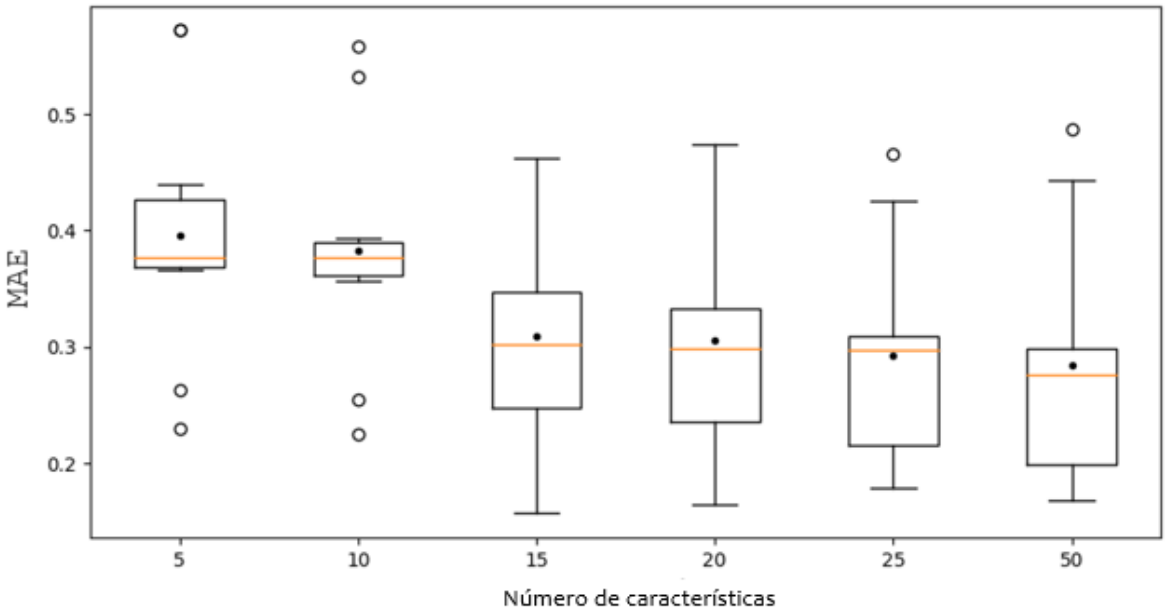
Os grupos de caracteristicas escolhidos atráves do método RFE estão graficamente demostrados na Figura 25. 50 características foram selecionadas para o coeficente angular com o moldeo RF (Figura 25 (a)) e com o modelo CB (Figura 25(c)). Para o coeficiente linear 10 foram selecionadas com o modelo RF(Figura25 (b)) e 5 com o modelo CB (Figura 25 (d)). Também estao presentes nas tabelas C5, C6, C7 e C8 no Anexo 3 as características pertencentes

a cada grupo escolhido da figura 25.

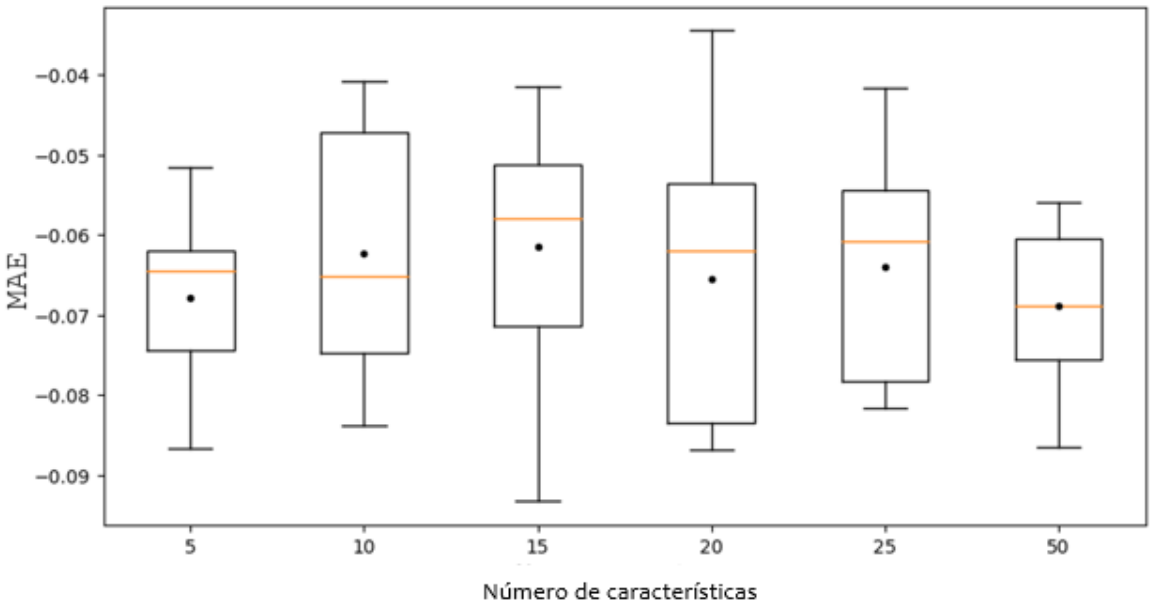
*Figura 25 - - RFE Boxplot (a) Boxplot para o Coeficiente angular com o modelo RF (b) Boxplot para o Coeficiente linear com o modelo RF (c) Boxplot para o Coeficiente angular com o modelo CB (d) Boxplot para o Coeficiente linear com o modelo CB*



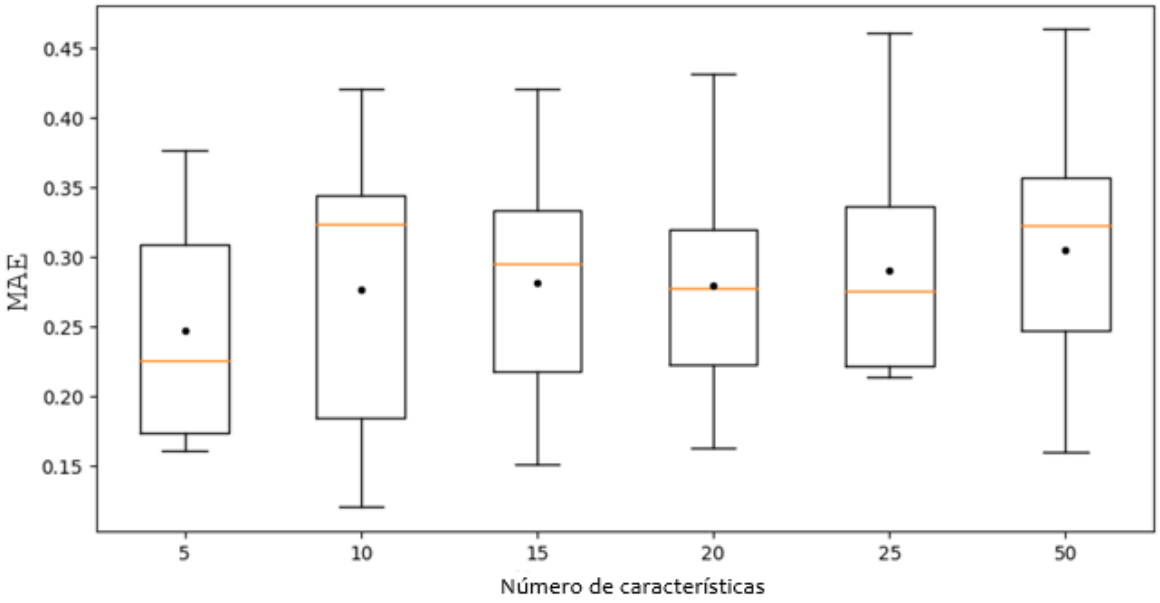

(a)




(b)




(c)




(d)

No anexo C também se encontram as tabelas C9, C10 e C11, onde as características que o método PS selecinou com os dois modelos para os dois coeficientes, sendo elas 69 para o coeficiente angular com o modelo RF, 47 para o coeficiente linear com o modelo RF e 4 para este mesmo coeficiente com o modelo CB.

As tabelas 3 e 4 apresentam uma síntese dos resultados obtidos com os três métodos propostos de seleção de características (MDI, RFE e PS) combinados com os dois modelos (RF

e CB) que foram analisados com a três métricas sugeridas (MSE, RMSE e MAE) para os dois coeficientes, assim como a quantidade de características que cada metodologia selecionou, os valores de cada métrica e a acurácia atingida em cada caso.

*Tabela 3 - Acurácias alcançadas para os modelos RF e CB para predizer o coeficiente angular usando as três metodologias de seleção de características (MDI, RFE e PS) e três métricas (MSE, RMSE e MAE).*

| | | RF | | | CB | |
|---|---|---|---|---|---|---|
| | | MDI | RFE | PS | MDI | RFE |
| | N°características | 50 | 10 | 6 | 12 | 10 |
| MSE | Resultado | $9{,}27 \times 10^{-3}$ | $1{,}12 \times 10^{-2}$ | $6{,}67 \times 10^{-3}$ | $7{,}68 \times 10^{-3}$ | $8{,}36 \times 10^{-3}$ |
| | Acurácia | 0,68 ± 0,05 | 0,62 ± 0,05 | **0,77 ± 0,05** | 0,70 ± 0,05 | 0,68 ± 0,05 |
| | N°características | 42 | 5 | 17 | 18 | 50 |
| RMSE | Resultado | 0.099 | 0,102 | 0,097 | 0.092 | 0,092 |
| | Acurácia | 0,68 ± 0,05 | 0,63 ± 0,05 | 0,63 ± 0,05 | 0.63±0,05 | 0,66 ± 0,05 |
| | N°características | 66 | 50 | 69 | 10 | 50 |
| MAE | Resultado | 0,050 | 0,051 | 0.050 | 0,060 | 0,049 |
| | Acurácia | 0,66 ± 0,05 | 0,61 ± 0,05 | 0,63 ± 0,05 | 0,61 | 0,68 ± 0,05 |

*Tabela 4- Acurácias alcançadas para os modelos RF e CB para predizer o coeficiente linear usando as três metodologias de seleção de características (MDI, RFE e PS) e três métricas (MSE, RMSE e MAE).*

| | | RF | | | CB | | |
|---|---|---|---|---|---|---|---|
| | | MDI | RFE | PS | MDI | RFE | PS |
| | Características | 42 | 5 | 28 | 12 | 10 | 30 |
| MSE | Resultado | 0,141 | 0,12 | 0,073 | 0.27 | 0,18 | 0,084 |
| | Acurácia | 0,56± 0,05 | 0,64±0,05 | **0,80 ± 0,05** | 0,60 ± 0,05 | 0,60 ± 0,05 | 0,74± 0,05 |
| | Características | 48 | 20 | 16 | 8 | 16 | 5 |
| RMSE | Resultado | 0,24 | 0,27 | 0,27 | 0,47 | 0,36 | 0,32 |
| | Acurácia | 0,70± 0,05 | 0,54±0,05 | 0,58± 0,05 | 0,50± 0,05 | 0,60± 0,05 | 0,63± 0,05 |
| | Características | 38 | 10 | 47 | 10 | 25 | 4 |
| MAE | Resultado | 0,10 | 0,11 | 0.17 | 0,19 | 0,21 | 0,09 |
| | Acurácia | 0,58± 0,05 | 0,53±0,05 | 0.54± 0,05 | 0,43± 0,05 | 0,54± 0,05 | 0,74± 0,05 |

Como é possível observar, a métrica MSE mostrou resultados superiores em relação ao RMSE e MAE.

A combinação que apresentou melhor desempenho para a predição do coeficiente angular e linear foi a RF selecionando as melhores características com o PS. As características selecionadas podem ser visualizadas nas tabelas (5 e 6).

*Tabela 5-Características utilizada para o desenvolvimento do modelo de regressão para o coeficiente angular*

| |
|---|
| *gelatin composition* |
| *wavelet_LHL_original_first_order_mean* |
| *wavelet_HHL__original_first_order_10Percentile* |
| *wavelet_HHL__original_firstorder_Minimum* |
| *wavelet_LLL_original_first_order_energy* |
| *Wavelet_LLL_original_first_order_totalEnergy* |

*Tabela 6 - Características utilizada para o desenvolvimento do modelo de regressão para o coeficiente linear*

