UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS MATEMÁTICAS E DE COMPUTAÇÃO

Alexandre Henrique Farah Dias

Estudo e implementação de modelos de aprendizado de máquina para predição de propriedades metalúrgicas em aços especiais

São Carlos

2021

Alexandre Henrique Farah Dias

Estudo e implementação de modelos de aprendizado de máquina para predição de propriedades metalúrgicas em aços especiais

Trabalho de conclusão de curso apresentado ao Centro de Ciências Matemáticas Aplicadas à Indústria do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, como parte dos requisitos para conclusão do MBA em Ciências de Dados.

Área de concentração: Ciências de Dados

Orientador: Prof. Dr. Adriano Kamimura Suzuki

São Carlos

2021

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi, ICMC/USP, com os dados fornecidos pelo(a) autor(a)

S856m

Dias, Alexandre Henrique Farah
Estudo e implementação de modelos de aprendizado de máquina para predição de propriedades metalúrgicas em aços especiais / Alexandre Henrique Farah Dias ; orientador Adriano Kamimura Suzuki. – São Carlos, 2021.
88 p. : il. (algumas color.) ; 30 cm.

Monografia (MBA em Ciências de Dados) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2021.

1. Ciência de Dados. 2. Bioinformática. 3. Single-cell RNA-seq. 4. Classificação celular. I. Suzuki, Adriano Kamimura, orient. . II. Título.

Alexandre Henrique Farah Dias

Estudo e implementação de modelos de aprendizado de máquina para predição de propriedades metalúrgicas em aços especiais

Trabalho de conclusão de curso apresentado ao Centro de Ciências Matemáticas Aplicadas à Indústria do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, como parte dos requisitos para conclusão do MBA em Ciências de Dados.

Data de defesa: XX de janeiro de 2022

Comissão Julgadora:

Prof. Dr. Adriano Kamimura Suzuki
Prof. Dr. Adriano Kamimura Suzuki

Professor
Convidado1

Professor
Convidado2

São Carlos
2021

À minha esposa Keika e à minha filha Maria Paula pela compreensão quanto ao sacrifício do tempo de convívio para a elaboração desse trabalho.

"A vida é como andar de bicicleta. Para manter o equilíbrio, é preciso se manter em movimento."

(Albert Einstein)

AGRADECIMENTOS

Agradeço, primeiramente, a Deus por me conceder as oportunidades que recebi da vida, à minha família pelo apoio e à Aperam South America por me escolher para a realização do curso.

Agradecimentos em especial são direcionados ao meu orientador, prof. Dr. Adriano Kamimura Suzuki cuja paciência, indicações, observações e apontamentos foram fundamentais desde o início dessa jornada para a produção do presente trabalho acadêmico.

Agradecimentos são também direcionados a todos os professores e monitores do curso de MBA em Ciência em Dados da USP, cujos ensinamentos foram de suma importância para o desenvolvimento do presente trabalho.

RESUMO

DIAS, A. H. F. **Estudo e implementação de modelos de aprendizado de máquina para predição de propriedades metalúrgicas em aços especiais**. 2021. 88p. Monografia (MBA em Ciências de Dados) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2021.

As propriedades metalúrgicas dos aços especiais têm forte dependência da sua composição química, bem como com as variáveis de processo nas diversas fases de produção. Pode-se portanto observar que entender a correlação entre composição química, condições de processamento e propriedades metalúrgicas seja fundamental na otimização da composição da liga metálica de forma a se obter as combinações desejadas das propriedades do aço. No entanto, em especial nos aços de alta liga, as influências da composição química e do processo de produção nas suas respectivas propriedades do aço são complexas de serem modeladas matematicamente.

Neste cenário, o objetivo principal deste trabalho consiste em estudar diversos modelos de machine learning capazes de prever com acurácia as propriedades metalúrgicas de aços especiais a fim de suportar a otimização dos projetos de desenvolvimento de novos aços e processos por meio de simulações que não envolvam custos e prazos elevados de experiências em escala industrial.

Para tanto, o trabalho utilizou o banco de dados de domínio público do *National Institute of Material Science* (NIMS) referente à medição do stress à fadiga em aços especiais, com aplicação e comparação de alguns dos principais algoritmos de machine learning utilizados para classificação e regressão. Os modelos estudados apresentaram boas métricas de desempenho, atingindo um coeficiente de correlação R^2 de até 99%.

Palavras-chave: ciência de dados, aços especiais, propriedades metalúrgicas, predição.

ABSTRACT

DIAS, A. H. F. **Study and implementations of machine learning models to predict metallurgical properties in special steel grades** . 2021. 88p. Monografia (MBA em Ciências de Dados) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2021.

The metallurgical properties of special steels are strongly dependent on their chemical composition, as well as on process variables in the different stages of production. It can therefore be seen that understanding the correlation between chemical composition, processing conditions and metallurgical properties is fundamental in optimizing the alloy composition in order to obtain the desired combinations of steel properties. However, especially in high-alloy steels, the influences of chemical composition and production process on their respective steel properties are complex to model mathematically.

In this scenario, the main objective of this work is to study several machine learning models capable of accurately predicting the metallurgical properties of special steels in order to support the optimization of new steel and process development projects through simulations that do not involve costs and long periods of experience on an industrial scale. This work used the public domain database of the National Institute of Material Science (NIMS) regarding the measurement of fatigue stress in special steels, with application and comparison of some of the main machine learning algorithms used for classification and regression. The models studied showed good performance metrics, reaching an R^2 of up to 99%.

Keywords: data science, special steel grades, metallurgical properties, prediction.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

ICMC	Instituto de Ciências Matemáticas e de Computação
IFSC	Instituto de Física de São Carlos
IQSC	Instituto de Química de São Carlos
TCC	Trabalho de Conclusão de Curso
USP	Universidade de São Paulo
USPSC	Campus USP de São Carlos
MIC	Maximal Information Coefficient
NMI	Normalized Mutual Information
PCA	Principal Components Analysis
MSVR	Support Vector Machine Multidimensional
LRC	Logistic Regression Classifier
RFC	Random Forest Classifier
CAT	CatBoost Classifier
LGB	Light Gradient Boosting Machine Classifier
XGB	Extreme Gradient Boosting Classifier
ETC	Extra Trees Classifier
DTC	Decision Tree Classifier
KNN	KNeighbors Classifier
GBC	Gradient Boosting Classifier
MLP	MLP Classifier
SVC	Support Vector Classifier
SGD	Stochastic Gradient Descent Classifier
ABC	AdaBoost Classifier
RID	Ridge Classifier

LDA	Linear Discriminant Analysis Classifier
YS	Yield strength
UTS	Ultimate Tensile Strength
RNA	Redes Neurais Artificiais

SUMÁRIO

1 INTRODUÇÃO

1.1 Um breve descritivo da Aperam South America

Fundada em 1944 com o nome de ACESITA, a Aperam South America é a única produtora integrada de aços planos elétricos, inoxidáveis e carbono especiais da América Latina. Com mais de 4.000 funcionários, possui capacidade instalada de mais de 900.000 toneladas de aço líquido. A sua principal planta industrial localizada na cidade de Timóteo-MG possui dois altos-fornos que utilizam apenas carvão vegetal como energia renovável fornecida pela sua subsidiária Aperam Bioenergia.

O processo de elaboração do aço produzido pela Aperam South America possui quatro grandes etapas, a saber: a transformação do minério de ferro em gusa nos alto-fornos usando carvão vegetal como redutor, a adição de ligas na Aciaria com o lingotamento de placas de aço, o processo de laminação a quente das placas convertendo em bobinas a quente, e finalmente o processo de laminação a frio do material.

A Aperam BioEnergia é uma subsidiária integral da Aperam South America, que tem como objetivo principal o fornecimento de biomassa na forma de carvão vegetal para os altos-fornos da planta industrial localizada em Timóteo (MG). A Aperam Bioenergia está equipada com tecnologias sustentáveis e inovadoras como o FAP 2000, maior forno de carvão vegetal do mundo (BIOENERGIA, 2022).

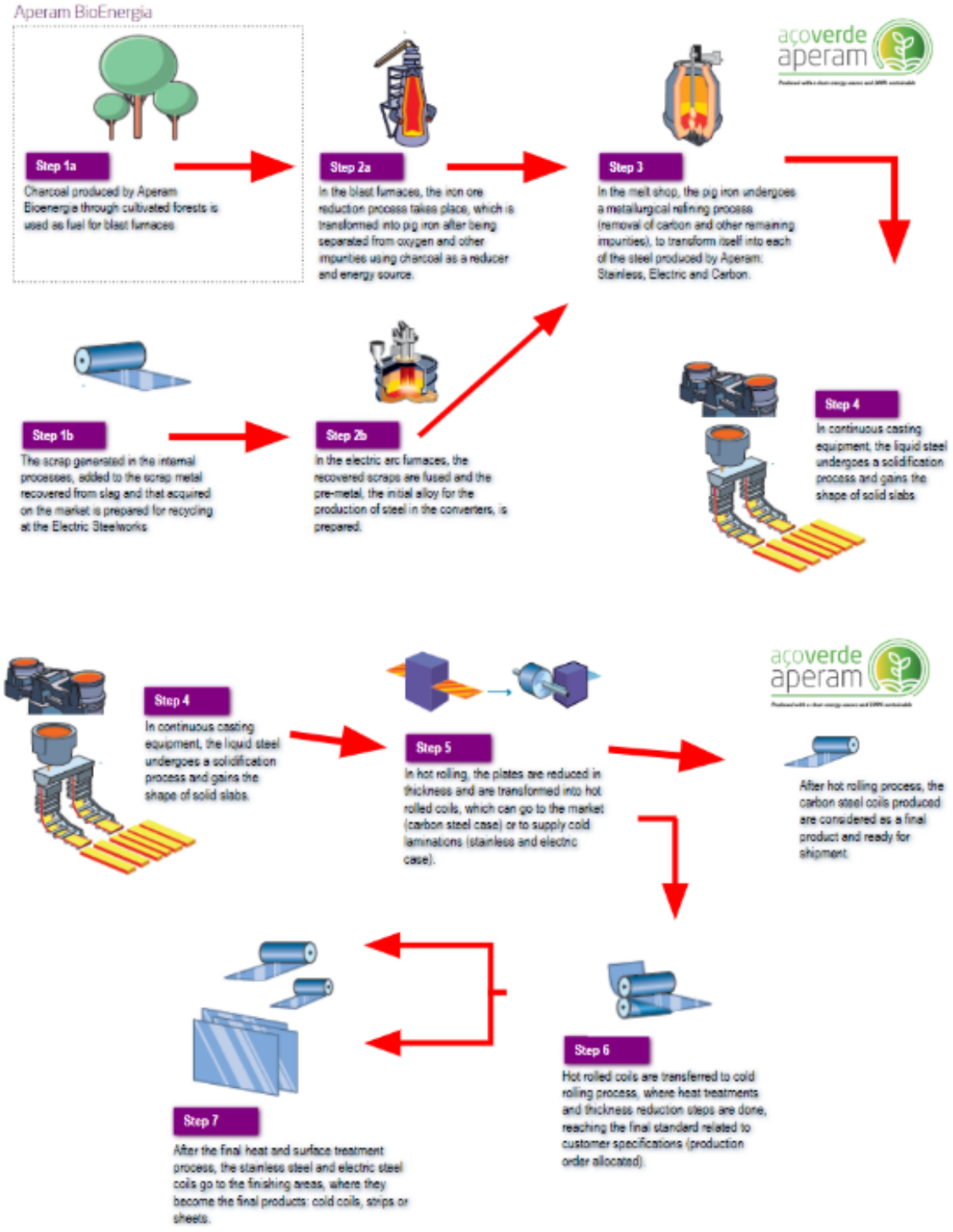
Atuando na região do Vale do Jequitinhonha, a Aperam BioEnergia iniciou suas atividades na cidade de Itamarandiba em 1974, e posteriormente expandiu suas atividades para os municípios de Capelinha, Turmalina, Veredinha e Minas Novas, no estado de Minas Gerais. O fato da Aperam BioEnergia produzir seus próprios clones de árvores e ter uma base de conhecimento de mais de 30 anos de experimentação, torna-se um forte *user case* para a aplicação das técnicas de ciência de dados no futuro.

1.2 Motivações para o uso da ciências de dados na Aperam South America

Na produção de aço, devido à alta complexidade técnica do negócio e à enorme quantidade de dados gerados diariamente nas diversas etapas de produção, as estratégias convencionais de análise exploratória de dados, em sua maioria, não oferecem soluções eficazes em tempo satisfatório durante as campanhas de produção.

Os problemas podem estar relacionados a fatores relativos à sua composição química, aos processos físicos e químicos que o material é submetido, ou o combinado dessas etapas ao longo do fluxo produtivo (figura 1). Um dos grandes desafios das áreas de metalurgia e controle de processos é entender os fatores que estão associados à qualidade do aço.

Figura 1 – Fluxo de produção resumido da Aperam South America



Fonte: Elaborada pelo autor.

O portfólio de produtos da Aperam South America ainda se mostra mais desafiador por fazer parte de um grupo classificado como aços especiais, requerendo padrões de qualidade extremamente elevados: exigência de grande apelo superficial e resistência mecânica no caso dos aços inoxidáveis bem como perdas magnéticas extremamente baixas, no caso dos aços elétricos de grão orientado e não-orientado.

Com o interesse e adoção progressiva da siderurgia nos últimos tempos por novas tendências como a inteligência artificial, indústria 4.0 e internet das coisas (IoT), claramente é possível visualizar uma jazida enorme de oportunidades a se explorar sob a ótica de ciência de dados.

Para evitar perdas e manter a competitividade no mercado atual, é fundamental agilizar e aprimorar as tomadas de decisões pelas equipes de qualidade e controle de processos da Aperam South America. Com o volume cada vez maior de dados, maior instrumentação nos equipamentos e sistemas de TA e TI mais capazes de armazená-los, não há mais espaço para intuição ou experimentação apenas, levando os negócios à serem cada vez mais orientados à exploração e interpretação dos dados de processo e produção.

Esse trabalho tem como objetivos específicos:

- Explorar técnicas baseadas em aprendizado de máquina supervisionado que sejam capazes de predizer algumas das características metalúrgicas e de qualidade do aço produzido, construindo modelos que sejam interpretáveis pelos técnicos e engenheiros de qualidade e processo das áreas de produção.
- Permitir que os modelos de aprendizado de máquina sejam capazes de apoiar o conhecimento empírico metalúrgico existente acerca dos fatores de sucesso envolvidos ao longo do processo de fabricação dos aços especiais, e assim identificar quais atributos possuem maior nível de influência na qualidade do aço.

Com esse projeto existe uma grande expectativa de mostrar que a ciência de dados não seja apenas tratada como um viés acadêmico, mas sim como uma área de conhecimento onde toda a Aperam South America possa tirar proveito. Esse trabalho faz parte de um movimento de capacitação de engenheiros, criando-se uma equipe de cientistas de dados nas diversas áreas da empresa (produção, finanças, suprimentos, RH, TI, entre outros).

O restante deste texto está organizado da seguinte maneira: no **capítulo 2** foi feito uma extensiva revisão bibliográfica as aplicações de aprendizado de máquina aplicados à predição de propriedades metalúrgicas em aços especiais. O **capítulo 3** contém uma breve revisão sobre os modelos de aprendizado de máquinas é apresentada. No **capítulo 4** os principais resultados obtidos aplicando-se as técnicas de aprendizado de máquinas são apresentados.Por fim no **capítulo 5** apresenta-se as considerações finais juntamente com algumas pespectivas futuras para pesquisa.

2 REVISÃO BIBLIOGRÁFICA

Esta seção buscou-se contemplar uma revisão da literatura sobre a aplicação da ciência de dados na engenharia de materiais, um breve resumo entre as diversas contribuições acadêmicas sobre o tema.

2.1 Ciência de dados e a sua utilização na indústria

Nas últimas décadas, a capacidade da indústria em geral de gerar e armazenar dados tem sido observada em praticamente todos os domínios científicos, e a ciência dos materiais não é exceção. Esse movimento global levou ao surgimento do quarto paradigma da ciência: a orientação por dados e desenvolvimento de modelos preditivos e estratégias de mineração de grandes volumes de dados de uma maneira mais abrangente e eficiente (TOLLE; TANSLEY; HEY, 2011).

O setor de manufatura é uma das maiores indústrias que continuam a crescer de maneira uniforme. Em 2016, representou cerca de 16% do PIB global e contribuiu com mais de 2 trilhões de dólares para a economia dos EUA. O que está impulsionado esse desenvolvimento acelerado? Entre os principais fatores pode-se destacar o avanço da tecnologia, onde mais recentemente o uso da ciência de dados contribuiu consideravelmente para esse desempenho. Como exemplo pode-se citar aumento da capacidade de produção, menores custos operacionais, melhor qualidade dos produtos e dos níveis de serviço, melhoria do processo de manutenção corretiva e preventiva, entre outros.

Com o advento do aumento da capacidade computacional, seja local ou em nuvem, as técnicas mais avançadas neste campo da ciência da computação - computação de alto desempenho, aprendizado de máquina e algoritmos de mineração de dados - têm-se tornado uma realidade fora do setor puramente acadêmico. A modelagem matemática por meio de inteligência artificial tem atraído um número cada vez maior de pesquisadores, engenheiros e entusiastas, nos mais diversos campos de atuação como controle de processos industriais, reconhecimento de padrões e sequências, diagnósticos médicos, ciência do clima, astronomia e cosmologia, detecção de intrusão e muitos outros exemplos.

Empresas como a Amazon, Netflix, Google e Walmart usam modelagem preditiva para recomendações de compra, notícias personalizadas para seus clientes, previsões mais assertivas de demanda e oferta, proporcionando aumento nas vendas e nos níveis de satisfação dos clientes (AGRAWAL et al., 2014; AGRAWAL; CHOUDHARY, 2016).

Cientistas de dados podem usar mineração de grandes volumes de dados para tomada de decisão de forma antecipada. Um exemplo bem famoso é o caso do furação Françês *Charley* que estava avançando pelo Caribe e se aproximando da costa da Flórida. Diante disso, os executivos da *Walmart* viram uma excelente oportunidade de utilizar os dados de consumo para enxergar padrões, como a demanda por itens incomuns, ajustar o estoque para os produtos mais vendidos nas lojas que estavam na rota do furacão Charles ocorrido semanas antes, e com isso realizar o abastecimento das lojas na rota do *Charley* de forma mais eficiente (HACKERS, 2021).

A digitalização mais intensa dos processos produtivos de fabricação de aço bem como a capacidade de armazenamento e processamento de grande volume de dados permitem a criação de modelos que ajudam a analisar as correlações entre as propriedades metalúrgicas e as variáveis de produção. A integração de IoT, *bigdata*, computação em nuvem e tecnologias de inteligência artificial ajudam as indústrias de manufatura a implementar soluções mais automatizadas e inteligentes, uma das características mais evidentes da Indústria 4.0 (SANTOS et al., 2018).

