UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA DE MATERIAIS

Utilização de *machine learning* e algoritmo genético no *design* de ligas de titânio para aplicações biomédicas

Luís Guilherme Santagnelo Nogueira

São Carlos - SP

2025

Utilização de *machine learning* e algoritmo genético no *design* de ligas de titânio para aplicações biomédicas

Trabalho de conclusão de curso apresentado ao Departamento de Engenharia de Materiais da Universidade Federal de São Carlos, como requisito para obtenção do título de bacharel em Engenharia de Materiais.

Orientador: Lucas Barcelos Otani

Coorientadora: Caroline Binde Stoco

São Carlos - SP

2025

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA DE MATERIAIS
Telefones: 16 –3351-8244 e 3351-8246
Endereço eletrônico: demachef@ufscar.br
Rodovia Washington Luís, km 235 – Caixa Postal 676
CEP 13565-905 – São Carlos – SP - Brasil

DEMa
UFSCar

ATA DE DEFESA DE TRABALHO DE CONCLUSÃO DE CURSO (TCC)

NOME: Luis Guilherme Santagnelo Nogueira

RA: 769745

TÍTULO: Utilização de machine learning e algoritmo genético no design de ligas de titânio para aplicações biomédicas

ORIENTADOR(A): Prof. Dr. Lucas Barcelos Otani

CO-ORIENTADOR(A): Be. Caroline Binde Stoco

DATA/HORÁRIO: 24/01/2025, 15h

BANCA – NOTAS:

	Monografia	**Defesa**
Prof. Dr. Lucas Barcelos Otani	9	10
Prof. Dr. Andre Luiz Vidilli	9	10
Média	9	10

BANCA – ASSINATURAS:

Prof. Dr. Lucas Barcelos Otani

gov.br Documento assinado digitalmente
LUCAS BARCELOS OTANI
Data: 24/01/2025 16:32:01-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Andre Luiz Vidilli

gov.br Documento assinado digitalmente
ANDRE LUIZ VIDILLI
Data: 24/01/2025 16:37:07-0300
Verifique em https://validar.iti.gov.br

DEDICATÓRIA

Dedico este trabalho à minha família e aos amigos que me auxiliaram durante esta trajetória, tanto aos que ainda estão aqui, quanto a aqueles que já se foram.

AGRADECIMENTO

Agradeço imensamente ao meu orientador Lucas Otani pelo suporte e orientação durante este trabalho, também agradeço à minha coorientadora Caroline Stoco pela imensa ajuda na construção do mesmo. Por fim, agradeço a todos os docentes da UFSCar que me proporcionaram o conhecimento necessário para concluir este e consequentemente minha graduação.

RESUMO

A produção de ligas de titânio beta metaestáveis para aplicações biomédicas tem crescido na última década, buscando atender à demanda por materiais com módulo de elasticidade mais compatível com o tecido ósseo, contribuindo de forma positiva para a biocompatibilidade e melhorando o desempenho mecânico. Ligas como Ti–29Nb–13Ta–4.6Zr (TNZT), Ti–12Mo–6Zr–2Fe (TMZF) e Ti–35Nb–7Zr–5Ta (TiOsteum) são amplamente usadas por suas propriedades, como elevada resistência mecânica específica e resistência à corrosão, além de apresentarem alguns dos elementos estabilizantes da fase beta e biocompatibilizantes como Zr, Ta, Nb e Mo, evitando elementos com toxicidade ao ser humano, como V, Co, Cr e Cu. Um desafio fundamental na utilização dessas ligas em implantes ortopédicos é reduzir a diferença entre o módulo de elasticidade e o do osso humano, minimizando o fenômeno de *stress shielding*, que pode causar fragilidade óssea ao longo do tempo. Essa discrepância é regida pela microestrutura final formada no processamento da liga e no tratamento térmico, na qual apresenta parâmetros como tamanho médio de grãos, fração do volume de fase alfa, fração do volume de martensita e fração do volume de fase ômega, que afetam as propriedades mecânicas. A integração de *machine learning*, especialmente com o uso de algoritmos genéticos, pode otimizar o desenvolvimento dessas ligas. Esta metodologia de combinação pode ajudar a encontrar composições que minimizem o módulo de elasticidade, ajustando parâmetros e identificando combinações mais eficientes. Assim, este trabalho visa aplicar o *machine learning* juntamente com algoritmo genético para obter composições químicas de ligas de titânio que possuam microestrutura predominantemente de fase beta, com o menor módulo de elasticidade possível. Para isso, utilizou-se tanto dados experimentais descritos na literatura, quanto modelos matemáticos para a otimização de parâmetros, como o E, Mo_{eq} (molibdênio equivalente), o Bo (*bond order*, ou ordem de ligação) e o Md (*mean d orbital energy level*, nível médio de energia do orbital d). A partir da utilização de códigos de algoritmo genético, foi possível então obter uma lista de composições químicas potencialmente com microestrutura e propriedades desejadas.

Palavras-chave: Liga de titânio beta; Módulo elástico; Machine learning; Algoritmo genético.

ABSTRACT

The production of metastable beta titanium alloys for biomedical applications has grown over the last decade, aiming to meet the demand for materials with an elastic modulus more compatible with bone tissue. This contributes positively to biocompatibility and improves mechanical performance. Alloys such as Ti–29Nb–13Ta–4.6Zr (TNZT), Ti–12Mo–6Zr–2Fe (TMZF), and Ti–35Nb–7Zr–5Ta (TiOsteum) are widely used for their properties, such as high specific mechanical strength and corrosion resistance. These alloys also incorporate beta-phase stabilizing and biocompatible elements, such as Zr, Ta, Nb, and Mo, avoiding elements toxic to humans, such as V, Co, Cr, and Cu. A fundamental challenge in using these alloys for orthopedic implants is reducing the difference between their elastic modulus and that of human bone, minimizing the phenomenon of stress shielding, which can cause bone fragility over time. This discrepancy is governed by the final microstructure formed during alloy processing and heat treatment, which includes parameters such as average grain size, alpha phase volume fraction, martensite volume fraction, and omega phase volume fraction, all of which affect mechanical properties. The integration of machine learning, especially through genetic algorithms, can optimize the development of these alloys. This methodology helps identify compositions that minimize the elastic modulus by adjusting parameters and identifying more efficient combinations. Thus, this study aims to apply machine learning combined with genetic algorithms to obtain chemical compositions of titanium alloys with predominantly beta-phase microstructures and the lowest possible elastic modulus. To this end, both experimental data described in the literature and mathematical models were used to optimize parameters such as E, Mo_{eq} (molybdenum equivalent), Bo (bond order), and Md (mean d orbital energy level). Using genetic algorithm codes, a list of chemical compositions with potentially desirable microstructures and properties was obtained.

Keywords: Beta titanium alloy; Elastic modulus; Machine learning;. Genetic algorithm.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE SIGLAS

ML - *Machine Learning*

AG - Algoritmo Genético

GB - *Gradient Boosting*

GBR - *Gradient Boosting Regressor*

RF - *Random Forest*

RFR - *Random Forest Regressor*

SV - *Support Vector*

SVR - *Support Vector Regressor*

MAD - *Mean Absolute Deviation*

VEC - Concentração média de elétrons de valência

RMSE - *Root Mean Square Error*

SS - *Stress Shielding*

LISTA DE SÍMBOLOS

a_m - Parâmetro de rede médio

T_m - Temperatura de fusão média

T - Diferença na temperatura de fusão ponderada pela composição

a - Diferença no parâmetro de rede ponderada pela composição

χ - Diferença na eletronegatividade de Pauling ponderada pela composição

Mo_{eq} - Molibdênio Equivalente

C - Cromossomo

C_p - Probabilidade de cruzamento

O - Prole

M_P - Probabilidade de mutação

Bo - Ordem de ligação

Md – Energia média do orbital d

E_0 – Energia de Ligação

Sumário

1. Introdução

As ligas monofásicas de titânio beta têm como característica principal a presença de elementos estabilizadores da fase beta: Vanádio (V), Tungstênio (W), Molibdênio (Mo), Tântalo (Ta), Nióbio (Nb) e Cromo (Cr), de modo a reter a estrutura cristalina cúbica de corpo centrado (CCC) quando resfriada, evitando assim a transformação martensítica. Tipicamente trabalhadas a frio, estas ligas são amplamente utilizadas no âmbito da medicina, visto que possuem propriedades que propiciam sua aplicação em uma diversa gama de implantes, como por exemplo: elevada resistência mecânica específica, resistência à corrosão e baixo módulo de elasticidade. As ligas comercialmente utilizadas (TNZT, TMZF) consistem nas ligas que contém elementos como nióbio, tântalo, zircônio e molibdênio (1,2).

A despeito das ligas de titânio de pureza comercial e ligas com microestrutura alfa/beta manterem sua predominância como componentes principais de titânio empregados nas aplicações biomédicas contemporâneas, nota-se que na última década, tenha ocorrido um aumento considerável na produção de ligas de titânio beta metaestáveis, desenvolvidas especificamente para estas aplicações (1). Entretanto, quando se analisa o módulo de elasticidade, é notável que a grande maioria das ligas de titânio disponíveis no mercado apresenta valores que se situam em uma faixa oscilando entre 80 a 110 GPa, o que representa cerca da metade dos valores típicos observados em aços (3). Inicialmente concebidos para atender às demandas de apresentar um módulo de elasticidade mais compatível com o tecido ósseo e garantir uma maior biocompatibilidade, as ligas titânio beta vêm sendo objeto de consideração em outras modalidades de aplicação na ortopedia, como em regiões da coluna vertebral e no tratamento de traumas (1). Sua biocompatibilidade é preservada, enquanto se aprimora seu desempenho mecânico mediante a aplicação de processos de envelhecimento artificial. Três ligas deste sistema se destacam, Ti–29Nb–13Ta–4.6Zr (TNZT), Ti–12Mo–6Zr–2Fe (TMZF) e Ti–35Nb–7Zr–5Ta (*TiOsteum*), foram desenvolvidas essencialmente de forma simultânea em empreendimentos conjuntos no Japão e nos Estados Unidos (1).

Um dos principais dilemas na utilização de ligas de titânio em implantes ortopédicos é a aproximação do valor de seu módulo de elasticidade com o da estrutura óssea humana, de modo a serem suficientemente próximos a fim de evitar o fenômeno de *stress shielding* (blindagem a tensão). Fenômeno que ocorre devido

a uma diferença significativa entre os módulos de elasticidade do implante e do osso, de modo que o implante suporte uma parcela muito maior das cargas mecânicas aplicadas, reduzindo significativamente o carregamento natural do osso. Como consequência, o osso deixa de ser submetido ao estímulo mecânico necessário para a manutenção de sua densidade e estrutura, ocasionando perda óssea, comprometendo a integração do implante e sua estabilidade a longo prazo, além de aumentar o risco de falhas e complicações clínicas (4).

No que se refere ao *Machine Learning (ML)*, consiste no desenvolvimento de sistemas de computador que têm a capacidade de aprender e melhorar a partir de dados, permitindo que máquinas tomem decisões e executem tarefas sem serem explicitamente programadas. Os modelos podem ser de diferentes tipos, como de aprendizado por reforço, supervisionado ou não supervisionado. No tipo supervisionado, as técnicas de ML forçam a criar uma relação entre a entrada fornecida (variável independente) e um atributo de destino (variável dependente da entrada). Consequentemente, na modalidade não supervisionada, as conclusões dos conjuntos de dados consistem em dados de entrada sem respostas rotuladas, implicando que a saída não é conhecida previamente pelo usuário, mas definida pelo algoritmo. No aprendizado por reforço, o modelo recebe o estado atual do sistema, um objetivo e uma lista de ações permitidas com suas restrições. Em vez de pares de entrada e saída, o modelo aprende por tentativa e erro, experimentando ações para alcançar o objetivo e maximizar uma recompensa. Os algoritmos genéticos são uma classe de algoritmos de otimização que se baseiam na teoria da evolução natural, usando conceitos como seleção, recombinação e mutação para encontrar soluções eficazes em problemas multiobjetivos. Além disso, algoritmos genéticos também podem ser usados para resolver problemas de otimização em tarefas complexas, onde é difícil encontrar soluções ótimas de forma tradicional (5–7).

Portanto, aliando-se a possibilidade e necessidade da otimização do módulo elástico de próteses de titânio para aplicações biomédicas, este trabalho tem como objetivo a aplicação de um método de *machine learning* associado a um algoritmo genético para a determinação de uma composição química, de liga de titânio beta, na qual apresente um valor mínimo de seu módulo de elasticidade dentre as composições analisadas.

2. Revisão bibliográfica

2.1. Titânio beta

As ligas de titânio beta são as mais versáteis dentro das classes de ligas de titânio. Estas oferecem as maiores relações resistência/peso, elevada tenacidade e resistência à fadiga em grandes seções transversais. Todavia, em comparação com as ligas de titânio convencionais, algumas desvantagens se fazem evidentes, tais como o aumento da densidade, uma janela de processamento relativamente restrita devido aos elementos presentes em sua composição e um custo mais elevado (8). Processadas geralmente por fundição/lingotamento, as ligas titânio beta apresentam desenvolvimentos recentes em relação à sua produção pela metalurgia do pó e manufatura aditiva (2).

A liga Ti-13V-13Cr-3Al representa um marco significativo como a primeira liga de titânio beta desenvolvida e comercializada com sucesso. Foi amplamente aplicada na aeronave SR-71 "Blackbird" devido à sua excecional resistência específica, resultando em benefícios significativos frente à economia de peso. Ademais, suas propriedades estáveis em temperaturas elevadas contribuíram para sua preferência e adoção nesse contexto de aplicação aeroespacial (9).

Em relação à aplicação médica, pelo fato de conterem estabilizantes da fase beta (principalmente Mo, Ta e Zr), as ligas de titânio beta não apresentam apenas módulos elásticos menores, mas também possuem biocompatibilidade superior em comparação com outros tipos de ligas de titânio. Entretanto, esses elementos estabilizadores possuem um preço elevado, deste modo, o custo destas ligas também aumenta (2).

Como forma de contornar a situação, novas ligas de titânio beta de baixo custo foram desenvolvidas nos últimos anos com base no método de equivalência de molibdênio (Equação 1), que contém principalmente elementos de liga de baixo custo, como Fe, Mn, Sn e Cr (2).

$$\begin{aligned} MO_{eq} = {} & 1,0\,[Mo] + 1,25[V] + 0,59[W] + 0,3[Nb] + 0,25[Ta] \\ & + 1,93[Fe] + 1,84[Cr] + 2,46[Ni] + 2,26[Mn] + 2,67[Co] \\ & + 1,51[Cu] + 0,3[Sn] + 0,31[Zr] + 3,01[Si] - 1,47[Al] \end{aligned} \quad (1)$$

O molibdênio equivalente (Mo_{eq}) é utilizado para avaliar o efeito equivalente de estabilizadores beta em ligas de Ti, na Equação 1 a contribuição do alumínio no Mo_{eq} é negativa devido ao fato deste elemento ser alfa-estabilizante (10).

O limite crítico inferior de estabilização beta é determinado como Mo_{eq} próximo a 11,8 %p., indicando que qualquer liga com um valor de Mo_{eq} acima deste valor exibiria uma estrutura única CCC de titânio beta desde que seja resfriada a partir do campo de fase beta a uma taxa de resfriamento suficiente (2,11)

Tratando-se de uma liga titânio beta correspondente a aplicada no SR-71 "Blackbird" (Ti-13V-11Cr-3Al), pode-se observar em sua microestrutura (Figura 1) a presença de grãos equiaxiais de fase beta, com tamanho médio de 220 μm, fundida usando um forno de fusão a arco a vácuo, homogeneizada a 1.100 °C por 3h, posteriormente laminada a quente a 1.100 °C e temperada em água (12).