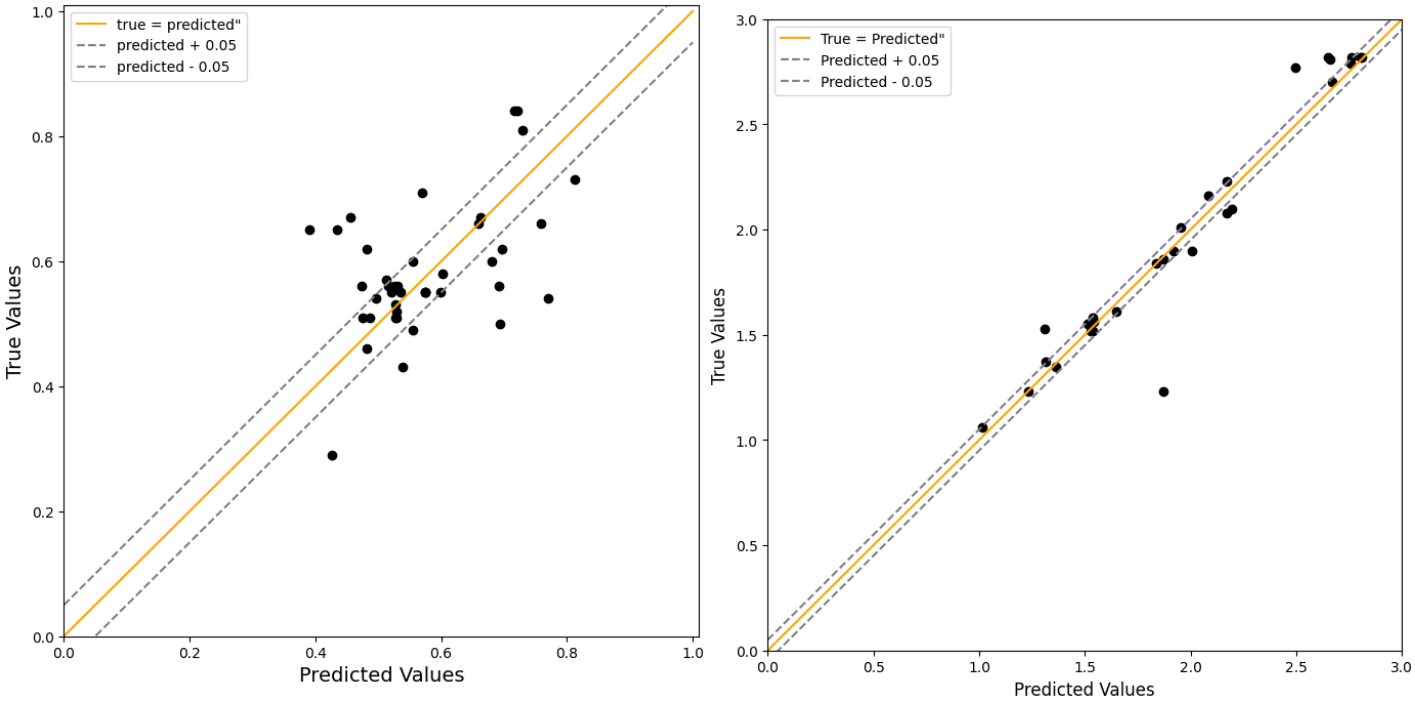
| |
|---|
| *wavelet-LHL_original_firstorder_Maximum* |
| *original_firstorder_Maximum* |
| *original_firstorder_Mean* |
| *original_firstorder_Media* |
| *original_firstorder_Minimum* |
| *original_firstorder_RootMeanSquared* |
| *wavelet-LHL_original_firstorder_Variance* |
| *wavelet-HLL_original_firstorder_Maximum* |
| *wavelet-LHL_original_firstorder_10Percentile* |
| *original_firstorder_10Percentile* |
| *original_firstorder_90Percentile* |
| *wavelet-LLL_original_firstorder_MeanAbsoluteDeviation* |
| *wavelet-LHL_original_firstorder_Minimum* |
| *wavelet-LLL_original_firstorder_Variance* |
| *wavelet-LLL_original_firstorder_RootMeanSquared* |
| *wavelet-LLL_original_firstorder_90Percentile* |
| *wavelet-LLL_original_firstorder_Maximum* |
| *wavelet-LLL_original_firstorder_Mean* |
| *wavelet-LLL_original_firstorder_Median* |
| *wavelet-LLL_original_firstorder_Minimum* |
| **R2_tubo_0gy* |
| *wavelet-LLL_original_firstorder_Energy* |
| *wavelet-HLL_original_firstorder_Minimum* |
| *wavelet-HLL_original_firstorder_Variance* |
| *wavelet-HHL_original_firstorder_MeanAbsoluteDeviation* |
| *wavelet-HHL_original_firstorder_Minimum* |
| *wavelet-HHL_original_firstorder_Variance* |
| *wavelet-LLL_original_firstorder_10Percentile* |

Portanto, esses grupos de características foram usados como dados de entradas nos

modelos de regressão finais para predição dos dois coeficientes. Assim, foi possível treinar o modelo no subconjunto de dado separado para esta finalidade e, após isso o modelo foi testado em um outro subcojunto teste, separado inicialmente e que não foi empregado no desenvolvimento do modelo.

A Figura 26 mostra a relação entre os valores reais e os preditos nos dados dos coeficientes angular (a) e linear (b) para o *dataset* 1. A linha central representa a curva exata dos valores reais e preditos e as linhas pontilhadas laterais indica o desvio de incerteza de ±5% Para o modelo de predição do coeficiente angular 77% dos dados estão dentro de ±5%, equanto que para o coeficiente linear esse percentual é de 80%.

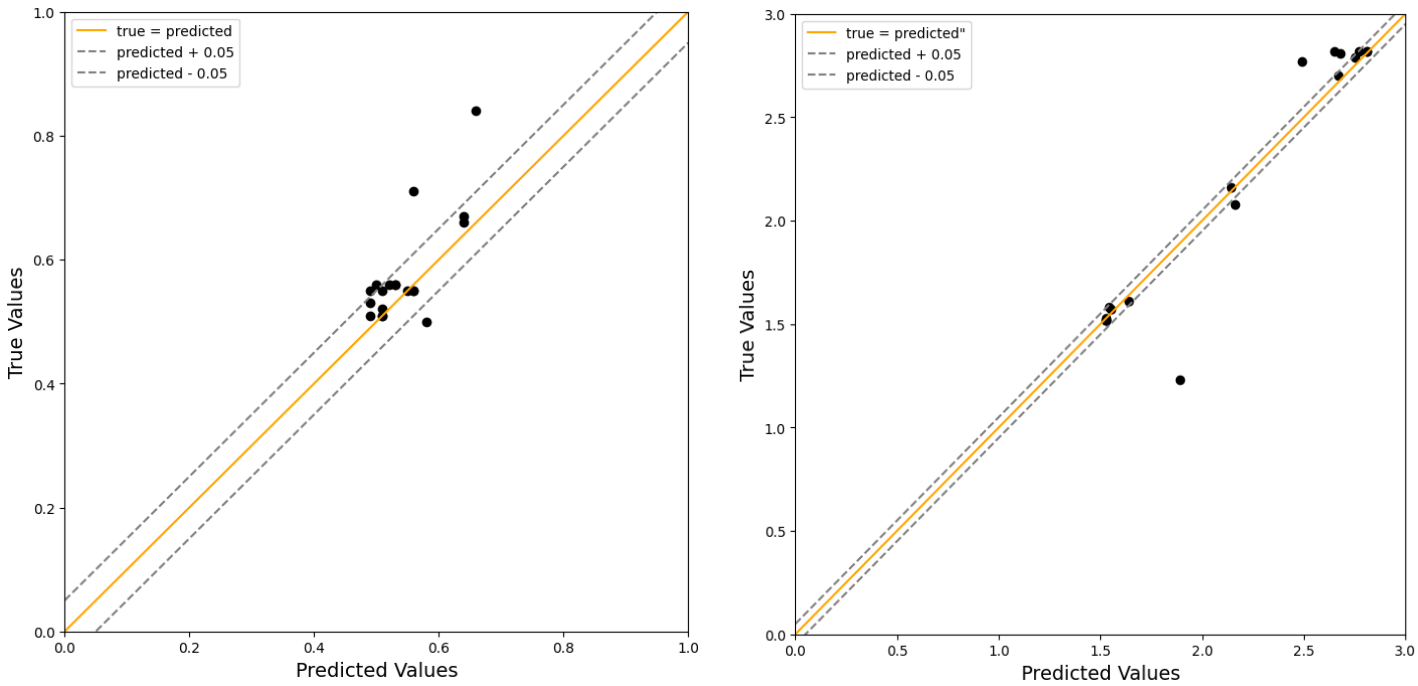
*Figura 26 - Gráfico dos valores reais x preditos para o Coeficiente angular (esquerda) e Coeficiente linear (direita) para o dataset 1*



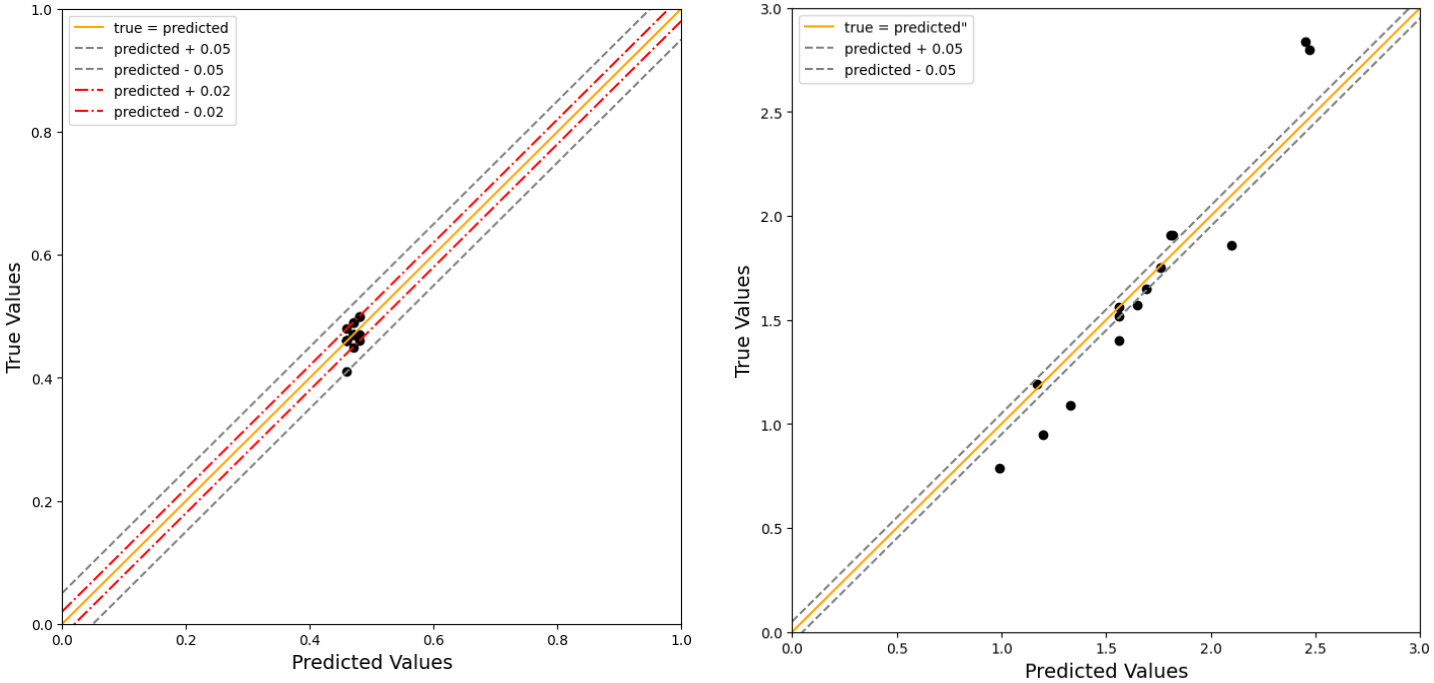

A aplicabilidade do modelo foi analisada utilizado o *dataset* 2. Este *dataset* é composto por amostras de géis que seguem o padrão de formulação dentro do precesso dosimétrico e não há variação nas sequências de aquisições das IRM. A fim de obter uma avaliação justa, foi comparado os resultados obtidos pelo *dataset* 1 com os dados equivalentes ao *dataset* 2, aplicando um filtro nos dados de teste do *dataset* 1 para obter um subconjunto composto apenas por amostras de géis que seguem o padrao de formulação dosimétrico e de aquisição das IRM. Para esse segundo grupo selecionado do *dataset* 1, o modelo obteve uma predição de 84% com

desvio de incerteza de ±5% para o coeficiente angular e 74% com desvio ±5% para o linear. A figura 28 mostra a comparação entre os valores preditos e reais desse subgrupo do *dataset* 1 assim como as predições do *dataset* 2. Para o grupo utilizado como forma de analisar a performance do modelo desenvolvido para predizer os coeficientes da curva de calibração, o resultado encontrado foi efetivamente satisfatório, todos os valores ficaram dentro de uma incerteza de ±5% para o coeficiente angular. Em uma análise mais restrita, um valor de desvio de incerteza de ±2% foi encontrado para mais de 94% dos dados, sendo esses 6% restantes, referente a um único ponto (Figura 27 (c)). O valor de MSE calculado para o dataset 2 foi de 2,84 x $10^{-4}$, menor valor comparado ao dataset 1(6,67 x $10^{-3}$). Já para o coeficiente linear, o *dataset* 1 obteve uma porcentagem de 74% dentro de um desvio de ±5%. Tendo desempenho um pouco menor quando comparado ao subgrupo do deste mesmo conjunto de dados para suas amostras semelhantes. Para a predição deste coeficiente, o modelo encontrou um MSE de 0.15, maior quando comparado ao encontrado no dataste 1 (0.073).

*Figura 27 - Gráficos dos valores reais x preditos para o subgrupo do dataset 1 para o coeficiente angular (a) e linear (b) e para as predições do dataset 2 (c) coeficiente angular e (d) coeficiente linear*




(a) (b)




(c) (d)

Fonte: O autor (2023)

Essa baixa performance apresentada pela aplicação da predição do Coeficiente linear no *dataset* 2 pode ser explicada pela distribuição de dados deste coeficiente. Como é possível visualizar na figura 18, esta distribuição não apresenta um perfil de distribuição normal mesmo depois do pré processamento dos dados, e a faixa dos dados do *dataset* 2 (figura 19 (b)) pertence a intervalos de pequenos do dataset 1, o que pode ter causado impacto no aprendizado do modelo para esse range. O que ocorre de maneira contraria para o coeficiente angular, uma vez que o intervalo de valores presentes no *dataset* 2 pertence à faixa que possui maior quantidade de amostras no *dataset* 1.

### *5.3 Resultados para os modelos de classificação*

Com o objetivo de complementar os resultados deste trabalho, foram desenvolvidos três modelos de classificação para predizer as características de composição das amostras de géis que saem do padrão de fabricação do dosímetro MAGIC-f.

Dentre as características de fabricação de géis que diferem dentro das amostras do conjunto de dados do dataset 1, apresentados na tabela 1, o considerado padrão apresenta:

gelatina com poder de gelificação de 250 bloom, ácido ascórbico e sulfato de cobre como agentes oxidantes e nenhuma alteração de luminosidade no ambiente. Assim, foram desenvolvidos três modelos para classificar se uma determinada amostra de gel está nessas condições e, caso contrário, se a amostra a ser analisada não apresentar essas características, pode se abrir espaço para discutir a razão do valor de sensibilidade encontrado para este lote (valor do coeficiente angular), servindo como um auxílio nos estudos e análises dos dosímetros químicos.

Para essa análise complementar, este trabalho optou por usar o modelo classificador de RF (*Random Forest Classifier)* e selecionar as melhores características com a biblioteca PS, uma vez que essas duas metodologias se adaptaram para esses dados, como é visto na seção 5.1 deste capítulo e obtiveram também uma boa performance para a classificação como será possível ver nos resultados desta seção.

O desenvolvimentos desses modelos seguiram a metodologia proposta para a predição dos coeficientes da curva de calibração dos dosímetros químicos, adaptando os parâmetros para um modelo de classificação. Em resumo, foram realizados o pré processamento dos dados, a tunagem de hiperparâmetros da floresta de decisão, seleção das melhores *características*, validação cruzada com 10 *folds*, divisão dos dados em 70% formando o conjunto de treino e, 30% o conjunto de teste, mas ao fim sendo o modelo avaliado com as métricas acurácia, precisão, revocação e *F1-score*, todos esses procedimentos descritos no capítulo 4.

Esta análise é uma continuação do desenvolvimento das predições dos coeficientes, portanto, ela é feita com a mesma divisão de dados (*random state*) feita para os modelos de regressão. Os três modelos desenvolvidos foram:

Modelo 1 – natureza do agente oxidante: hidroquinona ou ácido ascórbico e sulfato de cobre.