O desenvolvimento de novos algoritmos e teorias relacionados ao aprendizado de máquina permite que pesquisadores e desenvolvedores lidem com as demandas de análise de dados de manufatura. Além disso, têm grande potencial para descobrir conhecimento a partir de grandes quantidades de dados com o aumento sustentável de repositórios de dados históricos de manufatura (ALPAYDIN, 2004).

2.2 Desafios da indústria do aço

A indústria siderúrgica no Brasil e no mundo enfrenta enormes desafios. Excesso de capacidade instalada e de barreiras comerciais, redução do consumo aparente e o impacto de novos players como China, Indonésia, Taiwan entre outros no mercado global de aço são algumas das ameaças que põem em risco a longevidade e a sobrevivência das empresas.

É um consenso geral na indústria do aço, em especial no Brasil, que a adoção de novas tecnologias relacionadas à indústria 4.0 (sensores inteligentes, robôs, inteligência artificial, *advanced analytics* entre outros) é fundamental para a manutenção da competitividade da siderurgia brasileira. Na opinião de diversos especialistas do setor resta à indústria do aço modernizar as plantas já existentes e provavelmente fechar as unidades que estão muito defasadas tecnologicamente e que necessitam de muito capital para se tornarem financeiramente competitivas.

Lindström et al. (2019) cita que a indústria 4.0 impulsionou o crescimento gradual do uso de novas tecnologias de informação e comunicação de dados que levaram à uma ruptura do modelo mental anteriormente adotado. Para a ciência de dados significou

a transformação de informação em conhecimento relevante dos processos industriais monitorados, através de processos eficientes de agregação de grandes volumes de dados, maior capacidade de computação (*on-premise*, *cloud* ou híbrida) e a utilização de métodos de inteligência artificial e aprendizado de máquinas.

Nesse cenário, segundo Diez-Olivan et al. (2019) o conceito de data-driven, também conhecido por orientação por dados, tem conquistado a atenção dos diversos segmentos da indústria. A complexidade natural do processo de fabricação do aço associado com mais recentemente a crescente competição global criaram uma demanda por novos métodos e técnicas, sendo que o aprendizado de máquina acabou por desempenhar um papel essencial no monitoramento, controle e otimização do processo de fabricação.

O setor de produção de aços sempre foi considerado como conservador na adoção de novas tecnologias, até por não querer correr o risco de acidentes ou perdas substanciais de produção ao experimentar algo ainda pouco conhecido no âmbito industrial. Porém diversos exemplos mostram que essa tendência tem mudado:

- A Big River Steel, siderúrgica americana baseada em aciaria elétrica criada em 2014, define a si mesma como "uma empresa de tecnologia que, por acaso, produz aço". O projeto de seus equipamentos incorporou completamente o conceito da indústria 4.0 para coleta, transferência, armazenamento e análise inteligente de todos os seus dados industriais, aplicando em larga escala os conceitos de inteligência artificial em seus processos (BIGRIVERSTEEL, 2021).

- A Voest Alpine iniciou em 2018 a construção de uma nova planta de aços longos para atender o mercado aeroespacial e automotivo com elevado grau de digitalização. Um *roadmap* de ações voltadas para o conceito de Indústria 4.0 está em andamento em sua planta industrial de Linz (Áustria) (GORNI, 2021).

- De acordo com Gorni (2021) a siderúrgica coreana Posco implantou um grande centro de dados que centraliza todas as informações digitais que são coletadas nas várias linhas de sua planta industrial, anteriormente armazenadas de forma dispersa e descentralizada. O objetivo é treinar as equipes para extrair conhecimento a partir desses dados.

- O grupo TATA Steel iniciou há cerca de 3 anos uma jornada de transformação digital com foco em ciência de dados, na sua planta industrial de Kalinganagar (índia), com ações que vão desde a definição da arquitetura de TI baseada em *cloud* pública para armazenamento de dados, capacitação técnica das equipes envolvidas e identificação de cases relevantes até a demonstração da viabilidade de cada proposta (MCKINSEY, 2021).

2.3 Ciência de dados na previsão de propriedades metalúrgicas dos aços

Os modelos tradicionais de previsão de propriedades mecânicas são baseados principalmente na experiência e no mecanismo que modela as relações lineares e não-lineares entre os parâmetros do processo. Em escala industrial trata-se de um problema extremamente complexo, sujeito a diversas variáveis desde interferências externas até mesmo o *design* do equipamento de produção.

Diversos estudos de caso ilustram os benefícios potenciais da aplicação do Aprendizado de Máquina no campo da ciência de materiais. Em especial a Inteligência Artificial tem sido largamente utilizada para otimizar o projeto de novos materiais e estruturas na indústria do aço. As vantagens dos novos materiais e estruturas são obtidas principalmente devido à capacidade dos modelos de IA em encontrar o bom equilíbrio entre microestruturas racionais e alta precisão (JIAO; ALAVI, 2021).

Mandal et al. (2009) destaca que redes neurais artificiais (RNA) têm mostrado desempenho notável para construir relacionamentos complexos em ciência e engenharia de materiais. Diversas pesquisas baseadas nessas abordagens têm sido apresentadas ao longo dos anos recentes, algumas delas listadas a seguir.

Takahashi H. J. e Teixeira (2008) apresenta o desenvolvimento e a implantação, na planta industrial do grupo Usiminas de uma ferramenta de apoio à decisão baseada nas técnicas de inteligência computacional, para a predição de propriedades mecânicas de aços de alta resistência microligados, laminados a frio e revestidos por imersão a quente. As técnicas investigadas (RNA com regularização, RNA *ensemble* e *Neuro-Fuzzy*) para a predição de propriedades mecânicas mostraram-se como alternativas viáveis para melhoria na agilidade nas respostas às consultas de produtos não padronizados, desenvolvimento de novas ligas e condições de processo.

Segundo Jones, Watton e Brown (2005) entre os benefícios da utilização de técnicas de aprendizado de máquinas para predição das propriedades mecânicas de aços de alta liga destacam-se a possibilidade de se confirmar as propriedades mecânicas em tempo real em cada fase do processo e também a possibilidade da otimização da composição química e dos parâmetros de processo. Em seu trabalho, uma comparação da aplicação de regressão múltipla linear, regressão múltipla não linear e redes neurais não lineares é feita para várias famílias de aço usando dados retirados do laminador de tiras a quente da planta industrial de *Corus Port Talbot*.

Rajan, Suh e Mendez (2009) aplicou-se o método de Análise de Componentes Principais (PCA) em um banco de dados consistindo de 600 compostos de super condutores de alta temperatura para identificar padrões e fatores que governam esta propriedade importante. Observou-se que o conjunto de dados se agrupa de acordo com a valência média, um critério que tem sido relatado na literatura como de extrema importância para

propriedades supercondutoras.

Myllykoski, Larkiola e Nylander (1996) utilizaram RNA para predição das propriedades mecânicas de bobinas de aço laminadas a frio. Utilizando a composição química real das bobinas e variáveis de processo (temperaturas, parâmetros de recozimento e laminação), os modelos obtidos previram com precisão as propriedades mecânicas: resistência ao escoamento, resistência à tração e deformação à tração. Os autores citam o suporte para realização de ações de *feedforward* nos diversos processos da cadeia de produção como uma das vantagens desse tipo de modelagem.

Wu, Yan e Lv (2021) construiram um modelo baseado na regressão do vetor de suporte multidimensional (MSVR) combinado com o método de seleção de recursos usando o conceito de coeficiente de informação máxima (MIC). Os resultados de predição obtidos mostram que a estrutura proposta é capaz de boa precisão de predição com menor tempo computacional se comparado com modelos tradicionais.

Modelos baseados em RNA desenvolvidos para a análise e simulação da correlação entre o endurecimento de aços inoxidáveis especiais e suas condições de composição, processamento e trabalho são propostos por Guo e Sha (2004). Os modelos obtidos mostraram ser capazes de calcular as propriedades dos aços como funções da composição da liga, parâmetros de processamento e condição de trabalho, em concordância com os dados experimentais da literatura.

Agrawal et al. (2014) propuseram um *framework* para determinação das relações causais entre variáveis de processo de determinadas classes de aços especiais, suas composições químicas e suas resistências à fadiga, incluindo a aplicação de uma gama de métodos de aprendizado de máquina e análise de dados Uma previsão mais precisa da resistência à fadiga de aços é de particular importância em ciência dos materiais devido ao alto custo (e tempo) despendido dos testes de fadiga. A resistência à fadiga é considerado o dado mais importante no projeto e análise de falha de componentes mecânicos.

Singh et al. (1998) desenvolveram um modelo de RNA no qual o escoamento e a resistência à tração do aço foram estimados como uma função de cerca de 108 variáveis, incluindo a composição química e uma série de parâmetros de laminação. Fujii, Mackay e Bhadeshia (1996) aplicaram modelagem semelhante para a previsão da taxa de crescimento de trincas por fadiga de superligas à base de níquel, modelada em função de 51 variáveis de processo. Foi demonstrado a capacidade de tais métodos de investigação nos casos em que as informações não podem ser acessadas experimentalmente.

Patel e Jokhakar (2016) propuseram uma abordagem de aprendizagem de máquina e a metodologia para diagnóstico de defeitos de desvio de temperatura de resfriamento que consiste em quatro fases a saber: estruturação de dados, identificação de associação, derivação estatística e classificação. Resultados comparativos obtidos com vários algoritmos

de mineração de dados, como árvores de decisão, redes neurais e florestas aleatórias foram apresentados em termos de parâmetros de desempenho, alcançando uma precisão de 95%.

Wang et al. (2020) utilizaram RNA para prever as propriedades de tração, incluindo resistência ao escoamento (YS) e resistência à tração final (UTS) em aço inoxidável austenítico em função da composição química, tratamento térmico e temperaturas de processo. Os modelos obtidos apresentam bom desempenho de predição para YS e UTS com valores de R2 acima de 93%, compatíveis com outros dados publicados na literatura.

Narayana et al. (2020a) aplicaram modelos de RNA para correlacionar as relações entre composição química, temperaturas de processo e propriedades mecânicas do aço inoxidável austenítico 18Cr-12Ni-Mo. A resposta efetiva dos elementos químicos nas propriedades mecânicas em temperatura ambiente e também em temperaturas elevadas foi estimada quantitativamente com o auxílio da metodologia do índice de importância relativa. Em um estudo semelhante feito pelos mesmos autores, diversas topologias de RNA foram treinadas para correlacionar as relações entre composição química, temperatura e propriedades mecânicas do aço inoxidável austenítico 25Cr-20Ni-0,4C (NARAYANA et al., 2020b).

Sourmail, Bhadeshia e MacKay (2002) capturaram corretamente a influência das variáveis estudadas usando os modelos de RNA com base na banco de dados de limite de escoamento de aços especiais. Além das propriedades de fadiga e limite de escoamento, alguns modelos de aprendizado de máquina foram estabelecidos para obter as correlações entre as propriedades de tração e variáveis de processo. Fragassa et al. (2019) escolheram os fatores metalográficos como recursos de entrada e projetou três tipos de métodos de aprendizado de máquina para modelar as propriedades mecânicas do ferro fundido.

No entanto, para o aço inoxidável austenítico (ASS), mais atenção é dada ao limite de escoamento, resistência à fadiga e à corrosão (HODGSON; KONG; DAVIES, 1999). Além da resistência à corrosão, as propriedades mecânicas como a resistência à torção e à deformação também são importantes para sua aplicação e têm chamado muita atenção nas últimas décadas dos pesquisadores.

Palla et al. (2006) desenvolveram um modelo de rede neural artificial para correlacionar a composição da liga e a temperatura de teste às propriedades de tração de ASS modificado 15Cr-15Ni-2.2Mo-Ti. Nesse trabalho s propriedades de tração do ASS foram extensivamente estudadas experimentalmente e teoricamente.

Desu et al. (2016) utilizaram a temperatura de teste e as taxas de deformação como descritores para prever as propriedades de tração de ASS 304L e 316L usando o modelo baseado em RNA. Mandal et al. (2009) discutiram a aplicação de modelagem de RNA na pesquisa de aços inoxidáveis austeníticos, incluindo: (I) previsão do limite de escoamento em aços inoxidáveis austeníticos como uma função da composição química e parâmetros

de processo, (II) previsão do comportamento de deformação à quente do aço inoxidável AISI tipo 304L em função das variáveis do processo.

A corrosão do metal em sistemas de engenharia apresenta sérios problemas. Estruturas e componentes que operam em ambiente úmido e marinho sofrem com corrosão severa e falhas prematuras. Em Wen et al. (2009) foi aplicado o método de *Support Vector Regressor* para previsão da taxa de corrosão de aços em diferentes ambientes de água do mar. O estudo sugeriu que o método estudado pode ser uma metodologia promissora e prática para conduzir um rastreamento de corrosão em tempo real de aço em condições severas.

Stoll e Benner (2021) estudaram uma aplicação de métodos de ML em dados de teste de punção para a determinação da resistência à tração de vários materiais. Foi encontrada uma forte correlação entre os dados do SPT e os dados do teste de tração, o que, em última análise, permite a substituição de testes mais caros por testes simples e rápidos em combinação com as técnicas e métodos de aprendizado de máquinas.

Em Feng e Yang (2016) um algoritmo genético baseado em RNA foi proposto para otimizar os parâmetros de processamento termomecânico. Neste modelo, uma rede neural de retropropagação (BPNN) foi estabelecida para mapear e otimizar a relação entre parâmetros de processamento termomecânico, como nível de deformação, temperatura de recozimento e tempo de processo. Os resultados indicaram que o modelo proposto é um meio eficaz e confiável para a otimização dos parâmetros citados, o que resultou na melhora da resistência à corrosão intergranular do aço inoxidável austenítico 304 estudado.

Gautham et al. (2011) trabalharam com a previsão de força de fadiga usando o banco de dados de domínio público do National Institute of Material Science (NIMS). O estudo aplicou a técnica de PCA nos dados e, posteriormente, realizou regressão mínima quadrada parcial (PLSR) nos diferentes grupos identificados por PCA a fim de fazer as previsões desejadas. Deshpande et al. (2013) utilizaram o mesmo banco de dados empregando vários métodos de regressão com redução de dimensionalidade para a previsão das propriedades de fadiga de aços com uma avaliação dos erros residuais de cada método estudado.

Estudo similar foi realizado por Dobrzanski, Kowalski e Madejski (2005), no qual um método baseado em RNA foi proposto para prever o limite de escoamento e a resistência à tração final para aços especiais, baseado na composição química dos aços e nos parâmetros de processo. Um *software* foi desenvolvido pelos autores para auxiliar na pesquisa da melhor composição química dos aços de forma a minimizar o risco de fabricação de produtos que possam não atender aos padrões determinados pelos clientes finais.

Agrawal et al. (2014) exploraram a aplicação de diferentes técnicas de ciência de dados, incluindo seleção de recursos e modelagem preditiva, às propriedades de fadiga de

aços, utilizando o banco de dados NIMS. Os resultados demonstram que várias técnicas avançadas de análise de dados, como redes neurais, árvores de decisão e regressão polinomial multivariada podem alcançar uma melhoria significativa na precisão da previsão em relação à literatura, com valores de R2 acima de 97%.

A complexa interação entre as variáveis de entrada tem dificultado as tentativas convencionais, abrindo espaço para que técnicas avançadas de análise de dados possam liderar o caminho da próxima revolução no domínio da ciência dos materiais. Também há uma percepção geral na comunidade de mineração de dados sobre a modelagem preditiva de que é mais útil saber sobre um conjunto de técnicas de bom desempenho para um determinado problema do que identificar um único modelo ou método.

Em Agrawal et al. (2014) é mostrado que uma série de abordagens diferentes produzem ligações altamente confiáveis, algumas delas significativamente melhores do que o que foi relatado anteriormente na literatura (GAUTHAM et al., 2011)

3 METODOLOGIA

3.1 Uma breve revisão sobre os modelos de aprendizado de máquinas

Arthur Samuel (1901-1990), pioneiro americano no campo de jogos de computador e inteligência artificial, cunhou o termo de *machine learning* ou aprendizado de máquina em 1959. Ele o definiu como um *campo de estudo que dá aos computadores a capacidade de aprender sem ser explicitamente programado.*

Na sua forma mais básica, o aprendizado de máquina deve satisfazer a condição básica de se obter uma curva que melhor se ajusta aos dados em uma representação matemática multidimensional desses dados. Com o passar do tempo essa área de estudo ocupou praticamente todos os campos científicos como uma verdadeira tempestade, reduzindo a dependência de resultados experimentais ao criar um ambiente experimental virtual quase perfeito, permitindo usar os dados gerados por esses modelos para criar novas teorias, testar ideias e validar suposições.

Os modelos de ML são classificados em três categorias distintas com base no tipo de conjunto de dados em que estão associados:

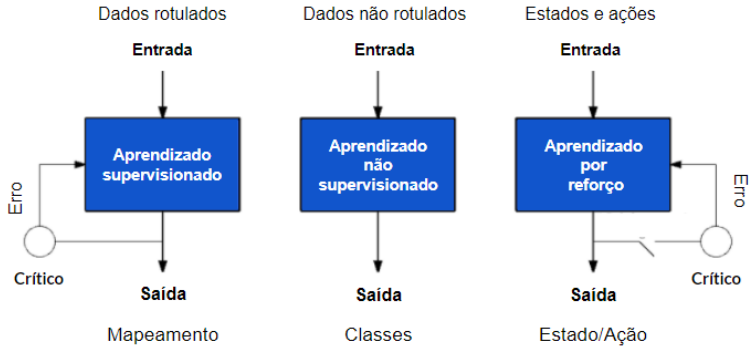
- Se o conjunto de dados consiste em variáveis (atributos) de entrada e uma (ou mais) variável objetivo o modelo é chamado de modelo de aprendizado supervisionado.
- Se o conjunto de dados consiste apenas em recursos sem nenhuma variável objetivo, o modelo em execução neste conjunto de dados é chamado de modelo de aprendizado não supervisionado.
- Se não houver um conjunto de dados previamente fornecidos, o modelo executado neste ambiente é chamado de modelo de aprendizagem por reforço (KAELBLING; LITTMAN; MOORE, 1996).

Neste trabalho vários modelos de aprendizado de máquina foram usados para ajustar as diversas bases de dados estudadas. Os modelos foram escolhidos com base em sua complexidade e capacidade de preservar a variância do conjunto de dados. Do método de regressão linear às redes neurais artificiais, os modelos foram otimizados para o melhor ajuste dos dados.

3.1.1 Tipos de aprendizado de máquinas

O foco da área de aprendizado de máquinas é justamente a geração de conhecimento, ou seja, a aquisição de habilidades ou conhecimento com a experiência. Mais comumente, isso significa sintetizar conceitos úteis de dados históricos. Os algoritmos de aprendizagem de máquina podem ser divididos em 3 categorias amplas: aprendizado supervisionado, aprendizado sem supervisão e aprendizado de reforço, conforme pode ser visto na figura 2.