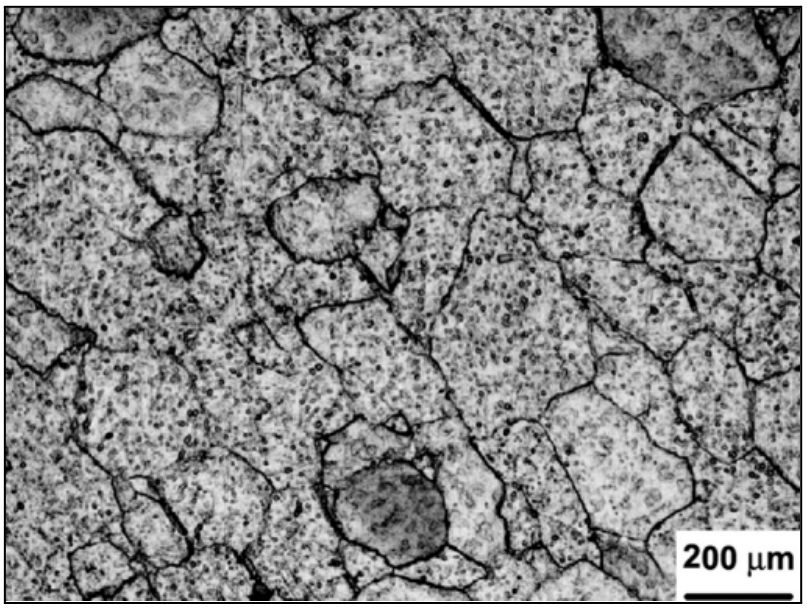

Figura 1. Microestrutura típica (Ti-β) da liga Ti-13V-11Cr-3Al fundida, homogeneizada, laminada a quente e temperada (12).

Sabe-se que na proximidade do limite inferior de estabilização beta (Figura 2), algumas segundas fases de ω, α'' e α' podem ser inevitavelmente precipitadas, uma vez que são sensíveis ao processamento e tratamentos térmicos realizados, potencialmente resultando em um aumento no módulo de elasticidade (2,13).

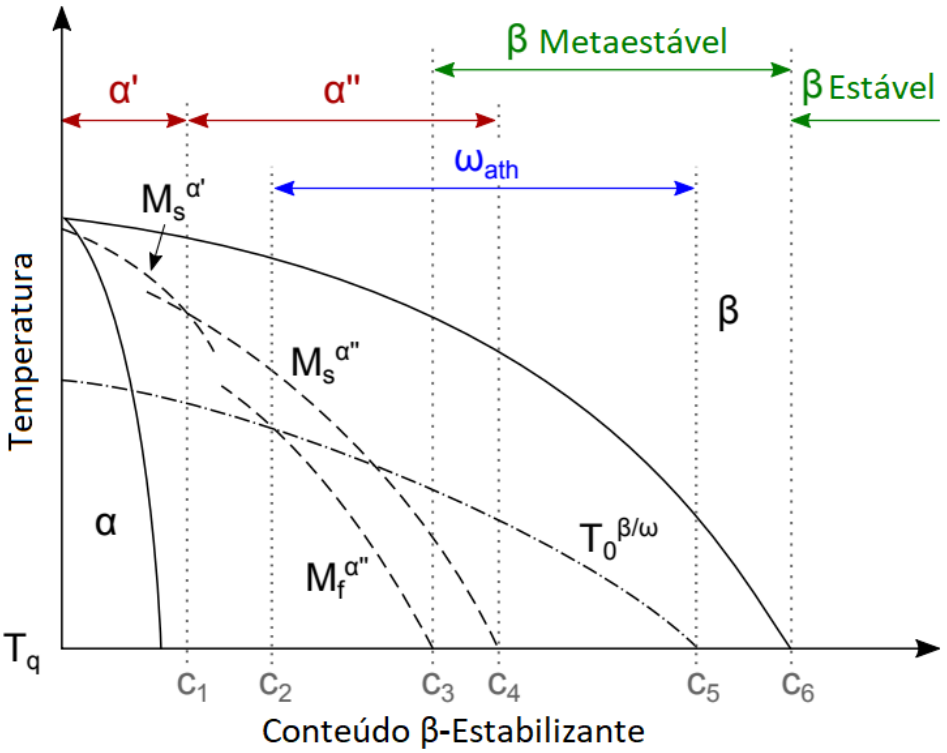

Figura 2. Diagrama de fases genérico para o titânio em função do teor de elemento estabilizador de fase beta. Adaptado de (12).

A sensibilidade da formação dessas fases é observada na Figura 3, na qual representam diferentes condições de resfriamento da liga Ti–6Al–4V. Observou-se que a espessura e o comprimento da fase alfa diminuem com o aumento da taxa de resfriamento, porém com este aumento, houve a formação de uma segunda fase α' (Figura 3).

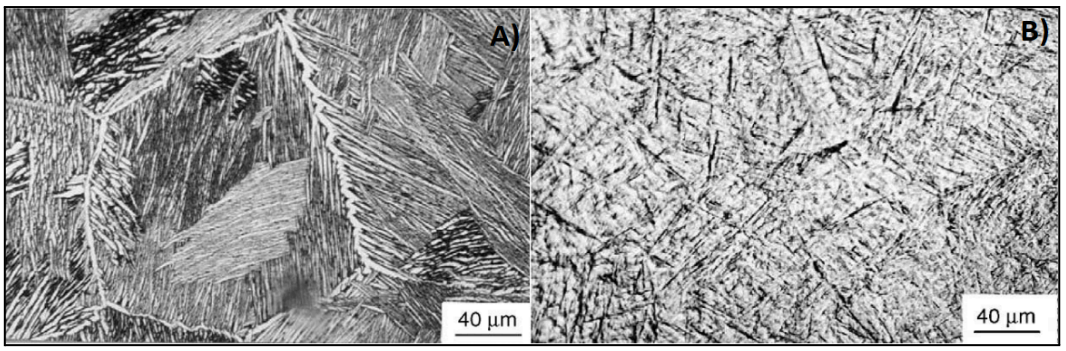

Figura 3. Microestrutura da liga Ti–6Al–4V em diferentes condições de resfriamento. a) Resfriamento ao ar, lamelas da fase α presentes na matriz beta e b) Resfriamento em água, apresentando placas martensíticas da fase α'. Adaptado de (14).

Experimentalmente, apenas quando o valor de Mo_{eq} é maior que 13,0 %p., essas segundas fases podem ser evitadas e o módulo de elasticidade pode alcançar um valor mínimo (2,13,15).

O menor valor de módulo de elasticidade relatado para uma liga de titânio policristalina do tipo beta, Ti-35Nb-4Sn ou Ti-24Nb-4Zr-7,9Sn, é em torno de 40 GPa (16). Outras ligas beta tipicamente utilizadas no âmbito médico estão elencadas na Tabela 1 (17).

Tabela 1. Módulo elástico de ligas de titânio beta utilizadas no âmbito médico (15).

Liga	E (GPa)	Liga	E (GPa)
Ti–29Nb–6Ta–5Zr	43	Ti–29Nb–13Ta–4.6Sn	66
Ti–35Nb	45	Ti–29Nb–13Ta–4.6Sn	66
Ti–35Nb–7Zr–5Ta	45	Ti–29Nb–13Ta–4.6Sn	66
Ti–10Zr–5Ta–5Nb	51.97	Ti–29Nb–13Ta–4.6Sn–0.4O	66
Ti–(18–20)Nb–(5–6)Zr	52.5	TLM Alloy	67
Ti–25Ta–25Nb	55	Ti–12Mo–6Zr–2Fe	74.5
Ti–25Ta–25Nb	55	Ti–15Mo–2.5Zr–2Fe	74.8
Ti–35Nb–7Zr–7Ta	55	Ti–15Mo–2.8Nb–3Al	75
Ti–35Nb–5Ta–7Zr–0.4O	60	Ti–7.5Mo–3Fe	85

2.1.1. Aplicação em Implantes

Durante o desenvolvimento de próteses, um dos principais pontos a serem questionados é a diferença entre o módulo de elasticidade do osso e da prótese, de maneira a minimizar o fenômeno de *Stress Shielding (*SS*)* (16). O SS acontece quando existe uma diferença pronunciada entre os módulos de elasticidade do osso e da prótese, tornando-a responsável por suportar as cargas exercidas pelo corpo humano ao invés do que naturalmente seria atribuído ao osso humano. Deste modo, ocorre uma redução na densidade óssea e, consequentemente, uma fragilização do sistema de sustentação natural humano, uma vez que este se renova e remodela em resposta às cargas sob as quais é submetido (1,15).

Como visto anteriormente (Tabela 1), as ligas beta são mais vantajosas para o desenvolvimento de ligas para aplicações biomédicas. Seus módulos de elasticidade ficam, em geral, abaixo de 80 GPa ao serem submetidas a um recozimento (16,17). Por exemplo, nota-se que a liga comercialmente utilizada *TiOsteum* apresenta um módulo de elasticidade de 55 GPa, enquanto a liga SUS 316L de aço inoxidável apresenta um módulo em torno de 200 GPa, e o osso humano apresenta apenas 30 GPa. Esta diferença reforça a necessidade da busca por uma liga que minimize esta discrepância (1,2,16).

Outro fator importante a ser considerado no desenvolvimento de implantes, é a biocompatibilidade, definida como a interação bem-sucedida entre tecidos vivos e materiais inertes, essencial para o sucesso de implantes (19).

Diversas características do implante podem influenciar na resposta do sistema do usuário, e. g., porosidade, topografia da superfície e sua composição. Assim, os elementos químicos presentes no metal, influenciam tanto em possíveis reações adversas, quanto na promoção de integração óssea, influenciando fortemente em sua biocompatibilidade. Alguns elementos se apresentam como biocompatíveis como o Ti, Nb, Zr e Ta, enquanto outros são considerados tóxicos, como o Ni, Pb, V e Co (19).

2.2. Machine Learning

O *Machine Learning* (ML) tem como objetivo principal adquirir automaticamente relações e padrões relevantes a partir de exemplos e observações por métodos computacionais. Avanços nessa área possibilitaram o desenvolvimento de sistemas inteligentes, capazes até de simular habilidades cognitivas humanas.

Esses sistemas impactam tanto o setor empresarial, quanto o pessoal, sendo utilizados para melhorar a tomada de decisões corporativas e para criar assistentes que se adaptam às preferências dos usuários, por exemplo (6,20).

A eficácia desses sistemas se baseia em modelos analíticos que geram previsões, regras, recomendações e resultados semelhantes. Esses modelos, cada vez mais desenvolvidos com ML, são impulsionados pela disponibilidade de estruturas de programação, grandes volumes de dados e acesso a maior capacidade computacional. O ML elimina a necessidade de traduzir conhecimento humano em formatos compreensíveis para máquinas, acelerando o desenvolvimento de sistemas inteligentes (6).

Um algoritmo de ML é um conjunto de instruções computacionais usadas pelo sistema para aprender a partir de um banco de dados. Ele define como o modelo é treinado para produzir previsões ou classificações. Existem diferentes tipos de algoritmos, como supervisionados, não supervisionados e de aprendizado por reforço, cada um adequado a diferentes tarefas e tipos de dados (6,20).

As abordagens de ML variam de acordo com o problema e os dados disponibilizados previamente:

1. Aprendizado supervisionado: Utiliza conjuntos de treinamento contendo exemplos de entrada e suas respectivas saídas rotuladas. Esses pares são usados para ajustar os parâmetros do modelo. Após o treinamento, o modelo pode prever saídas com base em novos dados de entrada não vistos (6,20);
2. Aprendizado não supervisionado: Identifica padrões em dados sem rótulos ou respostas predefinidas. O objetivo é encontrar estruturas de interesse, como grupos de elementos com propriedades comuns (6,20);
3. Aprendizado por reforço: Não utiliza pares de entrada e saída. O sistema é informado sobre o estado atual, o objetivo, as ações permitidas e restrições ambientais. Ele aprende por tentativa e erro, buscando maximizar recompensas para atingir o objetivo (6,20);

Os modelos de ML são escolhidos com base no desempenho durante o treinamento. Eles apresentam uma "aprendizagem" do sistema, identificando padrões nos dados para prever ou decidir sobre novas entradas. Esses modelos são representações matemáticas ou estatísticas e variam conforme a tarefa: em

problemas de classificação, o modelo mapeia entradas para classes; em problemas de regressão, ele correlaciona entradas com variáveis contínuas (20).

Deste modo, o funcionamento dos modelos de ML variam em diferentes tipos de representações. Por exemplo, o Bayesian Ridge usa probabilidade bayesiana para ajustar coeficientes com regularização automática, a Regressão Linear Múltipla modela relações lineares entre uma variável dependente e várias independentes, o *Support Vector Regression* utiliza margens de erro controladas e kernels para capturar relações não lineares, o *Gradient Boosting Regression* combina iterativamente árvores de decisão para corrigir erros residuais e lidar com padrões complexos e o *Random Forest Regressor* combina múltiplas árvores de decisão construídas aleatoriamente e faz previsões pela média (21–25).

2.2.1. Machine Learning na Ciência dos Materiais

O uso de ML no desenvolvimento de materiais permite analisar grandes volumes de dados experimentais e simulados, acelerando o processo de descoberta de novos materiais com propriedades desejadas, sendo possível prever o comportamento sob diferentes condições, otimizar suas propriedades e reduzir significativamente o tempo e os custos associados a métodos experimentais tradicionais (26)

A Tabela 2 exemplifica alguns estudos realizados no campo da ciência dos materiais utilizando diversos modelos de ML, diferentes tipos de conjuntos de dados, bem como em diferentes tópicos (materiais e previsões de propriedades moleculares, modelagem e simulações de materiais, *design* de materiais, descoberta e aprendizagem ativa, e caracterização de materiais e aplicações em imagens) (26). Vale destacar que na Tabela 2 também consta os métodos de ML descritos, de forma que os mesmos serão discutidos mais profundamente posteriormente no presente texto.

Tabela 2. Diferentes estudos realizados no campo da ciência dos materiais. Adaptado de (26).

Referência	Métodos de ML	Conjunto de Dados	Resumo dos Tópicos Abordados
Kalidindi, et al.	Tutorial GPR	Simulado e experimental	Conceitos fundamentais para sistemas de conhecimento baseados em materiais
Velli, et al.	Múltiplos (k-NN, SVM, GBR, etc.)	Experimental	Efeito dos parâmetros do laser na estrutura e propriedades dos materiais
Vanpoucke, et al.	Regressão linear e múltipla	Simulado e experimental	Formação e energia de entalpia para ligas metálicas binárias
Zhang, et al.	GPR	DFT	Previsão da formação de entalpia para ligas metálicas ternárias
Honrao, et al.	SVM	DFT	Representações configuracionais para previsões de energia de formação
Huang e Ling	Múltiplos (NN, RFR, etc.)	DFT	Previsão de energia de ligação em compostos inorgânicos
Magedov, et al.	NN	DFT	Previsão de ordem de ligação em moléculas
Sharma, et al.	RFR	DFT	ML para defeitos substitucionais e mudanças em propriedades de íons
Sadat e Wang	NN, RFR	Simulado	Previsão de banda proibida em cristais fotônicos
Chen, et al.	NN	Simulado	Resposta termo-mecânica de materiais para compostos unidimensionais
Parker, et al.	Múltiplos (clustering, classificação e regressão)	Simulado	Relações estrutura-propriedade de nanopartículas de Pt para catalisadores
Zhuo, et al.	Extreme GBR	Experimental	Previsão de transição Ce dopado para fosfatos inorgânicos
Costine, et al.	MDS, k-NN, RF	Experimental	Previsão de crescimento em monocamadas de dissulfeto metálico
Lightstone, et al.	GPR	Experimental	Previsão de reatividade de polímeros
Salahuddin, et al.	SVM	Experimental	Previsão de densidade em nanofluídos de etileno glicol
Alade, et al.	SVM, NN	Experimental	Previsão baseada em ML de viscosidade de nanofluídos
Gurgen, et al.	NN, SVM	Experimental	Previsão de perda de massa para materiais revestidos com magnésio
Alade, et al.	SVR	Experimental	Previsão de parâmetros de rede para compostos cúbicos de A3X6
Jacobs, et al.	RFR	Simulado	Detecção de falha em materiais cimentícios
Santos, et al.	NN, extreme GBR e regressão ridge	Simulado	Previsão de propriedades térmicas em cerâmicas refratárias
Zagaceta, et al.	NN	DFT	Potenciais de rede neural para ligas de Ni-Mo
Mangold, et al.	NN	DFT	Previsão de fonons e propriedades térmicas em ligas Mn-Ge

Zeledon, et al.	NN e GBR	DFT	Previsão de potencial de desenvolvimento usando representações otimizadas de características
Mazhnik e Oganov	NN	DFT	Triagem de materiais superduros baseada em ML
Dieb, et al.	NN	Simulado	Design inverso de materiais estruturados em gradientes para óticas de raio-x
Zheng, et al.	Modelo Gauss–Bayesiano	Simulado e experimental	Design de metamateriais para alta absorção sonora em baixas frequências
Tian, et al.	SVM	Experimental	Papel da estimativa de incerteza em aprendizado ativo eficiente
Ma, et al.	Agrupamento e NN	Experimental	Ligação de microestrutura às condições de processamento em ligas de urânio-molibdênio
Ziatdinov, et al.	GPR	Experimental	Processos Gaussianos para alcançar super-resolução em Microscopia de Força Kelvin de Contato
Vasudevan, et al.	Seleção de modelo	Experimental	Inferência Bayesiana em excitação de banda na microscopia de varredura por sonda para imagem dinâmica

No campo da ciência dos materiais, os dados podem ser aproximadamente classificados em quatro tipos principais: propriedades dos materiais derivadas de experimentos e/ou simulações (como propriedades físicas, químicas, estruturais, termodinâmicas e dinâmicas), dados de reações químicas (como taxa de reação e temperatura de reação), dados de imagem (incluindo imagens de microscopia eletrônica de materiais e fotos de superfícies de materiais), entre outros (27).