Modelo 2 - grau de gelificação (*bloom*) da gelatina utilizada no processo de produção do gel (25, 270 e 300 *bloom*).

Modelo 3 – se houve alguma alteração da luminosidade do ambiente durante algum processo desde a fabricação do gel até sua leitura.

Novamente, o *dataset* 1 é composto por 145 amostras como é descrito na tabela 1 e o *dataset* 2 contém 19 amostras e será utilizado como forma de aplicação dos modelos desenvolvidos, sendo esses géis todos dentro do considerado padrão de fabricação.

### 5.3.1 Modelo 1: agente oxidante hidroquinona ou ácido ascórbico e sulfato de cobre

As amostras que apresentam a hidroquinona como agente oxidante foram representados pela classe 1 e os géis que pertencem ao padrão de confecção foram representados pela classe 0.

Assim, o conjunto de dados do *dataset* 1 para esse modelo após o pré processamento, é composto por 99 amostras da classe 0 e 35 amostras da classe 1. Esse desbalanceamento foi tratado através da biblioteca *SMOTE,* resultando em 68 amostras para cada classe.

A biblioteca PS selecionou 12 características contendo peso suficiente para impactar a classificação do modelo, sendo elas exibidas na tabela 7. Alimentando o modelo com essas características e fazendo a tunagem dos hiperparâmetros da RFC, a acurácia do modelo é de 95% (tabela 8).

*Tabela 7 - Características selecionadas para o Modelo 1*

| |
|---|
| original_glrlm_GrayLevelNonUniformity |
| wavelet-HHL_original_glszm_SmallAreaLowGrayLevelEmphasis |
| wavelet-HHL_original_glszm_SmallAreaEmphasis |
| wavelet-LLH_original_firstorder_Skewness |
| wavelet-LLL_original_glrlm_GrayLevelNonUniformity |
| R2_tubo0 |
| original_firstorder_RobustMeanAbsoluteDeviation |
| original_firstorder_Variance |
| original_firstorder_InterquartileRange |
| original_firstorder_Range |
| wavelet-HLL_original_firstorder_Skewness |
| original_firstorder_MeanAbsoluteDeviation |

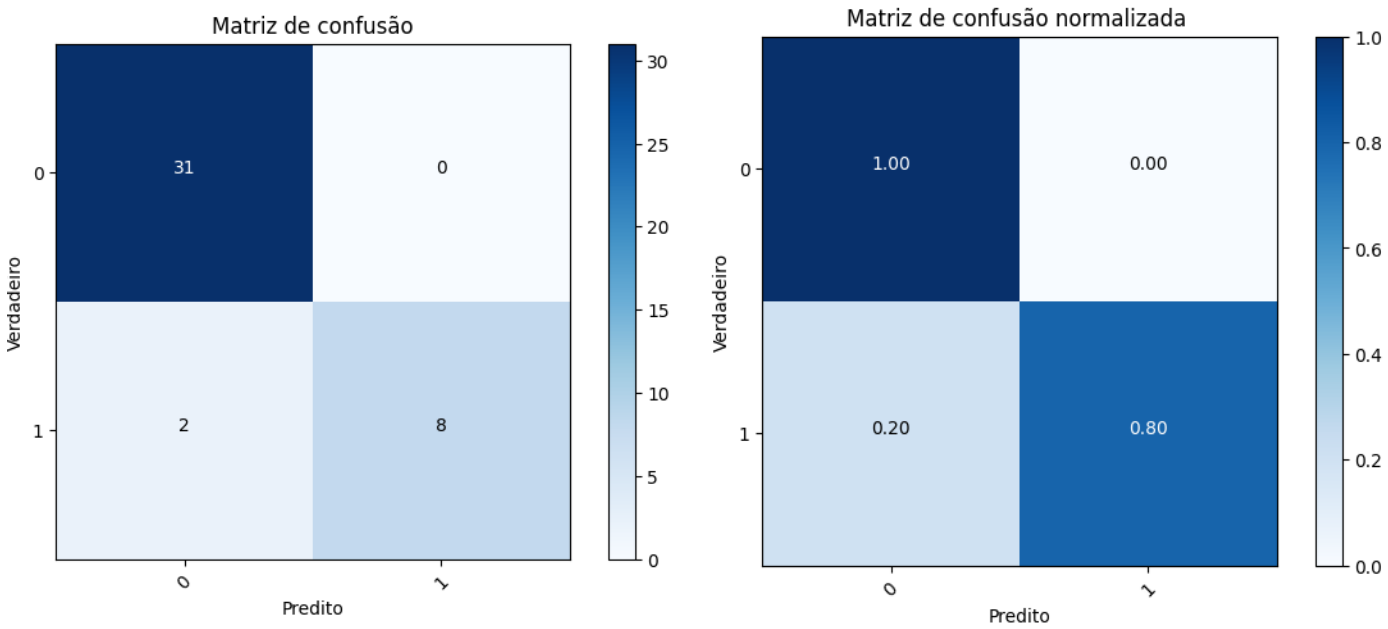
*Tabela 8 – Métricas obtidas para o Modelo 1*

| Acurácia | Parâmetros de avaliação | | |
|---|---|---|---|
| 0,95 | Precision | Revocação | F1-score |
| Classe 0 | 0,94 | 1,0 | 0,97 |
| Classe 1 | 1,0 | 0,80 | 0,89 |

A figura 28 traz a matriz de confusão (a) e sua versão normalizada (b) para as classificações feitas para o modelo. Como é possível notar, o modelo obteve êxito em classificar

os agentes oxidantes das amostras do *dataset* 1, classificando de maneira errônea apenas duas amostras da classe 1.

*Figura 28 - Matriz de confusão (esquerda) e matriz de confusão normalizada (direita) para o modelo 1*




(a) (b)

Fonte: O autor (2023)

5.3.2 Modelo 2: grau de gelificação da gelatina

Para este modelo, as amostras que seguem o padrão de fabricação (250 *bloom*) serão representadas pela classe 0 e as amostras que diferem disso (270 e 300 *bloom*) serão representadas pela classe 1, compondo um conjunto de dados, posteriormente ao pré-processamento, de 93 amostras da classe 0 e 60 da classe 1. A biblioteca *SMOTE* também foi utilizada para este modelo, mesmo não apresentando um desbalanceamento tão discrepante. Ao final, cada classe foi composta por 54 amostras

Com a seleção das características feita através do PS, 24 características foram selecionadas e estão apresentadas na tabela 9. As métricas utilizadas para avaliar este modelo, após o seu desenvolvimento, encontram-se na tabela 10, apresentando uma acurácia de 78%.

*Tabela 9- Características selecionadas para o modelo 2*

| |
|---|
| wavelet-LLH_original_glrlm_GrayLevelNonUniformity |

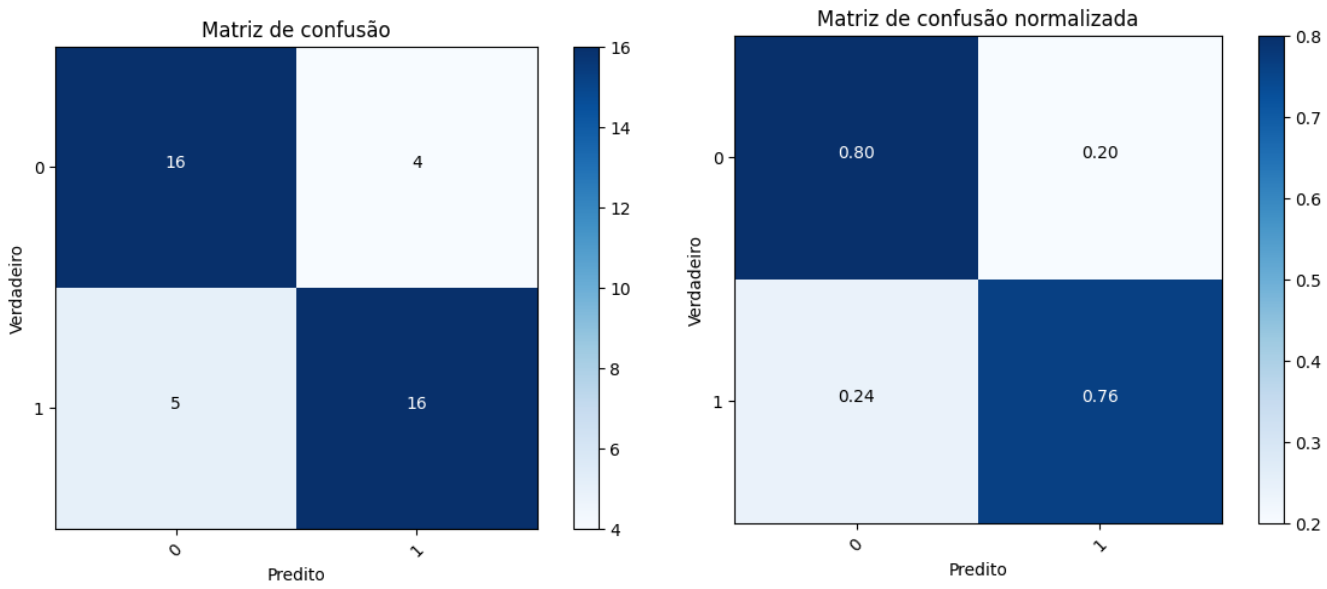
| |
|---|
| wavelet-LHL_original_gldm_DependenceNonUniformityNormalized |
| wavelet-LHL_original_glrlm_LongRunEmphasis |
| wavelet-LHL_original_glrlm_RunPercentage |
| wavelet-LHH_original_gldm_DependenceNonUniformityNormalized |
| wavelet-LHH_original_gldm_LargeDependenceEmphasis |
| wavelet-LHH_original_glrlm_LongRunEmphasis |
| wavelet-LHH_original_glrlm_RunPercentage |
| wavelet-LHH_original_glrlm_RunVariance |
| wavelet-HLL_original_gldm_HighGrayLevelEmphasis |
| wavelet-HLL_original_gldm_LargeDependenceLowGrayLevelEmphasis’ |
| wavelet-HLL_original_gldm_LowGrayLevelEmphasis |
| wavelet-HLL_original_glcm_Autocorrelation |
| wavelet-HLL_original_glcm_ClusterShade |
| wavelet-HLL_original_glcm_JointAverage |
| wavelet-HLL_original_glcm_SumAverage |
| wavelet-HLL_original_ngtdm_Busyness |
| wavelet-HLH_original_gldm_LowGrayLevelEmphasis |
| wavelet-HLH_original_glcm_Autocorrelation |
| avelet-HLH_original_glcm_ClusterShade |
| wavelet-HLH_original_glcm_JointAverage |
| wavelet-HLH_original_ngtdm_Busyness |
| wavelet-LLL_original_gldm_DependenceNonUniformityNormalized |
| wavelet-HLL_original_firstorder_Kurtosis |

*Tabela 10 – Métricas obtidas para o Modelo 2*

| Acurácia | Parâmetros de avaliação | | |
|---|---|---|---|
| 0,78 | Precision | Revocação | F1-score |
| Classe 0 | 0,76 | 0,80 | 0,78 |
| Classe 1 | 0,80 | 0,76 | 0,78 |

Através da figura 30 é possível concluir que o modelo conseguiu aprender de maneira eficiente a separar a gelatina padrão das outras duas com poder de gelificação diferente, obtendo um valor igual e acima de 75% para todas as métricas utilizadas nessa avaliação.

*Figura 30 - Matriz de confusão (esquerda) e matriz de confusão normalizada (direita) para o modelo 2*




(a) (b)

Fonte: O autor (2023)

5.3.3 Modelo 3: interferência de luz durante a fabricação e/ou leitura do gel

As amostras que não sofreram nenhum tipo de interferia de luz durante qualquer etapa de produção dos dosímetros serão chamadas de classe 1 e aquelas que sofreram algum tipo de mudanças na luminosidade, fugindo da normalidade, pertencerão à classe 0. Dessa forma, a classe 1 contém 103 amostras enquanto a classe 0 possui 31, após a etapa de pré-processamento. Essa diferença marcante entre as classes foi solucionada aplicando a técnica de balanceamento *SMOTE* com 84 amostras em cada uma das classes.

Para este terceiro modelo de classificação, a biblioteca PS selecionou 34 características com maior influência predição, sendo elas encontradas na tabela 11. Este modelo obteve uma acurácia de 80% (tabela 12).