Figura 2 – Diferentes tipos de aprendizado de máquinas



Fonte: BHATTAD (2019) (modificada)

3.1.1.1 Aprendizado supervisionado (SL)

Aplicações nas quais os dados de treinamento compreendem exemplos de vetores de entrada (atributos) junto com seus vetores de destino (objetivos) correspondentes são conhecidas como problemas de aprendizagem supervisionada (BISHOP, 2006).

Os modelos são ajustados em dados de treinamento consistindo em entradas e saídas e usados para fazer previsões em conjuntos de teste onde apenas as entradas são fornecidas e as saídas do modelo são comparadas com as variáveis de destino retidas e usadas para estimar a habilidade do modelo. Neste caso, o processo de aprendizagem é uma busca através do espaço de hipóteses possíveis para aquele modelo que terá o melhor desempenho, mesmo considerando novos exemplos além do conjunto de dados conhecido (BREWKA, 1996).

Existem dois tipos principais de problemas de aprendizagem supervisionada: classificação, que envolve a previsão de um dado categorizado e regressão, que envolve a previsão de um dado numérico. Ambos podem ter uma ou mais variáveis de entrada e as variáveis de entrada podem ser de qualquer tipo de dados, como numérico ou categórico. Alguns algoritmos podem ser projetados especificamente para problemas de classificação (como regressão logística) ou problemas de regressão (como regressão linear), bem como podem ser usados para ambos os tipos de problemas com pequenas modificações (como as redes neurais artificiais). Exemplos populares de modelos de aprendizagem supervisionada incluem: árvores de decisão e máquinas de vetores de suporte (SVM).

Na indústria de uma forma geral as técnicas de ML supervisionadas são aplicadas principalmente devido à riqueza de dados e geração de conhecimento, bem como devido à natureza complexa dos problemas. Além disso, o ML supervisionado se beneficia da coleta de dados para o controle estatístico do processo, o que usualmente são devidamente rotulados e categorizados.

3.1.1.2 Aprendizado não-supervisionado (UL)

O aprendizado não supervisionado descreve uma classe de problemas que envolve o uso de um modelo para descrever ou extrair relacionamentos nos dados. Comparado ao aprendizado supervisionado, o aprendizado não supervisionado opera apenas com os dados de entrada (atributos), sem saídas ou variáveis objetivo. Assim, a aprendizagem não supervisionada não tem um professor corrigindo o modelo, como ocorre no caso da aprendizagem supervisionada (GOODFELLOW; BENGIO; COURVILLE, 2016).

Existem muitos tipos de aprendizagem não supervisionada, embora existam dois problemas principais que são frequentemente encontrados: agrupamento (*clustering*), que consiste em encontrar grupos nos dados e estimativa de densidade (*density estimation*) que envolve resumir a distribuição dos dados. Um exemplo de algoritmo de agrupamento é o k-Means, em que o parâmetro k se refere ao número de clusters a serem descobertos nos dados. Um exemplo de algoritmo de estimativa de densidade é *Kernel Density Estimation*, que envolve o uso de pequenos grupos de amostras de dados intimamente relacionadas para estimar a distribuição de novos pontos no espaço do problema.

O agrupamento e a estimativa de densidade podem ser realizados para aprender sobre os padrões nos dados. Métodos não supervisionados adicionais também podem ser usados, como a visualização que envolve a representação gráfica ou plotagem de dados de diferentes maneiras e métodos de projeção que envolvem a redução da dimensionalidade dos dados. Um exemplo de técnica de visualização seria uma matriz de gráfico de dispersão que cria um gráfico de dispersão de cada par de variáveis no conjunto de dados.

Um exemplo de método de projeção seria o método de PCA, que envolve reduzir a dimensionalidade do problema, trazendo o conjunto de dados originais em n variáveis independentes que capturam a variação do conjunto de dados original com grande precisão (ABDI; WILLIAMS, 2010).

O objetivo em tais problemas de aprendizagem não supervisionados pode ser descobrir grupos de exemplos semelhantes dentro dos dados, onde é chamado de agrupamento, ou determinar a distribuição de dados dentro do espaço de entrada, conhecido como estimativa de densidade, ou para projetar os dados de um alto espaço dimensional em até duas ou três dimensões para fins de visualização (BISHOP, 2006).

3.1.1.3 Aprendizado por reforço (RL)

A aprendizagem por reforço consiste basicamente em aprender o que fazer - como mapear situações em ações - de modo a maximizar um sinal numérico de recompensa. O aluno não é informado sobre quais ações tomar, mas, em vez disso, deve descobrir quais ações geram mais recompensas experimentando-as (SUTTON; BARTO, 2018).

Alguns algoritmos de aprendizado de máquina não experimentam apenas um conjunto de dados fixo. Por exemplo, algoritmos de aprendizagem por reforço interagem com um ambiente, então há um ciclo de feedback entre o sistema de aprendizagem e suas experiências (GOODFELLOW; BENGIO; COURVILLE, 2016).

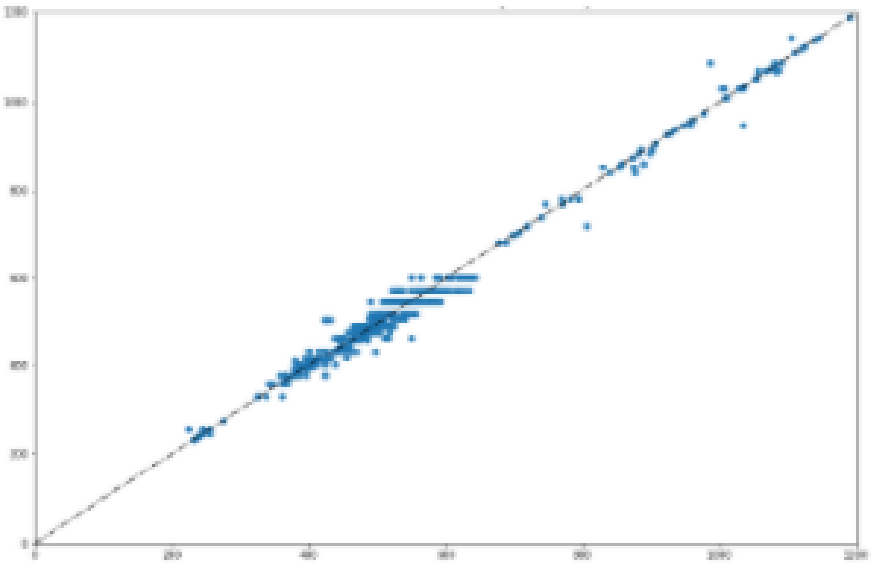
É semelhante à aprendizagem supervisionada no sentido de que o modelo tem alguma resposta a partir da qual possa aprender, embora o feedback possa ser estatisticamente ruidoso, tornando difícil para o modelo conectar causa e efeito. Um exemplo de problema de reforço é jogar um jogo em que o agente tem o objetivo de obter uma pontuação alta e pode fazer jogadas no jogo e receber feedback em termos de punições ou recompensas.

Em muitos domínios complexos, o aprendizado por reforço é a única maneira viável de treinar um programa para um desempenho em níveis elevados. Por exemplo, um modelo de aprendizado por reforço pode ser informado quando ganhou ou perdeu e pode usar essas informações para aprender uma função de avaliação que forneça estimativas razoavelmente precisas da probabilidade de vitória em qualquer posição. O AlphaGo do Google (originalmente desenvolvido pela startup inglesa DeepMind Technologies) é um bom exemplo da aplicação dessa técnica, sendo capaz de superar, em 2015, o 18-vezes campeão mundial do jogo *Go*, o chinês *Lee Sedol*, em uma competição épica por um placar final de 4 x 1 (Deepmind Technologies, 2017).

3.1.2 **Regressão Linear**

A regressão linear é uma equação matemática linear usada para encontrar a relação entre as variáveis independentes (atributos) e a variável dependente (objetivo). Ele determina uma relação estatística entre os dois, que não é determinística. No fundo, a regressão linear tenta encontrar a melhor linha ou plano que se ajusta aos dados fornecidos, minimizando o erro. Um exemplo gráfico pode ser visto na figura 3.

Figura 3 – Exemplo de Regressão Linear



Fonte: Elaborada pelo autor

O erro do modelo de regressão é definido como a distância entre o valor previsto na linha de regressão e o valor real. A regressão linear múltipla assume que a variável de resposta é uma função linear dos parâmetros do modelo e que há mais de uma variável independente no modelo.

A forma geral do modelo de regressão linear múltipla pode ser dada pela expressão 3.1:

$$\varphi = \hat{Y}_i = x_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \epsilon_i, \tag{3.1}$$

em que φ é a variável dependente, β_0, β_1, β_2,...,β_p são os coeficientes de regressão linear, x_1, x_2,...,x_p são as variáveis independentes no modelo e erro aleatório ϵ_i modela o erro associado a cada observação $\hat{Y}_i$, para i=1,2,...,n. Os valores dos coeficientes devem ser escolhidos com o objetivo de minimizar o erro do modelo. Isso é feito considerando a soma dos quadrados como a função de erro, conforme apresentado na equação 3.2:

$$Error = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3.2}$$

Assim, a regressão linear é um bom modelo de regressão para ajustar o conjunto de dados atual, mas não é o melhor devido à natureza simples falhando quando existe uma correlação mais complexa entre as variáveis preditoras (YAN; YAN; SU, 2009).

A regressão linear pode ser considerada como o método mais simples aplicado na engenharia de materiais. Em Jones, Watton e Brown (2005) métodos de aprendizado de máquina foram usados para prever a resistência à fadiga com alta precisão. Usando regressão linear como um dos modelos estudados, uma seleção de modelos foi conduzida a partir de todas as combinações possíveis de variáveis explicativas com base na técnica de validação cruzada.

3.1.3 Regressão Polinomial

A regressão polinomial é uma forma de regressão linear em que a relação entre a variável independente x a variável dependente y é modelada como um polinômio de grau n. A regressão polinomial se ajusta a uma relação não linear entre o valor de x e a média condicional correspondente à y, definida como sendo E(y|x).

Embora a regressão polinomial se ajuste a um modelo não linear para os dados, o problema de estimativa é linear, no sentido de que a função de regressão E(y|x) é linear nos parâmetros desconhecidos que são estimados a partir dos dados. Por esta razão, a regressão polinomial pode ser considerada um caso especial de regressão linear múltipla, conforme representado na expressão 3.3:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 + ... + \beta_n x_n^n + \epsilon_i, \tag{3.3}$$

A regressão polinomial é basicamente usada para definir ou descrever fenômenos não lineares como a taxa de crescimento dos tecidos, progressão da epidemia de doenças e distribuição de isótopos de carbono em sedimentos de lagos. A regressão polinomial é usada para evitar subajuste e aumenta a complexidade do modelo nos casos em que linear simples modelos de regressão são ineficazes.

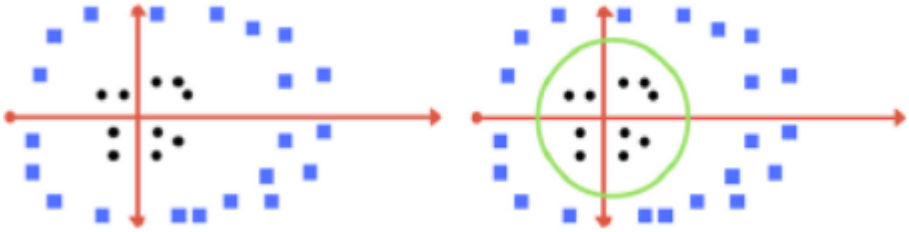
Embora a regressão polinomial reduza o subajuste, ela corre o risco de *overfiting* dos dados, causando grandes problemas ao modelo pois o método é muito sensível a *outliers*. Este problema é ainda mais afetado pela falta de algoritmos eficientes para detectar e eliminar *outliers* do conjunto de dados com sucesso.

3.1.4 *Support Vector Machine (SVM)*

Support-vector machine são modelos de aprendizado supervisionado frequentemente usados para análise de problemas de classificação, mas também podem ser usados, com bons resultados, para problemas de regressão. Nesse caso o SVM se torna o *Support Vector Regression* (SVR), que basicamente usa os mesmos princípios do modelo para classificação (CORTES; VAPNIK, 1995).

O algoritmo SVM é projetado de forma a procurar pontos no gráfico que estão localizados diretamente na linha divisória mais próxima. Esses pontos são chamados de vetores de suporte. Em seguida, o algoritmo calcula a distância entre os vetores de referência e o plano divisor, valor conhecido como lacuna. O melhor hiperplano é aquele hiperplano para o qual essa lacuna é a maior possível. A complexidade por trás do método pode ser entendido usando a figura 4 abaixo.

Figura 4 – Support Vector Machine



Fonte: Bishop (2006)

O hiperplano é a linha que nos ajudará a prever o valor contínuo ou valor alvo. Existem duas linhas além do hiperplano que criam uma margem. Estas são as linhas de fronteira. Os vetores de suporte podem estar nas linhas de limite ou fora dela. Mantendo essa definição em mente, aqui estão os principais pontos de diferença entre os métodos SVM e SVR:

- Como a saída é um número real e não uma classificação, torna-se muito difícil prever as informações disponíveis, que têm possibilidades infinitas.
- No caso de regressão, uma margem de tolerância (épsilon) é definida como critério de convergência da resposta do modelo.

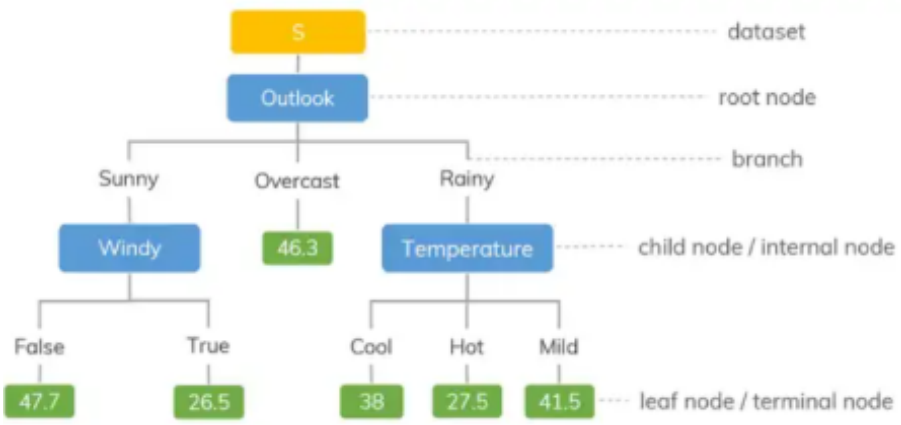
O SVR pode ser facilmente aplicado em diversas aplicações industriais. Wang et al. (2020) aplicou o método para predizer propriedades mecânicas do aço. O resultado obtido a partir da aplicação prática em algumas siderúrgicas mostrou que o algoritmo proposto melhorou o desempenho da modelagem anteriormente existente.

3.1.5 Árvore de Decisão (*Decision Trees*)

Como o próprio nome sugere, o regressor baseado em árvores de decisão (mais comumente reconhecido na literatura como *Decision Trees*) usa uma estrutura em cascata para tomar decisões conforme "desce" na estrutura da árvore. O nó folha ou os nós mais externos das árvores consistem em valores ou categorias atribuídas conforme você cascateia

por um dos ramos da árvore de decisão. O conceito básico do modo de funcionamento pode ser visto na figura 5.

Figura 5 – Conceitos básicos da árvore de decisão



Fonte: Quinlan (1986)

A árvore de decisão divide o conjunto de dados em subconjuntos cada vez menores e tenta ajustar as árvores de decisão a esses subconjuntos. Isso torna o modelo extremamente robusto e genérico. O resultado final é uma árvore com ramificações e possíveis decisões no ramo final ou nó folha (*leaf node*).

O nó raiz de cada árvore de decisão, responsável pela decisão que está sendo tomada por aquela árvore, é chamado de nó de decisão. A partir dos nós de decisão, duas ou mais ramificações aparecem, com cada ramificação espelhando os valores ou intervalo de valores para um atributo específico.

Embora a descrição acima aponte para uma aplicação clara para problemas de classificação, as árvores de decisão são extremamente eficientes e robustas ao serem implementadas em problemas de regressão. Os nós internos desenvolvem uma gama de valores à medida que o valor verifica os recursos e leva a um grande número de nós folha devido ao aumento da sofisticação.

Um ponto importante que impede o uso da árvore de decisão é o quão instável ela pode se tornar, mesmo para pequenas mudanças no banco de dados. Como o método usa todo o intervalo ou classes de uma atributo, qualquer mudança nesse atributo pode desestabilizar a árvore de decisão.

As árvores de decisão são transparentes por natureza, com toda a árvore acessível ao desenvolvedor. Isso o torna fácil de entender, por isso é classificado como um "modelo caixa branca". A tomada de decisão também é precisa com valores / categorias específicos atribuídos no final de um ramo no nó folha. Em última análise, o regressor de árvore de decisão é uma boa opção para aplicações industriais (QUINLAN, 1986)

3.1.6 **Florestas Aleatórias (*Random Forest*)**

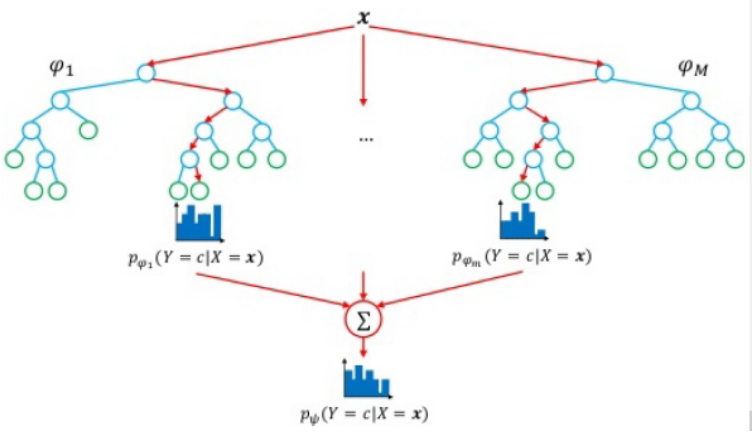
O método de Florestas Aleatórias é um tipo de modelo aditivo que faz previsões combinando decisões a partir de uma sequência de modelos básicos. Mais formalmente, podemos escrever esta classe de modelos como sendo:

$$g(x) = f_0(x) + f_1(x) + ... \tag{3.4}$$

em que o modelo final g é a soma dos modelos de base simples. Aqui, cada classificador base é uma árvore de decisão simples. Esta técnica de usar vários modelos para obter melhor desempenho preditivo é chamado de agrupamento de modelos (*model ensembling*).

Neste método todos os modelos básicos são construídos de forma independente usando uma subamostra diferente dos dados. O tamanho da subamostra é sempre igual ao tamanho da amostra de entrada original. O uso de múltiplas árvores de decisão para criar o modelo completo é conhecido como agregação *Bootstrap* ou *Bagging*. O método *Bagging* envolve o treinamento de cada árvore de decisão de forma individual com uma subamostra diferente de conjunto de dados. O conceito básico do método pode ser visto na figura 6.