A coleta de informações e o processamento de dados na ciência dos materiais consistem em duas partes: seleção de dados e engenharia de características (27).

Na seleção dos dados, a abordagem adotada envolve uma criteriosa seleção, levando em consideração diversos aspectos, como o tipo dos dados, sua qualidade e formato. O uso de dados de alta qualidade é crucial para evitar a inclusão de informações imprecisas, ausentes ou redundantes, sendo minimizada devido a existência de diversos bancos de dados de alta qualidade para materiais (*Open Quantum Material Database*, *Material Project*, *Computational Materials Repository*, etc.), os quais já têm sido utilizados na área da ciência computacional de materiais (27).

Após a etapa de seleção dos dados, é essencial realizar a extração de características apropriadas para o objetivo de previsão, processo conhecido como engenharia de características. A engenharia de características envolve a extração

de características dos dados brutos para torná-los adequados à aplicação de algoritmos de ML (27).

Nos métodos tradicionais de ML, a seleção de características é realizada manualmente. Entretanto, a engenharia manual de características possui suas limitações. A experiência humana muitas vezes tem dificuldade em identificar as características mais representativas para a tarefa de previsão. Recentemente, o avanço da aprendizagem profunda eliminou a necessidade de engenharia manual de características, o que pode se tornar uma tendência no campo da aprendizagem de máquina aplicada à ciência dos materiais (27).

Um estudo realizado por (28) ressalta que atualmente a descoberta de perovskitas estáveis experimentalmente desenvolvidas e de alto desempenho é lenta e ineficiente quando realizada apenas por experimentos e/ou cálculos. Assim, o uso de ML desempenha um papel importante na evolução deste cenário.

Portanto, estudos anteriores mostraram que as propriedades descritas na Tabela 3 têm um efeito direto no módulo de elasticidade de qualquer material (29).

Tabela 3. Principais propriedades que influenciam no módulo de elasticidade dos materiais (29).

Equação	Descrição
$\text{VEC} = \sum_{i=1}^{n} C_i (VEC)_i$	Concentração média de elétrons de valência calculada pela regra da mistura
$a_m = \sum_{i=1}^{n} C_i a_i$	Parâmetro de rede médio calculado pela regra da mistura
$T_m = \sum_{i=1}^{n} C_i T_i$	Temperatura de fusão média calculada pela regra da mistura
$\Delta T = \sqrt{\sum_{i=1}^{n} C_i (T_i - T_m)^2}$	Diferença na temperatura de fusão T_i, ponderada pela composição Ci para cada elemento i
$\Delta a = \sqrt{\sum_{i=1}^{n} C_i (a_i - a_m)^2}$	Diferença no parâmetro de rede a_i , ponderada pela composição Ci para cada elemento i
$\Delta \chi = \sqrt{\sum_{i=1}^{n} C_i (\chi_i - \chi_m)^2}$	Diferença na eletronegatividade de Pauling χ_i, ponderada pela composição C_i para cada elemento i

$Bo = \sum_{i=1}^{n} C_i (Bo)_i$	Ordem de ligação calculada pela regra da mistura
$Md = \sum_{i=1}^{n} C_i (Md)_i$	Energia média do orbital d calculada pela regra da mistura

De forma geral, as propriedades descritas na Tabela 3 estão relacionadas com a Energia de Ligação (E_0) pois descrevem características dos átomos e da estrutura cristalina de ligas. E_0 está associada ao mínimo da curva de energia potencial entre dois átomos em função da distância interatômica (Figura 4). Quanto maior for E_0 (em módulo), maior será o módulo elástico (30).

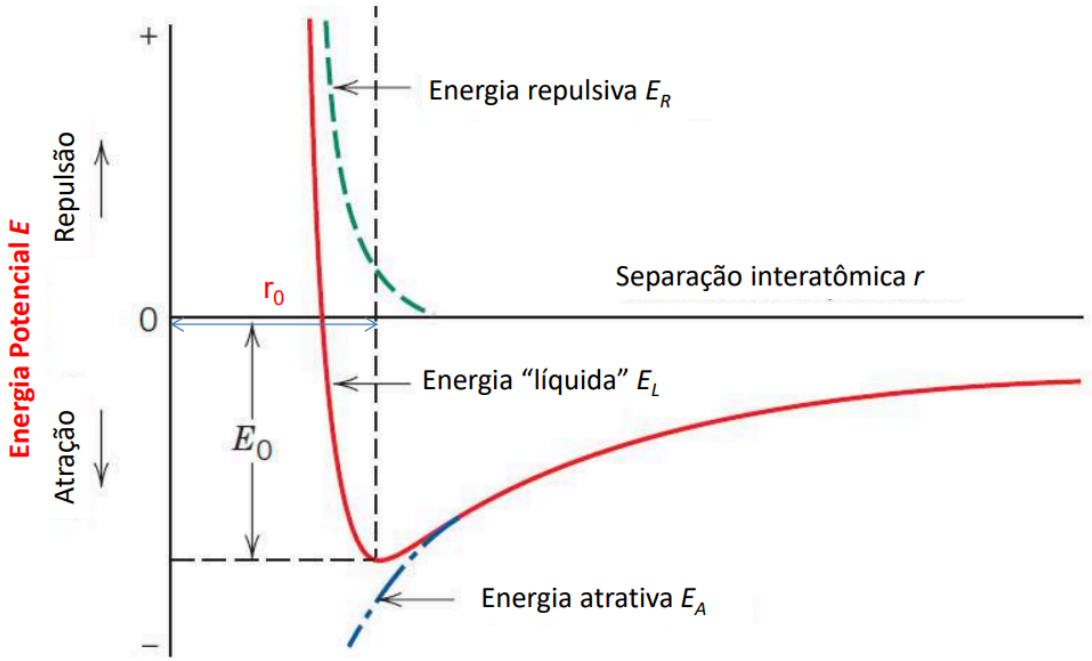

Figura 4. Energia potencial (Eo) em função da distância interatômica (r) para dois átomos isolados. Adaptado de (30).

Ademais, além da importância na escolha do conjunto de dados de treinamento com base no tipo de liga que está sendo estudada, também é necessário validar contra dados experimentais de alta qualidade de proveniência conhecida, uma vez que ao treinar modelos de ML para prever características de ligas, é vantajoso incluir ligas de composição semelhante no conjunto de dados de treinamento.

Em um estudo realizado por (29) indica que quando os modelos foram treinados com dados contendo apenas ligas refratárias, as previsões foram mais próximas dos valores experimentais, comparando-se com o treinamento utilizando dados contendo liga refratárias e não-refratárias.

2.2.2. Dimensionamento do modelo

A principal preocupação ao construir um modelo de ML a partir de dados é a capacidade de generalização do modelo resultante. Se o algoritmo de aprendizado não for aplicado corretamente, o modelo pode se ajustar demais aos dados, o que significa que ele prevê os próprios dados de treinamento (*overfitting*), ao invés de captar a relação funcional entre as variáveis de entrada e de resposta. Essas preocupações estão diretamente relacionadas ao procedimento de ajuste, que deve equilibrar o desempenho preditivo do modelo resultante (31).

Um dos parâmetros importantes a serem considerados no dimensionamento dos modelos de *machine learning*, especialmente em problemas de regressão, é a Raiz do Erro Quadrático Médio (*Root Mean Square Error*, RMSE), descrita pela Equação 2. Este parâmetro mede o desvio médio entre os valores preditos pelo modelo e os valores reais, sendo uma das formas mais comuns de avaliar a precisão de um modelo (21,32).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{m} (y^{i} - y^{-})^{2}}{m}} \quad (2)$$

Consequentemente, quanto menor for o valor do erro encontrado, melhor será a precisão do modelo frente aos dados estudados (21,32).

2.2.3. *Gradient Boosting* (GB)

O modelo *Gradient Boosting* (GB) é um modelo integrado que apresenta alto desempenho e melhor estabilidade dentre os modelos existentes, pode ser usado tanto em problemas de regressão quanto de classificação. Proposto por Friedman, estende o algoritmo de *Boosting* para resolver problemas de regressão, sendo denominado *Gradient Boosting Regressor (GBR)* (21).

O termo *Boosting* se refere a uma família de algoritmos que convertem aprendizes fracos combinados de forma sequencial em aprendizes fortes, onde cada aprendiz seguinte tenta corrigir os erros do anterior, aumentando a precisão geral do modelo.

Um aprendiz fraco é um modelo simples, apresentando um desempenho ligeiramente melhor do que o acaso, enquanto um aprendiz forte é um modelo complexo e robusto que pode fazer previsões altamente precisas, sendo capaz de resolver problemas de aprendizado de maneira eficaz.

O GBR pode entender que aprendizes fracos são ligeiramente melhores do que uma escolha aleatória, enquanto aprendizes fortes têm um desempenho perfeito. Esta abordagem pode produzir um modelo de aprendizado em conjunto (*ensemble learning*) a partir de modelos preditivos fracos (33).

O modelo GBR utiliza os gradientes negativos da função de perda para encontrar o valor mínimo, sendo que, estes gradientes representam a direção de maior redução no erro da função de perda durante a previsão, ou seja, minimizando o erro e melhorando o ajuste aos dados. A queda do gradiente no espaço da função é usado em etapas para construir o conjunto, sendo mais confiável e mais fácil quando comparado a outros algoritmos de ML (21,33).

Em um estudo realizado por (21), em comparação com outros métodos como: Bayesian Ridge, Regressão Linear Múltipla e SVR, o GBR apresentou um melhor desempenho na predição estatística entre a metilação do DNA e a idade, apresentando o menor desvio médio absoluto (MAD) e raiz do erro quadrático médio (RMSE) dentre os modelos analisados (21).

O GBR tem sido amplamente utilizado em pesquisas biológicas devido à sua capacidade de lidar com dados não lineares e ruidosos, além de suportar diferentes funções de perda (21).

As árvores de decisão, conhecidas como árvores de regressão, são métodos que consistem em dividir o espaço dos parâmetros de entrada em regiões distintas e não sobrepostas, seguindo um conjunto de regras "if-then". As regras identificam regiões que têm uma resposta que melhor corresponda ao preditor, e dentro de cada região, assim como uma constante, é ajustado. Tratando-se da regressão, a constante tende a representar a média da saída do conjunto de dados de treinamento da região correspondente (34).

Exemplificando o funcionamento de uma árvore, na Figura 5 se pode notar dois parâmetros de entrada T e t, que representam a temperatura externa do ar e o tempo da semana, respectivamente. A saída E corresponde ao consumo de energia de um edifício. T_1 e T_2 são os pontos de divisão da temperatura, t_1 e t_2 são os pontos de divisão do tempo, E_1, E_2, E_3, E_4 e E_5 são os nós terminais, também chamados de folhas da árvore (saídas) (34).

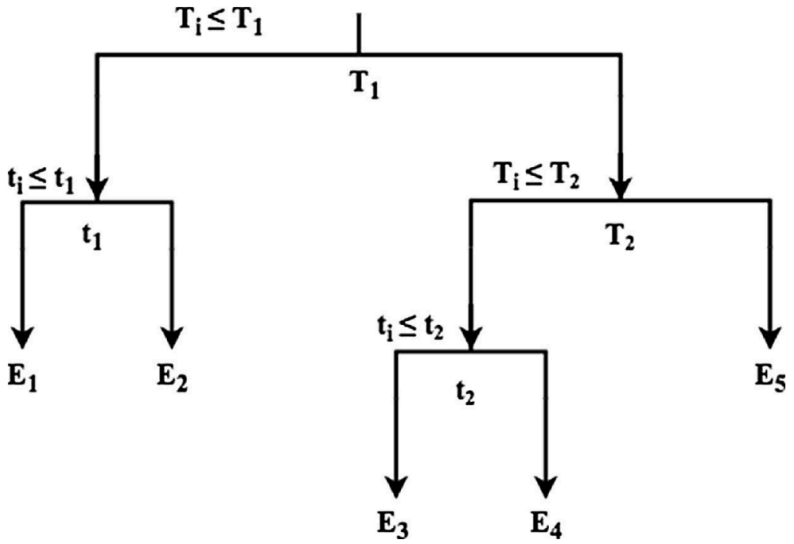

Figura 5. Árvore de decisão com dois parâmetros de entrada: A temperatura externa T e o horário da semana t. Adaptado de (34).

Os pontos de divisão são escolhidos para minimizar uma função de perda, que no caso de árvores de regressão é geralmente o erro quadrático médio (MSE). As divisões continuam até que um critério de parada seja atingido, por exemplo, o número de pontos de treinamento dentro de uma região atinge um determinado limite (34).

Esses diferentes passos de divisão correspondem à profundidade da árvore. Para fazer uma previsão para novos pontos de dados, estes são divididos seguindo os pontos de divisão treinados, e as mesmas constantes nos nós terminais são usadas para fazer as previsões (34).

Pode-se limitar a profundidade das árvores ou o número mínimo de instâncias necessário para dividir um nó. Ao contrário da *random forest*, os valores padrão para esses parâmetros no *Gradient Boosting* são configurados para limitar

severamente o poder expressivo das árvores (por exemplo, a profundidade é geralmente limitada a cerca de 3 a 5) (35).

Os parâmetros comumente testados para os modelos de *Gradient Boosting* são:

- Taxa de aprendizado: *learning_rate* ou *shrinkage*;
- Profundidade máxima da árvore (*max_depth):* Número mínimo de instâncias necessário para dividir um nó;
- Taxa de Amostragem (*subsample*): Tamanho das amostras aleatórias. Ao contrário da *random forest*, isso geralmente é feito sem reposição;
- Número de Atributos (*max_features)*: Considerados ao buscar a melhor divisão;
- Número Mínimo de Amostras (*min_samples_split)*: que são necessárias para dividir um nó interno.

O procedimento de regularização mais simples introduzido para o GB é o *subsample*. Este procedimento mostrou melhorar as propriedades de generalização do modelo, ao mesmo tempo que reduz o esforço computacional necessário. A ideia por trás deste método é introduzir certa aleatoriedade no processo de ajuste dos parâmetros (31).

Para controlar a complexidade do modelo, o mesmo é regularizado através de *shrinkage*. Este método reduz o tamanho dos passos incrementais e penaliza a importância de cada iteração subsequente. Esta técnica baseia-se no princípio de que é melhor aperfeiçoar um modelo em muitos pequenos passos, em vez de poucos passos grandes. Se uma das iterações do *boosting* for incorreta, seu impacto negativo pode ser facilmente corrigido em etapas posteriores (31).

2.2.4. *Random Forest* (RF)

Leo Breiman (2001) introduziu o algoritmo de floresta aleatória (*Random Forest*, RF), fundamentado em preditores baseados em árvores de decisão (22,36). Este algoritmo é amplamente utilizado em tarefas de classificação e regressão (*Random Forest Regressor*, RFR), pois combina os resultados das árvores individuais em uma única saída: utilizando “votação” para classificação e média para regressão (22,36).

A aleatoriedade no algoritmo ocorre de duas maneiras. Primeiro, o conjunto de dados é redimensionado com reposição, processo conhecido como *bagging*. Diferentemente do *boosting*, que transforma aprendizes fracos em fortes de forma sequencial, no *bagging* o aprendizado ocorre de maneira paralela (36).