*Tabela 11 - Características selecionadas para o modelo 3*

| |
|---|
| wavelet-HLL_original_glcm_JointEnergy |

wavelet-HHH_original_gldm_HighGrayLevelEmphasis

original_firstorder_Median

wavelet-LLL_original_firstorder_RootMeanSquared

wavelet-HLL_original_glcm_MaximumProbability

wavelet-HLL_original_glcm_ClusterShade

wavelet-LLL_original_firstorder_MeanAbsoluteDeviation

wavelet-HHH_original_glcm_MCC

wavelet-LHL_original_firstorder_Range

wavelet-HLL_original_firstorder_Range

original_firstorder_10Percentile

wavelet-HHH_original_glcm_ClusterShade

wavelet-HLL_original_glcm_SumSquares

wavelet-HHH_original_glcm_Imc2

wavelet-HHL_original_firstorder_Range

wavelet-HHL_original_firstorder_Minimum

original_firstorder_90Percentile

wavelet-LLL_original_firstorder_Minimum

wavelet-HHL_original_firstorder_Median

wavelet-HLL_original_firstorder_Energy

wavelet-HLL_original_firstorder_TotalEnergy

original_firstorder_Mean

wavelet-LHL_original_firstorder_Variance

wavelet-HHL_original_firstorder_Maximum

wavelet-HLH_original_glcm_ClusterShade

wavelet-LHL_original_firstorder_Energy

original_firstorder_Energy

wavelet-LLL_original_firstorder_Median

wavelet-LLL_original_firstorder_90Percentile

wavelet-LLL_original_firstorder_TotalEnergy

wavelet-HHL_original_firstorder_TotalEnergy

original_firstorder_TotalEnergy

wavelet-LLL_original_firstorder_Energy

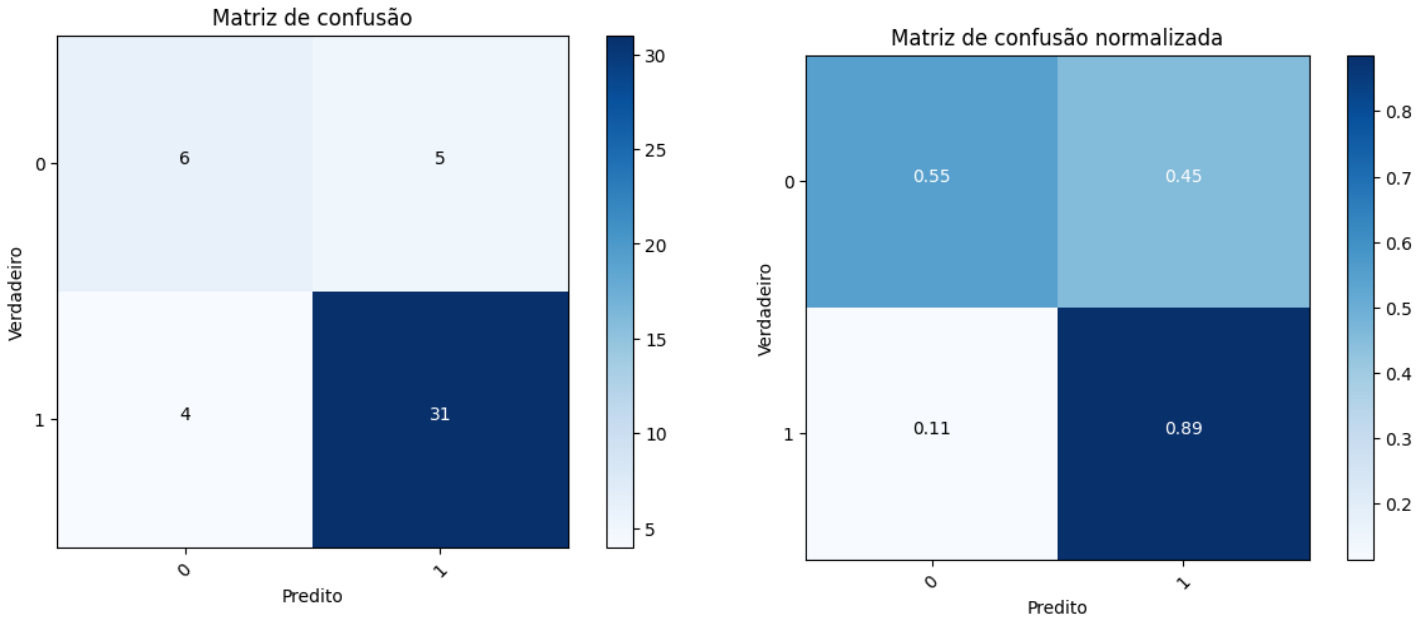
wavelet-LHL_original_firstorder_TotalEnergy

As métricas apresentadas na tabela 12 mostram que o modelo teve uma tendencia de aprendizado para a classe 1, apresentando valores mais baixos de precisão, revocação e *f1 – score* para a classe 0. Analisando a matriz de confusão da figura 31, é possível visualizar o motivo dos baixos valores para as métricas citadas, pois o modelo prediz bem próximo à metade dos dados erroneamente para a classe 0, podendo ser interpretado que o modelo não aprendeu muito bem a separar essas classe.

*Tabela 12 - Métricas obtidas para o Modelo 3*

| Acurácia | Parâmetros de avaliação | | |
|---|---|---|---|
| 0,80 | Precision | Revocação | F1-score |
| Classe 0 | 0,60 | 0,55 | 0,57 |
| Classe 1 | 0,86 | 0,89 | 0,87 |

*Figura 31 - Matriz de confusão (a) e matriz de confusão normalizada (b) para o modelo 3*



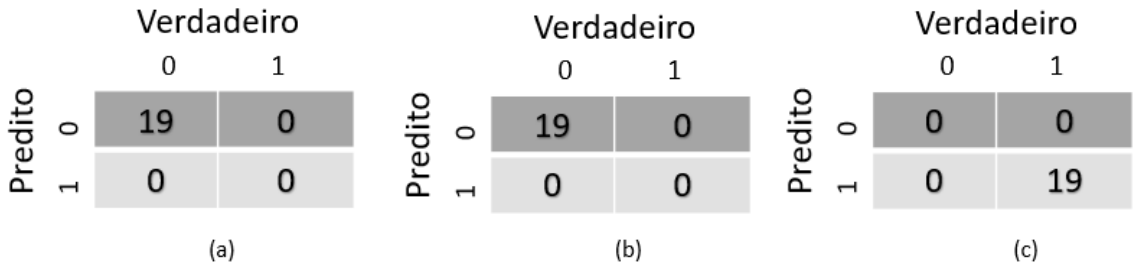

(a) (b)

Fonte: O autor (2023)

Após o desenvolvimentos dos 3 modelos de classificação, o *dataset* 2 foi utilizado novamente a fim de confirmar a aplicabilidade destes estudos. Com isso, os três modelos foram capazes de predizer todos as 19 amostras pertencentes às classes corretas, ressaltando que estes

dados correspondem apenas ao padrão de fabricação (Figura 32). Os modelos apresentaram aprendizado eficaz para as classes, com a exceção do modelo 3 para a classe 0, mas, mesmo assim, ele foi capaz de acertar o grupo onde foi aplicado. É possível ter uma tendência a predizer géis padrão, uma vez que a maioria das amostras dos conjuntos onde o modelo foi treinado apresentam esse comportamento, mas, por outro lado, as métricas mostram a capacidade destes modelos a lidarem com essas classes minoritárias em relação a falsos positivos e falsos negativos, com exceção, novamente, do modelo 3.

*Figura 32 - Resultado da aplicação dos modelos de classificação no dataset 2. (a) Matriz de confusão para o modelo1 (b) para o modelo 2 e (c) para o modelo 3*




Fonte: O outro (2023)

*5.4 Considerações finais*

Este trabalho apresenta uma metodologia para predizer a curva de calibração de um dosímetro químico através dos coeficientes angulares e lineares utilizando modelos de regressão de AM, além de trazer uma metodologia complementar para classificar as diferenças presentes nas amostras que compõe o conjunto de dados deste estudo. Dois modelos de regressão foram inicialmente considerados, RF e CB e três maneiras diferentes de selecionar o grupo de características que mais influenciam no aprendizado foram analisadas (MDI, RFE e PS). O modelo RF e o método PS foram a combinação em que o modelo mostrou a melhor performance alcançada para a predição de ambos os coeficientes (Tabela 3 e 4). Observa-se que o modelo CB obteve melhores resultados do que o modelo RF com os métodos de seleção de características MDI e RFE para o coeficiente angular, onde a biblioteca PS combinada com o modelo de regressão CB não foi eficaz em selecionar nenhum grupo de entrada e, talvez se essa combinação fosse eficiente em selecionar um grupo de características, poderia alcançar uma performance maior do que o resultado obtido pela RF, seguindo a expectativa que a literatura traz quando uma técnica de *boosting* é aplicada. Para o Coeficiente linear, o modelo CB

também atingiu resultados melhores do que a RF com as técnicas MDI e RFE, mas, não superou o modelo com a biblioteca PS. É possível observar também que, para os dois coeficientes, a biblioteca PS selecionou as características "*wavelet-LLL_original_firstorder_Energy*" e "*wavelet-HHL_original_firstorder_Minimum*" como importantes para suas predições.

No geral, os modelos apresentaram performance maior para o coeficiente angular quando comparado ao coeficiente linear. Esse fato pode ser explicado pela distribuição de dados dos dois coeficientes em questão (Figura 18). A distribuição de dados do coeficiente angular apresenta caráter mais gaussiano e um intervalo de valores com pouca dispersão (0.3 a 0.85 $s^{-1}$ $Gy^{-1}$ após o pré processamento). Já a distribuição do coeficiente linear se distancia evidentemente de um perfil de distribuição normal, além de apresentar um intervalo mais amplo ( 1 a 9.7Gy) onde a quantidade de dados pertencente a cada faixa deste intervalo apresenta grandes variações, o que pode influenciar o aprendizado do modelo.

Os modelos foram avaliados e desenvolvidos com base na métrica MSE pois, entre as três calculadas (MSE, RMSE e MAE) foi a que apresentou melhores valores e acurácia na porcentagem de acerto das predições. Para essa métrica, no dataset 1, o modelo obteve um valor de 6,67 x $10^{-3}$ para o coeficiente angular e 0,073 para o coeficiente linear, apresentando uma porcentagem de acerto de 77% com um desvio de ±5% e 80% destro do mesmo desvio, respectivamente (Figura 26).

Ao se desenvolver um modelo de aprendizado de máquina, todos os passos do procedimento a ser ensinado para a máquina, devem ser evidenciados e todas particularidades e individualidade do ambiente em que se trabalha devem ser atenciosamente descritos. A mesma ideia é aplicada para reprodução desses modelos, onde somente em condições similares às que foi treinado, o modelo apresentará resultados parecidos.

Neste trabalho, o modelo foi desenvolvido em um conjunto de dados heterogêneo, com variações na formulação dos géis e na sequência de aquisições das imagens de ressonância (Tabela 2) podendo ser aplicado de uma maneira mais genérica. Com essa variação, o *Dataset* 1 possui valores de sensibilidades entre 0.30 e 0.85 $s^{-1}$ $Gy^{-1}$. Entretanto, é necessário ter formulações, processos de produção dos dosímetros e sequencias de aquisição das IRM fixos para uma dosimetria otimizada. Este procedimento específico faz com que os géis apresentem valores de sensibilidades contantes em todas as aplicações dos dosímetros, que para o dataset 1, corresponde a faixa central dos valores de Coeficiente angular (Figura 19). Analisando a performance do modelo no *dataset* 2 é possível perceber que o modelo mostrou um bom desempenho com dados de outra máquina de RM para predizer o coeficiente angular

(sensibilidade dos géis), sendo este resultado superior ao resultado obtido pelo conjunto de teste do *dataset* 1 e para o subconjunto incluindo amostras formuladas e adquiridas dentro do padrão (Figura 28 - c). Esse resultado pode ser explicado uma vez que que os dados têm origem de um processo dosimétrico padrão de onde as amostras foram coletadas (laboratório DART – 3D do departamento de física da faculdade de filosofia, Ciências e Letras de Ribeirão Preto – USP), que resultou em sensibilidades dos géis pertencentes à faixa central dos valores (0.41 a 0.50 $s^{-1}Gy^{-1}$) usados no conjunto de treino do desenvolvimento do modelo.

A aplicação de modelos de AM para predizer a curva de calibração com base em amostras de géis não irradiadas pode poupar tempo durante experimentos de irradiação e aquisição de imagens. O modelo desenvolvido neste trabalho pode ser aplicado na dosimetria, mas deve ser considerado os desvios de incertezas apresentados. Contudo, assim como em todos os estudo de AM, a performance dos modelos pode ser melhorada se um conjunto de dados mais amplo for aplicado para desenvolver o algoritmo.

Os modelos de classificação foram desenvolvidos a partir da RF e com suas características selecionadas usando o PS. O modelo 1 e 2 mostrou aprendizado satisfatório para predizer as classes com acurácia de 95% e 78%, respectivamente, e métricas que analisam predições que podem ter sidos feitas para classes erradas ( tabela 8 e 10) bem balanceadas entre as classes, o que mostra que o modelo realmente aprendeu a separá-las. Já o modelo 3, mesmo apresentando uma acurácia de 80%, quando se analisa as métricas *Precision* e *recall* (tabela 12) nota-se que o modelo se confundiu bastante para classificar as amostras, tendo um aprendizado um pouco inferior aos outros dois modelos.

Aplicando esses modelos no dataset 2, pôde-se comprovar a eficácia dos três modelos em predizer todas as 19 amostras do dataset de maneira correta em todas as classes a que esses dados pertencem.

# Capítulo 6 - Conclusão

O modelo de regressão RF desenvolvido neste trabalho foi capaz de predizer os coeficientes Coeficiente angular e Coeficiente linear com valores de MSE de 6,67 x $10^{-3}$ englobando 77% das predições em um intervalo de incerteza de ±5% para o coeficiente angular e de 0,073 contendo 80% das predições dentro do mesmo desvio para o coeficiente linear. Este modelo foi aplicado para um segundo conjunto de dados composto de amostras de géis adquiridos em uma máquina diferente, géis estes feito a partir do processo de dosimetria padrão, apresentando valores de MSE de 2,84 x $10^{-2}$ e 0.15 para os coeficientes angular e linear, respectivamente., mostrando uma melhora das predições para esta faixa central dos valores de sensibilidade (0.41 a 0.50 $s^{-1}Gy^{-1}$). Mesmo a predição do coeficiente linear não tendo apresentado esta melhora para o dataset 2, a sensibilidade dos dosímetros é analisada em relação ao coeficiente angular, o que permite que o modelo desenvolvido seja aplicado para predizer amplas faixas de sensibilidades com performances melhores nos valores centrais, correspondendo ao padrão do processo dosimétrico.

Os modelos de classificação desenvolvidos também apresentaram boa performance, com acurácia de 95%, 78% e 80% para os três modelos, respectivamente. Quando aplicados nos *dataset* 2, eles foram capazes de acertar a classe de todas as amostras. Essa classificação pode auxiliar a análise dos dosímetros químicos, uma vez que ela pode identificar nas amostras algum componente fora do padrão de fabricação que pode influenciar a sensibilidade dos géis.

A dosimetria gel é uma técnica dosimétrica tridimensional que embora apresente bons resultados, ainda é pouco utilizada na prática clínica. O desenvolvimento de ferramentas desse tipo, que tem a capacidade de acelerar o processo de dosimetria gel são bem vindas e podem ajudar a ampliar a empregabilidade dessa técnica na rotina.

## Capítulo 7 – Referências Bibliográficas


Acurio, E. S. R., Lizar, J. C., Arruda, G. V., & Pavoni, J. F. (2021). Technical Note: Three-dimensional QA of simultaneous integrated boost radiotherapy treatments by a dose-volume histogram methodology and its comparison with 3D gamma results. *Medical Physics*, *48*(6), 3208–3215. https://doi.org/10.1002/mp.14859

Amadasun, M., & King, R. (1989). Textural features corresponding to textural properties. *IEEE Transactions on Systems, Man, and Cybernetics*, *19*(5), 1264–1274. https://doi.org/10.1109/21.44046

Baldock, C., Burford, R. P., Billingham, N., Wagner, G. S., Patval, S., Badawi, R. D., & Keevil, S. F. (1998). Experimental procedure for the manufacture and calibration of polyacrylamide gel (PAG) for magnetic resonance imaging (MRI) radiation dosimetry. *Physics in Medicine and Biology*, *43*(3), 695–702. https://doi.org/10.1088/0031-9155/43/3/019

Baldock, C., De Deene, Y., Doran, S., Ibbott, G., Jirasek, A., Lepage, M., McAuley, K. B., Oldham, M., & Schreiner, L. J. (2010a). Polymer gel dosimetry. *Physics in Medicine and Biology*, *55*(5), R1–R63. https://doi.org/10.1088/0031-9155/55/5/R01

Berg, A., Ertl, A., & Moser, E. (2001). High resolution polymer gel dosimetry by parameter selective MR-microimaging on a whole body scanner at 3 T. *Medical Physics*, *28*(5), 833–843. https://doi.org/10.1118/1.1358304

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning* (1st ed.). Springer.