Figura 6 – Conceitos básicos do *Random Forest*



Fonte: Breiman (2001a)

O resultado final do modelo é decidido por meio de uma média ponderada por meio de um mecanismo de votação para classificar as árvores de decisão propostas. Isso garante que a variação do conjunto de dados original seja contabilizada juntamente com a redução do *overfiting* no modelo levando também em consideração os resultados das árvores de decisão mal ajustadas.

Comparado com outras técnicas baseadas em árvores de decisão, o método das Florestas Aleatórias tem duas vantagens principais: leva em consideração as correlações e dependências entre as variáveis do projeto e apresenta boa robustez com relação à presença do ruído (BREIMAN, 2001b).

3.1.7 *Gradient Boosting*

O método *Gradient Boosting* é bastante semelhante ao método das Florestas Aleatórias, pois o primeiro também é um conjunto de *weak learnes*, na sua maioria das vezes árvores de decisão. Este conjunto recebe pesos individuais e o resultado final observado é o aumento de gradiente. O regressor obtido é uma média ponderada de resultados individuais dessas árvores de decisão. O grande ponto de diferença entre o método visto anteriormente e *Gradient Boosting* é a a sua flexibilidade de escolher a função de perda. No caso do método das Florestas Aleatórias é definido um função de perda fixa, conhecida como Impureza Gini (*Gini Impurity)* (BREIMAN, 2001b).

Especificamente, em cada iteração, uma subamostra dos dados de treinamento é desenhada aleatoriamente (sem substituição) do conjunto de dados de treinamento completo. Cada subamostra selecionada aleatoriamente é então usada no lugar da amostra completa para ajustar o aluno básico e calcular a atualização do modelo para a iteração atual. Essa abordagem aleatória também aumenta a robustez do modelo contra efeito do *overfiting*.

Algumas das funções de perda que podem ser usadas estão: soma dos mínimos quadrados, menor desvio absoluto ou uma combinação de soma de quadrados mínimos e desvio absoluto mínimo. Além disso, funções de perda personalizadas que são diferenciáveis também pode ser aplicadas para otimizar o regressor de aumento de gradiente.

3.1.8 *Adaptative Boosting*

O método *ADABoost*, abreviação de *Adaptive Boosting*, é um modelo de reforço de conjunto único baseado em árvores de decisão para trabalhar em problemas de regressão. A principal diferença entre *ADABoost* e outros regressores baseados em *boosting* é o uso de árvores de decisão com apenas um único nível, sendo que o número de árvores de decisão usadas é um parâmetro fundamental do modelo (ZOU et al., 2009).

O algoritmo foi desenvolvido originalmente para classificação e envolve a combinação das previsões feitas por todas as árvores de decisão no conjunto. Uma abordagem semelhante também foi desenvolvida para problemas de regressão em que as previsões são feitas usando a média das árvores de decisão. A contribuição de cada modelo para a previsão de conjunto é ponderada com base no desempenho do modelo no conjunto de dados de treinamento (FREUND; SCHAPIRE, 1995).

De acordo com Zou et al. (2009) o algoritmo de treinamento envolve começar com uma árvore de decisão, encontrar os exemplos no conjunto de dados de treinamento que foram classificados incorretamente e adicionar mais peso a esses exemplos. Outra árvore é portanto treinada com os mesmos dados, embora agora ponderada pelos erros de classificação. Este processo é repetido até que o número de árvores seja adicionado.

3.1.9 *Light Gradient Boosting Machine (LightGBM)*

O método *LightGBM*, abreviação de *Light Gradient Boosting Machine*, apresenta uma arquitetura baseado na técnica de *gradient boosting* que usa algoritmos de aprendizagem baseados em árvores. Originalmente desenvolvido pela Microsoft em 2017, esse método foi cuidadosamente projetado para ser eficiente computacionalmente e com diversas vantagens sobre outros modelos de aprendizado de máquinas: maior eficiência e velocidade de treinamento mais rápida, melhor precisão e capacidade de lidar com grandes volumes de dados.

De forma a ser mais eficiente o *LightGBM* não usa o algoritmo de aprendizagem de árvore de decisão baseado em classificação amplamente usado, como por exemplo o método *XGBoost*. Em vez disso, o método implementa um algoritmo de aprendizagem de árvore de decisão baseado em histograma altamente otimizado, que oferece grandes vantagens tanto na eficiência quanto no consumo de memória (MEHTA; AGRAWAL; RISSANEN, 1996).

De acordo com Ke et al. (2017), uma das grandes desvantagens do algoritmo baseado em árvores é o risco de *overfitting* no caso de conjuntos de dados relativamente pequenos. Em particular no método *LightGBM* isso pode ser controlado fixando a profundidade máxima da árvore, reduzindo assim o problema.

3.1.10 *Extreme Gradient Boosting (XGBoost)*

O método *XGBoost*, abreviação para *Extreme Gradient Boosting*, cujo termo se origina do artigo de 1999 *Greedy Function Approximation: A Gradient Boosting Machine*, proposto por Jerome H. Friedman. O método é essencialmente uma biblioteca desenvolvida para ser eficiente e flexível computacionalmente, implementando algoritmos de aprendizado de máquina sob a arquitetura do *Gradient Boosting*.

O *XGBoost* funciona tendo técnicas de árvores de classificação e regressão (CART), classificando os pontos de dados em categorias apropriadas e então pontuando-os para problemas de regressão. O funcionamento básico é muito semelhante aos outros algoritmos de *boosting* nos quais os modelos são ajustados usando qualquer função de perda diferenciável e arbitrária através de um algoritmo de otimização de gradiente descendente. Isso dá à técnica o nome de "reforço de gradiente", pois o gradiente de perda é minimizado conforme o modelo é ajustado, de forma semelhante à uma rede neural. A importância das variáveis de entrada (influência no modelo) pode ser calculada usando XGBoost ou usando um modelo híbrido combinando com outras técnicas de aprendizado de máquina.

O *XGBoost* tem uma ampla gama de aplicativos, usados para resolver tanto problemas de regressão como classificação. O método é atualmente um dos melhores algoritmos de *boosting*, identificando e capturando a correlação complexa entre recursos e preserva a variância do conjunto de dados (CHEN; GUESTRIN, 2016).

3.1.11 *Redes Neurais Artificiais (RNA)*

Basicamente pode-se citar que uma RNA é um modelo de aprendizado de máquina feito para imitar o funcionamento do cérebro humano que tem milhões de células cerebrais, chamadas neurônios. Uma rede neural artificial é uma replicação simplificada disso, pois esquematicamente é um complexo de nós interconectados que agem como neurônios.

Os neurônios são as unidades computacionais básicas e essenciais de uma rede neural artificial. Eles pegam uma determinada entrada, executa alguns cálculos específicos e produz uma determinada saída. Um neurônio tem duas componentes essenciais: pesos e viés. Além dos neurônios, uma RNA tem as seguintes partes básicas de sua topologia: camada de entrada, uma ou mais camada ocultas e uma camada de saída.

Cada camada é composta por um determinado número de neurônios. O treinamento de uma RNA será nada mais do que a atualização constante dos pesos e tendências associadas a cada neurônio até que eles atinjam um valor ideal que ajude a ajustar o conjunto de dados. Este processo de atualização é feito por meio da propagação à frente (*forward*) do conjunto de dados através do modelo e da propagação para trás (*backward*) do erro associado às previsões por meio do modelo.

A camada de entrada é uma camada única que contém um número de neurônios normalmente igual ao número de atributos de entrada no conjunto de dados, pois assim cada neurônio na camada de entrada representa um atributo desse conjunto de dados. As camadas ocultas são as camadas intermediárias da rede neural. Cada camada pega a entrada de uma camada anterior, realiza os cálculos e passa a saída para a próxima camada.

Antes que a saída dos neurônios seja transmitida, uma função de ativação é aplicada para isso. Esta função de ativação garante que nenhum neurônio individual em uma determinada camada domina a produção geral dessa camada. Ele restringe a saída de um neurônio dentro de um faixa particular, aplicando uma faixa uniforme para toda a camada. Obviamente esseintervalo depende de o tipo de função de ativação usada.

A camada de saída é considerada a camada final da RNA. O número de neurônios na camada de saída depende do tipo de problema que está sendo trabalhado: um problema de regressão exigirá um único neurônio, enquanto que um problema de classificação binária requer dois neurônios e um problema multiclasse k requer k neurônios.

As RNA são um modelo dito caixa preta, complexo e muito eficiente, pois pode capturar a correlação complexa entre recursos que alguns dos modelos mais simples não são capazes de fazê-lo. Sem dúvida as RNA são um componente importante entre as técnicas de aprendizado de máquina BHATTAD (2019).

3.2 Motivação para a seleção de atributos

A grande expectativa é que os modelos de aprendizado de máquinas preditores projetados capturem uma relação preditiva com o resultado objetivado. Alguns problemas de modelagem preditiva têm um grande número de variáveis que podem retardar o desenvolvimento e o treinamento de modelos e requerem uma grande quantidade de memória do sistema. Além disso, o desempenho de alguns modelos pode degradar ao incluir variáveis de entrada que não são relevantes para a variável de destino.

Pode-se perceber que existe uma necessidade real de selecionar as variáveis preditoras (atributos) de forma adequada para modelagem dos problemas estudados, ou seja, têm como objetivo reduzir o número de variáveis de entrada para aquelas que se acredita serem mais úteis para um modelo, a fim de prever a variável objetivo do problema em estudo.

Os métodos de seleção de recursos podem ser classificados em supervisionados e não supervisionados. Basicamente a diferença está relacionada com o fato das variáveis preditores serem selecionadas com base na variável objetivo ou não:

- As técnicas de seleção de recursos não supervisionados ignoram a variável objetivo, tais como métodos que removem variáveis redundantes usando correlação.
- As técnicas de seleção de recursos supervisionados usam a variável objetivo, tais como métodos que removem variáveis irrelevantes.

Existem ainda algoritmos de aprendizado de máquina que realizam a seleção de recursos automaticamente como parte do aprendizado do modelo. Podemos nos referir a essas técnicas como métodos de seleção de recursos intrínsecos.

Algumas técnicas podem inclusive ter o processo de seleção de atributos de forma integrada, o que significa que o modelo incluirá apenas preditores que ajudam a maximizar a precisão. Nesses casos, o modelo pode selecionar e escolher qual representação dos dados é a melhor. Isso inclui algoritmos como modelos de regressão penalizados como Lasso, *Decision Tree* e *Random Forest*.

A seleção de recursos também está relacionada à técnicas de redução de dimensionalidade em que ambos os métodos buscam um menor número de atributos para um modelo preditivo. A diferença é que a seleção de recursos seleciona atributos para manter ou remover do conjunto de dados, enquanto a redução de dimensionalidade cria uma projeção dos dados, resultando em atributos inteiramente novos.

3.2.1 Seleção univariada de atributos

A seleção univariada de atributos funciona selecionando os melhores utilizando testes estatísticos univariados. Essencialmente comparamos cada atributo com a variável objetivo, para ver se há alguma relação que seja estatisticamente significativa entre eles, ignorando assim os outros atributos na análise (por isso o método é chamado de seleção univariada). Para a medição dessa relação pode-se usar o coeficiente de *Pearson*, coeficiente de informação máxima ou correlação de distância.

Portanto pode-se calcular uma pontuação para cada um dos atributos do problema e, ao final do processo, todas as pontuações são então comparadas e os recursos com os maiores valores são selecionados.

3.2.2 *Recursive Feature Elimination* (RFE)

Guyon et al. (2002) define a técnica RFE como uma seleção retroativa das variáveis preditores (atributos). Essa técnica começa construindo um modelo em todo o conjunto de preditores e calculando uma medida de pontuação que classifica os preditores do mais importante ao menos importante. Em cada estágio da pesquisa, os preditores menos importantes são eliminados interativamente antes de reconstruir o modelo. As pontuações podem ser determinadas usando o modelo de aprendizado de máquina (por exemplo alguns algoritmos como árvores de decisão oferecem pontuações de importância) ou usando algum método estatístico.

3.2.3 *Permutation Importance* (FI)

O método *Permutation Importance* fornece uma maneira de calcular as importâncias dos atributos para qualquer estimador a partir de qualquer métrica de interesse (acurácia, F1, R^2, etc), que basicamente diminui quando um atributo não está disponível. O método é também conhecido na literatura como *Mean Decrease Accuracy* (MDA).

3.2.4 *Mutual Information (MI)*

O método *Mutual Information* mede as quedas de entropia sob a condição da variável objetivo. A pontuação do parâmetro MI cairá sempre no intervalo de 0 a 1, de forma que quanto maior o valor, mais estreita é conexão entre esse atributo e a variável objetivo, sugerindo portanto que esse atributo esteje no conjunto de dados de treinamento. Se a pontuação do MI for 0 ou muito baixa, sugere-se uma conexão fraca entre essa feature e o objetivo, nesse caso deve-se considerar não inserir o atributo nos dadps de treinamento.

3.3 PyCaret

Nesse trabalho foi usado o pacote PyCaret para as simulações dos modelos de machine learning estudados. PyCaret é uma biblioteca de aprendizado de máquina de código aberto em Python cuidadosamente projetada para facilitar a execução de tarefas padrão em um projeto de aprendizado de máquina. Diversas etapas de um projeto de aprendizado de máquina são automatizadas, conforme descrito nas subseções a seguir.

3.3.1 Setup do modelo

O pacote PyCaret, utilizando a função *setup*, orquestra automaticamente todas as dependências em um pipeline onde as tarefas de pré-processamento e preparação de dados são realizadas tais como: armazenagem das fontes dos dados, *hold-out*, amostragem, tratamento de dados faltantes, *encoding*, transformação dos dados, PCA, *feature importance* entre outros.

3.3.2 Criação do modelo

A criação de um modelo de aprendizado de máquina no PyCaret é feito utilizando a função *create model*, retornando uma tabela com métricas de desempenho obtidas através de validação cruzada:

- Classificação: precisão, AUC, revocação, precisão, F1, Kappa,..
- Regressão: MAE, MSE, RMSE, R2, RMSLE, MAPE...

Diversos modelos estão disponíveis, entre eles: Regressão Logística, Regressão linear e Polinomial, *Ridge*, *Lasso*, KNN, *Naives Bayes*, SVM, Multi Level Perceptron (MLP), Árvores de Decisão, Florestas Aleatórias, *AdaBoost*, *Catboost*, *Gradient Boosting*, *Extreme Gradient Boosting* (Xgboost) e *Light Gradient Boosting* (Lightgbm).

Existe também a possibilidade da criação de agrupamento de modelos no PyCaret utilizando a função *ensemble model*. Esta função retorna uma tabela com a lista dos modelos usados no agrupamente com as suas respectivas pontuações obtidas por validação cruzada. O método *bagging* é usado por padrão no agrupamento, podendo ser alterado para *boosting*. A biblioteca PyCaret também fornece a possibilidade de combinar modelos através de um sistema de votação utilizando a função *blend model*.

3.3.3 Ajuste do modelo

Quando um modelo é criado o PyCaret usa os valores padrões dos hiperparâmetros para treiná-lo. Para fazer um ajuste de hiperparâmetros, pode-se usar a função *tune model*. Esta função ajusta automaticamente os hiperparâmetros do modelo usando a estratégia

Random Grid Search em um espaço de pesquisa predefinido, tendo como resposta um grid de pontuação do modelo com os seguintes valores:

- Classificação: acurácia, recall, precisão, F1, curva ROC e MCC.
- Regressão: MAE, MSE, RMSE, R2, RMSLE e MAPE.

3.3.4 Visualização do modelo

A função *plot model* pode ser usada para analisar o desempenho do ajuste dos hiérparâmetros em diferentes aspectos. Existem mais de 15 gráficos diferentes disponíveis na biblioteca PyCaret tais como resíduos e erro de predição, curvas de aprendizagem e validação, curva ROC, curva precisão-revocação, matriz de confusão, fronteiras de decisão, *feature importance* e outros.

3.3.5 Interpretação do modelo

Outra função interessante é a *interpret model*, pois trata-se de uma ferramenta bastante útil para explicar as previsões de um determinado modelo utilizando a biblioteca *SHAP* do python.

4 RESULTADOS

Neste trabalho visou-se estabelecer um *framework* para explorar as técnicas de aprendizado de máquina no campo da ciência dos materiais. A partir da metodologia adotada buscou-se estabelecendo ligações causais e confiáveis entre variáveis de processo, composição química e propridades metalúrgicas como resistência à fadiga de diversas classes de aços. O *framework* proposto é composto essencialmente de 4 passos:

- Pré-processamento para descrição consistente de dados, que pode incluir ações como preencher dados ausentes ou faltantes sempre que possível, com a ajuda de conhecimento prévio do problema.
- Seleção de recursos para classificação de atributos e/ou identificação do melhor subconjunto de atributos para estabelecer uma determinada ligação com o que se deseja predizer.
- Modelagem preditiva usando várias estratégias estatísticas e avançadas baseadas em dados para o estabelecimento das ligações desejadas.
- Avaliação crítica dos diferentes métodos usando métricas apropriadas e avaliação do ajuste dos modelos de forma a evitar o *overfitting*.

A previsão precisa da resistência à fadiga de aços é de particular importância na ciência dos materiais para várias aplicações de tecnologia avançada por causa do custo (e tempo) extremamente alto dos testes de fadiga e, muitas vezes, das consequências debilitantes das falhas por fadiga. A resistência à fadiga é o dado mais importante e básico necessário para o projeto e análise de falha de componentes mecânicos. É relatado que a fadiga é responsável por mais de 90% de todas as falhas mecânicas de componentes estruturais (AGRAWAL et al., 2014; AGRAWAL; CHOUDHARY, 2016).

Portanto, a previsão da vida em fadiga é de extrema importância para as comunidades de ciência de materiais e engenharia mecânica. O uso de um grande número de parâmetros de processo de tratamento térmico, composição para prever propriedades de valor extremo, como resistência à fadiga, motivou à trabalhar neste problema.

4.1 Base de dados NIMS

O conjunto de dados de fadiga de aços do Instituto Nacional de Ciência de Materiais (NIMS) MatNavi foi usado neste trabalho, sendo reconhecido na literatura como um dos maiores bancos de dados do mundo com detalhes sobre a composição química, propriedades mecânicas oriundas do processo de laminação e parâmetros relacionados ao processo de tratamento térmico. O banco de dados inclui aços carbono e de baixa liga.

Os dados de vida de fadiga, que dizem respeito aos testes de fadiga por flexão rotativa em condições de temperatura ambiente, foram a propriedade alvo para a qual objetivamos construir modelos preditivos. Os atributos do problema em questão podem ser categorizados da seguinte forma:

- Composição química: %C, %Si, %Mn, %P, %S, %Ni, %Cr, %Cu, %Mo (% em peso).
- Dados da laminação: tamanho do lingote e taxa de redução.
- Dados de tratamento térmico: temperatura e tempo de processo para as fases de normalização, carburização, difusão, endurecimento e têmpera, medidas em laboratório.
- Propriedade mecânica: resistência à fadiga (Mpa).