A segunda etapa de randomização ocorre nos nós de decisão: em cada nó, um subconjunto aleatório de preditores é selecionado (geralmente, a raiz quadrada do número total de preditores). O algoritmo testa todos os possíveis limiares para as variáveis selecionadas, escolhendo aquele que proporciona a melhor divisão dos dados, resultando em subconjuntos mais homogêneos (36).

Este processo de seleção aleatória e teste continua até que os nós "puros" sejam alcançados — isto é, quando todos os dados em um nó pertencem à mesma classe (classificação) ou têm o mesmo valor (regressão), ou até que um critério de parada seja atingido. A construção da floresta envolve a repetição deste procedimento para formar várias árvores de decisão (geralmente de 100 a 1000), compondo a floresta aleatória (36).

O modelo consiste em um conjunto de árvores de decisão (Figura 6), cada uma construída a partir de uma versão do conjunto de treinamento. Cada árvore segue o princípio de partição sucessiva: começando no nó raiz, o mesmo processo de divisão é repetido até que os critérios de parada sejam atendidos (37).

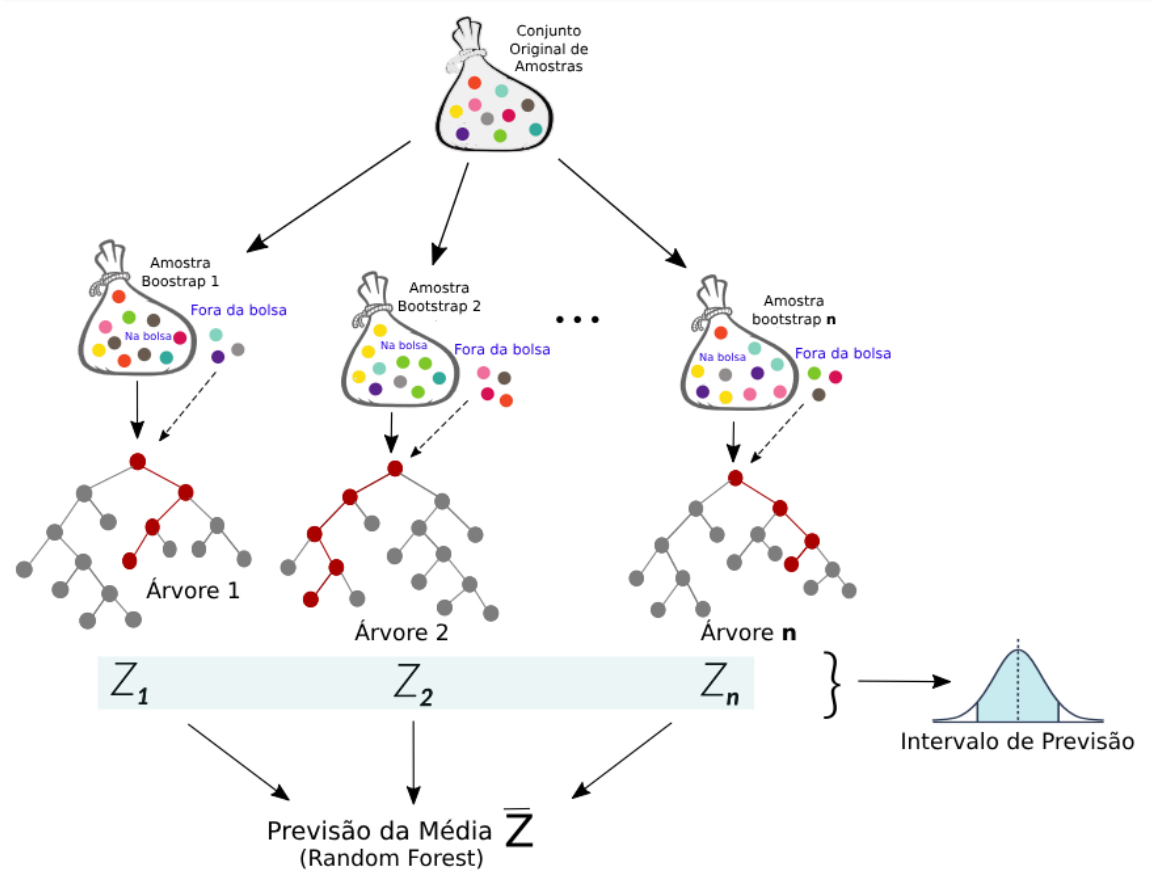

Figura 6. Diagrama esquemático da *Random Forest*. Adaptado de (38).

A força preditiva da floresta advém da combinação de várias árvores fracas. O desempenho é particularmente bom quando as correlações entre as árvores são baixas, ou seja, quando elas produzem predições suficientemente distintas. A maior vantagem do algoritmo está em sua capacidade de identificar interações e efeitos não lineares entre preditores sem necessidade de especificá-los previamente (37).

Após o treinamento do modelo, ele pode ser usado para previsões. Cada novo caso é avaliado por todas as árvores na floresta, e o resultado final é determinado pela classe mais votada (classificação) ou pela média das previsões (regressão). No entanto, a interpretação do modelo é mais complexa do que a de modelos lineares. Com tantas árvores, é difícil identificar as variáveis mais influentes (36). Para isso, utilizam-se métricas conhecidas como “importância das variáveis”, que avaliam o impacto de cada variável com base no desempenho do modelo na ausência dela. Essas métricas indicam a relevância de uma variável, mas não

especificam como ela influencia o resultado. Isso torna o algoritmo particularmente adequado para análise de grandes conjuntos de dados (22,36).

Um exemplo prático de aplicação é descrito por C. Carranza (38), que utilizou três anos de medições diárias em campos agrícolas para modelar uma floresta aleatória de duas formas: interpolando valores em pontos aleatórios da série temporal e extrapolando para prever estados futuros com base nos dados anteriormente obtidos.

Como outros modelos, a floresta aleatória possui parâmetros ajustáveis para melhor adequação aos dados. Entre os principais, destacam-se:

- ***mtry***: Define o número de variáveis a serem testadas em cada divisão. Geralmente, é a raiz quadrada do número total de preditores (36);
- ***ntree***: Especifica o número de árvores. Mais árvores geralmente aumentam a precisão, mas também a demanda computacional. Entre 100 e 1000 árvores são comuns, embora florestas menores também possam ser eficazes (36);
- ***nsplit***: Limita o número máximo de divisões testadas em cada variável, útil para conjuntos com muitos preditores contínuos. Isso pode acelerar o processamento significativamente, especialmente em grandes conjuntos de dados (36).

Esses parâmetros podem ter diferentes nomenclaturas em pacotes estatísticos distintos, mas seus princípios permanecem os mesmos (36).

2.2.5. *Support Vector* (SV)

O *Support Vector Machine* (SVM) é um algoritmo de aprendizado supervisionado amplamente utilizado para classificação e regressão. Em sua forma básica, é um classificador linear binário que identifica um hiperplano ótimo para separar duas classes, maximizando a margem de separação (23).

Para isso, utiliza os vetores de suporte, que são os pontos de treinamento mais próximos da fronteira de decisão e têm impacto direto na sua definição (Figura 7). Esse processo iterativo, chamado de aprendizado, busca minimizar erros de classificação e é especialmente eficaz quando os dados são linearmente separáveis (23).

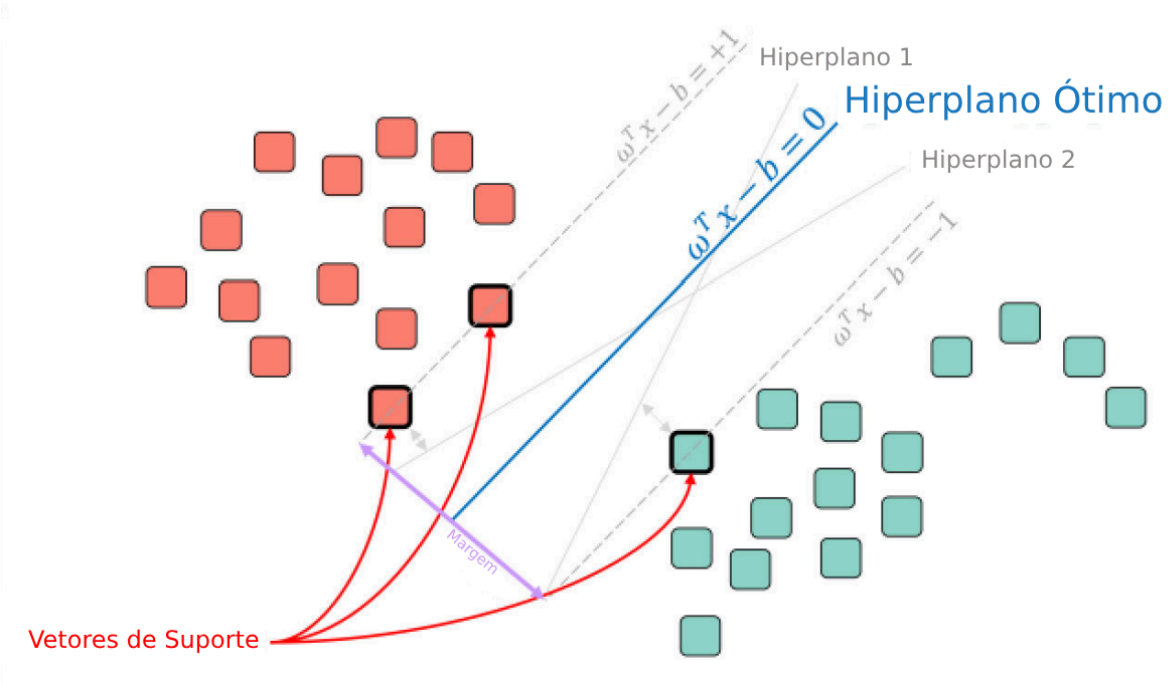

Figura 7. Exemplo de separação de dados pelo SVM. Adaptado de (23).

No entanto, em situações em que as classes não são linearmente separáveis, o SVM utiliza estratégias como a margem suave (que incorpora variáveis de folga para lidar com dados sobrepostos) e o truque do kernel (que mapeia os dados para um espaço de maior dimensionalidade tornando-os separáveis linearmente, Figura 8). Funções kernel, como RBF, polinomial e linear, são usadas para projetar os dados nesse novo espaço, sendo sua escolha essencial para o desempenho do modelo (23).

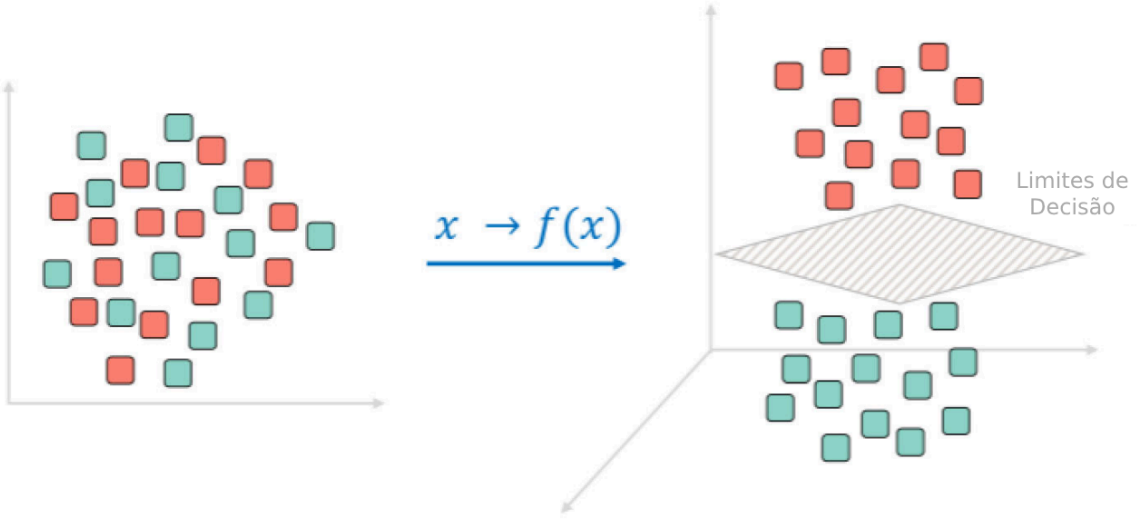

Figura 8. Exemplo de dados não linearmente separáveis com o truque do kernel pelo SVM. Adaptado de (23).

Na regressão, o algoritmo é conhecido como *Support Vector Regression* (SVR), onde um modelo é ajustado para prever a relação entre as variáveis independentes e dependentes. Nesse caso, a relação entre a variável dependente (y) e as variáveis independentes (x) é modelada por uma função f(x), descrita pela Equação 3:

$$f(x) = \omega^{*}\varphi(x) + b \tag{3}$$

Onde ω é o vetor peso, φ(x) é uma função não linear que mapeia o vetor do espaço de entrada x para um espaço de características de alta dimensionalidade, e b é o termo de viés (39).

O SVR realiza uma regressão linear nesse espaço de alta dimensionalidade, que é gerado a partir do mapeamento não linear dos dados de entrada (39).

Como método supervisionado, o SVR utiliza uma função de perda simétrica, que penaliza igualmente tanto as superestimações quanto as subestimações. Além disso, erros cujos valores absolutos sejam menores que um determinado limiar (ϵ) são ignorados, tanto acima quanto abaixo da estimativa.

Uma das principais vantagens do SVR é que sua complexidade computacional não depende da dimensionalidade do espaço de entrada. Além

disso, possui excelente capacidade de generalização, com alta precisão de predição (40).

2.3. Algoritmo genético

O Algoritmo Genético (AG) é uma técnica de busca e otimização inspirada no processo biológico de seleção natural que serve, principalmente, para a resolução de problemas complexos de otimização de múltiplas variáveis. Ele opera sobre uma população inicial de soluções aleatórias, aplicando os princípios de "sobrevivência do mais apto". Os componentes fundamentais do AG incluem representação cromossômica, seleção, cruzamento, mutação e cálculo da função de aptidão. Essas características tornam o AG robusto e eficiente para a resolução de problemas complexos em engenharia (5,7,27,41).

O processo se inicia com a geração aleatória de uma população X, composta por n cromossomos. Cada cromossomo representa uma solução codificada, composta por genes que correspondem aos parâmetros do problema a ser otimizado. A aptidão de cada cromossomo é avaliada, determinando sua adequação à solução do problema (5,7,27,41).

A seleção ocorre com base nos valores de aptidão, escolhendo dois cromossomos, C1 e C2, para cruzamento. Um operador de cruzamento de ponto único, com probabilidade de cruzamento (CP), gera uma prole (O). Após o cruzamento, um operador de mutação é aplicado à prole, com uma probabilidade de mutação (MP), resultando na nova prole ('O′). Este processo continua até a formação de uma nova população completa onde seu término é forçado, conforme ilustrado na Figura 9 (5,7,27,41).

Figura 9. Processo geral de algoritmos evolutivos. Adaptado de (27).

O AG ajusta dinamicamente as probabilidades de cruzamento e mutação, buscando soluções ótimas. Essa capacidade de busca global permite ao AG explorar diversas soluções simultaneamente, aumentando sua eficiência (5,42).

Assim, alguns indivíduos gerados estarão mais próximos de uma solução possível do que outros, mesmo que não atendam a todos os requisitos do problema, baseados em seu valor de aptidão (*fitness*) (5,42).

Embora tenham sido adotados tardiamente no campo dos materiais, os estudos baseados em algoritmos genéticos estão se tornando cada vez mais proeminentes em diversas áreas dessa área, sendo recentemente aplicados com sucesso em uma ampla gama de problemas, abrangendo desde o *design* de ligas, processamento de polímeros, compactação e sinterização de pós, metalurgia de produção de ferrosos, fundição contínua, laminação de metais, corte de metais, soldagem e muitos outros (7).

Também vale ressaltar que existem algumas configurações utilizadas no AG que regem os passos realizados anteriormente (seleção, cruzamento e mutação), como por exemplo: Número de Gerações, Tamanho da População, Número de Repetições e Tamanho do Pódio.