Borra, S., & Di Ciaccio, A. (2010). Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods. *Computational Statistics & Data Analysis*, *54*(12), 2976–2989. https://doi.org/10.1016/j.csda.2010.03.004

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2

Carbonell, J. G., Michalski, R. S., & Mitchell, T. M. (1983). AN OVERVIEW OF MACHINE LEARNING. In *Machine Learning* (pp. 3–23). Elsevier. https://doi.org/10.1016/B978-0-08-051054-5.50005-4

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Cho, B. (2018). Intensity-modulated radiation therapy: a review with a physics perspective. *Radiation Oncology Journal*, *36*(1), 1–10. https://doi.org/10.3857/roj.2018.00122

Dablain, D., Krawczyk, B., & Chawla, N. V. (2022). DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data. *IEEE Transactions on Neural Networks and Learning Systems*, 1–15. https://doi.org/10.1109/TNNLS.2021.3136503

Dhiviyaraj Kalaiselven, S. K., & Emmanvel Raj, J. J. S. (2012). Polymer Gel Dosimetry for Radiation Therapy. In *Modern Practices in Radiation Therapy*. InTech. https://doi.org/10.5772/33828

Fernandes, J., Carneiro, A., Araújo, D., Elias, J., Ribeiro, L., Santos, A., & Baffa, O. (2003). Desenvolvimento de Softwares para o Estudo da Relaxometria em Imagens por Ressonância Magnética. *Anais Do VIII Congresso Brasileiro de Física Médica*.

Fernandes, J. P., Pastorello, B. F., de Araujo, D. B., & Baffa, O. (2008). Formaldehyde increases MAGIC gel dosimeter melting point and sensitivity. *Physics in Medicine and Biology*, *53*(4), N53–N58. https://doi.org/10.1088/0031-9155/53/4/N04

Gallo, S., & Locarno, S. (2023). Gel Dosimetry. *Gels*, *9*(4), 311. https://doi.org/10.3390/gels9040311

Galloway, M. M. (1975). Texture analysis using gray level run lengths. *Computer Graphics and Image Processing*, *4*(2), 172–179. https://doi.org/10.1016/S0146-664X(75)80008-6

Gore, J. C., & Kang, Y. S. (1984). Measurement of radiation dose distributions by nuclear magnetic resonance (NMR) imaging. *Physics in Medicine and Biology*, *29*(10), 1189–1197. https://doi.org/10.1088/0031-9155/29/10/002

Gulin Andrey, & et al. (n.d.). *CatBoost for Apache Spark API documentation*.

Haibo, H., Bai, Y., & Garcia, E. A. (2013). *Imbalanced Learning: Foundations, Algorithms, and Applications* (1st ed.). Wiley-IEEE Press.

Haralick, R. M., Shanmugam, K., & Dinstein, I. (1973). Textural Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-3*(6), 610–621. https://doi.org/10.1109/TSMC.1973.4309314

Jabeur, S. Ben, Gharib, C., Mefteh-Wali, S., & Arfi, W. Ben. (2021). CatBoost model and artificial intelligence techniques for corporate failure prediction. *Technological Forecasting and Social Change*, *166*, 120658. https://doi.org/10.1016/j.techfore.2021.120658

Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis*, *53*(11), 3735–3745. https://doi.org/10.1016/j.csda.2009.04.009

Larue, R. T. H. M., Defraene, G., De Ruysscher, D., Lambin, P., & van Elmpt, W. (2017). Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures. *The British Journal of Radiology*, *90*(1070), 20160665. https://doi.org/10.1259/bjr.20160665

Liu, C. H. B., Chamberlain, B. P., Little, D. A., & Cardoso, Â. (2017). *Generalising Random Forest Parameter Optimisation to Include Stability and Cost* (pp. 102–113). https://doi.org/10.1007/978-3-319-71273-4_9

Liu, H., & Motoda, H. (2013). *Feature Selection for Knowledge Discovery and Data Mining*. Springer.

Lizar, J. C., Volpato, K. C., Brandão, F. C., da Silva Guimarães, F., Arruda, G. V., & Pavoni, J. F. (2021). Tridimensional dose evaluation of the respiratory motion influence on breast radiotherapy treatments using conformal radiotherapy, forward IMRT, and inverse IMRT planning techniques. *Physica Medica*, *81*, 60–68. https://doi.org/10.1016/j.ejmp.2020.11.036

Mallat, S. G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *11*(7), 674–693. https://doi.org/10.1109/34.192463

Maryanski, M. J., Gore, J. C., Kennan, R. P., & Schulz, R. J. (1993). NMR relaxation enhancement in gels polymerized and cross-linked by ionizing radiation: A new

approach to 3D dosimetry by MRI. *Magnetic Resonance Imaging*, *11*(2), 253–258. https://doi.org/10.1016/0730-725X(93)90030-H

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press.

Olsson, L. E., Westrin, B. A., Fransson, A., & Nordell, B. (1992). Diffusion of ferric ions in agarose dosimeter gels. *Physics in Medicine and Biology*, *37*(12), 2243–2252. https://doi.org/10.1088/0031-9155/37/12/006

Pavoni, J. F., & Baffa, O. (2012). An evaluation of dosimetric characteristics of MAGIC gel modified by adding formaldehyde (MAGIC-f). *Radiation Measurements*, *47*(11–12), 1074–1082. https://doi.org/10.1016/j.radmeas.2012.10.004

Pavoni, J. F., Neves-Junior, W. F. P., da Silveira, M. A., Haddad, C. M. K., & Baffa, O. (2017). Evaluation of a composite Gel-Alanine phantom on an end-to-end test to treat multiple brain metastases by a single isocenter VMAT technique. *Medical Physics*, *44*(9), 4869–4879. https://doi.org/10.1002/mp.12400

Pedregosa FABIAN, F., Michel, V., Grisel OLIVIERGRISEL, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Pedregosa, F., Varoquaux, G., Gramfort, A., Thirion, B., Grisel, O., Dubourg, V., Passos, A., Brucher, M., Perrot andÉdouardand, M., Duchesnay, andÉdouard, & Duchesnay EDOUARDDUCHESNAY, Fré. (2011). Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot. In *Journal of Machine Learning Research* (Vol. 12). http://scikit-learn.sourceforge.net.

Schreiner, L. J. (2004). Review of Fricke gel dosimeters. *Journal of Physics: Conference Series*, *3*, 9–21. https://doi.org/10.1088/1742-6596/3/1/003

Segura, D., Khatib, E. J., & Barco, R. (2022). Dynamic Packet Duplication for Industrial URLLC. *Sensors*, *22*(2), 587. https://doi.org/10.3390/s22020587

Shelke, M., Desmukh, P., & Sahndilya Vijaya. (2017). A Review on Imbalanced Data Handling Using Undersampling and Oversampling Technique. *International Journal of Recent Trends in Engineering and Research*, *3*(4), 444–449. https://doi.org/10.23883/IJRTER.2017.3168.0UWXM

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees,

bagging, and random forests. *Psychological Methods*, *14*(4), 323–348. https://doi.org/10.1037/a0016973

Su, X., Xu, Y., Tan, Z., Wang, X., Yang, P., Su, Y., Jiang, Y., Qin, S., & Shang, L. (2020). Prediction for cardiovascular diseases based on laboratory data: An analysis of random forest model. *Journal of Clinical Laboratory Analysis*, *34*(9). https://doi.org/10.1002/jcla.2342

Sun, C., & Wee, W. G. (1983). Neighboring gray level dependence matrix for texture classification. *Computer Vision, Graphics, and Image Processing*, *23*(3), 341–352. https://doi.org/10.1016/0734-189X(83)90032-4

Tan, P.-N., Steinbach, M., & Kumar, V. (2007). *Introduction to Data Mining*. Pearson Education.

Thibault, G., Angulo, J., & Meyer, F. (2014). Advanced Statistical Matrices for Texture Characterization: Application to Cell Classification. *IEEE Transactions on Biomedical Engineering*, *61*(3), 630–637. https://doi.org/10.1109/TBME.2013.2284600

van Timmeren, J. E., Cester, D., Tanadini-Lang, S., Alkadhi, H., & Baessler, B. (2020). Radiomics in medical imaging—"how-to" guide and critical reflection. *Insights into Imaging*, *11*(1), 91. https://doi.org/10.1186/s13244-020-00887-2

Vandecasteele, J., & De Deene, Y. (2013). On the validity of 3D polymer gel dosimetry: II. Physico-chemical effects. *Physics in Medicine and Biology*, *58*(1), 43–61. https://doi.org/10.1088/0031-9155/58/1/43

Varoquaux, G., Buitinck, L., Louppe, G., Grisel, O., Pedregosa, F., & Mueller, A. (2015). Scikit-learn. *GetMobile: Mobile Computing and Communications*, *19*(1), 29–33. https://doi.org/10.1145/2786984.2786995

Verhaeghe, J., Van Der Donckt, J., Ongenae, F., & Van Hoecke, S. (2023). *Powershap: A Power-Full Shapley Feature Selection Method* (pp. 71–87). https://doi.org/10.1007/978-3-031-26387-3_5

Yan, K., & Zhang, D. (2015). Feature selection and analysis on correlated gas sensor data with recursive feature elimination. *Sensors and Actuators B: Chemical*, *212*, 353–363. https://doi.org/10.1016/j.snb.2015.02.025

# ANEXO A

Características selecionada através da métrica MSE

*A 1- Características selecionadas com o MDI para o coeficiente angular com o modelo CB*

| |
|---|
| *wavelet-LLH_original_firstorder_Range* |
| *wavelet-HLH_original_firstorder_RootMeanSquared* |
| *wavelet-HLH_original_firstorder_Variance* |
| *original_firstorder_TotalEnergy* |
| *wavelet-HHL_original_gldm_SmallDependenceLowGrayLevelEmphasis* |
| *wavelet-HHH_original_gldm_SmallDependenceLowGrayLevelEmphasis* |
| *wavelet-HHH_original_firstorder_RobustMeanAbsoluteDeviation* |
| *wavelet-HHL_original_firstorder_TotalEnergy* |
| *wavelet-HLH_original_firstorder_Minimum* |
| *wavelet-HHH_original_firstorder_10Percentile* |
| *wavelet-LLL_original_firstorder_TotalEnergy* |
| *composição_gelatina* |

*A 2- Características selecionadas com o MDI para o coeficiente angular com o modelo RF*

| |
|---|
| *wavelet-HLL_original_firstorder_Mean* |
| *wavelet-LLL_original_firstorder_Range* |
| *wavelet-HHL_original_glrlm_RunLengthNonUniformityNormalized* |
| *wavelet-LHL_original_firstorder_Range* |
| *wavelet-LHL_original_firstorder_RootMeanSquared* |
| *composição_gelatina* |
| *wavelet-HHL_original_firstorder_Range* |
| *wavelet-HHH_original_glrlm_LongRunEmphasis* |
| *wavelet-HLL_original_ngtdm_Busyness* |
| *wavelet-LHH_original_glcm_SumEntropy* |
| *wavelet-LHL_original_glcm_ClusterProminence* |
| *wavelet-HHL_original_firstorder_Variance* |
| *wavelet-HHL_original_firstorder_TotalEnergy* |
| *wavelet-LHL_original_firstorder_Minimum* |
| *wavelet-LHL_original_firstorder_Kurtosis* |
| *wavelet-LLL_original_firstorder_RobustMeanAbsoluteDeviation* |
| *wavelet-LHL_original_firstorder_Energy* |
| *wavelet-HHH_original_glszm_LargeAreaEmphasis* |
| *wavelet-LHH_original_glcm_ClusterProminence* |
| *wavelet-LHL_original_gldm_DependenceNonUniformityNormalized* |
| *original_firstorder_90Percentile* |
| *wavelet-LHH_original_glcm_ClusterTendency* |
| *original_firstorder_Mean* |
| *wavelet-HLL_original_firstorder_Range* |
| *original_firstorder_10Percentile* |

| |
|---|
| *wavelet-HHL_original_firstorder_MeanAbsoluteDeviation* |
| *wavelet-LHL_original_firstorder_Variance* |
| *wavelet-HLL_original_firstorder_MeanAbsoluteDeviation* |
| *original_firstorder_RootMeanSquared* |
| *wavelet-LLL_original_firstorder_MeanAbsoluteDeviation* |
| *wavelet-LHL_original_firstorder_Mean* |
| *wavelet-HHL_original_firstorder_RootMeanSquared* |
| *wavelet-LLL_original_firstorder_Maximum* |
| *original_firstorder_Median* |
| *wavelet-HHL_original_firstorder_Minimum* |
| *wavelet-LLL_original_firstorder_10Percentile* |
| *original_firstorder_Maximum* |
| *wavelet-LLL_original_firstorder_Mean* |
| *R2_tubo0* |
| *wavelet-LLL_original_firstorder_90Percentile* |
| *wavelet-LLL_original_firstorder_TotalEnergy* |
| *wavelet-LLL_original_firstorder_Median* |
| *original_firstorder_TotalEnergy* |
| *wavelet-LLL_original_firstorder_RootMeanSquared* |
| *original_firstorder_Energy* |
| *wavelet-LLL_original_firstorder_Energy* |
| *wavelet-LLL_original_firstorder_Minimum* |
| *wavelet-HHL_original_firstorder_Energy* |
| *wavelet-HHL_original_firstorder_10Percentile* |
| *original_firstorder_Minimum* |

*A 3 - Características selecionadas com o MDI para o coeficiente linear t com o modelo CB*

| |
|---|
| *wavelet-HHL_original_firstorder_Range* |
| *original_firstorder_Range* |
| *wavelet-LLL_original_firstorder_Variance* |
| *original_firstorder_10Percentile* |
| *R2_tubo0* |
| *wavelet-LLL_original_firstorder_90Percentile* |
| *wavelet-LLH_original_firstorder_Variance* |
| *wavelet-HHL_original_firstorder_Variance* |
| *wavelet-LHH_original_firstorder_Variance* |
| *original_firstorder_Energy* |
| *wavelet-LLL_original_firstorder_RootMeanSquared* |
| *wavelet-LLH_original_firstorder_RobustMeanAbsoluteDeviation* |