Os dados usados neste trabalho possuem 437 instâncias, 25 atributos e 1 variável objetivo (resistência à fadiga). As 437 instâncias de dados incluem 371 referentes à aços carbono e de baixa liga, 48 outros tipos de aço. Esses dados referem-se a vários tipos de tratamento térmico de cada tipo de aço e diferentes condições de processamento (AGRAWAL et al., 2014).

Este conjunto de dados foi processado, incluindo limpeza de dados para remover pontos de dados nulos, escalonamento de dados para trazer todos os atributos para uma faixa uniforme e remoção de assimetria para garantir uma curva gaussiana para todas as variáveis preditoras.

Esses dados são então executados por meio de uma série de modelos baseados em regressão para predição da resistência à fadiga (Mpa). A eficácia desses modelos é avaliada dividindo o conjunto de dados em treinamento e dados de teste, permitindo o cálculo da precisão do modelo, As técnicas de regressão estudadas são listadas abaixo.

- Regressão Linear, *Lasso* e *Ridge*.
- Regressão baseada em árvores de decisão (*Decision Tree*, *Extra Tree*).
- Técnicas baseadas em *Boosting* como *Gradient Boosting*, *ADA*, *LightGBM* e *XGBoost*.

- Técnicas baseadas em *Bagging* como *Random Forest* e MLP.

A acurácia de cada um dos modelos estudados foi medida usando os indicadores R, R^2, *Mean Absolute Error* (MAE) e *Root Mean Square Error* (RMSE). A definição de cada um desses critérios é apresentada a seguir nas equações 4.1, 4.2 e 4.3 respectivamente.

$$R = \frac{\sum_{i=1}^{n} (y_i - \bar{y}) \left(\hat{y}_i - \bar{\hat{y}}\right)}{sqrt \sum_{i=1}^{n} (y_i - \bar{y})^2 \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2} \tag{4.1}$$

$$MAE = \bar{e} = \frac{1}{n} \sum_{n} |y - \hat{y}| \tag{4.2}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{n} (y - \hat{y})^2} \tag{4.3}$$

em que y define os valores atuais da resistência à fadiga (Mpa), $\hat{y}$ define o valor da predição da variável objetivo, $\bar{e}$ definido como o erro médio absoluto (MAE) e n sendo o número de instâncias obtidas do banco de dados NIMS.

O quadrado do coeficiente de correlação, R^2, representa a variância explicada pelo modelo (nesse caso quanto maior, melhor) e é considerada uma das métricas mais relevantes para avaliar a precisão dos modelos de regressão.

Tabela 1 – **Detalhes da base de dados NIMS**

Abreviação	**Detalhes**
C	% Carbono
Si	% Silício
Mn	% Manganês
P	% Fósforo
S	% Enxofre
Ni	% Niquel
Cr	% Cromo
Cu	% Cobre
Mo	% Mobilibidênio
NT	Temperatura da fase de normalização
THT	Temperatura da fase de endurecimento
THt	Tempo da fase de endurecimento
THQCr	Taxa de resfriamento na fase de endurecimento
CT	Temperatura da fase de de Carburização
Ct	Tempo da fase de carburização
DT	Temperatura da fase de difusão
Dt	Tempo da fase de difusão
QmT	Temperatura média de têmpera (para carburização)
TT	Temperatura da fase de têmpera
Tt	Tempo da fase de têmpera
TCr	Taxa de resfriamento para têmpera
RedRatio	Taxa de redução (lingote para barra laminada)
dA	Proporção de área de inclusões formadas por deformação plástica
dB	Proporção de área de inclusões formadas por descontinuidade
dC	Proporção de área de inclusões isoladas
Fatigue	Resistência à fadiga (Mpa)

4.2 Análise exploratória dos dados (EDA)

4.2.1 *Describe* da base de dados

A atributo *RedRatio* apresenta registros com valores muito superiores ao limite máximo do *bloxplot*, caracterizando um provável *outlier* a ser excluido da base de dados. Com relação aos outros atributos, Como não se tem muita informação adicional dos tipos (ou classes) de aços utilizandos na construção da base de dados NMIS, recomenda-se que os possíveis *outliers* não sejam eliminados em uma primeira análise.

Tabela 2 – **Análise estatística resumida da base de dados**

Atributo	**Contagem**	**média**	**desvio-padrão**	**Mínimo**	**25%**	**50%**	**75%**	**Máximo**
NT	437,0	872,29	26,21	825,00	865,00	870,00	870,00	930,00
THT	437,0	737,64	280,03	30,00	845,00	845,00	855,00	865,00
THt	437,0	25,94	10,26	0,00	30,00	30,00	30,00	30,00
THQCr	437,0	10,65	7,84	0,00	8,00	8,00	8,00	24,00
CT	437,0	128,85	281,74	30,00	30,00	30,00	30,00	930,00
Ct	437,0	40,50	126,92	0,00	0,00	0,00	0,00	540,00
DT	437,0	123,69	267,12	30,00	30,00	30,00	30,00	903,33
Dt	437,0	4,84	15,70	0,00	0,00	0,00	0,00	70,20
QmT	437,0	35,49	19,41	30,00	30,00	30,00	30,00	140,00
TT	437,0	536,84	164,10	30,00	550,00	600,00	650,00	680,00
Tt	437,0	65,08	21,47	0,00	60,00	60,00	60,00	120,00
TCr	437,0	20,81	8,07	0,00	24,00	24,00	24,00	24,00
% C	437,0	0,388	0,096	0,170	0,340	0,400	0,430	0,630
% Si	437,0	0,299	0,246	0,160	0,240	0,260	0,290	2,050
% Mn	437,0	0,823	0,279	0,370	0,700	0,760	0,800	1,600
% P	437,0	0,015	0,005	0,002	0,012	0,016	0,019	0,031
% S	437,0	0,014	0,006	0,003	0,010	0,015	0,019	0,030
% Ni	437,0	0,517	0,852	0,010	0,020	0,060	0,460	2,780
% Cr	437,0	0,570	0,411	0,010	0,120	0,710	0,980	1,170
% Cu	437,0	0,067	0,049	0,010	0,020	0,060	0,100	0,260
% Mo	437,0	0,069	0,088	0,000	0,000	0,000	0,170	0,240
RedRatio	437,0	923,62	576,61	240,00	590,00	740,00	1228,00	5530,00
dA	437,0	0,047	0,031	0,000	0,020	0,040	0,070	0,130
dB	437,0	0,00339	0,00824	0,000	0,000	0,000	0,000	0,050
dC	437,0	0,00771	0,01041	0,000	0,000	0,000	0,010	0,058
Fatigue	437,0	552,90	186,63	225,00	448,00	505,00	578,00	1190,00

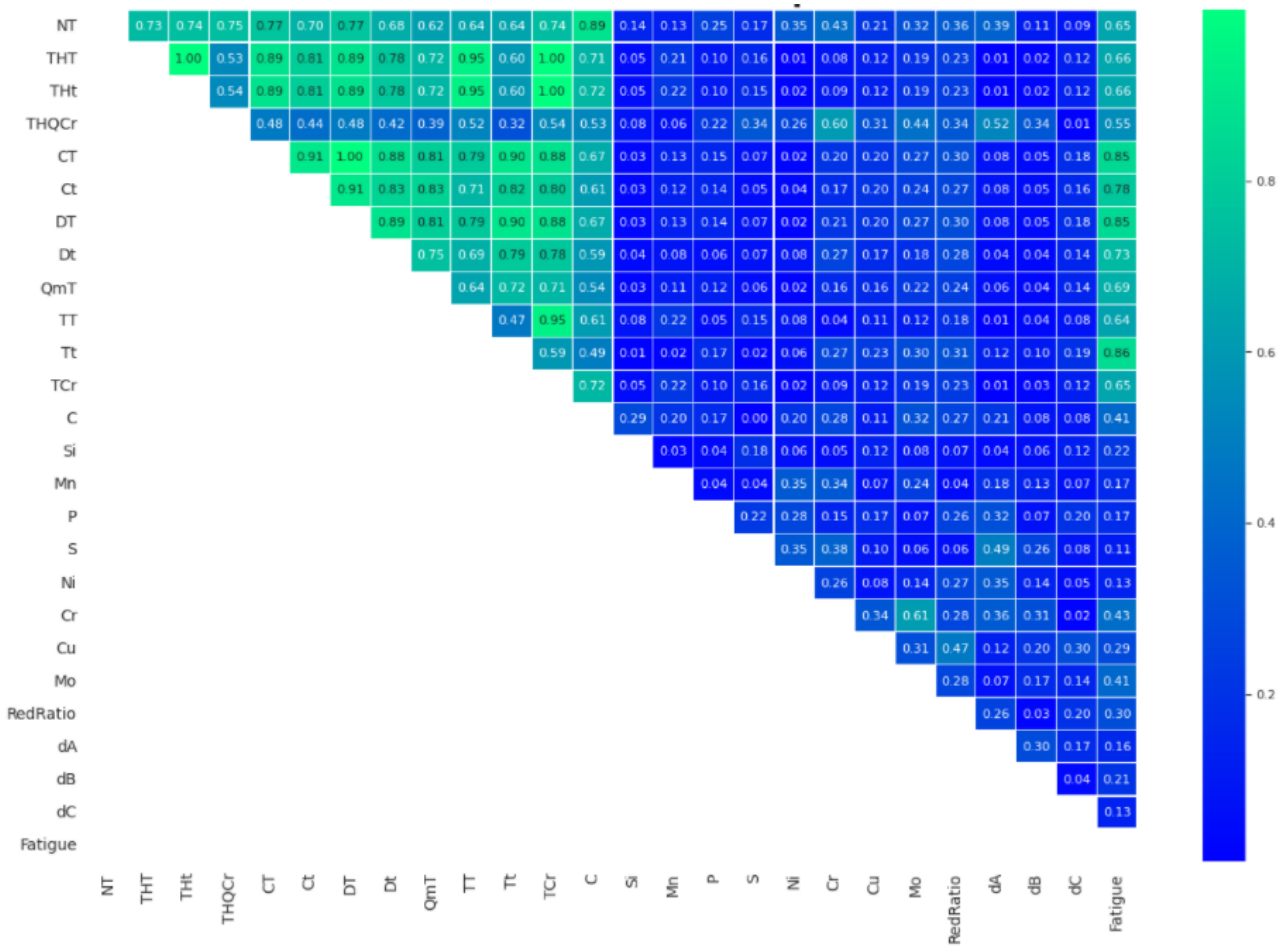
4.2.2 Análise da correlação dos dados

Uma análise da correlação dos dados dos atributos de entrada com a variável objetivo resistência à fadiga pode ser visualizado através do gráfico de *Heatmap* representado na Figura 7.

Pode-se observar claramente que existe um grupo específico de atributos que apresentam alta correlação entre si, referentes às medições de temperatura, tempo e taxa de resfriamento dos processos de normalização, endurencimento, carburização, difusão e têmpera. Por outro lado os atributos referentes à composição química dos aços e proporção de área de incluões mostram ter baixa correlação entre si e com a variável objetivo.

Na Figura 8 pode-se observar que a variável CT (Temperatua da fase de carburização) tem uma alta correlação (100%) com a variável DT (Temperatura da fase de difusão).

Figura 7 – Heatmap da base de dados NMIS



Fonte: Elaborada pelo autor.

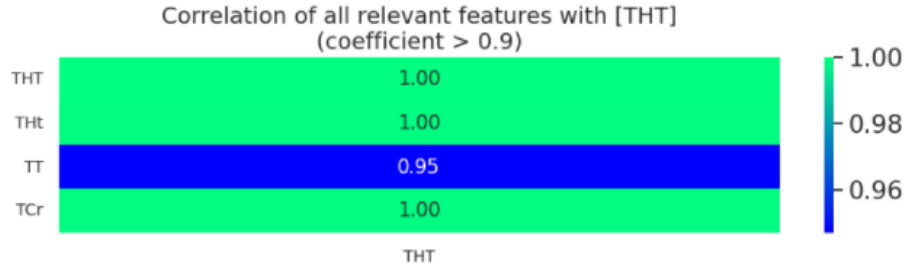
Portanto o atributo DT será eliminado da lista de atributos para as próximas fases desse estudo.

Figura 8 – Correlação do atributo CT

Fonte: Elaborada pelo autor.

Com o atributo THT (Temperatura da fase de endurecimento) observa-se na Figura 9 caso similar de alta correlação (100%) com os atributos THt (Tempo da fase de endurecimento) e TCr (Taxa de resfriamento para a fase de têmpera). Esses dois atributos serão também eliminados para as próximas fases desse estudo.

Figura 9 – Correlação do atributo THT



Fonte: Elaborada pelo autor.

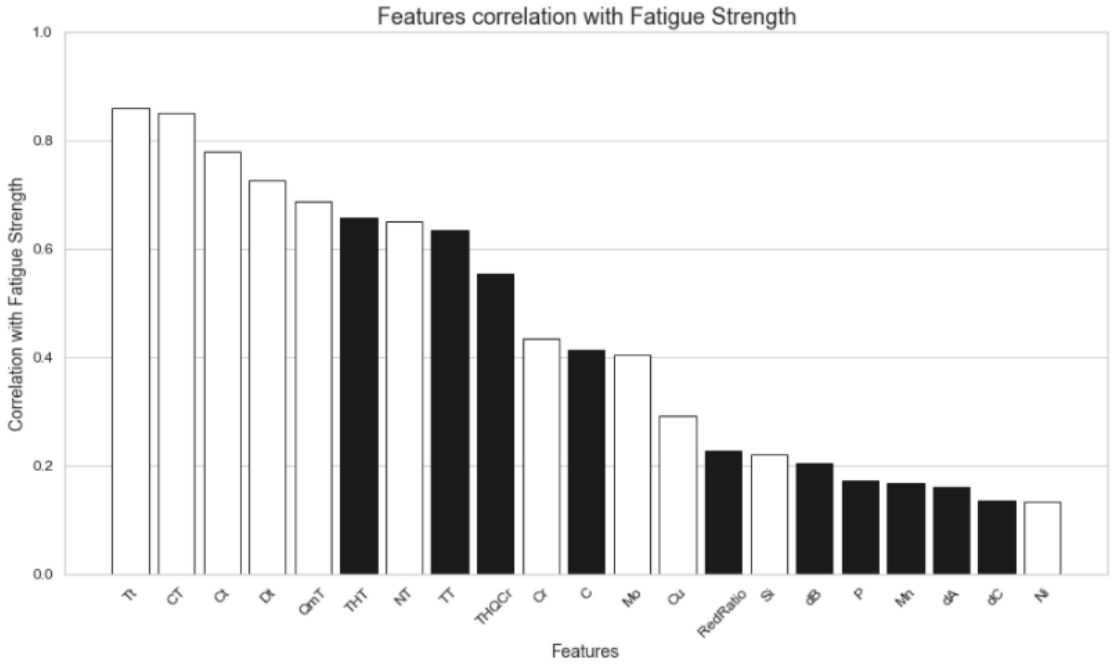
Pode-se observar claramente que existe um grupo específico de atributos que apresentam alta correlação entre si, referentes às medições de temperatura, tempo e taxa de resfriamento dos processos de normalização, endurencimento, carburização, difusão e têmpera. Por outro lado os atributos referentes à composição química dos aços e proporção de área de incluões mostram ter baixa correlação entre si e com a variável objetivo resistência à fadiga.

A Figura 10 apresenta a correlação entre os atributos de entrada remanescentes com o a variável objetivo resistência à Fadiga. Barras na cor preta indicam correlação negativa.

Uma análise estatística da variável objetivo pode ser visualizada na Figura 11. Note que a curva de distribuição sinaliza três possíveis distribuições de probabilidade distintas. Isso faz sentido uma vez que existem famílias de aço muito distintas em termos de comportamento à resistência por fadiga: aços carbono, aços de baixa liga e outros tipos de aço.

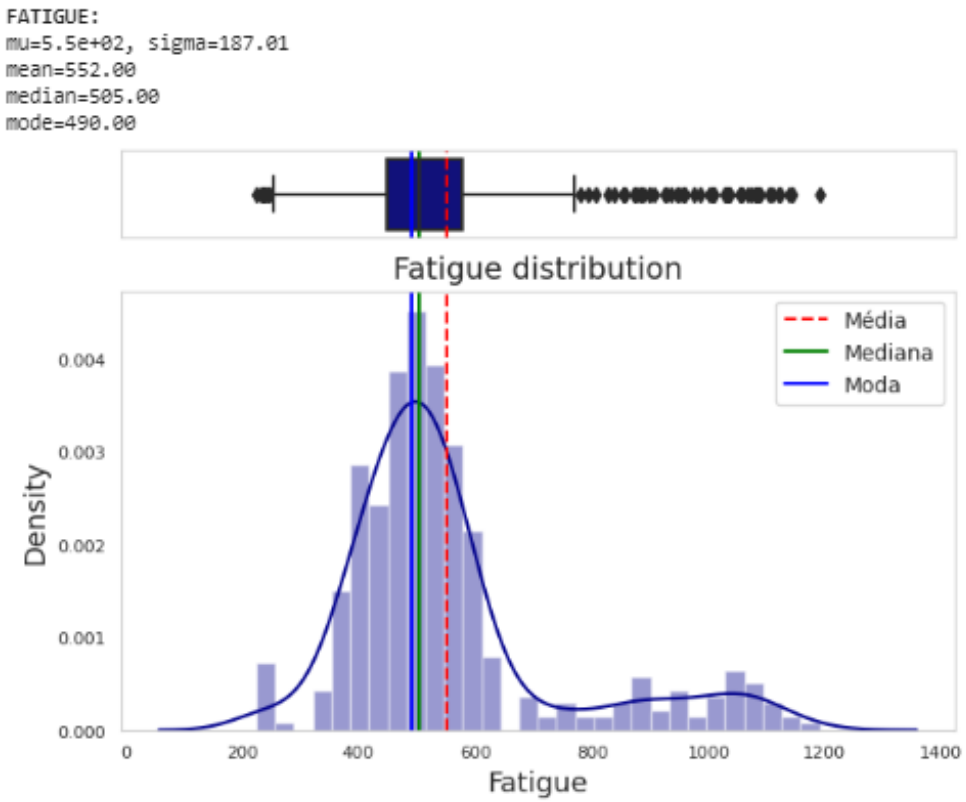
Para fins de simulação e comparação com as referências bibliográficas não será feito a separação por famílias de aços.

Figura 10 – Ccorrelação dos atributos remanescentes com a resistência à fadiga



Fonte: Elaborada pelo autor.