2.3.1. Esquema de Seleção

A seleção é crucial nos Algoritmos Genéticos (AG), determinando quais sequências participam da reprodução. A pressão de seleção (intensidade com que indivíduos com maior aptidão são favorecidos durante o processo de seleção para a próxima geração) impacta diretamente a taxa de convergência do AG. Técnicas como roleta, classificação, torneio, Boltzmann e amostragem universal estocástica são amplamente utilizadas (Figura 10) (5,7)

Na seleção por roleta, cada indivíduo recebe uma porção proporcional à sua aptidão em uma “roda de roleta”, que é girada para seleção. Apesar de eficiente, apresenta limitações relacionadas a flutuações aleatórias nos dados de treinamento ou por incertezas nos parâmetros. A seleção por classificação usa *rankings* em vez de aptidão, reduzindo o risco de convergência prematura (5,7)

Já a seleção por torneio compara pares de indivíduos, selecionando os mais aptos para a próxima geração. A amostragem universal estocástica adota intervalos equidistantes na seleção, garantindo maior equidade entre os indivíduos. A seleção

por Boltzmann, baseada em entropia, favorece a escolha de sequências ótimas com menor tempo de execução (5,7).

2.3.2. Operadores de Cruzamento

Os operadores de cruzamento são essenciais para gerar novas soluções em AG, combinando informações genéticas de dois ou mais genitores. Entre os métodos mais usados estão o cruzamento de ponto único, de dois pontos, de ponto k, uniforme, parcialmente mapeado, ordenado, com preservação de precedência, aleatório, substituto reduzido e de ciclo, como ilustrado na Figura 10 (5,7).

No cruzamento de ponto único, um ponto aleatório é escolhido para trocar a informação genética dos pais a partir desse ponto (Figura 11a). Nos cruzamentos de dois pontos e de ponto k, a troca ocorre em segmentos definidos por múltiplos pontos (Figura 11b).

O cruzamento uniforme, utilizado neste trabalho, trata cada gene individualmente, decidindo aleatoriamente se ele será trocado com o outro cromossomo (Figura 11c) (5,7).

O cruzamento parcialmente mapeado utiliza dois pais, combinando materiais genéticos de ambos e preenchendo os alelos faltantes com genes do segundo genitor, sendo altamente eficiente em exploração (5,7). O cruzamento ordenado copia partes do pai para a prole com base em pontos de corte e preenche o restante com valores não repetidos (5,7). Já o cruzamento com preservação de precedência mantém a ordem das soluções originais, enquanto o cruzamento aleatório embaralha os valores para evitar vieses durante a operação (5,7). O cruzamento substituto reduzido busca diversidade genética entre os pais, mas é menos eficaz se as representações genéticas forem muito semelhantes (5,7). Por fim, o cruzamento de ciclo distribui os genes em ciclos alternados entre os pais (5,7).

2.3.3. Operadores de Mutação

A mutação é essencial para preservar a diversidade genética e garantir a evolução da população, desempenhando um papel crucial na formação da próxima geração. Entre os operadores mais utilizados estão os de deslocamento, inversão e mutação embaralhada, ilustrados na Figura 10 (5,7).

A mutação de deslocamento consiste em deslocar uma subsequência escolhida aleatoriamente dentro da solução. Podendo variar em dois tipos:

substituindo uma parte da solução por outra (mutação de troca) ou insere uma parte da solução em um local diferente (mutação de inserção) (5,7).

A mutação de inversão inverte uma subsequência entre dois pontos aleatórios na solução, realocando-a em outra posição também escolhida aleatoriamente (5,7)

Por fim, a mutação embaralhada reorganiza os elementos de uma subseção da solução em ordem aleatória. Após a mutação, a nova solução é avaliada para verificar melhorias no valor de aptidão (5,7).

Os operadores de mutação podem ser combinados com diversas técnicas de codificação e cruzamento. Por exemplo, o cruzamento parcialmente mapeado é frequentemente associado à mutação de inversão, enquanto o esquema de codificação por permutação demonstra desempenho superior na busca pela solução ideal (5,7).

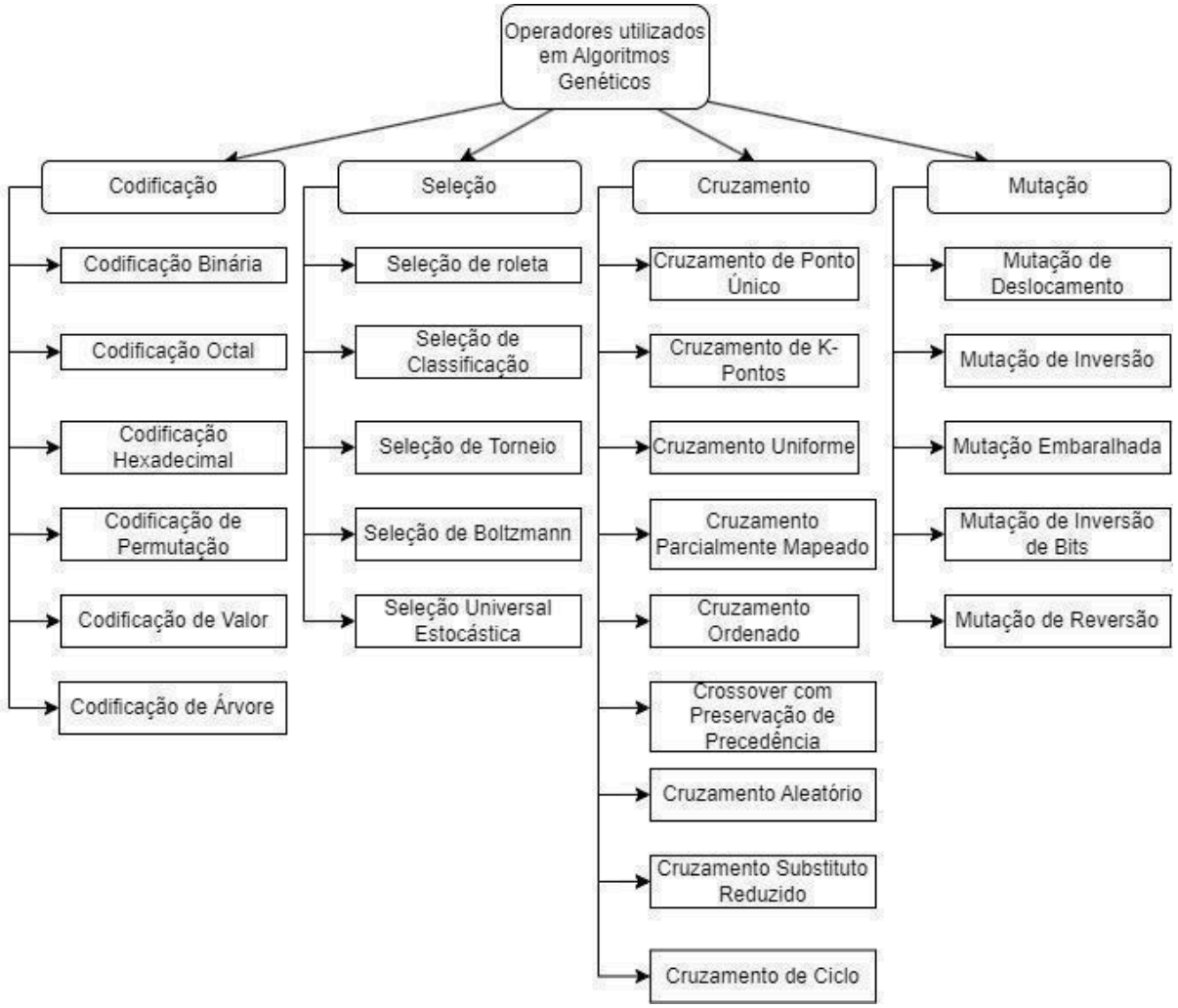

Figura 10. Operadores utilizados no algoritmo genético. Adaptado de (5).

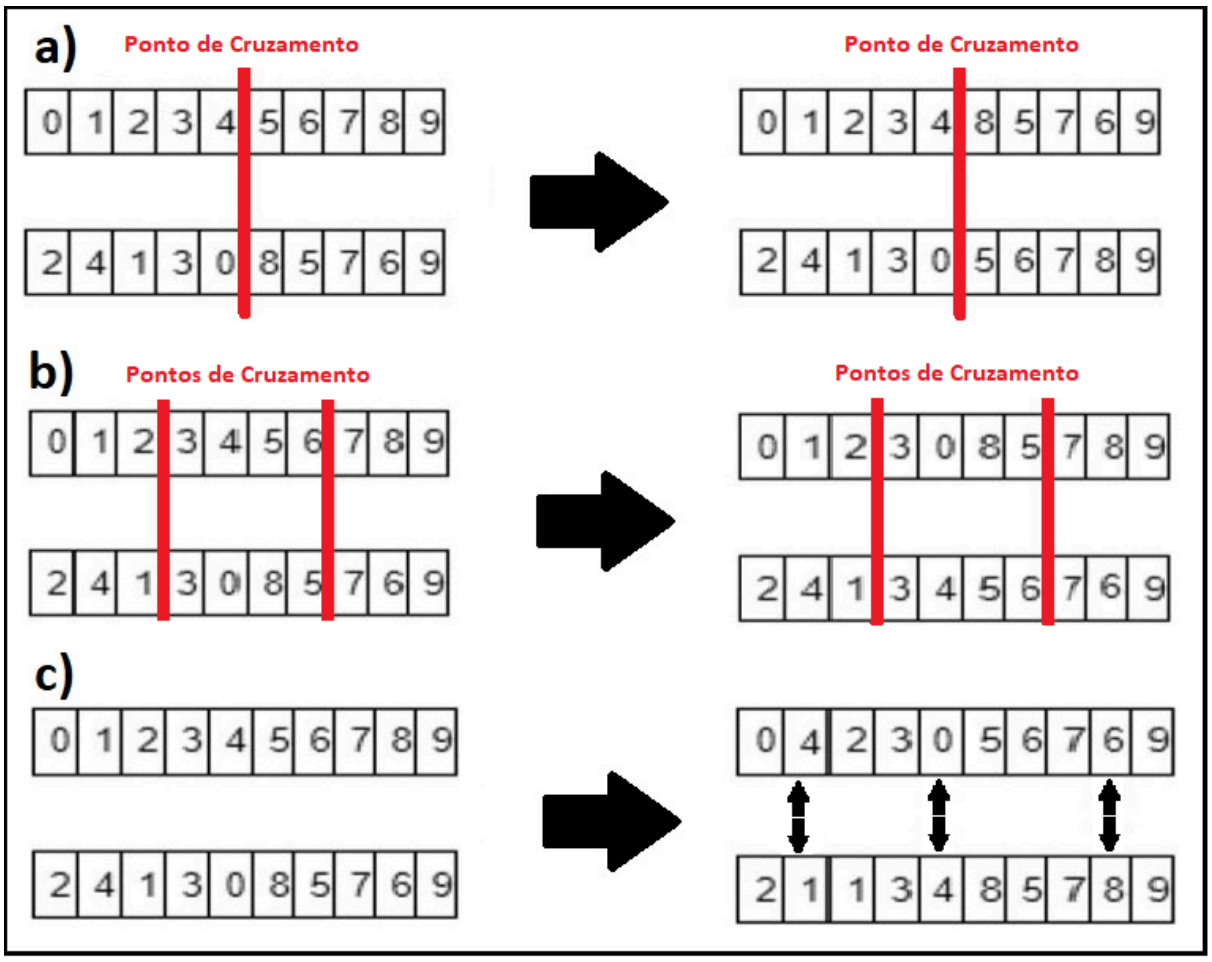

Figura 11. Esquema de diferentes cruzamentos. a) Troca de informações genéticas após um ponto de cruzamento, b) Troca de informações genéticas entre pontos de cruzamento e c) Troca de genes individuais. Adaptado de (5).

3. Materiais e métodos

3.1. Base de dados

Previamente ao trabalho realizado, construiu-se uma base de dados de diversas ligas de titânio, proveniente de diversos estudos anteriores publicados em revistas como: *Elsevier*, *Nature*, MDPI e *SpringerLink*, sendo elas majoritariamente beta estável. Devido a restrições encontradas na obtenção dessas ligas, também foram incluídas ligas de alta entropia e ligas titânio alfa, totalizando 460 ligas diferentes.

O conteúdo da base de dados consiste na composição das ligas, descritas em peso atômico por elementos como: Cromo (Cr), Cobre (Cu), Ferro (Fe), Manganês (Mn), Molibdênio (Mo), Nióbio (Nb), Silício (Si), Estanho (Sn), Tântalo (Ta), Vanádio (V), Zircônia (Zr) e, por fim, Titânio (Ti) como balanço, além do Módulo de elasticidade (E) de cada uma dessas ligas. A base de dados utilizada neste trabalho pode ser encontrada em (44).

3.2. Saídas dos Algoritmos

Por conseguinte, utilizando o conteúdo anterior, foi realizada a engenharia de características, isto é, os modelos de aprendizado supervisionado de ML citados foram construídos e alimentados.

Assim, reuniu-se as seguintes propriedades dos elementos químicos (38) em análise: eletronegatividade, raio atômico, parâmetro de rede, temperatura de fusão e número de valência, ordem de ligação, energia média do orbital d e o molibdênio equivalente, bem como a composição de cada liga.

Utilizando linguagem Python (44), as características foram calculadas de acordo com a regra da mistura para cada liga, obtendo-se os valores médios das propriedades, bem como seus respectivos desvios padrão, de acordo com a Tabela 4. Deste modo, valores de entrada para o ML foram gerados, referentes às propriedades citadas na Tabela 3, anteriormente ditas como de extrema importância para determinação do módulo de elasticidade.

Tabela 4. Características estudadas.

Propriedade	Característica	Cálculo
Desvio padrão da energia média do orbital d	*Md_std_devs_w*	$\sqrt{\sum_{i=1}^{n} C_i((Md)_i - \bar{Md})^2}$
Desvio padrão do raio atômico	*AtomicRadius_std_devs_w*	$\sqrt{\sum_{i=1}^{n} C_i(r_i - \bar{r})^2}$
Desvio padrão da temperatura de fusão	*MeltT_std_devs_w*	$\sqrt{\sum_{i=1}^{n} C_i(T_i - \bar{T})^2}$
Desvio padrão da ordem de ligação	*Bo_std_devs_w*	$\sqrt{\sum_{i=1}^{n} C_i((Bo)_i - \bar{Bo})^2}$
Concentração média de elétrons de valência	*VEC_avgs_w*	$\sum_{i=1}^{n} C_i(VEC)_i$
Desvio padrão do molibdênio equivalente	*Meq_std_devs_w*	$\sqrt{\sum_{i=1}^{n} C_i((Mo_{eq})_i - \bar{Mo_{eq}})^2}$
Desvio padrão da concentração elétrons de valência	*VEC_std_devs_w*	$\sqrt{\sum_{i=1}^{n} C_i((VEC)_i - \bar{VEC})^2}$
Média do molibdênio equivalente	*Meq_avgs_w*	$\sum_{i=1}^{n} C_i(Mo_{eq})_i$
Média da temperatura de fusão	*MeltT_avgs_w*	$\sum_{i=1}^{n} C_i T_i$
Desvio padrão da eletronegatividade	*Electronegativity_std_devs_w*	$\sqrt{\sum_{i=1}^{n} C_i(\chi_i - \bar{x})^2}$
Média do parâmetro de rede	*LatticeConstant_avgs_w*	$\sum_{i=1}^{n} C_i a_i$
Média do raio atômico	*AtomicRadius_avgs_w*	$\sum_{i=1}^{n} C_i r_i$
Média da ordem de ligação	*Bo_avgs_w*	$\sum_{i=1}^{n} C_i(Bo)_i$

Desvio padrão do parâmetro de rede	*LatticeConstant_std_devs_w*	$\sqrt{\sum_{i=1}^{n} C_i(a_i - \bar{a})^2}$
Média da energia média do orbital d	*Md_avgs_w*	$\sum_{i=1}^{n} C_i (Md)_i$
Média da eletronegatividade	*Electronegativity_avgs_w*	$\sum_{i=1}^{n} C_i \chi_i$

Posteriormente, utilizando linguagem de programação em Python, os modelos RFR, GBR e SVR foram avaliados, a fim de comparar o RMSE obtido entre eles, valor no qual mede a diferença entre os valores preditos e os valores reais dos dados, fornecendo uma estimativa de quão bem o modelo está ajustado ao conjunto de dados, de modo a utilizar 80% dos dados para treinamento e 20% para teste, proporção padrão presente na literatura (43). O código utilizado pode ser encontrado no repositório (44).