*A 4 - Características selecionadas com o MDI para o coeficiente linear com o modelo RF*

| |
|---|
| *original_firstorder_10Percentile* |
| *original_firstorder_Range* |
| *original_firstorder_90Percentile* |
| *wavelet-LLL_original_firstorder_Variance* |

| |
|---|
| *wavelet-HHL_original_firstorder_Minimum* |
| *wavelet-HHL_original_firstorder_90Percentile* |
| *wavelet-LLL_original_firstorder_Maximum* |
| *wavelet-HHL_original_firstorder_RootMeanSquared* |
| *wavelet-LHL_original_firstorder_Range* |
| *original_firstorder_TotalEnergy* |
| *wavelet-LHL_original_firstorder_10Percentile* |
| *wavelet-HLL_original_firstorder_10Percentile* |
| *wavelet-HLL_original_firstorder_Variance* |
| *R2_tubo0* |
| *wavelet-HHL_original_firstorder_MeanAbsoluteDeviation* |
| *wavelet-HLL_original_firstorder_RootMeanSquared* |
| *wavelet-HLL_original_firstorder_MeanAbsoluteDeviation* |
| *wavelet-LHL_original_firstorder_RobustMeanAbsoluteDeviation* |
| *wavelet-LHL_original_firstorder_TotalEnergy* |
| *wavelet-LHL_original_firstorder_Energy* |
| *original_firstorder_Energy* |
| *wavelet-LLL_original_firstorder_RootMeanSquared* |
| *original_firstorder_Minimum* |
| *wavelet-LLL_original_firstorder_TotalEnergy* |
| *wavelet-LHL_original_firstorder_Minimum* |
| *wavelet-HLL_original_firstorder_Minimum* |
| *wavelet-LLL_original_firstorder_Mean* |
| *wavelet-HHL_original_firstorder_Maximum* |
| *original_firstorder_Maximum* |
| *wavelet-LHL_original_firstorder_Variance* |
| *wavelet-HLL_original_firstorder_TotalEnergy* |
| *wavelet-LLL_original_firstorder_90Percentile* |
| *wavelet-HHL_original_firstorder_Energy* |
| *wavelet-HHL_original_firstorder_Variance* |
| *original_firstorder_Median* |
| *wavelet-LLL_original_firstorder_Energy* |
| *original_firstorder_Mean* |
| *wavelet-LHL_original_firstorder_MeanAbsoluteDeviation* |
| *wavelet-LLL_original_firstorder_MeanAbsoluteDeviation* |
| *wavelet-LLL_original_firstorder_Median* |
| *wavelet-HLL_original_firstorder_Energy* |
| *wavelet-HLL_original_firstorder_Range* |

*A 5 - Tabela 17 - Características selecionadas com o RFE para o coeficiente angular com o modelo CB*

| |
|---|
| *composição_gelatina* |
| *hidroquinona* |
| *original_firstorder_10Percentile* |
| *original_firstorder_90Percentile* |
| *original_firstorder_Energy* |
| *original_firstorder_InterquartileRange* |
| *original_firstorder_Kurtosis* |
| *original_firstorder_Maximum* |

| |
|---|
| *original_firstorder_MeanAbsoluteDeviation* |
| *original_firstorder_Minimum* |

*A 6- Características selecionadas com o RFE para o coeficiente angular e com o modelo RF*

| |
|---|
| *original_firstorder_10Percentile* |
| *original_firstorder_90Percentile* |
| *original_firstorder_InterquartileRange* |
| *original_firstorder_Median* |
| *original_firstorder_Minimum* |
| *original_firstorder_RootMeanSquared* |
| *original_firstorder_TotalEnergy'* |
| *wavelet-LLH_original_glrlm_RunLengthNonUniformity* |
| *wavelet-LLH_original_glrlm_RunEntropy* |
| *original_firstorder_Energy* |

*A 7- Características selecionadas com o RFE para o coeficiente linear com o modelo CB*

| |
|---|
| *wavelet-LLH_original_firstorder_TotalEnergy* |
| *wavelet-LLH_original_firstorder_Variance* |
| *wavelet-LLH_original_gldm_DependenceNonUniformity* |
| *wavelet-LLH_original_gldm_DependenceVariance* |
| *wavelet-LLH_original_gldm_GrayLevelNonUniformity* |
| *wavelet-LLH_original_gldm_GrayLevelVariance* |
| *wavelet-LLH_original_gldm_SmallDependenceLowGrayLevelEmphasis* |
| *wavelet-LLH_original_glcm_InverseVariance* |
| *wavelet-LLH_original_glrlm_HighGrayLevelRunEmphasis* |
| *wavelet-LLH_original_glrlm_RunLengthNonUniformity* |

*A 8 - Características selecionadas com o RFE para o coeficiente linear com o modelo RF*

| |
|---|
| *original_firstorder_10Percentile* |
| *original_firstorder_90Percentile* |
| *original_firstorder_Maximum* |
| *original_firstorder_Mean'* |
| *original_firstorder_Median* |

*A 9- Características selecionadas com o PS para o coeficiente angular com o modelo RF*

| |
|---|
| *gelatin composition* |
| *wavelet_LHL_original_first_order_mean* |
| *wavelet_HHL__original_first_order_10Percentile* |
| *wavelet_HHL__original_firstorder_Minimum* |
| *wavelet_LLL_original_first_order_energy* |
| *Wavelet_LLL_original_first_order_totalEnergy* |

*A 10- Características selecionadas com o PS para o coeficiente linear com o modelo RF*

| |
|---|
| *wavelet-LHL_original_firstorder_Maximum* |
| *original_firstorder_Maximum* |
| *original_firstorder_Mean* |
| *original_firstorder_Media* |
| *original_firstorder_Minimum* |
| *original_firstorder_RootMeanSquared* |
| *wavelet-LHL_original_firstorder_Variance* |
| *wavelet-HLL_original_firstorder_Maximum* |
| *wavelet-LHL_original_firstorder_10Percentile* |
| *original_firstorder_10Percentile* |
| *original_firstorder_90Percentile* |
| *wavelet-LLL_original_firstorder_MeanAbsoluteDeviation* |
| *wavelet-LHL_original_firstorder_Minimum* |
| *wavelet-LLL_original_firstorder_Variance* |
| *wavelet-LLL_original_firstorder_RootMeanSquared* |
| *wavelet-LLL_original_firstorder_90Percentile* |
| *wavelet-LLL_original_firstorder_Maximum* |
| *wavelet-LLL_original_firstorder_Mean* |
| *wavelet-LLL_original_firstorder_Median* |
| *wavelet-LLL_original_firstorder_Minimum* |
| **R2_tubo_0gy* |
| *wavelet-LLL_original_firstorder_Energy* |
| *wavelet-HLL_original_firstorder_Minimum* |
| *wavelet-HLL_original_firstorder_Variance* |
| *wavelet-HHL_original_firstorder_MeanAbsoluteDeviation* |
| *wavelet-HHL_original_firstorder_Minimum* |
| *wavelet-HHL_original_firstorder_Variance* |
| *wavelet-LLL_original_firstorder_10Percentile* |

*A 11 - Características selecionadas com o PS para o coeficiente linear com o modelo CB*

| |
|---|
| *original_firstorder_10Percentile* |
| *original_firstorder_90Percentile* |
| *original_firstorder_Maximum* |
| *original_firstorder_Mean* |
| *original_firstorder_Median* |
| *original_firstorder_Minimum* |
| *original_firstorder_RootMeanSquared* |
| *wavelet-LHL_original_firstorder_10Percentile* |
| *wavelet-LHL_original_firstorder_Maximum* |
| *wavelet-LHL_original_firstorder_Minimum* |
| *wavelet-LHL_original_firstorder_Variance* |
| *wavelet-HLL_original_firstorder_Maximum* |
| *wavelet-HLL_original_firstorder_Minimum* |
| *wavelet-HLL_original_firstorder_Range* |
| *wavelet-HLL_original_glszm_ZoneVariance* |
| *wavelet-HHL_original_firstorder_Maximum* |
| *wavelet-HHL_original_firstorder_MeanAbsoluteDeviation* |
| *wavelet-HHL_original_firstorder_Minimum* |
| *wavelet-HHL_original_firstorder_Variance* |
| *wavelet-LLL_original_firstorder_10Percentile* |
| *wavelet-LLL_original_firstorder_90Percentile* |
| *wavelet-LLL_original_firstorder_Energy* |
| *wavelet-LLL_original_firstorder_Maximum* |
| *wavelet-LLL_original_firstorder_MeanAbsoluteDeviation* |
| *wavelet-LLL_original_firstorder_Mean* |
| *wavelet-LLL_original_firstorder_Median* |
| *wavelet-LLL_original_firstorder_Minimum* |
| *wavelet-LLL_original_firstorder_RootMeanSquared* |
| *wavelet-LLL_original_firstorder_Variance* |
| *R2_tubo0* |

# ANEXO B

Características selecionada através da métrica RMSE

*B 1 - Características selecionadas com o MDI para o coeficiente angular com o modelo RF*

| |
|---|
| wavelet-LLH_original_glrlm_RunLengthNonUniformity |
| wavelet-LHL_original_firstorder_90Percentile |
| wavelet-HLL_original_firstorder_Mean |
| wavelet-LHL_original_firstorder_10Percentile |
| wavelet-LHL_original_firstorder_Range |
| wavelet-HLL_original_firstorder_Minimum |
| wavelet-LHH_original_ngtdm_Contrast |
| wavelet-LLL_original_firstorder_RobustMeanAbsoluteDeviation |
| wavelet-LLL_original_firstorder_Mean |
| wavelet-LHL_original_firstorder_TotalEnergy |
| wavelet-LLL_original_firstorder_Variance |
| wavelet-HLL_original_firstorder_90Percentile |
| original_firstorder_90Percentile |
| wavelet-HHL_original_firstorder_MeanAbsoluteDeviation |
| wavelet-LHL_original_firstorder_Maximu |
| wavelet-HHL_original_firstorder_Minimum |
| wavelet-HLL_original_firstorder_RobustMeanAbsoluteDeviation |
| wavelet-LHL_original_firstorder_RobustMeanAbsoluteDeviatio |
| wavelet-LHL_original_firstorder_Mean |
| wavelet-LHL_original_firstorder_Energy |
| wavelet-HHL_original_firstorder_Maximum |
| wavelet-HHL_original_firstorder_90Percentile |
| wavelet-HLL_original_firstorder_RootMeanSquared |
| wavelet-HHL_original_firstorder_10Percentile |
| original_firstorder_Mean |
| original_firstorder_Maximum |
| wavelet-HHL_original_firstorder_RobustMeanAbsoluteDeviation |
| wavelet-LLL_original_firstorder_Maximum |
| R2_tubo0 |
| original_firstorder_Median |
| wavelet-LLL_original_firstorder_RootMeanSquared |
| wavelet-LLL_original_firstorder_90Percentile |
| wavelet-LLL_original_firstorder_Median |
| wavelet-LLL_original_firstorder_10Percentile |
| original_firstorder_Energy |
| wavelet-LLL_original_firstorder_TotalEnergy |
| wavelet-LLL_original_firstorder_Minimum |
| wavelet-LLL_original_firstorder_Energy |
| wavelet-LLL_original_firstorder_Energy |
| original_firstorder_TotalEnergy |
| original_firstorder_10Percentile |

*B 2- Características selecionadas com o MDI para o coeficiente linear com o modelo RF*

| |
|---|
| original_firstorder_10Percentile |
| riginal_firstorder_Range |
| original_firstorder_90Percentile |
| wavelet-LLL_original_firstorder_Variance |
| wavelet-HHL_original_firstorder_Minimum |
| avelet-HHL_original_firstorder_90Percentile |
| wavelet-LLL_original_firstorder_Maximum |
| wavelet-HHL_original_firstorder_RootMeanSquared |
| wavelet-LHL_original_firstorder_Range |
| original_firstorder_TotalEnergy |
| wavelet-LHL_original_firstorder_10Percentile |
| wavelet-HLL_original_firstorder_10Percentile |
| wavelet-HLL_original_firstorder_Variance |
| R2_tubo0 |
| wavelet-HHL_original_firstorder_MeanAbsoluteDeviation |
| wavelet-HLL_original_firstorder_RootMeanSquared |
| wavelet-HLL_original_firstorder_MeanAbsoluteDeviation |
| wavelet-LHL_original_firstorder_RobustMeanAbsoluteDeviation |
| wavelet-LHL_original_firstorder_TotalEnergy |
| wavelet-LHL_original_firstorder_Energy |
| original_firstorder_Energy |
| wavelet-LLL_original_firstorder_RootMeanSquared |
| original_firstorder_Minimum |
| wavelet-LLL_original_firstorder_TotalEnergy |
| wavelet-LHL_original_firstorder_Minimum |
| wavelet-HLL_original_firstorder_Minimum |
| wavelet-LLL_original_firstorder_Mean |
| wavelet-HHL_original_firstorder_Maximum |
| original_firstorder_Maximum |
| wavelet-LHL_original_firstorder_Variance |
| wavelet-HLL_original_firstorder_TotalEnergy |
| wavelet-LLL_original_firstorder_90Percentile |
| wavelet-HHL_original_firstorder_Energy |
| avelet-HHL_original_firstorder_Variance |
| original_firstorder_Median |
| wavelet-LLL_original_firstorder_Energy |
| original_firstorder_Mean |
| wavelet-LHL_original_firstorder_MeanAbsoluteDeviation |
| wavelet-LLL_original_firstorder_MeanAbsoluteDeviation |
| wavelet-LLL_original_firstorder_MeanAbsoluteDeviation |
| wavelet-HLL_original_firstorder_Energy |
| wavelet-HLL_original_firstorder_Range |

*B 3 - Características selecionadas com o MDI para o coeficiente angular com o modelo CB*

| |
|---|
| wavelet-HLL_original_glcm_InverseVariance |
| wavelet-LLH_original_gldm_DependenceVariance |
| wavelet-HHH_original_gldm_LargeDependenceLowGrayLevelEmphasis |
| wavelet-LHL_original_glcm_Idm |
| original_firstorder_Minimum |
| wavelet-LLL_original_firstorder_Media |
| wavelet-LLH_original_firstorder_Range |
| wavelet-LLL_original_firstorder_Minimum |
| wavelet-HHL_original_firstorder_10Percentile |
| wavelet-HHL_original_firstorder_TotalEnergy |
| original_firstorder_TotalEnergy |
| wavelet-HLH_original_firstorder_Energy |
| compGelatina |
| wavelet-HHH_original_gldm_SmallDependenceLowGrayLevelEmphasis |
| wavelet-HLH_original_firstorder_RootMeanSquared |
| wavelet-HLH_original_firstorder_Variance |
| wavelet-LLH_original_firstorder_Median |
| wavelet-LLL_original_firstorder_TotalEnergy |