Figura 11 – Análise estatística da variável objetivo resistência à fadiga



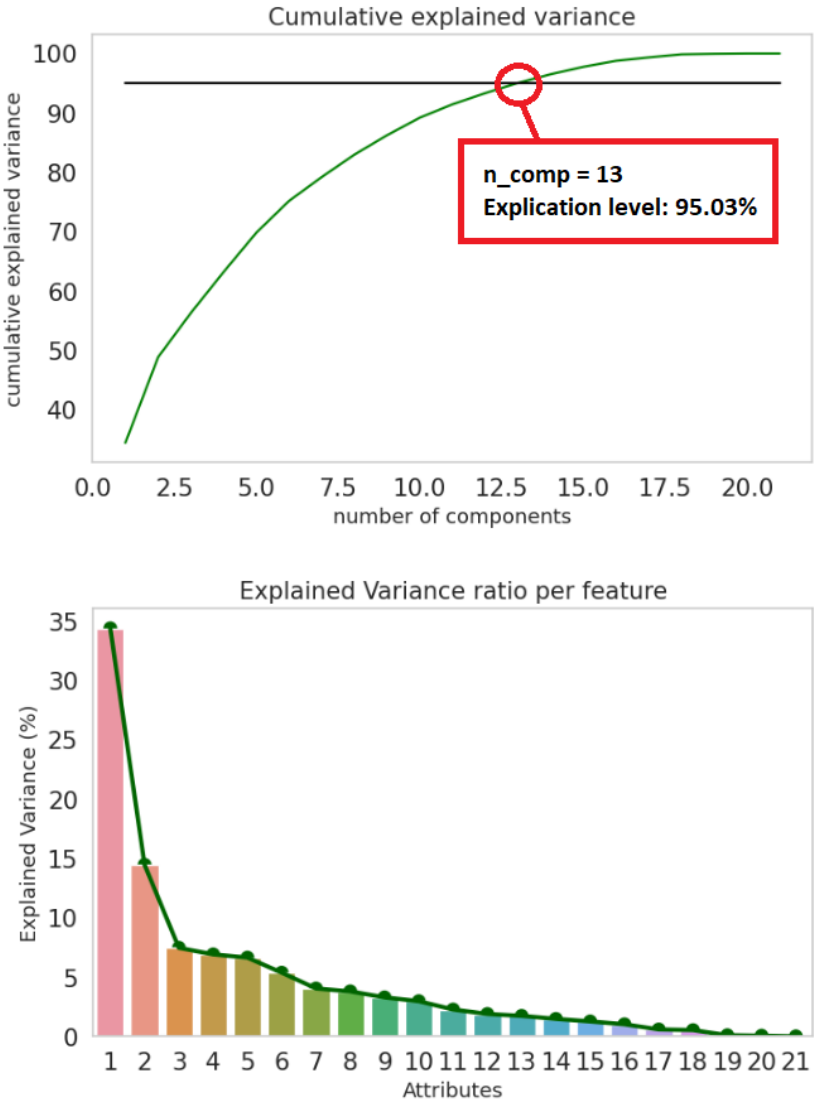
Fonte: Elaborada pelo autor.

4.2.3 Análise de Componentes Principais (PCA)

O PCA foi aplicado na base de dados revista após a análise de *outliers* e de alta correlação entres os atributos. Na Figura 12 pode-se notar que 13 componentes pincipais serão usadas para um nível de explicação acima de 95%, mostrando uma redução significativa da dimensionalidade do problema. Destaca-se que são 22 atributos restantes após a análise exploratória inicial.

Analisando a Figura 13 pode-se notar que o melhor valor do número de agrupamentos está definido para k = 3. Com o número de componentes principais definidos o próximo passo será uma análise de agrupamentos (*clustering*).

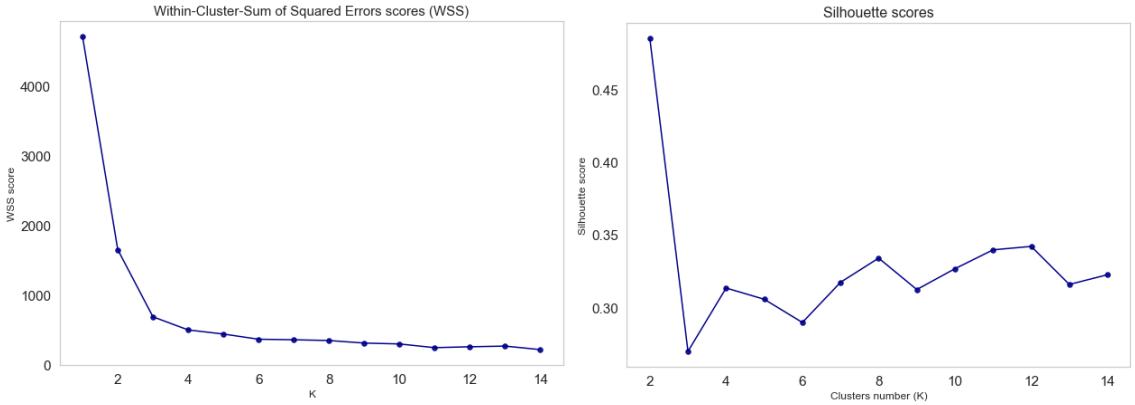
Figura 12 – Análise de componentes principais



Fonte: Elaborada pelo autor.

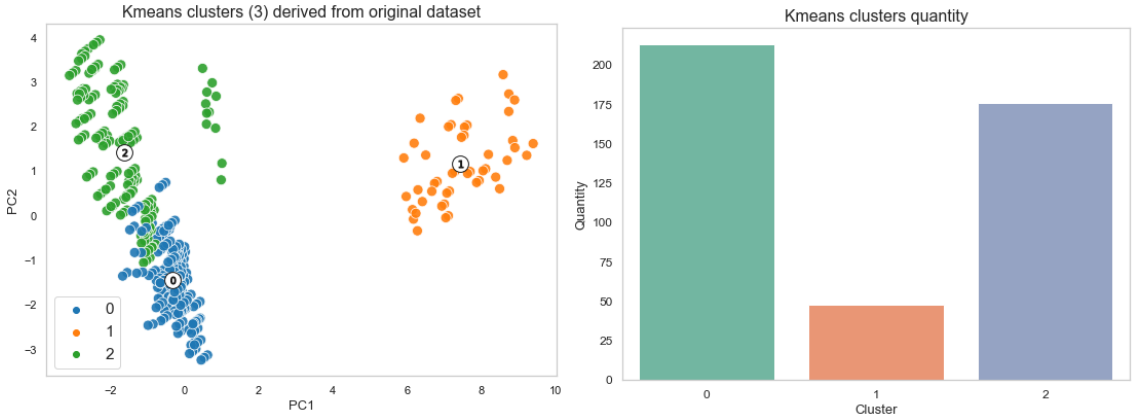
Utilizando o número de agrupamentos igual a 3 e usando as componentes do PCA como atributos tem-se a distribuição dos dados apresentada na Figura 14.

Figura 13 – *Within-Cluster Sum of Squared Errors* (WSS)



Fonte: Elaborada pelo autor.

Figura 14 – *Kmeans* - agrupamento das componentes principais



Fonte: Elaborada pelo autor.

Esses agrupamentos, no entanto, não oferecem pontos de dados suficientes para criar meta-modelos para cada agrupamento e, portanto, para todos os métodos usados, todo o conjunto de dados é usado para desenvolver modelos preditivos.

Além disso, por buscar uma interpretabilidade maior e mais direta dos atributos de entrada para predição da variável objetivo, optou-se por não usar o espaço de transformação dos atributos feito pela técnica PCA.

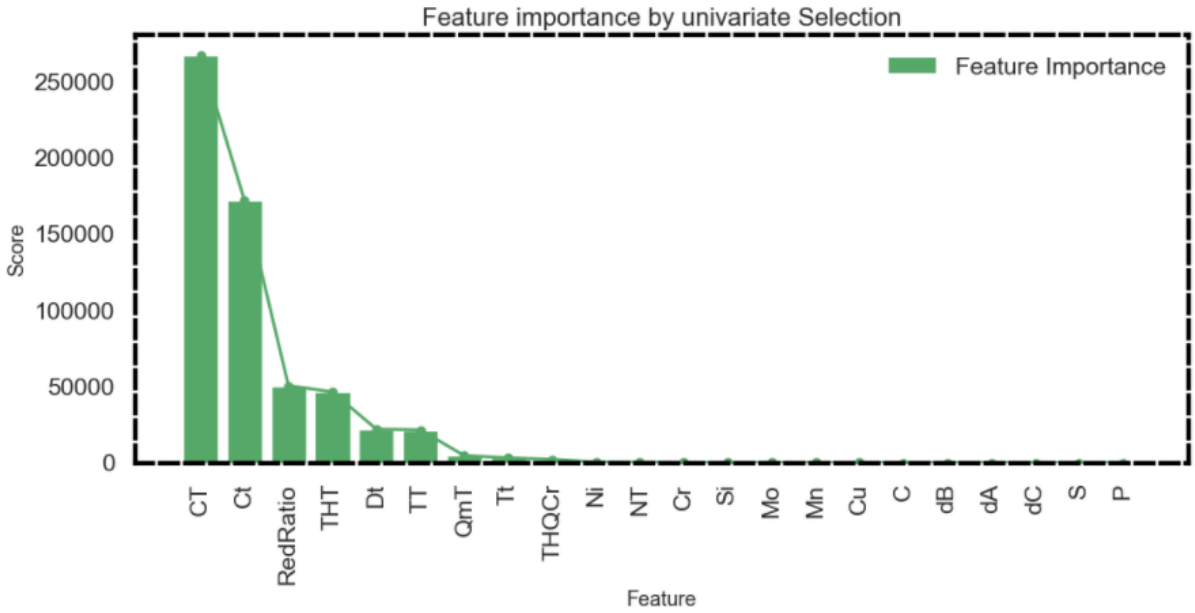
4.3 Seleção de Atributos

4.3.1 Seleção univariada de atributos

A seleção univariada de atributos funciona selecionando os melhores atributos com base em testes estatísticos univariados. Compara-se cada atributo com a variável objetivo, para ver se há alguma relação estatisticamente significativa entre eles, por isso sendo também chamada de análise de variância (ANOVA). Quando analisamos a relação entre um atributo e a variável objetivo, ignoramos os outros recursos. É por isso que é chamado de univariada, onde cada atributo individualmente tem sua pontuação de teste apurada. Finalmente todas as pontuações dos testes são comparadas e os atributos com as melhores pontuações são portanto selecionados.

Utilizando a técnica na base de dados NMIS podemos observar pelo Figura 15 que a dimensão do problema pode ser simplificada para apenas 6 atributos (CT, Ct, RedRatio, THT, Dt e TT). Note que nenhum elemento da composição química aparece como atributo de alta importância para descrever o problema da fadiga.

Figura 15 – Seleção univariada de atributos



Fonte: Elaborada pelo autor.

4.3.2 Eliminação recursiva de atributos (RFE)

Dado um estimador externo que atribui pesos aos atributos, o objetivo da eliminação recursiva de atributos (RFE) é selecioná-los considerando recursivamente conjuntos cada vez menores de atributos. Primeiro, o estimador é treinado no conjunto inicial e a importância de cada recurso é obtida. Em seguida, os atributos menos importantes são removidos do

conjunto atual, sendo que esse procedimento é repetido recursivamente no conjunto podado até que o número desejado de atributos a serem selecionados seja finalmente alcançado.

A biblioteca Scikit-learn tem uma função nativa onde é implementada a técnica RFE, cujos resultados são apresentados na Tabela 3. Nas simulações foi usado a função *StratifiedKFold* com *splits*=5, da mesma biblioteca.

Pode-se notar um R^2>94% com apenas 5 atributos selecionados: NT, CT, Dt, Cr e Tt. O melhor resultado, de R^2=98,21%, foi obtido com 14 atributos (NT, CT, Dt, Cr, Tt, QmT, Ct, TT, C, P, Mn, dA, THT e Mo)

Tabela 3 – Técnica RFE aplicada na seleção de atributos

N° atributos	**R^2**	**RMSE**	**MAE**	**MAPE**
1	0,8092	8647,47	71,10	12,24
2	0,8514	6733,80	61,25	10,59
3	0,8221	8061,37	66,77	11,09
4	0,8442	7060,29	63,47	10,82
5	0,9496	2283,64	38,34	7,02
6	0,9449	2496,30	39,78	7,36
7	0,9467	2413,86	39,05	7,05
8	0,9666	1513,18	26,78	4,45
9	0,9713	1302,04	23,56	3,64
10	0,9719	1273,97	22,69	3,46
11	0,9742	1169,13	20,95	3,20
12	0,9746	1151,96	21,60	3,34
13	0,9812	849,92	19,09	3,03
14	0,9821	810,51	19,68	3,20
15	0,9809	863,90	19,77	3,18
16	0,9806	881,54	20,24	3,24
17	0,9810	861,26	19,79	3,18
18	0,9797	918,15	20,78	3,35
19	0,9813	849,34	20,36	3,31
20	0,9803	892,17	19,99	3,21
21	0,9805	882,27	20,54	3,34

4.3.3 Seleção de atributos baseado em *Permutation Importance*

A técnica *Permutation Importance* está relacionada em como medir o impacto (redução) na pontuação de um modelo genérico quando um único recurso é embaralhado aleatoriamente. Este procedimento quebra a relação entre os atributos e a variável objetivo, onde a redução na pontuação do modelo é um indicativo de quanto o modelo depende de um determinado atributo. Esta técnica se beneficia por ser agnóstica de modelo de regressão ou classificação utilizado.

A biblioteca **eli5**, baseada na biblioteca Scikit-learn, tem a função *PermutationImportance* implantada e foi usada nesse estudo, cujos resultados são apresentados na Tabela 4. Nas simulações foi usado o regressor base *Random Forest* com parâmetros *max depth*=6 e *number of estimators*=10.

Tabela 4 – *Permutation Importance list*

Peso	Atributo
0,3580 ± 0,0568	Cr
0,0893 ± 0,0061	CT
0,0756 ± 0,0076	TT
0,0563 ± 0,0063	NT
0,0516 ± 0,0044	Ct
0,0472 ± 0,0075	QmT
0,0393 ± 0,0075	C
0,0297 ± 0,0099	P
0,0134 ± 0,0015	Tt
0,0125 ± 0,0012	Dt
0,0078 ± 0,0037	dA
0,0066 ± 0,0018	Mn
0,0028 ± 0,0010	THT
0,0012 ± 0,0005	Mo
0,0012 ± 0,0006	Si
0,0004 ± 0,0002	S
0,0004 ± 0,0002	Cu
0,0003 ± 0,0001	dC
0,0002 ± 0,0001	Ni
0,0001 ± 0,0001	dB
... 2 more ...	

Os valores na parte superior são os atributos mais importantes, e aqueles na parte inferior importam menos para a variável objetivo. O primeiro número em cada linha mostra o quanto o desempenho do modelo diminuiu com um embaralhamento aleatório (neste caso, usando a precisão como métrica de desempenho).

Ocasionalmente, pode-se observar valores negativos para importância de cada atributo. Nesses casos, as previsões sobre os dados embaralhados eram mais precisas do que os dados reais. Isso acontece quando o atributo realmente não importa (deveria ter uma importância próxima a 0), mas a chance aleatória faz com que as previsões nos dados embaralhados fiquem mais precisas. Considerando a base de dados NMIS os cinco atributos de maior impacto na variável objetivo são: Cr, CT, TT, NT e Ct.

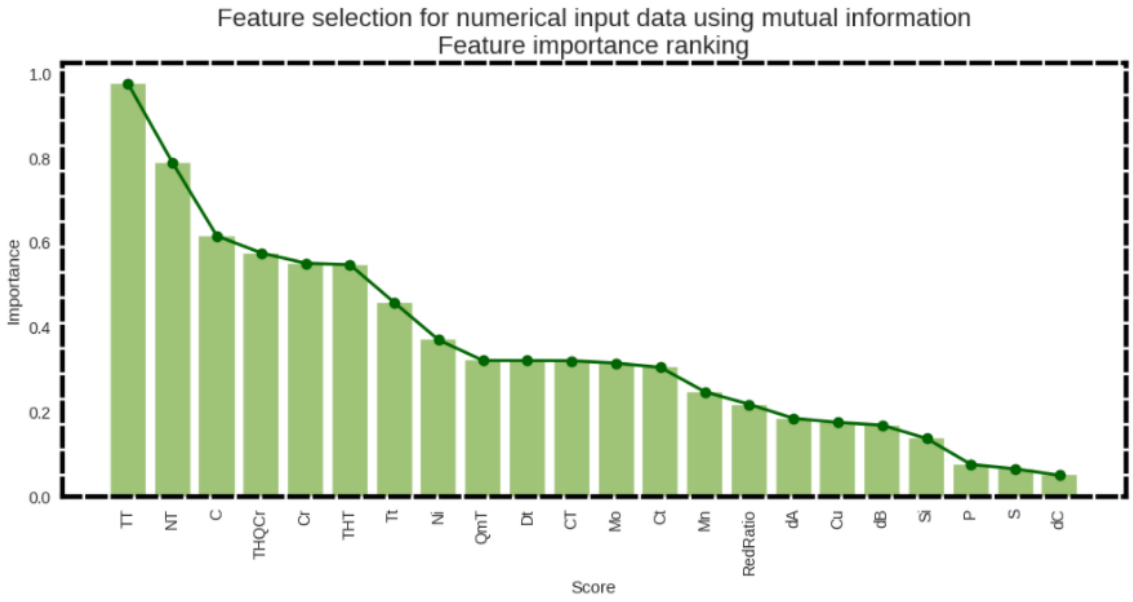
4.3.4 Seleção de atributos baseado em Informação Mútua *Mutual Information*

O conceito vem do campo da teoria da informação com a aplicação do conceito de ganho de informação (normalmente usado na construção de árvores de decisão) para a seleção de Atributos. A informação mútua é calculada entre duas variáveis e mede-se a redução na incerteza de uma variável dado um valor conhecido da outra variável.

A biblioteca scikit-learn fornece uma implementação de informações mútuas para seleção de atributos por meio da função *mutual_info_regression()*, onde a estratégia de seleção *SelectKBest* é a mais usual.

A resultado da seleção de atributos baseado em Informação Mútua sobre a base de dados NMIS pode ser visto na Figura 16 abaixo. Os cinco atributos de maior impacto na variável objetivo são: TT, NT, C, THQCr e Cr.

Figura 16 – Seleção de atributos baseado em Informação Mútua



Fonte: Elaborada pelo autor.

Tabela 5 – Seleção de atributos baseado em Informação Mútua

Atributo	**Score**
TT	0,975563
NT	0,789052
C	0,615859
THQCr	0,575693
Cr	0,551272
THT	0,547883
Tt	0,458242
Ni	0,369737
QmT	0,321213
Dt	0,321015
CT	0,320598
Mo	0,314944
Ct	0,304842
Mn	0,247303
RedRatio	0,217553
dA	0,184362
Cu	0,174754
dB	0,168136
Si	0,136778
P	0,075472
S	0,064796
dC	0,049516

4.3.5 Seleção de atributos em algoritmos baseados em *Boosting*

Um benefício de se usar algoritmos baseados em *boosting* está relacionado ao fato de que, depois que as árvores aumentadas (*boosted trees*) são construídas, é relativamente simples obter o valor de importância para cada atributo.