Também foram avaliadas as importâncias de cada característica (44), ou seja, sua contribuição para o módulo elástico da liga. Essas importâncias são calculadas de acordo com a redução de variância durante a divisão de características nas árvores de decisão, ou seja, características que resultam em maior redução de erro são consideradas mais importantes.

Logo, a partir dos valores de RMSE, o algoritmo que apresentou o menor valor de erro dentre os modelos estudados foi escolhido para dar prosseguimento no estudo, utilizando-se o mesmo no algoritmo genético.

Deste modo, algumas combinações de hiperparâmetros de otimização de ML da literatura (35,41) foram testadas também entre os modelos, representados na Tabela 5, a fim de apresentar uma melhor adequação aos dados em análise.

Por fim, foram testadas diferentes combinações destes hiperparâmetros, onde uma combinação em específica apresentou o decréscimo mais expressivo frente aos valores padrão obtidos de RMSE, ou seja, indicando uma combinação que proporcionaria o melhor cenário de otimização para cada modelo. Assim como proposto anteriormente, o modelo no qual apresentou o menor valor de RMSE foi utilizado no algoritmo genético com seus parâmetros otimizados.

Tabela 5. Hiperparâmetros estudados.

Algoritmo	Parâmetro	Padrão	Estudado	Otimizado
Gradient Boosting Regression	n_estimators	100	50, 100 e 200	200
	learning_rate	0.1	0.025, 0.05, 0.1, 0.2 e 0.3	0.1
	max_depth	3	2, 3, 5, 7 e 10	3
	min_samples_split	2	2, 5, 10 e 20	20
	min_samples_leaf	1	1, 2, 4 e 6	1
	random_state	None	1, 10, 54, 100 e 500	1
	max_features	None	sqrt ,0.25 e 1	0.25
Random Forest Regression	n_estimators	100	50, 100, 200, 500 e 1000	100
	max_depth	None	2, 3, 5, 7 e 10	10
	min_samples_split	2	2, 5, 10, 20 e 50	5
	min_samples_leaf	1	1, 2, 4, 6 e 8	1
	random_state	None	10, 45, 100 e 500	45
	max_features	1	Auto e Sqrt	Sqrt
	bootstrap	True	True e False	False
Support Vector Regression	epsilon	0.1	0.05, 0.1 e 0.2	0.2
	kernel	rbf	linear e rbf	rbf
	gamma	scale	scale, 0.05, 0.1 e 0.2	0.05
	C	1	1, 10, 100, 1000 e 10000	100
	degree	3	1, 2, 3 e 4	1

A consequente utilização do algoritmo genético aliado aos modelos em destaque foi adaptada de (42) no qual o autor utiliza da linguagem Python para projetar materiais cerâmicos com propriedades óticas otimizadas, com base em suas composições.

3.3. Restrições do Algoritmo Genético

Algumas restrições foram definidas na composição e no molibdênio equivalente das ligas resultantes, bem como nas configurações do algoritmo genético, representados na Tabela 6.

Tabela 6. Restrições do Algoritmo Genético.

Restrição	Magnitude
Cr, at.%	10 máx
Cu, at.%	10 máx
Fe, at.%	10 máx
Mn, at.%	10 máx
Mo, at.%	10 máx
Nb, at.%	30 máx
Si, at.%	-
Sn, at.%	10 máx
Ta, at.%	10 máx
Ti, at.%:	50 - 100
V, at.%	10 máx
Zr, at.%	30 máx
Molibdênio Equivalente	10 mín
Número de Gerações	1000
Tamanho da População	100
Tamanho do Pódio	1
Número de Repetições	100

O Número de Gerações delimita a quantidade de gerações de indivíduos, enquanto o Tamanho da População é o número de indivíduos na população de uma única execução do algoritmo genético. O Número de Repetições é a quantidade de vezes que o algoritmo será executado de forma independente. Por fim, o Tamanho

do Pódio evidencia o número de melhores indivíduos a cada execução, baseados em seu *fitness*, garantindo que as melhores soluções sejam registradas.

Deste modo, os valores para essas configurações do algoritmo genético foram embasados em estudos da literatura, a exemplo do estudo realizado por (45), onde aborda o uso de algoritmos genéticos para projetar polímeros com propriedades desejadas, além de propor melhorias que integram conhecimento químico para lidar com restrições práticas, como estabilidade e viabilidade.

Em relação a faixa de componentes de liga, também se optou por eliminar qualquer quantidade de Silício presente nas ligas geradas pelo algoritmo genético, a fim de minimizar a formação de silicatos na microestrutura final. Ademais, as restrições para as faixas especificadas estão em concordância com diversas composições de ligas presentes na literatura, uma vez que as restrições representam 432 (93%) ligas presentes na base de dados deste estudo.

Por fim, o molibdênio equivalente foi estimado em um valor que propicie a estabilização teórica da fase beta na microestrutura do material, baseada em sua composição (2).

3.4. Cálculo da função de aptidão

A função de aptidão utilizada neste estudo foi calculada pela distância euclidiana no espaço (Equação 4), onde valores menores indicam maior probabilidade de seleção (5,42).

$$f = \sqrt{(x - x_d)^2 + \varepsilon_1 + \varepsilon_2 + \varepsilon_3} \quad (4)$$

Na equação, x é o valor das propriedades de cada indivíduo, enquanto x_d é o valor desejados dessa propriedade, ε_1 ε_2 e ε_3 são fatores de penalidade aplicadas, cujas funções são penalizar os resultados que estão fora de um determinado intervalo estabelecido previamente no algoritmo genético (5,42).

Essa função retorna um valor numérico que reflete o quão bem o indivíduo atende aos objetivos do problema, que para o caso do presente trabalho, é a minimização do módulo de elasticidade.

4. Discussões e resultados

4.1. Parâmetros Não-Otimizados

Como citado anteriormente, foram testados os algoritmos frente a sua adequação aos dados em análise. Dentre os modelos estudados, o GBR apresentou um menor RMSE em comparação com seus concorrentes, em um valor de 12,67 GPa (Figura 14). Os algoritmos RF e SVR apresentaram um RMSE de 13,09 GPa (Figura 15) e 14,64 GPa (Figura 16) respectivamente.

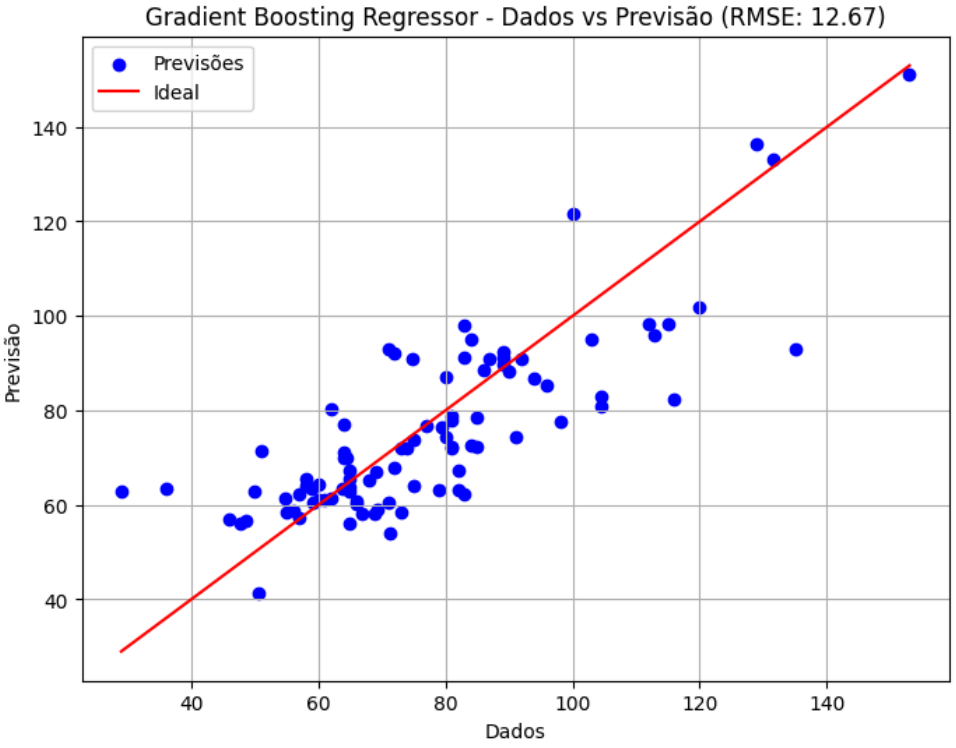

Figura 14. *Gradient Boosting Regression* Padrão - Dados versus Previsão.

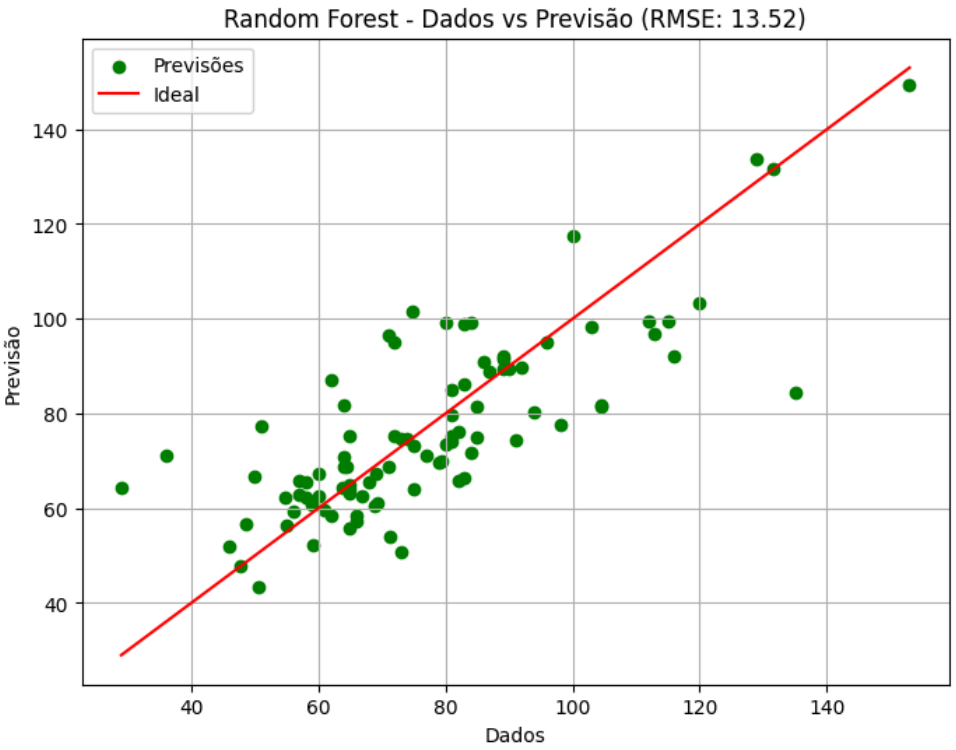

Figura 15. *Random Forest Regression* Padrão - Dados versus Previsão.

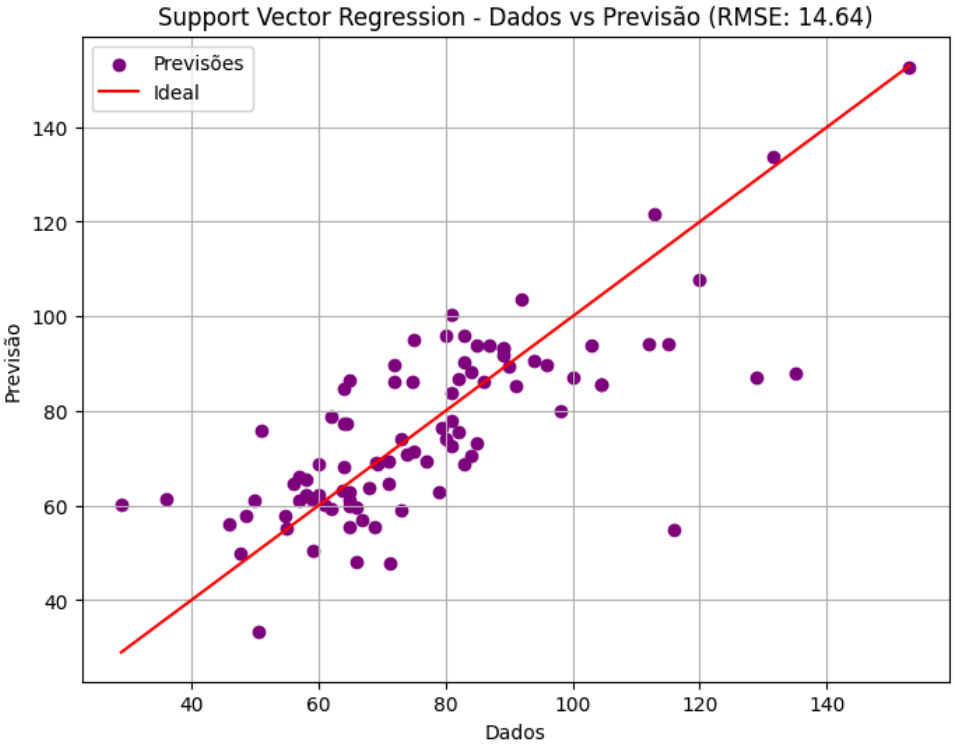

Figura 16. *Support Vector Regression* Padrão - Dados versus Previsão.

Assim, calculou-se a importância de cada característica na contribuição do módulo de elasticidade. A importância de todas as características foi evidenciada nas Figuras 11 e 12, para cada modelo analisado.

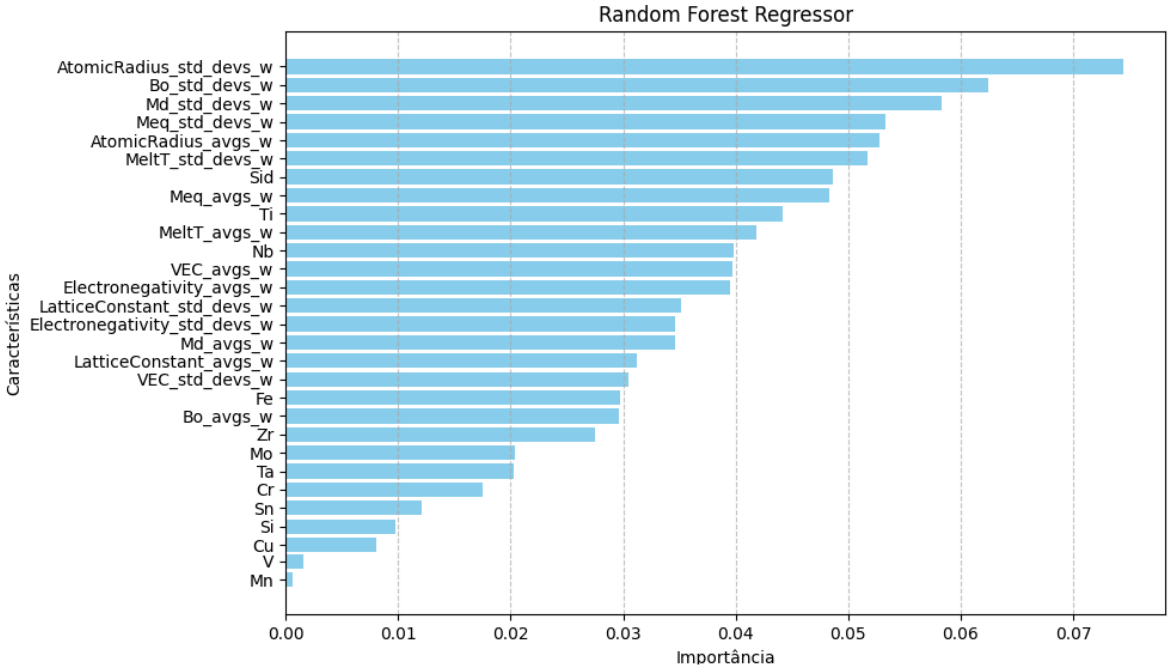

Figura 12. Importâncias por característica do RFR.