*B 4 - Características selecionadas com o MDI para o coeficiente linear com o modelo CB*

| |
|---|
| wavelet-HHL_original_firstorder_Range |
| wavelet-LLL_original_firstorder_90Percentile |
| wavelet-HHL_original_firstorder_Variance |
| wavelet-LLH_original_firstorder_Variance |
| wavelet-LHH_original_firstorder_Variance |
| original_firstorder_Energy |
| avelet-LLL_original_firstorder_RootMeanSquared |
| wavelet-LLH_original_firstorder_RobustMeanAbsoluteDeviation |

*B 5- Características selecionadas com o RFE para o coeficiente angular com o modelo RF*

| |
|---|
| compGelatina |
| original_firstorder_10Percentile |
| original_firstorder_90Percentile |
| original_firstorder_Energy |
| original_firstorder_Median |

*B 6-Características selecionadas com o RFE para o coeficiente linear com o modelo RF*

| |
|---|
| original_firstorder_10Percentile |
| original_firstorder_90Percentile |
| original_firstorder_Maximum |
| original_firstorder_Mean |
| original_firstorder_Median |

*B 7 - Características selecionadas com o RFE para o coeficiente angular com o modelo CB*

| |
|---|
| compGelatina |
| hidroquinona |
| original_firstorder_10Percentile |
| original_firstorder_Energy |
| original_firstorder_Kurtosis |
| original_firstorder_Maximum |
| original_firstorder_MeanAbsoluteDeviation |
| original_firstorder_Minimum |
| original_firstorder_TotalEnergy |
| original_gldm_LargeDependenceEmphasis |
| original_gldm_LargeDependenceLowGrayLevelEmphasis |
| original_glrlm_LongRunLowGrayLevelEmphasis |
| original_glrlm_RunLengthNonUniformity |
| original_glrlm_RunLengthNonUniformityNormalized |
| wavelet-LLH_original_firstorder_10Percentile |
| wavelet-LLH_original_firstorder_Energy |
| wavelet-LLH_original_firstorder_MeanAbsoluteDeviation |
| wavelet-LLH_original_firstorder_TotalEnergy |
| wavelet-LLH_original_gldm_DependenceVariance |
| wavelet-LLH_original_glcm_ClusterProminence |
| wavelet-LLH_original_glrlm_GrayLevelNonUniformityNormalized |
| wavelet-LLH_original_glrlm_RunEntropy |
| wavelet-LLH_original_glrlm_RunLengthNonUniformity |
| wavelet-LLH_original_ngtdm_Strength |
| wavelet-LHL_original_firstorder_Mean |
| wavelet-LHL_original_glrlm_RunLengthNonUniformity |
| wavelet-LHL_original_glrlm_ShortRunHighGrayLevelEmphasi |
| wavelet-LHL_original_glrlm_RunLengthNonUniformityNormalized |
| wavelet-LHH_original_glcm_Imc2 |
| wavelet-LHH_original_glrlm_ShortRunHighGrayLevelEmphasis |
| wavelet-HLH_original_firstorder_InterquartileRange |
| wavelet-HLH_original_firstorder_Kurtosis |
| wavelet-HLH_original_firstorder_Mean |
| wavelet-HLH_original_firstorder_RootMeanSquared |
| wavelet-HLH_original_firstorder_TotalEnergy |
| wavelet-HLH_original_firstorder_Uniformity |

| |
|---|
| wavelet-HLH_original_firstorder_Variance |
| wavelet-HLH_original_gldm_GrayLevelVariance |
| wavelet-HLH_original_gldm_LargeDependenceLowGrayLevelEmphasis |
| wavelet-HLH_original_gldm_SmallDependenceHighGrayLevelEmphasis |
| wavelet-HLH_original_glcm_ClusterShade |
| wavelet-HLH_original_glcm_InverseVariance |
| wavelet-HLH_original_glcm_MCC |
| wavelet-HLH_original_glcm_SumSquares |
| wavelet-HLH_original_glrlm_RunEntropy |
| wavelet-HLH_original_glszm_ZoneEntropy |
| wavelet-HLH_original_ngtdm_Busyness |
| wavelet-HHL_original_firstorder_10Percentile |
| wavelet-HHL_original_firstorder_RobustMeanAbsoluteDeviation |
| wavelet-HHL_original_gldm_DependenceNonUniformityNormalized |

*B 8- Características selecionadas com o RFE para o coeficiente linear com o modelo CB*

| |
|---|
| wavelet-HLL_original_glrlm_LowGrayLevelRunEmphasis |
| wavelet-HLL_original_glrlm_RunEntropy |
| wavelet-HLL_original_glszm_GrayLevelVariance |
| wavelet-HLL_original_glszm_ZoneVariance |
| wavelet-HLH_original_glcm_DifferenceVariance |
| wavelet-HLH_original_glcm_Idm |
| wavelet-HLH_original_glcm_Imc |
| wavelet-HLH_original_glcm_JointEntropy |
| wavelet-HLH_original_glrlm_RunLengthNonUniformity |
| wavelet-HHL_original_firstorder_Energy |
| wavelet-HHL_original_firstorder_Minimum |
| wavelet-HHL_original_firstorder_RootMeanSquared |
| wavelet-HHL_original_firstorder_Skewness |
| wavelet-HHL_original_firstorder_TotalEnergy |
| wavelet-HHL_original_firstorder_Variance |
| wavelet-HHL_original_glcm_DifferenceVariance |
| wavelet-HHL_original_glcm_InverseVariance |
| wavelet-HHL_original_glcm_JointAverage |
| wavelet-HHH_original_firstorder_Energy |
| wavelet-HHH_original_firstorder_RootMeanSquared |

*B 9- Características selecionadas com o PS para o coeficiente angular com o modelo RF*

| |
|---|
| wavelet-HLH_original_firstorder_Kurtosis |
| wavelet-HLH_original_firstorder_Mean |
| wavelet-HLH_original_firstorder_RootMeanSquared |
| wavelet-HLH_original_firstorder_TotalEnergy |

| |
|---|
| wavelet-HLH_original_firstorder_Uniformity |
| wavelet-HLH_original_firstorder_Variance |
| wavelet-HLH_original_gldm_GrayLevelVariance |
| wavelet-HLH_original_gldm_LargeDependenceLowGrayLevelEmphasis |
| wavelet-HLL_original_glszm_ZoneVariance |
| wavelet-HLH_original_glcm_DifferenceVariance |
| wavelet-HLH_original_glcm_Idm |
| wavelet-HLH_original_glcm_Imc |
| wavelet-HLH_original_glcm_JointEntropy |
| wavelet-HLL_original_glszm_ZoneVariance |
| wavelet-HLH_original_glcm_DifferenceVariance |
| wavelet-HLH_original_glcm_Idm |
| wavelet-HLL_original_firstorder_Maximum |

*B 10 - Características selecionadas com o PS para o coeficiente linear com o modelo RF*

| |
|---|
| wavelet-HHL_original_firstorder_RobustMeanAbsoluteDeviation |
| wavelet-LLL_original_firstorder_Maximum |
| wavelet-HLL_original_ngtdm_Busyness |
| original_firstorder_Median |
| wavelet-LLL_original_firstorder_RootMeanSquared |
| wavelet-LLL_original_firstorder_90Percentile |
| wavelet-LLL_original_firstorder_Median |
| wavelet-HLL_original_firstorder_Skewness |
| wavelet-LLL_original_firstorder_Mean |
| wavelet-LLL_original_firstorder_TotalEnergy |
| wavelet-LLL_original_firstorder_Minimum |
| original_firstorder_RootMeanSquared |
| original_firstorder_Maximum' |
| R2_tubo0 |
| wavelet-LLL_original_firstorder_Variance |
| wavelet-HHL_original_firstorder_Energy |

*B 11 - Características selecionadas com o PS para o coeficiente linear com o modelo CB*

| |
|---|
| original_firstorder_10Percentile |
| original_firstorder_90Percentile |
| original_firstorder_Maximum |
| original_firstorder_Mean |
| R2_tubo0 |

# ANEXO C

Características selecionada através da métrica MAE

*C 1 - Características selecionada com o MDI para o coeficiente linear com o modelo RF*

| |
|---|
| wavelet-HHL_original_firstorder_10Percentile |
| wavelet-HHL_original_firstorder_Energy |
| wavelet-HHL_original_firstorder_RobustMeanAbsoluteDeviation |
| avelet-HHL_original_firstorder_RootMeanSquared |
| wavelet-HHL_original_firstorder_Variance |
| wavelet-HHL_original_gldm_SmallDependenceLowGrayLevelEmphasis |
| wavelet-HHL_original_glcm_Imc2 |
| wavelet-HHH_original_gldm_HighGrayLevelEmphasis |
| wavelet-HHH_original_glcm_Imc2 |
| wavelet-HHH_original_glcm_MCC |
| wavelet-HHH_original_ngtdm_Busyness' |
| wavelet-LLL_original_firstorder_10Percentile |
| wavelet-LLL_original_firstorder_90Percentile |
| wavelet-LLL_original_firstorder_Energy |
| wavelet-LLL_original_firstorder_Maximum |
| wavelet-LLL_original_firstorder_Mean |
| wavelet-LLL_original_firstorder_Median |
| wavelet-LLL_original_firstorder_Minimum |
| wavelet-LLL_original_firstorder_RobustMeanAbsoluteDeviation |
| wavelet-LLL_original_firstorder_RootMeanSquared |
| wavelet-LLL_original_firstorder_TotalEnergy |
| wavelet-LLL_original_firstorder_Variance |
| wavelet-LHL_original_glrlm_ShortRunHighGrayLevelEmphasis |
| wavelet-LHH_original_glcm_Imc2 |
| wavelet-LHH_original_glrlm_ShortRunHighGrayLevelEmphasis |
| wavelet-HLH_original_firstorder_InterquartileRange |
| wavelet-HLH_original_firstorder_Kurtosis |
| wavelet-HLH_original_firstorder_Mean |
| wavelet-HLH_original_firstorder_RootMeanSquared |
| wavelet-HLH_original_firstorder_TotalEnergy |
| wavelet-HLH_original_firstorder_Uniformity |
| wavelet-HLH_original_firstorder_Variance |
| avelet-HLH_original_gldm_GrayLevelVariance |
| Sequencia_aquisição |
| R2_tubo0 |
| wavelet-LLH_original_firstorder_Range |
| wavelet-HHH_original_gldm_LargeDependenceLowGrayLevelEmphasis |
| wavelet-HHL_original_firstorder_TotalEnergy |
| wavelet-LHL_original_gldm_SmallDependenceLowGrayLevelEmphasis |
| wavelet-HLH_original_firstorder_RootMeanSquared |
| wavelet-HLH_original_firstorder_Variance |

| |
|---|
| hidroquinona |
| original_firstorder_TotalEnergy |
| original_firstorder_10Percentile |
| original_firstorder_Median |
| original_firstorder_Minimum |
| original_firstorder_90Percentile |
| original_firstorder_RootMeanSquared |
| original_firstorder_Variance |
| original_firstorder_Variance |
| original_firstorder_Energy |
| original_firstorder_Mean |
| original_firstorder_MeanAbsoluteDeviation |
| compGelatina |
| wavelet-LHL_original_glrlm_RunLengthNonUniformity |
| wavelet-LHL_original_glrlm_RunLengthNonUniformityNormalized |
| original_firstorder_Range |
| original_firstorder_Maximum |
| wavelet-HLL_original_ngtdm_Busyness |
| wavelet-HLL_original_ngtdm_Complexity |
| wavelet-HLH_original_ngtdm_Busyness |
| wavelet-HLH_original_ngtdm_Complexity |
| wavelet-LHH_original_glcm_ClusterProminence |
| wavelet-LHH_original_glcm_ClusterTendency |
| wavelet-HLH_original_glcm_JointEntropy |
| wavelet-HLH_original_glcm_ClusterShade |

*C 2- Características selecionada com o MDI para o coeficiente linear com o modelo RF*

| |
|---|
| wavelet-LLL_original_firstorder_TotalEnergy |
| wavelet-LHL_original_firstorder_RootMeanSquared |
| wavelet-LLL_original_firstorder_RootMeanSquared |
| wavelet-HHL_original_firstorder_Minimum |
| wavelet-LLL_original_firstorder_Median |
| wavelet-HLL_original_firstorder_Energy |
| wavelet-HLL_original_firstorder_Varianc |
| wavelet-LLL_original_firstorder_10Percentile |
| wavelet-HHL_original_firstorder_MeanAbsoluteDeviation |
| original_firstorder_TotalEnergy |
| wavelet-LHL_original_firstorder_MeanAbsoluteDeviation |
| R2_tubo0 |
| wavelet-LLL_original_firstorder_90Percentile |
| wavelet-LHL_original_firstorder_Range |
| wavelet-HHL_original_firstorder_Maximum |
| original_firstorder_90Percentile |
| wavelet-LLL_original_firstorder_MeanAbsoluteDeviation |
| original_firstorder_RootMeanSquared |
| wavelet-HHL_original_firstorder_Variance |

| |
|---|
| original_firstorder_Minimum |
| wavelet-LLL_original_firstorder_Minimum |
| wavelet-LHL_original_firstorder_Energy |
| wavelet-HLL_original_firstorder_TotalEnergy |
| original_firstorder_Maximum |
| wavelet-LLL_original_firstorder_Energy |
| wavelet-HHL_original_firstorder_Energy |
| wavelet-LHL_original_firstorder_TotalEnergy |
| original_firstorder_10Percentile |
| original_firstorder_Median |
| wavelet-LLL_original_firstorder_Maximum |
| original_firstorder_Mean |
| wavelet-HHL_original_firstorder_RootMeanSquared |
| wavelet-HLL_original_firstorder_Maximum |
| wavelet-HLL_original_firstorder_RootMeanSquared |
| wavelet-HLL_original_firstorder_Range |
| wavelet-LLL_original_firstorder_Mean |
| wavelet-LHL_original_firstorder_Variance |
| wavelet-LLL_original_firstorder_TotalEnergy |

*C 3- Características selecionada com o MDI para o coeficiente angular com o modelo CB*

| |
|---|
| wavelet-LLH_original_firstorder_Range |
| wavelet-HHH_original_gldm_LargeDependenceLowGrayLevelEmphasis |
| wavelet-HHL_original_firstorder_TotalEnergy |
| wavelet-HHL_original_gldm_SmallDependenceLowGrayLevelEmphasis |
| wavelet-HLH_original_firstorder_RootMeanSquared |
| wavelet-HLH_original_firstorder_Variance |
| original_firstorder_TotalEnergy |
| wavelet-LLH_original_firstorder_Median |
| wavelet-LLL_original_firstorder_TotalEnergy |
| wavelet-HHH_original_firstorder_10Percentile |