A medição da importância fornece uma pontuação que indica o quão útil ou valioso é cada atributo na construção das árvores de decisão do modelo. Quanto mais um atributo

é usado para tomar decisões importantes com árvores de decisão, maior a sua importância relativa. Essa importância é calculada explicitamente para cada atributo no conjunto de dados, permitindo que os atributos sejam classificados e comparados entre si.

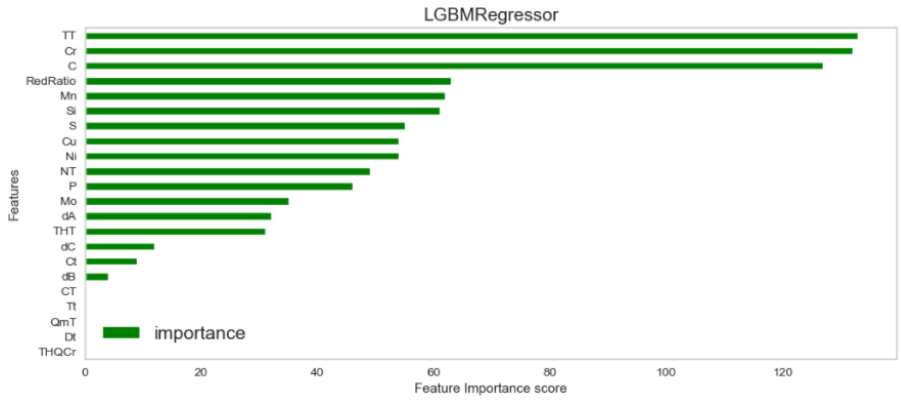
Para a base de dados NMIS foram estudados os seguintes regressores: *Light GBM* (LGB), *Extreme Gradient Boosting* (XGB), *Extra Tree* (ETR), *Decision Tree* (DTR), *Random Forest* (RFR), *Gradient Boosting* (GBR) e *ADA Boosting* (ADA), cujo resultado está sumarizado na Tabela 6.

Tabela 6 – Seleção de atributos em algoritmos baseados em *Boosting*

XGB	LGB	ETR	DTR	RFR	GBR	ADA
NT	TT	CT	Dt	NT	Tt	QmT
Cr	Cr	Tt	Cr	Tt	CT	Dt
Mo	C	NT	TT	Ct	NT	Cr
TT	RedRatio	Cr	NT	Cr	QmT	Ct
C	Mn	QmT	P	Dt	Cr	Tt
P	Si	TT	C	CT	Dt	CT
THT	S	Si	Mn	QmT	TT	C

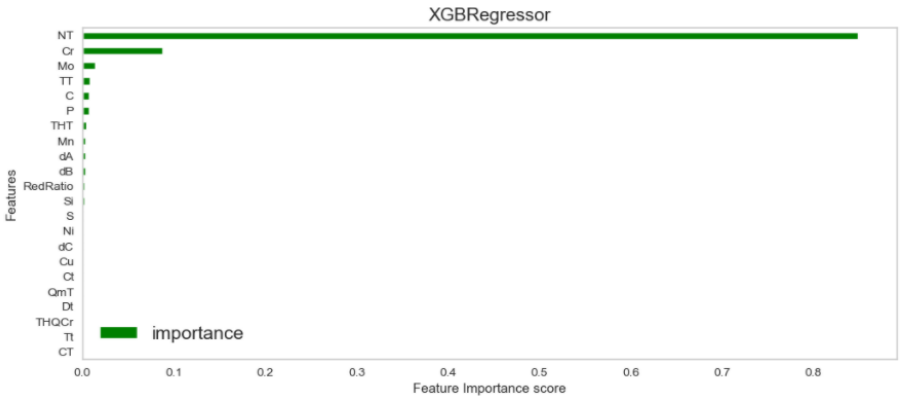
Nas Figuras 17, 18, 19, 20, 21, 22 e 23 a seguir pode-se visualizar de forma mais detalhada a análise de *Feature Importance* desses modelos.

Figura 17 – *Feature importance* usando regressor *LightGBM*



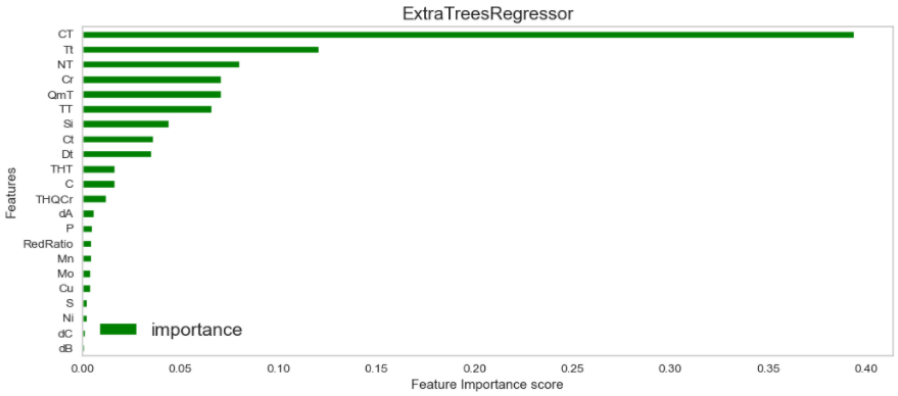
Fonte: Elaborada pelo autor.

Figura 18 – *Feature importance* usando regressor XGB



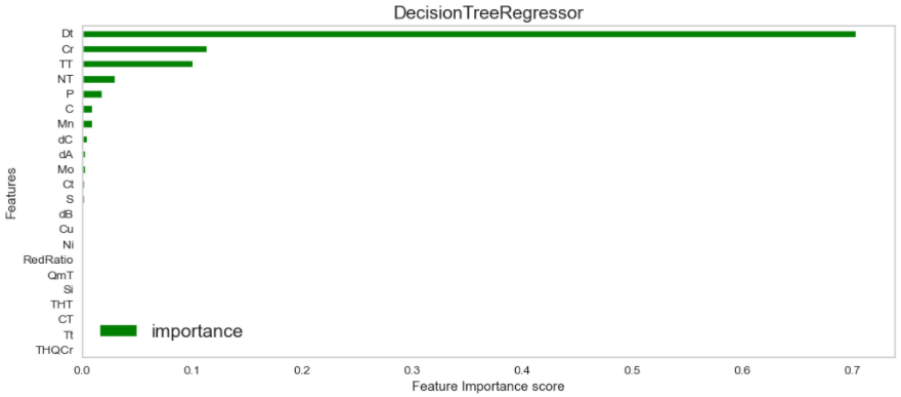
Fonte: Elaborada pelo autor.

Figura 19 – *Feature importance* usando regressor *ExtraTree*



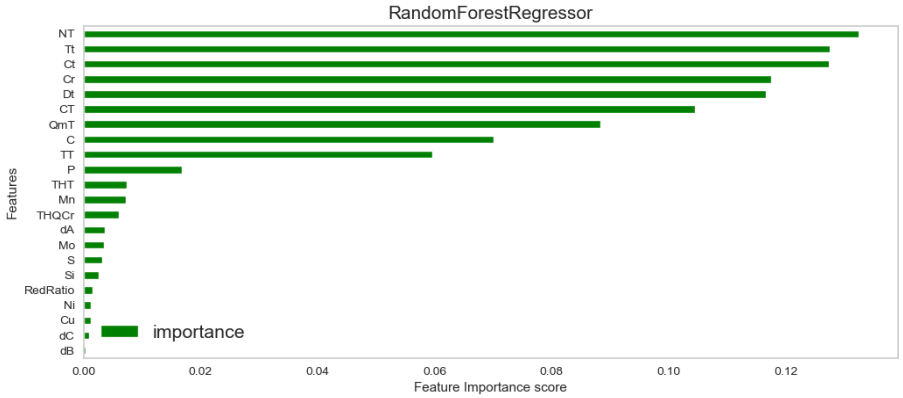
Fonte: Elaborada pelo autor.

Figura 20 – *Feature importance* usando regressor *Decision Tree*



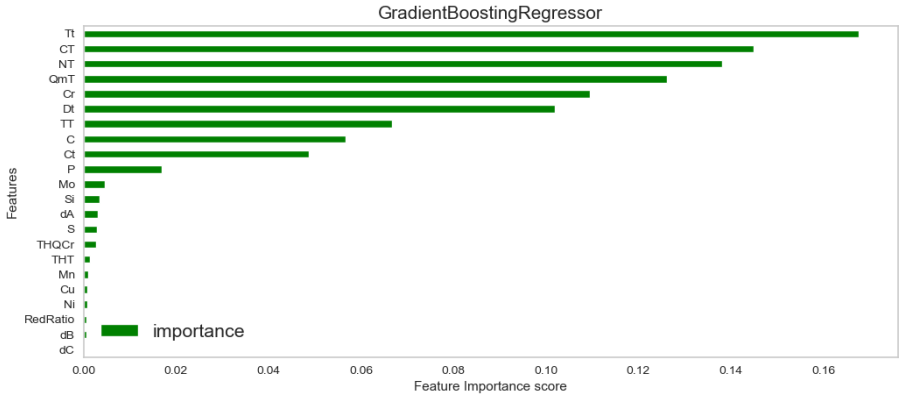
Fonte: Elaborada pelo autor.

Figura 21 – *Feature importance* usando regressor *Random Forest*



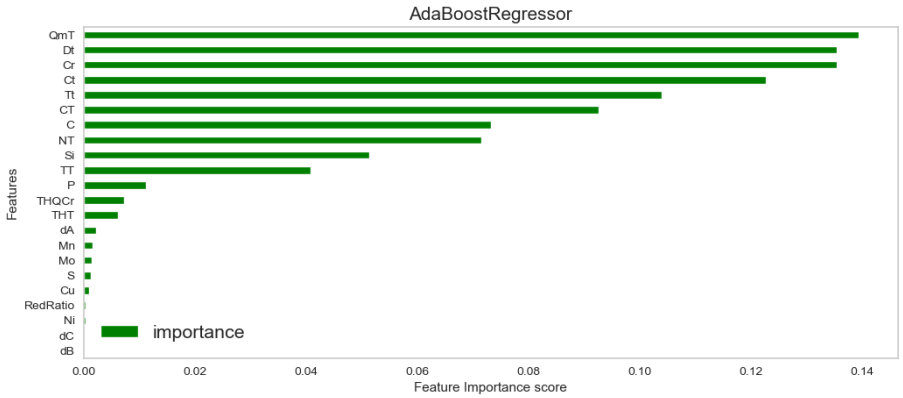
Fonte: Elaborada pelo autor.

Figura 22 – *Feature importance* usando regressor *Gradient Boosting*



Fonte: Elaborada pelo autor.

Figura 23 – *Feature importance* usando regressor *Adaptive Boosting*



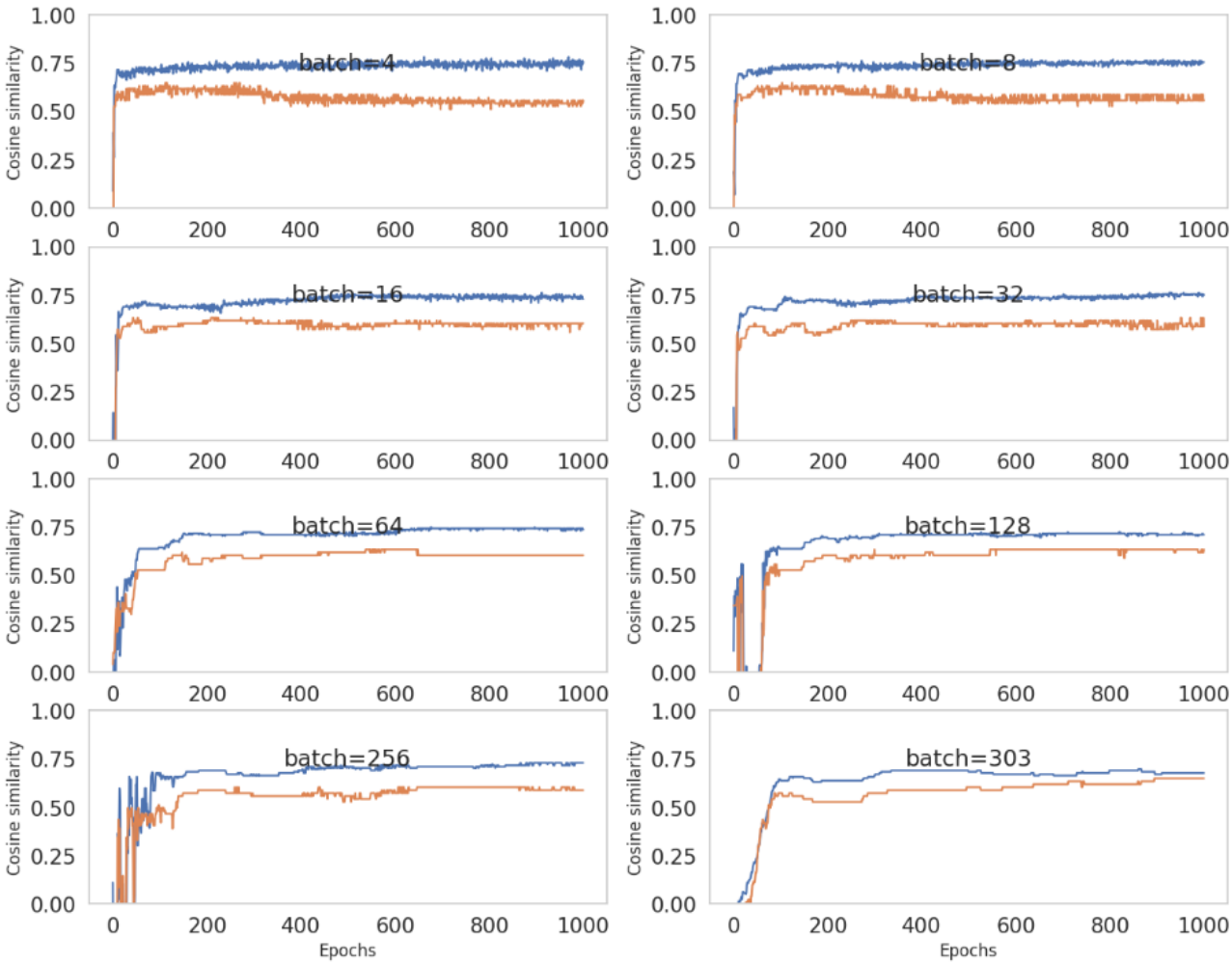
Fonte: Elaborada pelo autor.

4.4 Redes neurais profundas

4.4.1 Análise de diferentes valores de *batch size*

A partir de valores pre-estabelecidos de *batch size*, o resultado com 1.000 iterações e com métrica a similaridade cosseno entre base de teste e valores preditos é apresentado na Figura 24

Figura 24 – Simulação com diferente valores de *batch size*



Fonte: Elaborada pelo autor.

na Tabela 4.4.1 os indicadores de desempenho R^2, MSE, MAE e MAPE são apresentados calculados para cada valor de *batch size*. Pode-se notar que o parâmetro *batch size*=16 apresentou o melhor resultado com R^2=0,93 e MSE=0,05.

Tabela 7 – Análise do parâmetro *Batch size* utilizado KERAS

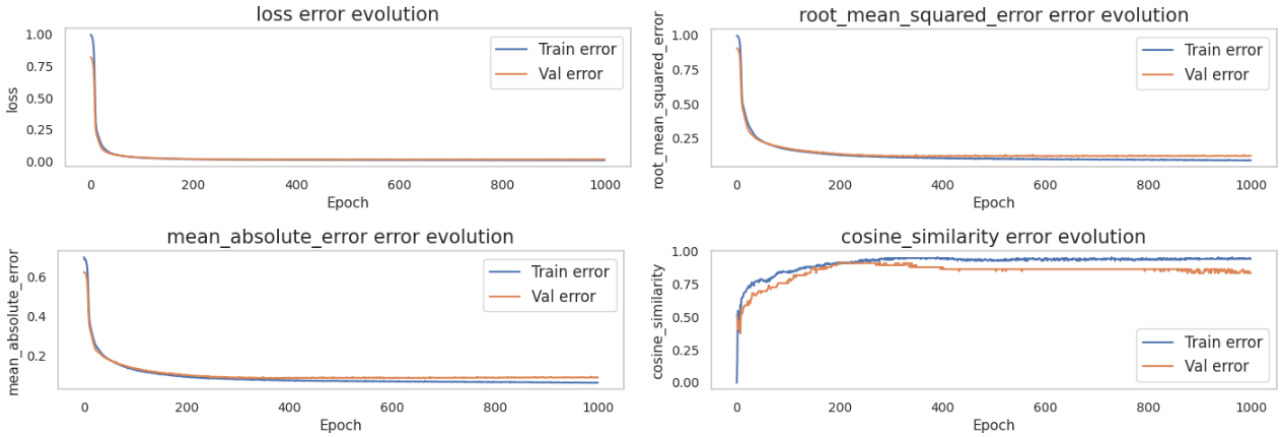
Simulação	**R2**	**MSE**	**MAE**	**MAPE**
batch=4	0,93	0,06	0,20	183,13
batch=8	0,93	0,06	0,19	162,38
batch=16	0,93	0,05	0,18	137,31
batch=32	0,93	0,06	0,19	170,14
batch=64	0,93	0,06	0,19	154,68
batch=128	0,93	0,06	0,19	143,69
batch=256	0,88	0,10	0,24	143,40
batch=303	0,79	0,17	0,32	274,44

Os resultados foram obtidos utilizando a seguinte topologia de rede: camada de entrada densa com 8 neurônios, camada oculta densa com 24 neurônios e camada de saída densa com apenas 1 neurônio, todas com função de ativação *Relu*

Utilizando o parâmetro *batch size*=16 os resultados detalhados são apresentados na Figura 25. As métricas observadas são *loss*, *mean absolute error*, *root mean squared error* e *cosine similarity*.