De acordo com a Figura 12, as 3 características mais importantes para o algoritmo de RF na influência do módulo de elasticidade são:

1. Desvios padrão do raio atômico (AtomicRadius_std_devs_w) = 0,0745 A;
2. Desvio padrão da ordem de ligação (Bo_std_devs_w) = 0,0624;
3. Desvio padrão da energia média do orbital d (Md_std_devs_w) = 0,0583 eV.

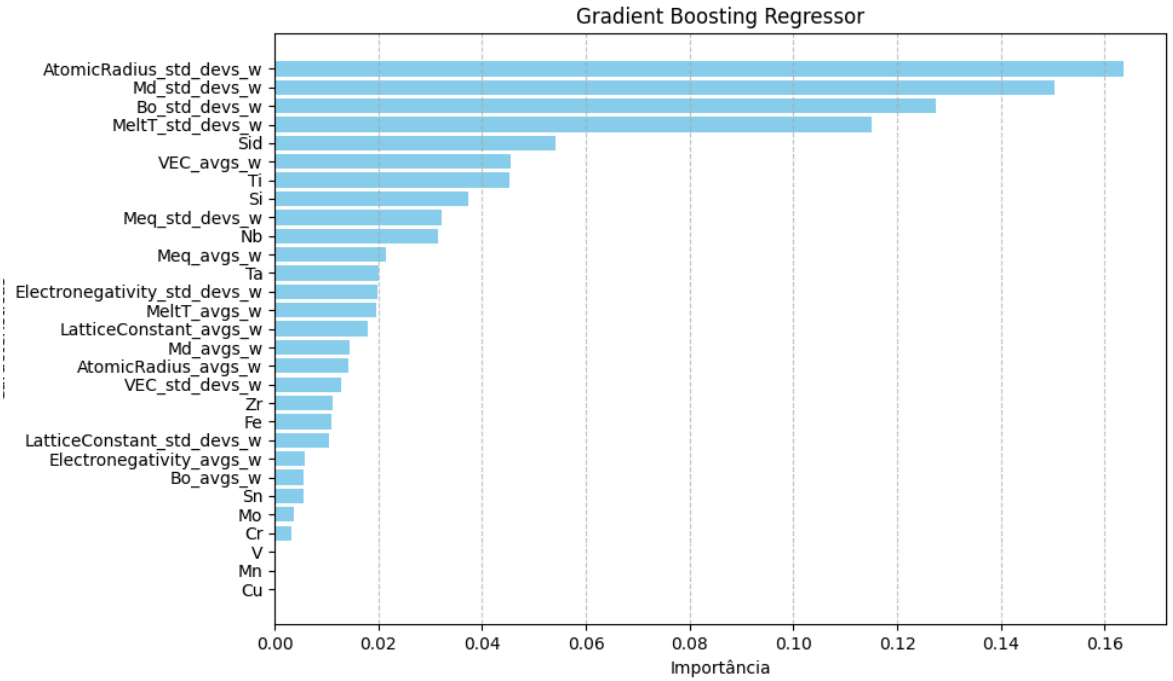

Figura 13. Importâncias por característica do GBR.

De acordo com a Figura 13, as 3 características mais importantes para o algoritmo de RF na influência do módulo de elasticidade são:

1. Desvios padrão do raio atômico (AtomicRadius_std_devs_w) = 0,0745 A;
2. Desvio padrão da energia média do orbital d (Md_std_devs_w) = 0,1552 eV;
3. Desvio padrão da ordem de ligação (Bo_std_devs_w) = 0,1237.

Tratando-se do SVR, por não ser um modelo embasado em árvores de decisão, a importância das características não pode ser determinada da mesma forma que os outros modelos, assim, para determinar a importância de suas características, é necessário uso de uma função que não é facilmente interpretável.

Logo, com a saída do código, conclui-se que as características mais influentes no módulo elástico de um material, dentre as analisadas, estão relacionadas ao raio atômico (r), a ordem de Ligação (Bo) e energia média do orbital d (Md).

Assim, aliou-se o GBR utilizando os hiperparâmetros padrão ao Algoritmo Genético, a fim de realizar a previsão do módulo elástico das diferentes ligas geradas pelo algoritmo de busca. Deste modo, iniciou-se um ciclo, onde cada indivíduo foi avaliado com base em uma função de aptidão (Equação 4), a qual determina a qualidade de sua solução para o problema. O valor da propriedade

desejada (x_d) é representada pelo menor valor de módulo de elasticidade encontrado no pódio.

Posteriormente, na seleção, indivíduos foram escolhidos para reprodução com base em sua aptidão, priorizando aqueles com melhor desempenho. O cruzamento foi realizado de maneira uniforme e ocorreu entre os indivíduos selecionados, combinando suas características genéticas para gerar novos indivíduos. Por fim, a mutação introduziu variações aleatórias no cromossomo dos indivíduos para evitar a estagnação e promover a exploração de novas áreas do espaço da solução. O indivíduo mais apto presente na população ao fim deste ciclo entrou para o “hall da fama”.

Este ciclo foi realizado em um total de 1000 vezes, utilizando uma população de 100 indivíduos, para cada uma das execuções do código, totalizando 100 execuções, conforme especificado nas configurações do AG, na Tabela 6.

Por fim, cada indivíduo das 100 execuções do código foram inseridos em uma tabela. As cinco melhores ligas de titânio foram exemplificadas na Tabela 7, ordenadas em ordem decrescente pelo parâmetro de *fitting*.

Tabela 7. Resultados obtidos utilizando os parâmetros padrão. As composições químicas constam em % em massa dos elementos.

%Cr	%Cu	%Fe	%Mn	%Mo	%Nb	%Si	%Sn	%Ta	%Ti	%V	%Zr	Fitting	$E_{Previsto}$	Mo_{eq}
0.806	0.000	0.000	2.419	0.806	0.000	0.000	0.806	0.806	76.613	4.032	13.710	0.245	48.999	17.492
4.972	3.315	4.972	5.525	6.630	2.210	0.000	0.000	0.000	52.486	9.392	10.497	0.260	51.987	58.591
6.667	2.222	4.444	5.556	7.222	1.667	0.000	0.556	0.000	51.667	8.889	11.111	0.260	51.987	59.250
0.000	8.000	2.000	8.000	9.000	5.000	0.000	1.000	0.000	50.000	6.500	10.500	0.261	52.254	56.350
3.261	7.609	0.543	6.522	9.783	4.348	0.000	1.630	0.000	52.717	2.717	10.870	0.261	52.254	51.750

Nota-se que o melhor resultado obtido nessa modalidade, referindo-se a precisão de previsão, deve-se ao fitness de 0,245 na qual alcançou um módulo elástico previsto de 48,999 GPa.

4.2. Parâmetros Otimizados

Foram testados os algoritmos frente a sua adequação aos dados em análise, com seus parâmetros otimizados.

Dentre os modelos estudados, o *Gradient Boosting Regression* apresentou um menor RMSE em comparação com seus concorrentes, em um valor de 12,71 GPa (Figura 19). Os algoritmos *Random Forest* e *Support Vector Regression* apresentaram um RMSE de 13,09 GPa (Figura 20) e 14,64 GPa (Figura 21) respectivamente.

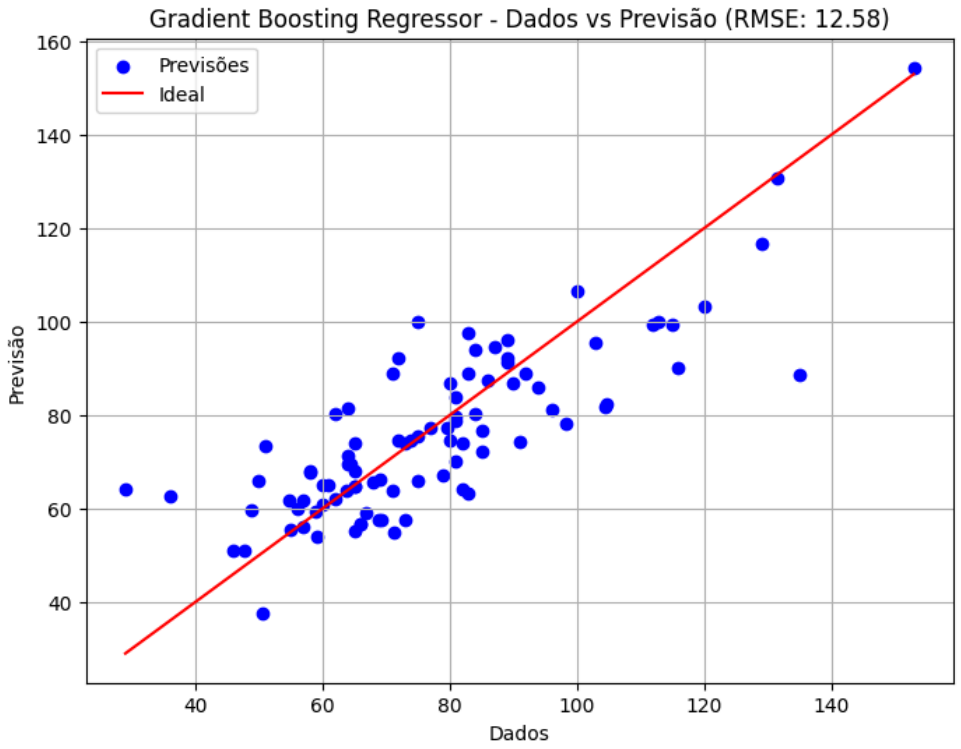

Figura 19. *Gradient Boosting Regression* Otimizado - Dados versus Previsão.

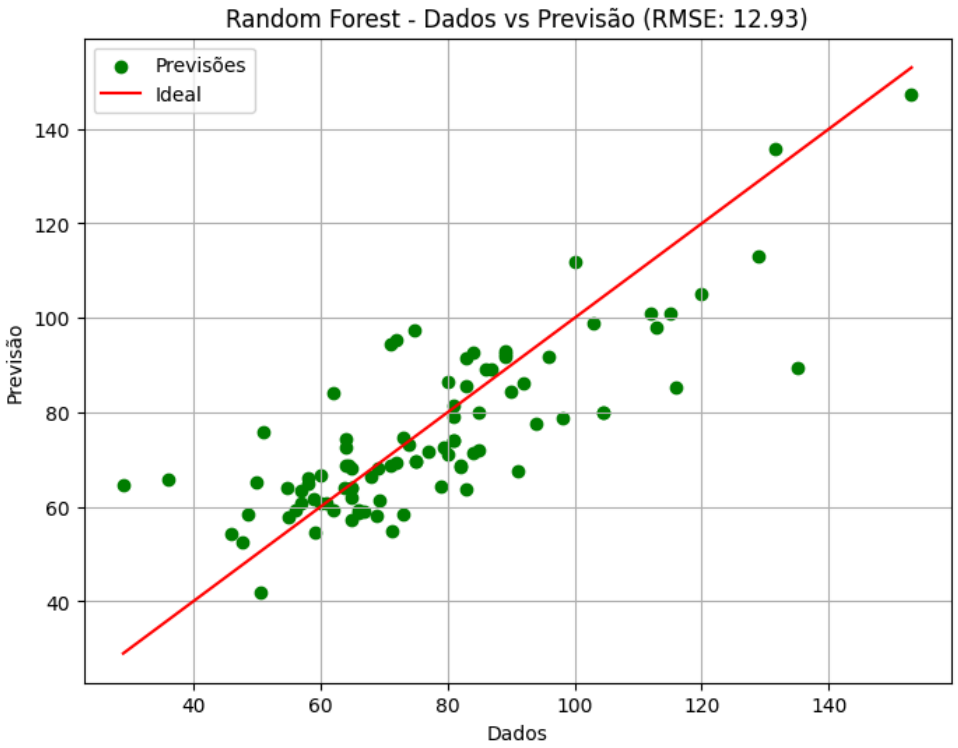

Figura 20. *Random Forest Regression* Otimizado - Dados versus Previsão

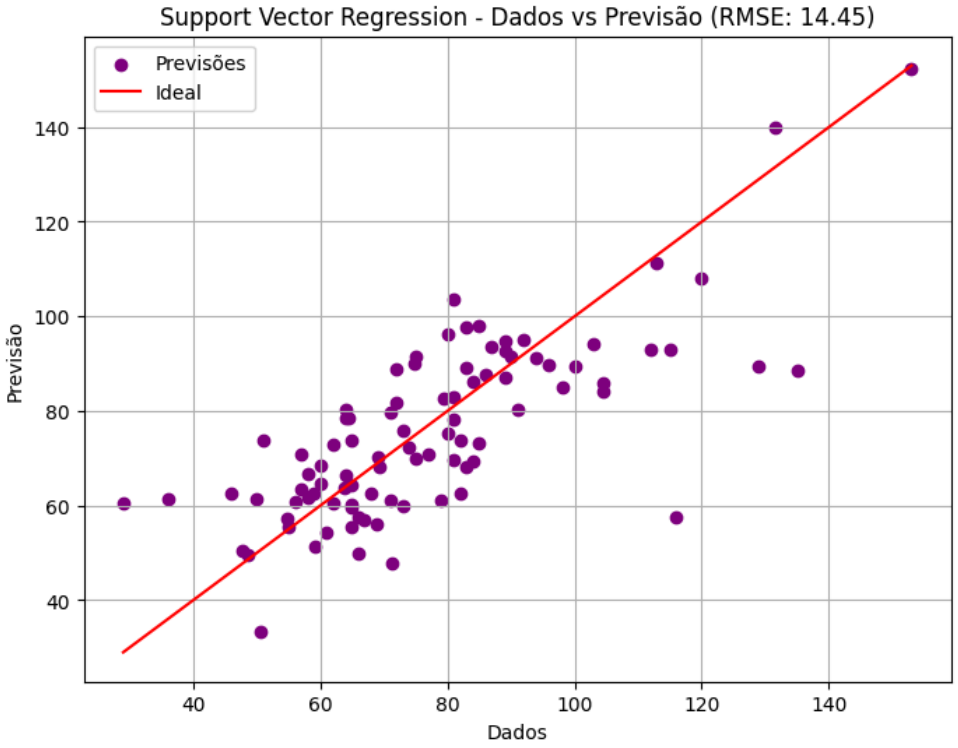

Figura 21. *Support Vector Regression* Otimizado - Dados versus Previsão

Assim, calculou-se a importância de cada característica na contribuição do módulo de elasticidade. A importância de todas as características foi evidenciada nas Figuras 16 e 17, para cada modelo analisado. Para o SVR, aplica-se o mesmo fator anteriormente discutido.

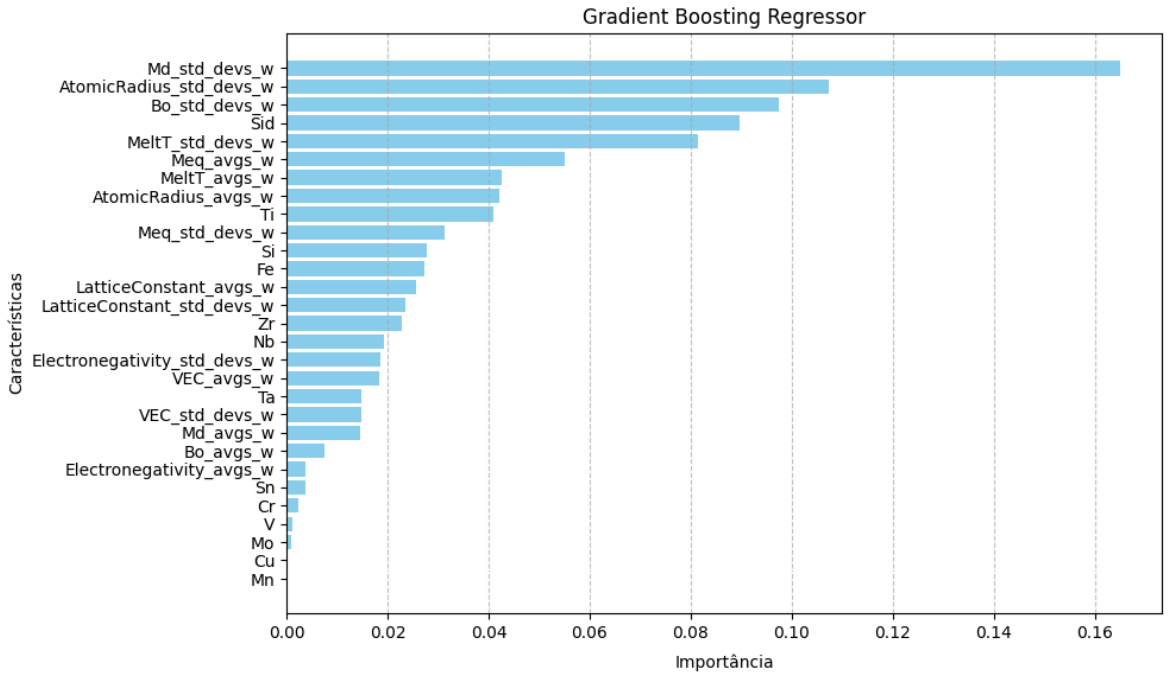

Figura 17. Importâncias por propriedade do GBR otimizado.