*C 4- Características selecionada com o MDI para o coeficiente linear com o modelo CB*

| |
|---|
| wavelet-LHL_original_firstorder_10Percentile |
| wavelet-HLH_original_firstorder_Maximum |
| wavelet-LHH_original_firstorder_Energy |
| wavelet-LHH_original_firstorder_Range |
| wavelet-LLH_original_firstorder_Energy |
| wavelet-HHL_original_firstorder_Maximum |
| original_firstorder_Mean |
| wavelet-HLL_original_firstorder_RootMeanSquared |
| original_firstorder_Maximum |

| |
|---|
| wavelet-LLH_original_firstorder_RobustMeanAbsoluteDeviation |

*C 5 - Características selecionada com o RFE para o coeficiente angular com o modelo RF*

| |
|---|
| compGelatina |
| original_firstorder_10Percentile |
| original_firstorder_Energy |
| original_firstorder_Median |
| original_firstorder_Minimum |
| original_firstorder_TotalEnergy |
| wavelet-LLH_original_glrlm_RunEntropy |
| wavelet-LLH_original_glrlm_RunLengthNonUniformity |
| wavelet-LHL_original_firstorder_Mean |
| wavelet-LHL_original_glrlm_RunLengthNonUniformity |
| wavelet-LHL_original_glrlm_RunLengthNonUniformityNormalized |
| wavelet-LHL_original_glrlm_RunPercentage |
| wavelet-LHL_original_glrlm_ShortRunHighGrayLevelEmphasis |
| wavelet-LHH_original_glcm_ClusterProminence |
| wavelet-LHH_original_glcm_ClusterTendency |
| wavelet-LHH_original_glrlm_RunPercentage |
| wavelet-HLL_original_firstorder_Energy |
| wavelet-HLL_original_firstorder_Minimum |
| wavelet-HLL_original_firstorder_RobustMeanAbsoluteDeviation |
| wavelet-HLL_original_firstorder_TotalEnergy' |
| wavelet-HLL_original_glrlm_LongRunLowGrayLevelEmphasis |
| wavelet-HLL_original_ngtdm_Busyness |
| wavelet-HLL_original_ngtdm_Complexity |
| wavelet-HLH_original_ngtdm_Busyness |
| wavelet-HLH_original_ngtdm_Complexity |
| wavelet-HHL_original_firstorder_10Percentile |
| wavelet-HHL_original_firstorder_Energy |
| wavelet-HHL_original_firstorder_RobustMeanAbsoluteDeviation |
| avelet-HHL_original_firstorder_RootMeanSquared |
| wavelet-HHL_original_firstorder_Variance |
| wavelet-HHL_original_gldm_SmallDependenceLowGrayLevelEmphasis |
| wavelet-HHL_original_glcm_Imc2 |
| wavelet-HHH_original_gldm_HighGrayLevelEmphasis |
| wavelet-HHH_original_glcm_Imc2 |
| wavelet-HHH_original_glcm_MCC |
| wavelet-HHH_original_ngtdm_Busyness' |
| wavelet-LLL_original_firstorder_10Percentile |
| wavelet-LLL_original_firstorder_90Percentile |
| wavelet-LLL_original_firstorder_Energy |
| wavelet-LLL_original_firstorder_Maximum |
| wavelet-LLL_original_firstorder_Mean |
| wavelet-LLL_original_firstorder_Median |
| wavelet-LLL_original_firstorder_Minimum |

| |
|---|
| wavelet-LLL_original_firstorder_RobustMeanAbsoluteDeviation |
| wavelet-LLL_original_firstorder_RootMeanSquared |
| wavelet-LLL_original_firstorder_TotalEnergy |
| wavelet-LLL_original_firstorder_Variance |
| wavelet-LLL_original_gldm_LargeDependenceHighGrayLevelEmphasis |
| R2_tubo0 |
| wavelet-HLL_original_firstorder_10Percentile |

*C 6 - Características selecionada com o RFE para o coeficiente linear com o modelo RF*

| |
|---|
| original_firstorder_10Percentile |
| original_firstorder_90Percentile |
| original_firstorder_Energy |
| original_firstorder_Maximum |
| original_firstorder_Mean |
| original_firstorder_Median |
| original_firstorder_Minimum |
| original_firstorder_Range |
| original_firstorder_RootMeanSquared |
| original_firstorder_TotalEnergy |

*C 7 - Características selecionada com o RFE para o coeficiente angular com o modelo CB*

| |
|---|
| compGelatina |
| hidroquinona |
| original_firstorder_10Percentile |
| original_firstorder_Energy |
| original_firstorder_Kurtosis |
| original_firstorder_Maximum |
| original_firstorder_MeanAbsoluteDeviation |
| original_firstorder_Minimum |
| original_firstorder_TotalEnergy |
| original_gldm_LargeDependenceEmphasis |
| original_gldm_LargeDependenceLowGrayLevelEmphasis |
| riginal_glrlm_LongRunLowGrayLevelEmphasis' |
| original_glrlm_RunLengthNonUniformity |
| original_glrlm_RunLengthNonUniformityNormalized |
| wavelet-LLH_original_firstorder_10Percentile |
| wavelet-LLH_original_firstorder_Energy |
| wavelet-LLH_original_firstorder_MeanAbsoluteDeviation |
| wavelet-LLH_original_firstorder_TotalEnergy |
| wavelet-LLH_original_gldm_DependenceVariance |
| wavelet-LLH_original_glcm_ClusterProminence |
| avelet-LLH_original_glrlm_GrayLevelNonUniformityNormalized |

| |
|---|
| wavelet-LLH_original_glrlm_RunEntropy |
| wavelet-LLH_original_glrlm_RunLengthNonUniformity |
| avelet-LLH_original_ngtdm_Strength |
| wavelet-LHL_original_firstorder_Mean |
| wavelet-LHL_original_glrlm_RunLengthNonUniformity |
| wavelet-LHL_original_glrlm_RunLengthNonUniformityNormalized |
| wavelet-LHL_original_glrlm_ShortRunHighGrayLevelEmphasis |
| wavelet-LHH_original_glcm_Imc2 |
| wavelet-LHH_original_glrlm_ShortRunHighGrayLevelEmphasis |
| wavelet-HLH_original_firstorder_InterquartileRange |
| wavelet-HLH_original_firstorder_Kurtosis |
| wavelet-HLH_original_firstorder_Mean |
| wavelet-HLH_original_firstorder_RootMeanSquared |
| wavelet-HLH_original_firstorder_TotalEnergy |
| wavelet-HLH_original_firstorder_Uniformity |
| wavelet-HLH_original_firstorder_Variance |
| avelet-HLH_original_gldm_GrayLevelVariance |
| wavelet-HLH_original_gldm_LargeDependenceLowGrayLevelEmphasis |
| wavelet-HLH_original_gldm_SmallDependenceHighGrayLevelEmphasis |
| wavelet-HLH_original_glcm_ClusterShade |
| wavelet-HLH_original_glcm_InverseVariance |
| wavelet-HLH_original_glcm_MCC |
| wavelet-HLH_original_glcm_SumSquares |
| wavelet-HLH_original_glrlm_RunEntropy |
| wavelet-HLH_original_glszm_ZoneEntropy |
| wavelet-HLH_original_ngtdm_Busyness |
| wavelet-HHL_original_firstorder_10Percentile |
| wavelet-HHL_original_firstorder_RobustMeanAbsoluteDeviation |
| wavelet-HHL_original_gldm_DependenceNonUniformityNormalized |

*C 8- Características selecionada com o RFE para o coeficiente linear com o modelo CB*

| |
|---|
| wavelet-HLL_original_glrlm_LowGrayLevelRunEmphasis |
| wavelet-HLL_original_glrlm_RunEntropy' |
| wavelet-HLL_original_glszm_GrayLevelVariance |
| wavelet-HLL_original_glszm_ZoneVariance |
| wavelet-HLH_original_glcm_DifferenceVariance |

*C 9- Características selecionada com o PS para o coeficiente angular com o modelo RF*

| |
|---|
| compGelatina |
| original_firstorder_TotalEnergy |
| wavelet-LHL_original_firstorder_Mean |

original_firstorder_10Percentile
original_firstorder_90Percentil
original_firstorder_Energy
original_firstorder_Maximum
original_firstorder_Mean
original_firstorder_Median
original_firstorder_Minimum
original_firstorder_Range
original_firstorder_RootMeanSquared
original_firstorder_TotalEnergy
wavelet-HLL_original_glrlm_LowGrayLevelRunEmphasis
wavelet-HLL_original_glrlm_RunEntropy'
wavelet-HLL_original_glszm_GrayLevelVariance
wavelet-HHL_original_firstorder_Variance
wavelet-LLL_original_firstorder_10Percentile
wavelet-LLL_original_firstorder_90Percentile
wavelet-LLL_original_firstorder_Energy
wavelet-LLL_original_firstorder_Maximum
wavelet-LLL_original_firstorder_MeanAbsoluteDeviation
wavelet-LLL_original_firstorder_Mean
wavelet-LLL_original_firstorder_Median
wavelet-LLL_original_firstorder_Minimum
wavelet-LLL_original_firstorder_Range
wavelet-LLL_original_firstorder_RootMeanSquared
wavelet-LLL_original_firstorder_TotalEnergy
wavelet-LHL_original_firstorder_10Percentile
wavelet-LLH_original_firstorder_90Percentile
wavelet-HLL_original_firstorder_Energy
wavelet-HHL_original_firstorder_Maximum
wavelet-HLL_original_firstorder_MeanAbsoluteDeviation
wavelet-HLL_original_firstorder_Mean
wavelet-LLL_original_firstorder_Median
wavelet-LHL_original_firstorder_Minimum
wavelet-LHL_original_firstorder_Range
wavelet-LLH_original_gldm_LargeDependenceLowGrayLevelEmphasis
wavelet-LLH_original_gldm_SmallDependenceHighGrayLevelEmphasis
wavelet-HHL_original_firstorder_RobustMeanAbsoluteDeviation
wavelet-HLL_original_glrlm_LowGrayLevelRunEmphasis
original_glrlm_RunEntropy
original_glszm_GrayLevelVariance
wavelet-HLL_original_glcm_ClusterShade
original_firstorder_RootMeanSquared
original_firstorder_TotalEnergy
original_firstorder_RobustMeanAbsoluteDeviation
wavelet-HLH_original_gldm_LargeDependenceLowGrayLevelEmphasis
wavelet-HLH_original_gldm_SmallDependenceHighGrayLevelEmphasis
wavelet-HLH_original_glcm_ClusterShade
wavelet-HLH_original_glcm_InverseVariance
wavelet-HLH_original_glcm_MCC
wavelet-HLH_original_glcm_SumSquares

| |
|---|
| wavelet-HLH_original_glrlm_RunEntropy |
| wavelet-HLH_original_glszm_ZoneEntropy |
| wavelet-HLH_original_firstorder_InterquartileRange |
| wavelet-HLH_original_glcm_DifferenceVariance |
| wavelet-HLH_original_glcm_Idm |
| wavelet-HHH_original_glcm_Imc |
| wavelet-HLH_original_glcm_JointEntropy |
| wavelet-HLH_original_glcm_ClusterShade |
| wavelet-HLL_original_glcm_MCC |
| wavelet-HLL_original_gldm_LargeDependenceLowGrayLevelEmphasis |
| wavelet-HLL_original_gldm_SmallDependenceHighGrayLevelEmphasis |
| hidroquinona |
| sequencia_aquisição |
| wavelet-LLH_original_glcm_JointEntropy |
| wavelet-HLL_original_firstorder_InterquartileRange |

*C 10 - Características selecionada com o PS para o coeficiente linear com o modelo RF*

| |
|---|
| original_firstorder_10Percentile |
| original_firstorder_90Percentil |
| original_firstorder_Energy |
| original_firstorder_Maximum |
| original_firstorder_Mean |
| original_firstorder_Median |
| original_firstorder_Minimum |
| original_firstorder_Range |
| original_firstorder_RootMeanSquared |
| original_firstorder_TotalEnergy |
| original_firstorder_Variance |
| wavelet-LHL_original_firstorder_10Percentile |
| wavelet-LHL_original_firstorder_Energy |
| wavelet-LHL_original_firstorder_Maximum |
| wavelet-LHL_original_firstorder_Minimum |
| wavelet-LHL_original_firstorder_Range |
| wavelet-LHL_original_firstorder_RootMeanSquared |
| wavelet-LHL_original_firstorder_TotalEnergy |
| wavelet-LHL_original_glszm_ZoneEntropy |
| wavelet-HLL_original_firstorder_10Percentile |
| wavelet-HLL_original_firstorder_Energy |
| wavelet-HLL_original_firstorder_Maximum |
| wavelet-HLL_original_firstorder_Minimum |
| wavelet-HLL_original_firstorder_Range |
| wavelet-HLL_original_firstorder_RootMeanSquared |
| wavelet-HLL_original_firstorder_TotalEnergy |
| wavelet-HLL_original_firstorder_Variance |
| wavelet-HHL_original_firstorder_Energy |
| wavelet-HHL_original_firstorder_Maximum |

| |
|---|
| wavelet-HHL_original_firstorder_Minimum |
| wavelet-HHL_original_firstorder_Range |
| wavelet-HHL_original_firstorder_RootMeanSquared |
| wavelet-HHL_original_firstorder_TotalEnergy |
| wavelet-HHL_original_firstorder_Variance |
| wavelet-LLL_original_firstorder_10Percentile |
| wavelet-LLL_original_firstorder_90Percentile |
| wavelet-LLL_original_firstorder_Energy |
| wavelet-LLL_original_firstorder_Maximum |
| wavelet-LLL_original_firstorder_MeanAbsoluteDeviation |
| wavelet-LLL_original_firstorder_Mean |
| wavelet-LLL_original_firstorder_Median |
| wavelet-LLL_original_firstorder_Minimum |
| wavelet-LLL_original_firstorder_Range |
| wavelet-LLL_original_firstorder_RootMeanSquared |
| wavelet-LLL_original_firstorder_TotalEnergy |
| wavelet-LLL_original_firstorder_Variance |
| R2_tubo0 |

*C 11 - Características selecionada com o PS para o coeficiente linear com o modelo CB*

| |
|---|
| Sequencia_aquisição |
| original_glrlm_RunLengthNonUniformity |
| wavelet-HHH_original_gldm_DependenceNonUniformityNormalized |
| R2_tubo0 |