De acordo com a Figura 17, as 3 características mais importantes para o algoritmo de GBR otimizado na influência do módulo de elasticidade são:

1. Desvio padrão da energia média do orbital d (Md_std_devs_w) = 0,165 eV;
2. Desvios padrão do raio atômico (AtomicRadius_std_devs_w) = 0,1073 A;
3. Desvio padrão da ordem de ligação (Bo_std_devs_w) = 0,1130.

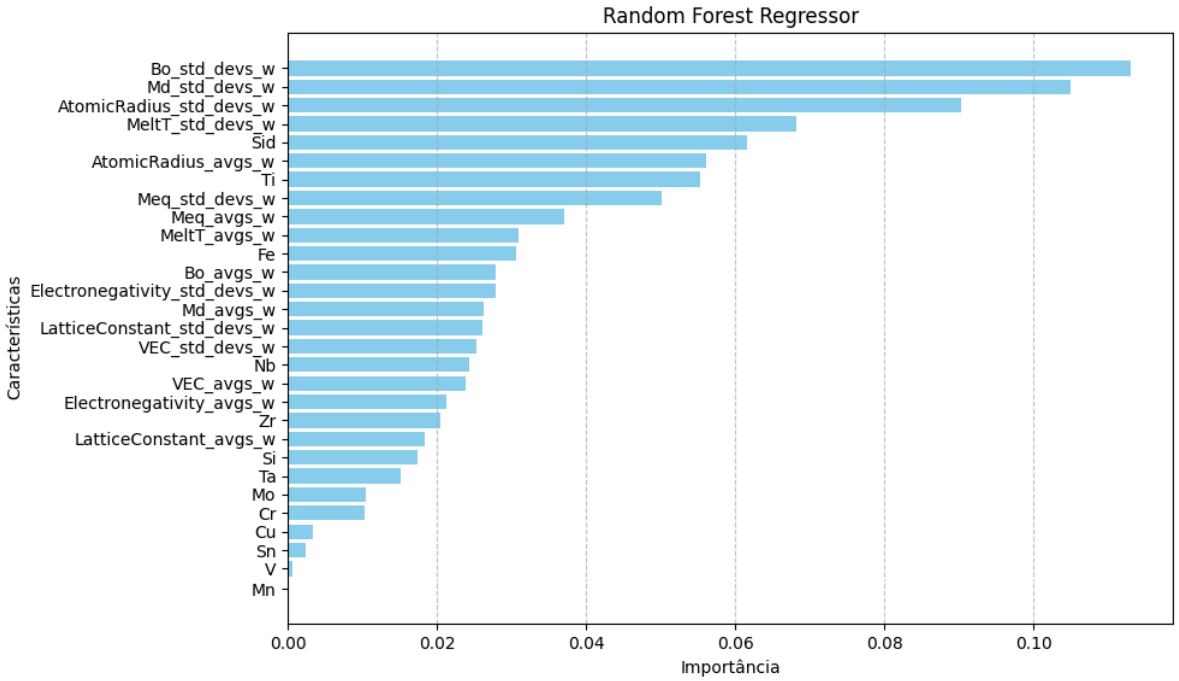

Figura 18. Importâncias por propriedade do RFR otimizado.

De acordo com a Figura 18, as 3 características mais importantes para o algoritmo de RFR otimizado na influência do módulo de elasticidade são:

1. Desvio padrão da ordem de ligação (Bo_std_devs_w) = 0,1130;
2. Desvio padrão da energia média do orbital d (Md_std_devs_w) = 0,105 eV;
3. Desvios padrão do raio atômico (AtomicRadius_std_devs_w) = 0,0973 A.

Deste modo, também se conclui que as características mais influentes no módulo elástico de um material, estão relacionadas ao raio atômico (r), a ordem de Ligação (Bo) e energia média do orbital d (Md), assim como no caso não-otimizado.

O procedimento de seleção, cruzamento e mutação foi realizado conforme descrito no item anterior, bem como todas as configurações anteriormente propostas para o AG. As cinco melhores ligas de titânio foram exemplificadas na Tabela 8, também ordenadas em ordem decrescente pelo parâmetro *fitting*.

Tabela 8. Resultados obtidos utilizando os parâmetros otimizados. As composições químicas constam em % em massa dos elementos.

% Cr	% Cu	% Fe	% Mn	% Mo	% Nb	% Si	% Sn	% Ta	% Ti	% V	% Zr	Fitting	$E_{Previsto}$	Mo_{eq}
0.877	0.000	0.877	1.754	7.895	3.509	0.000	7.895	0.000	75.439	0.000	1.754	0.214	44.975	19.237
2.206	0.000	0.000	0.735	9.559	6.618	0.000	5.147	0.735	72.794	1.471	0.735	0.232	46.498	21.257
0.769	0.769	0.000	1.538	7.692	3.077	0.000	6.923	0.000	76.154	1.538	1.538	0.233	46.696	19.238
0.578	0.000	0.000	0.000	9.827	27.168	0.000	2.312	0.000	57.803	0.578	1.734	0.247	49.326	21.809
0.000	0.000	1.316	2.632	0.658	19.737	0.000	6.579	0.000	65.789	0.658	2.632	0.252	50.489	19.270

Nota-se que o melhor resultado obtido nessa modalidade, referindo-se a precisão de previsão, deve-se ao fitness de 0,214 na qual alcançou um módulo elástico previsto de 44,975 GPa, demonstrando um resultado mais promissor, frente ao caso não-otimizado.

5. Conclusão

O avanço no controle e desenvolvimento das propriedades dos materiais possui um interesse em comum tanto para os diversos campos da indústria, quanto para a pesquisa científica. Diante das informações obtidas, pode-se notar que com a utilização dos dados reunidos, a combinação entre o *Gradient Boosting Regressor* e o Algoritmo Genético proporcionou o descobrimento de novas ligas de Ti com propriedades otimizadas, inclusive com valores de módulo de elasticidade inferiores aos reportados na literatura.

Portanto, a partir do presente trabalho foi possível observar o grande potencial da utilização de *machine learning* para o desenvolvimento de ligas metálicas, visto que a associação dos modelos de ML ao algoritmo genético é extremamente vantajosa na previsão e geração de novas ligas metálicas, destacando-se o GBR, uma vez que seu uso aliado ao AG proporcionou os indivíduos de maior assertividade, isto é, menores valores de RMSE frente aos outros modelos estudados, obtendo uma melhora na previsão a partir da otimização dos hiperparâmetros padrão utilizados, indicando a possibilidade da existência de uma liga de titânio com módulo elástico de 44,98 GPa.

Por fim, vale ressaltar então que somente a partir da utilização de dados experimentais presentes na literatura, foi possível fazer e descrever um método de *machine learning* eficaz que garante como resultado ligas promissoras a serem exploradas experimentalmente, tendo sido essa a maior contribuição do presente trabalho.

6. Sugestões para trabalhos futuros

Durante o desenvolvimento deste trabalho, notou-se a necessidade do envolvimento de mais informações acerca das características de processamento das ligas, uma vez que a rota escolhida para sua fabricação interfere fortemente em sua microestrutura final, e consequentemente, em suas propriedades mecânicas. Ademais, a projeção das fases formadas destas ligas durante solidificação, usando métodos CALPHAD, é de suma importância na verificação da assertividade da previsão, outro elemento do qual não pode ser abordado durante este trabalho, dois pontos sugeridos a trabalhos futuros.

7. Referências bibliográficas

1. Rack HJ, Qazi JI. Titanium alloys for biomedical applications. Mater Sci Eng C. 2006 Sep;26(8):1269–77.
2. Chen LY, Cui YW, Zhang LC. Recent development in beta titanium alloys for biomedical applications. Metals (Basel). 2020 Sep 1;10(9):1–29.
3. Ozaki T, Matsumoto H, Watanabe S, Hanada S. Beta Ti Alloys with Low Young's Modulus.
4. Ridzwan MIZ, Shuib S, Hassan AY, Shokri AA, Mohammad Ibrahim MN. Problem of stress shielding and improvement to the hip implant designs: A review. J Med Sci. 2007;7(3):460–7.
5. Katoch S, Chauhan SS, Kumar V. A review on genetic algorithm: past, present, and future. Multimed Tools Appl. 2021 Feb 1;80(5):8091–126.
6. Janiesch C, Zschech P, Heinrich K. Machine learning and deep learning. Available from: https://doi.org/10.1007/s12525-021-00475-2
7. Chakraborti N. Genetic algorithms in materials design and processing. Vol. 49, International Materials Reviews. IOM Communications Ltd.; 2004. p. 246–60.
8. G. Terlinde GF. Beta Titanium Alloys. Wiley;
9. Kolli RP, Devaraj A. A review of metastable beta titanium alloys. Vol. 8, Metals. MDPI AG; 2018.
10. Sidhu SS, Singh H, Gepreel MAH. A review on alloy design, biological response, and strengthening of β-titanium alloys as biomaterials. Mater Sci Eng C [Internet]. 2021;121(September 2020):111661. Available from: https://doi.org/10.1016/j.msec.2020.111661
11. Yang F, Li Z, Wang Q, Jiang B, Yan B, Zhang P, et al. Cluster-formula-embedded machine learning for design of multicomponent β-Ti alloys with low Young's modulus. npj Comput Mater. 2020 Dec 1;6(1).
12. Abbasi SM, Momeni A, Lin YC, Jafarian HR. Dynamic softening mechanism in Ti-13V-11Cr-3Al beta Ti alloy during hot compressive deformation. Mater Sci Eng A [Internet]. 2016;665:154–60. Available from: http://dx.doi.org/10.1016/j.msea.2016.04.040
13. Bönisch M. Structural properties, deformation behavior and thermal stability of martensitic Ti-Nb alloys. 2016;149.

14. Filip R, Kubiak K, Ziaja W, Sieniawski J. The effect of microstructure on the mechanical properties of two-phase titanium alloys. J Mater Process Technol. 2003;133(1–2):84–9.
15. Bishnoi S, Singh S, Ravinder R, Bauchy M, Gosvami NN, Kodamana H, et al. Predicting Young's modulus of oxide glasses with sparse datasets using machine learning. J Non Cryst Solids. 2019 Nov 15;524.
16. Niinomi M, Nakai M. Titanium-based biomaterials for preventing stress shielding between implant devices and bone. Int J Biomater. 2011;2011.
17. Mohammed MT. Beta Titanium Alloys: The Lowest Elastic Modulus for Biomedical Applications: A Review Zahid A Khan Jamia Millia Islamia [Internet]. 2014. Available from: https://www.researchgate.net/publication/265396160
18. Kuroda D, Niinomi M, Morinaga M, Kato Y, Yashiro T. Design and mechanical properties of new i type titanium alloys for implant materials. Vol. 243, Materials Science and Engineering. 1998.
19. Bandyopadhyay A, Mitra I, Goodman SB, Kumar M, Bose S. Improving biocompatibility for next generation of metallic implants. Prog Mater Sci [Internet]. 2023;133(July 2021):101053. Available from: https://doi.org/10.1016/j.pmatsci.2022.101053
20. Mahesh B. Machine Learning Algorithms - A Review. Int J Sci Res. 2020;9(1):381–6.
21. Li X, Li W, Xu Y. Human age prediction based on DNA methylation using a gradient boosting regressor. Genes (Basel). 2018 Sep 1;9(9).
22. Meenal R, Michael PA, Pamela D, Rajasekaran E. Weather prediction using random forest machine learning model. Indones J Electr Eng Comput Sci. 2021 May 1;22(2):1208–15.
23. Sheykhmousa M, Mahdianpari M, Ghanbari H, Mohammadimanesh F, Ghamisi P, Homayouni S. Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review. Vol. 13, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. Institute of Electrical and Electronics Engineers Inc.; 2020. p. 6308–25.
24. Saqib M. Forecasting COVID-19 outbreak progression using hybrid polynomial-Bayesian ridge regression model. Appl Intell. 2021;51(5):2703–13.

25. Uyanık GK, Güler N. A Study on Multiple Linear Regression Analysis. Procedia - Soc Behav Sci. 2013;106:234–40.
26. Vasudevan R, Pilania G, Balachandran P V. Machine learning for materials design and discovery. J Appl Phys. 2021;129(7).
27. Wei J, Chu X, Sun XY, Xu K, Deng HX, Chen J, et al. Machine learning in materials science. Vol. 1, InfoMat. Blackwell Publishing Ltd; 2019. p. 338–58.
28. Tao Q, Xu P, Li M, Lu W. Machine learning for perovskite materials design and discovery. npj Comput Mater [Internet]. 2021;7(1):1–18. Available from: http://dx.doi.org/10.1038/s41524-021-00495-8
29. Khakurel H, Taufique MFN, Roy A, Balasubramanian G, Ouyang G, Cui J, et al. Machine learning assisted prediction of the Young's modulus of compositionally complex alloys. Sci Rep. 2021 Dec 1;11(1).
30. Callister W. Materials Science and Engineering: An Introduction 7th Ed. Wiley. 2007.
31. Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobot. 2013;7(DEC).
32. Hodson TO. Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. Geosci Model Dev. 2022;15(14):5481–7.
33. Elango S, Natarajan E, Varadaraju K, Abraham Gnanamuthu EM, Durairaj R, Mohanraj K, et al. Extreme Gradient Boosting Regressor Solution for Defy in Drilling of Materials. Adv Mater Sci Eng. 2022;2022.
34. Touzani S, Granderson J, Fernandes S, Author C. Gradient boosting machine for modeling the energy consumption of commercial buildings. 2017.
35. Bentéjac C, Csörgő A, Martínez-Muñoz G. A Comparative Analysis of XGBoost. 2019 Nov 5; Available from: http://arxiv.org/abs/1911.01914
36. Steven J. Random Forest. J Insur Med. 2017;201–7.
37. Hu J, Szymczak S. A review on longitudinal data analysis with random forest. Brief Bioinform. 2023;24(2):1–11.
38. Carranza C, Nolet C, Pezij M, van der Ploeg M. Root zone soil moisture estimation with Random Forest. J Hydrol [Internet]. 2021;593(November 2020):125840. Available from: https://doi.org/10.1016/j.jhydrol.2020.125840
39. Ferreira LB, da Cunha FF, de Oliveira RA, Fernandes Filho EI. Estimation of reference evapotranspiration in Brazil with limited meteorological data using ANN and SVM – A new approach. J Hydrol [Internet].

2019;572(February):556–70. Available from: https://doi.org/10.1016/j.jhydrol.2019.03.028

40. Awad M. Efficient Learning Machines. 2015;6.
41. Üstün B, Melssen WJ, Oudenhuijzen M, Buydens LMC. Determination of optimal support vector regression parameters by genetic algorithms and simplex optimization. Anal Chim Acta. 2005;544(1-2 SPEC. ISS.):292–305.
42. Cassar DR, Santos GG, Zanotto ED. Designing optical glasses by machine learning coupled with a genetic algorithm. Ceram Int [Internet]. 2021;47(8):10555–64. Available from: http://dx.doi.org/10.1016/j.ceramint.2020.12.167
43. Tao H, Al-Sulttani AO, Salih Ameen AM, Ali ZH, Al-Ansari N, Salih SQ, et al. Training and Testing Data Division Influence on Hybrid Machine Learning Model Process: Application of River Flow Forecasting. Complexity. 2020;2020.
44. Nogueira. L. Repositório de Códigos e Base de Dados. Available from: https://github.com/Lnogueira2222/Reposit-rio-de-C-digos-e-Base-de-Dados
45. Venkatasubramanian V, Chan K, Caruthers JM. Evolutionary Design of Molecules with Desired Properties Using the Genetic Algorithm. J Chem Inf Comput Sci. 1995;35(2):188–95.