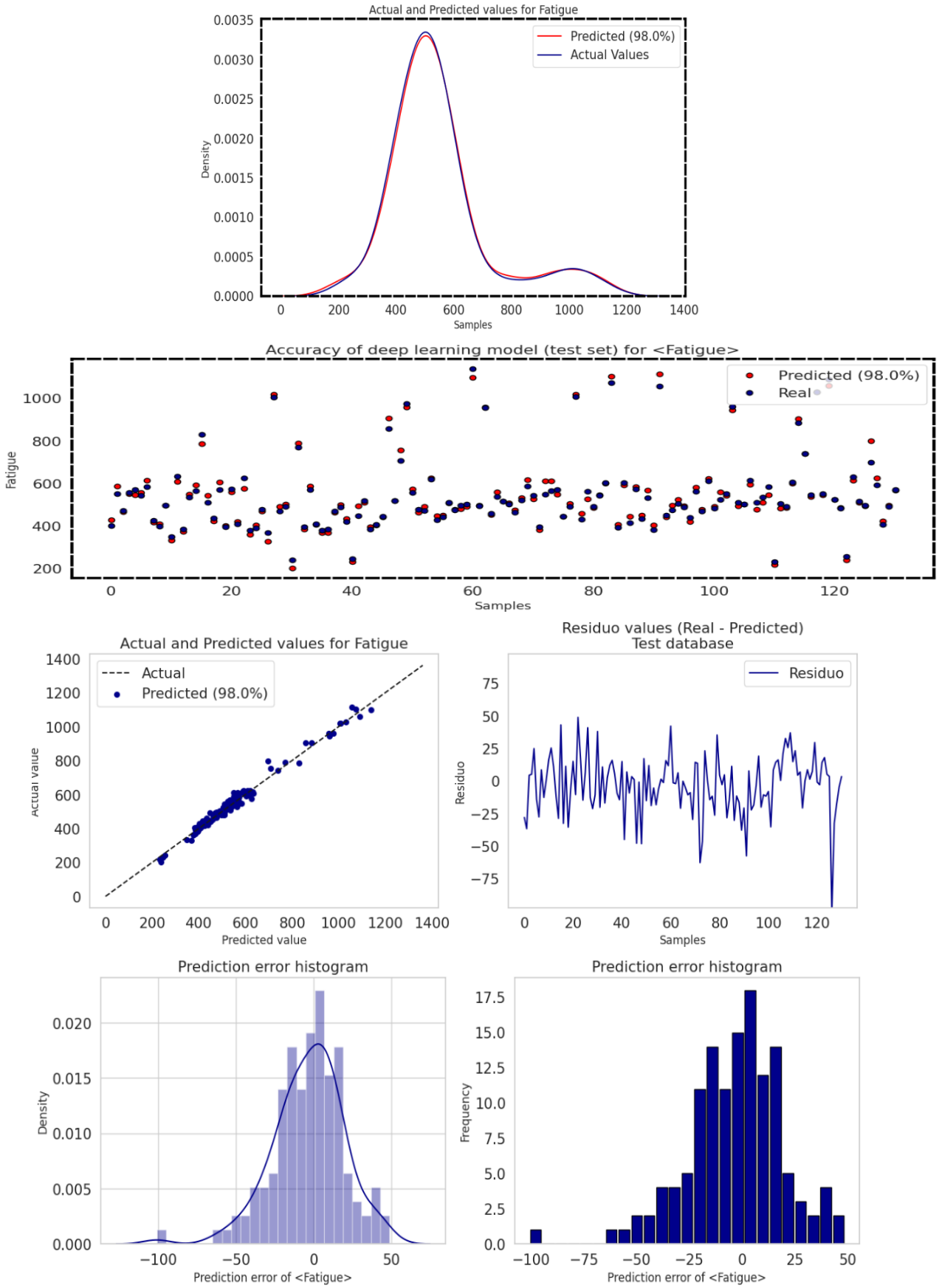
A predição comsiderando os dados de teste fornece R2 = 0,98243, superior às referências bibliográficas usadas nesse trabalho, apesar da dificuldade clássica da rede neural ser considerada um modelo caixa-preta. Na Figura 26 uma análise estatística mais detalhada da predição pode ser visualizada

Figura 25 – Simulação com *batch size*=16



Fonte: Elaborada pelo autor.

Figura 26 – Predição da resistência à fadiga com *batch size*=16



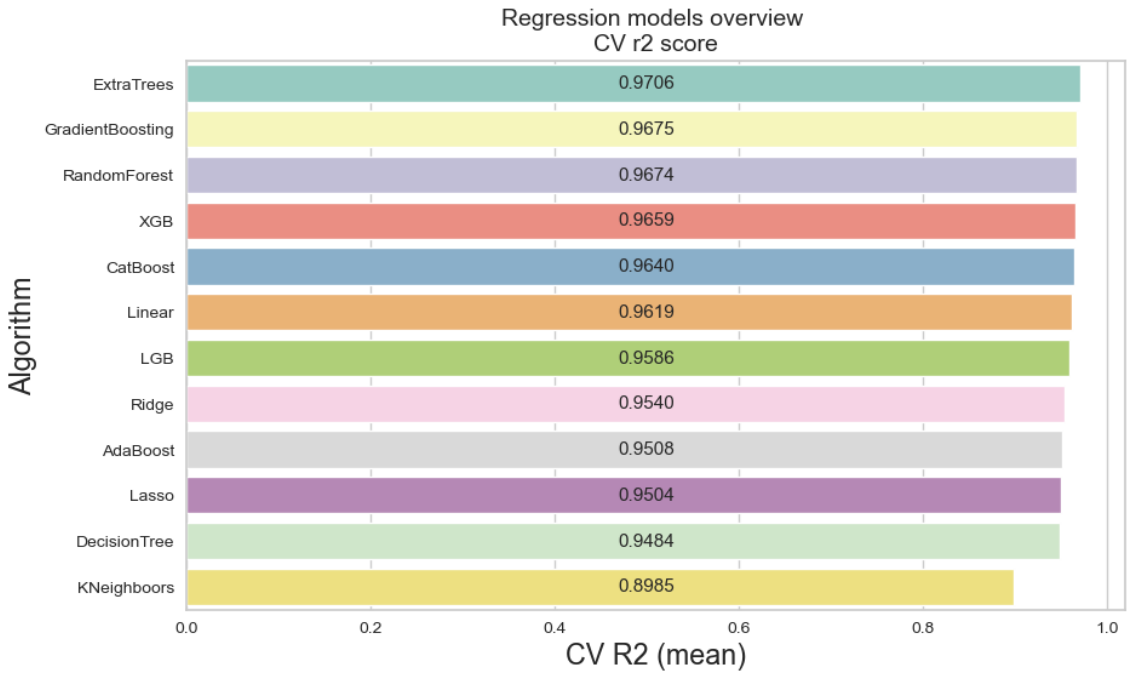
Fonte: Elaborada pelo autor.

4.5 Visão geral dos regressores via biblioteca *scikit-learn*

Utilizando a biblioteca *scikit-learn* do Python inicialmente tem-se uma visão geral do desempenho dos seguintes regressores: regressão linear, *Ridge*, *Lasso*, *KNeighbors*, *Random Forest*, *Gradient Boosting*, *Decision Tree*, *AdaBoost*, *CatBoost*, *Extra Trees*, *XGB* e *LGB*.

Nas simulações foi aplicado o processo de cross-validação com 5 *folders*, sem realizar nenhum ajuste dos seus respectivos parâmetros (*hypertuning*). Os resultados podem ser vizualidados na Figura 27.

Figura 27 – Desempenho dos regressores via *scikit-learn*

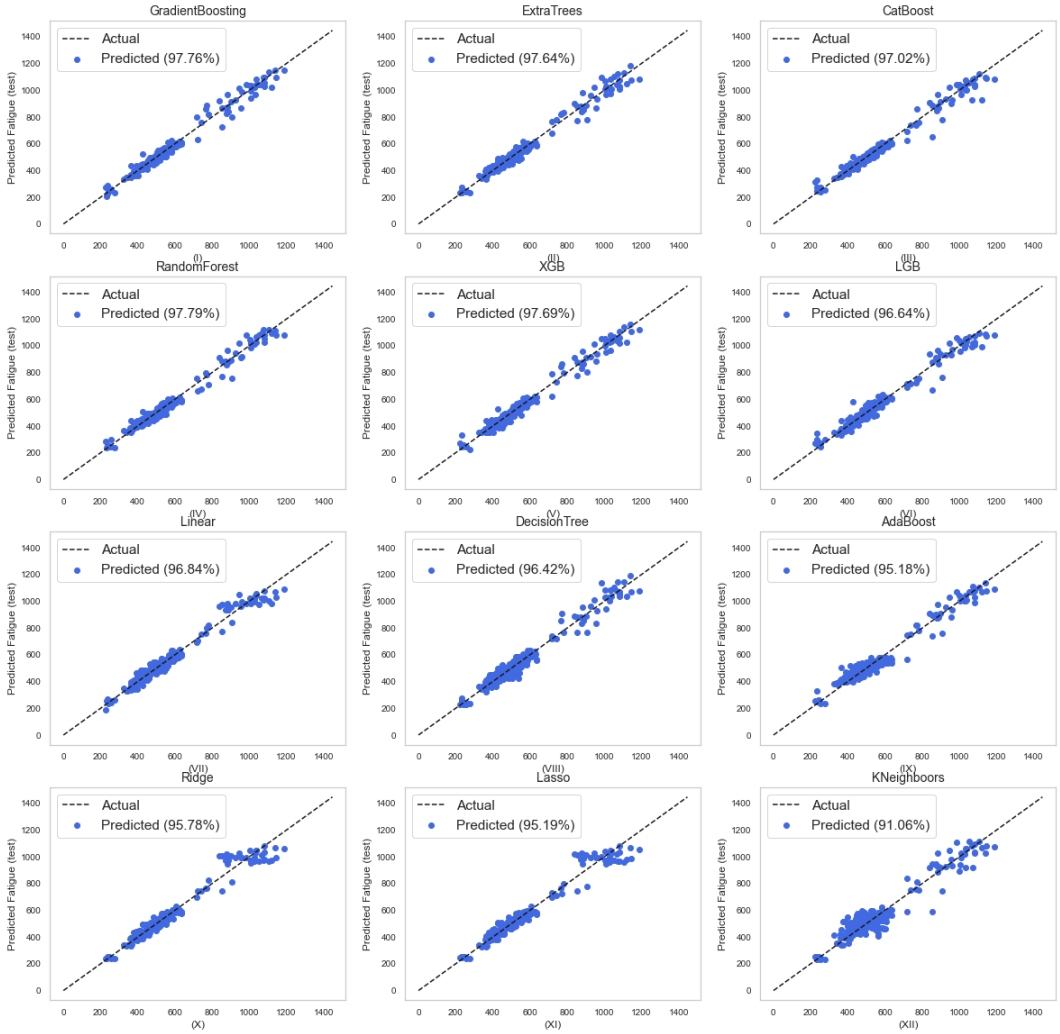


Fonte: Elaborada pelo autor.

XGB, *ExtraTrees*, *Random Forest* e regressores baseados em *Boosting* conseguem ajustar o conjunto de dados de treinamento com uma precisão superior à 0,96. Assim, esses regressores seriam um bom começo para ajustar o conjunto de dados de resistência à fadiga com precisão razoável.

A Figura 28 mostra a correlação entre os valores reais e predito para oa modelos de aprendizado de máquina inicialmente estudados para os conjuntos de treinamento e teste. Os eixos X e Y denotam os valores atuais e preditos para variável resistência à Fadiga (em MPa) respectivamente. (I) *Gradient Boosting*, (II) *ExtraTrees*, (III) *CatBoost*, (IV) *Random Forest*, (V) *XGB*, (VI) *LGB*, (VII) regressão linear, (VIII) *DecisionTree*, (IX) *ADABoost*, (X) regressão *RIDGE*, (XI) regressão *LASSO* e (XII) *KNeighboors*.

Figura 28 – Correlação entre valores reais e preditos (dados de treinamento)



Fonte: Elaborada pelo autor.

O melhor regressor base, em termos de desempenho no ajuste dos dados, foi o *ExtraTrees*. Uma visão geral do desempenho de todos os regressores base pode ser observado na tabela 8.

Tabela 8 – Visão geral dos regressores base

Modelo (tipo)	**R^2 treino**	**R^2 teste**	**MAE treino**	**MAE teste**	**RMSE treino**	**RMSE teste**
EtraTrees	1,00	0,99	0,00	0,03	0,00	24,42
CatBoost	1,00	0,98	0,01	0,03	5,11	29,46
Random Forest	1,00	0,98	0,01	0,03	12,07	29,38
Gradient Boosting	0,99	0,98	0,02	0,03	13,00	30,46
XGB	0,99	0,98	0,02	0,03	14,46	30,41
LGBM	0,99	0,98	0,02	0,04	19,54	30,02
Linear	0,97	0,97	0,04	0,04	30,90	37,51
Decision Tree	1,00	0,97	0,00	0,04	0,00	38,05
Ridge	0,96	0,96	0,04	0,05	34,94	42,64
Lasso	0,96	0,95	0,05	0,05	36,94	45,66
KNeighbors	0,98	0,95	0,03	0,05	24,17	47,09
AdaBoost	0,93	0,95	0,07	0,07	47,38	47,86

Para efeitos de comparação vamos utilizar um regressor baseado em votação (*voting regressor*). Trata-se de um metaestimador de conjunto que se ajusta a vários regressores de referência, cada um aplicado individualmente ao conjunto de dados de treinamento. O resultado é apresentado na tabela 9.

O regressor baseado em votação consegue ajustar o conjunto de dados de treinamento com precisão de 0,99 com redução marginal para 0,98 no conjunto de dados de teste, representando um resultado muito bom.

Tabela 9 – Regressor de votação (*voting regressor)*

Modelo (tipo)	R^2 treino	R^2 teste	MAE treino	MAE teste	RMSE treino	RMSE teste
Voting	0,99	0,98	0,02	0,03	15,14	28,34

Na próxima seção vamos utilizar o pacote **PyCaret** para os modelos de regressão disponívei na biblioteca.

4.6 PyCaret

Para o setup dos parâmetros gerais do modelo PyCaret os valores estão apresentados na Tabela 10.

Tabela 10 – Parâmetros gerais do modelo PyCaret

Parâmetro	**Valor**
train_size	0,70
fold_shuffle	*True*
normalize	*True*
normalize_method	*zscore*
feature_selection	*True*
feature_selection_threshold	0,4
numeric_imputation	*'mean'*
categorical_imputation	*'constant'*
imputation_type	*'iterative'*
ignore_low_variance	*True*

PyCaret versão 2,3,4

Após a inicialização dos parâmetros gerais todos os modelos de regressão disponíveis no pcatore PyCaret foram inicialmente treinados, cujos resultados ordenados pelo indicador R^2 podem ser visualizados na Tabela 11. O número de *folders* da validação cruzada pode ser definido pelo usuário, nas simulações foi usado o valor padrão de 10.

Tabela 11 – Desempenho dos regressores base

Id	**Nome**	**MAE**	**MSE**	**RMSE**	**R^2**	**RMSLE**	**MAPE**	**TP (Sec)**
catboost	*CatBoost*	18,44	903,50	28,80	0,9702	0,0487	0,0337	0,660
rf	*Random Forest*	20,82	876,73	29,12	0,9665	0,0502	0,0379	0,455
et	*Extra Trees*	21,12	898,96	29,26	0,9680	0,0483	0,0373	0,408
gbr	*Gradient Boosting*	20,50	926,97	29,67	0,9678	0,0495	0,0365	0,062
xgboost	*Extreme Gradient Boosting*	20,37	1028,29	31,05	0,9654	0,0510	0,0355	1,318
lightgbm	*Light Gradient Boosting Machine*	24,04	1308,81	34,79	0,9553	0,0628	0,0459	0,030
dt	*Decision Tree*	28,73	1579,23	39,01	0,9436	0,0671	0,0516	0,016
ada	*AdaBoost*	31,46	1620,57	39,95	0,9321	0,0730	0,0601	0,088
lr	*Linear*	28,38	1669,72	40,18	0,9424	0,0614	0,0489	0,015
br	*Bayesian Ridge*	28,50	1678,12	40,23	0,9425	0,0613	0,0490	0,017
ridge	*Ridge*	28,98	1722,49	40,67	0,9415	0,0622	0,0500	0,016
huber	*Huber*	27,19	1830,18	40,96	0,9416	0,0590	0,0453	0,031
lasso	*Lasso*	29,43	1820,69	41,55	0,9397	0,0635	0,0509	0,017
par	*Passive Aggressive*	30,43	2285,16	45,32	0,9301	0,0653	0,0512	0,018
knn	*K Neighbors*	39,03	2549,07	49,44	0,9062	0,0858	0,0711	0,060
llar	*Lasso Least Angle*	47,37	3965,66	61,69	0,8515	0,1073	0,0887	0,016
en	*Elastic Net*	54,44	5457,09	72,75	0,7978	0,1369	0,1081	0,016
omp	*Orthogonal Matching Pursuit*	71,24	8805,68	92,64	0,6131	0,2613	0,1482	0,015

O modelo ***Catboost*** teve o melhor desempenho inicial com R^2 igual a 0,9702 sem nenhum tipo de ajuste fino nos parâmetros do modelo. Vamos agora realizar a etapa de *hypertuning* para os cinco melhores modelos em termos de R^2, a saber: *Catboost*, *Randon Forest*, *Extra Trees*, *Gradient Boosting* e *Extreme Gradient Boosting*.

Proximo passo foi realizar o ajuste dos hiperparâmetros de cada modelo individualmente. Particularmente no Pycaret a função de ajuste (*tune*) é baseado no conjunto de treinamento enquanto que a função de predição (*predict*) utiliza a base de teste como *default*. Os resultados são apresentados na Tabela 12

Tabela 12 – Regressores ajustados (*hypertunning*)

Id	Nome	Tipo	MAE	MSE	RMSE	R^2	RMSLE	MAPE
catboost	Catboost	tuned	19,78	927,59	29,48	0,9684	0,0493	0,0359
		predicted	17,65	587,83	24,25	0,9808	0,0410	0,0316
rf	*Random Forest*	*tuned*	23,76	1220,67	34,26	0,9529	0,0586	0,0437
		predicted	18,87	605,18	24,60	0,9802	0,0458	0,0348
et	*Extra Trees*	*tuned*	27,43	1903,98	41,55	0,9332	0,0683	0,0498
		predicted	21,20	779,71	27,92	0,9745	0,0485	0,0378
gbr	*Gradient Boosting*	*tuned*	18,94	777,17	27,11	0,9740	0,0449	0,0335
		predicted	18,93	716,67	26,77	0,9765	0,0424	0,0328
xgboost	*Extreme Gradient Boosting*	*tuned*	19,24	866,21	28,33	0,9719	0,0467	0,0341
		predicted	17,26	652,22	25,54	0,9787	0,0413	0,0304
lightgbm	*Light Gradient Boosting Machine*	*tuned*	21,68	1006,27	30,85	0,9650	0,0544	0,0401
		predicted	20,22	730,32	27,02	0,9761	0,0435	0,0347

Somente regressores base com R2>95%, validação cruzada com folders=10

O PyCaret também permite realizar análises adicionais como modelamento tipo *ensembled* para cada modelo de aprendizado de máquina estudado (*bagging* ou *boosting*) ou até mesmo combinar modelos em um regressor baseado em voto (*blending*). Os resultados podem ser vistos nas tabelas 13 e 14

Tabela 13 – *Ensemble* e *Blend* dos modelos (base de treinamento)

Nome	Tipo	MAE	MSE	RMSE	R^2	RMSLE	MAPE
Ensembled Random Forest	boosting	22,59	998,81	30,99	0,9646	0,0525	0,0412
Ensembled Extra Trees	boosting	23,67	1095,99	32,47	0,9618	0,0548	0,0438
Ensembled Gradient Boost	bagging	21,10	1032,93	31,34	0,9630	0,0547	0,0390
	boosting	22,61	1082,80	32,07	0,9615	0,0566	0,0425
Ensembled Extreme Gradient Boosting	bagging	19,01	898,80	28,92	0,9708	0,0459	0,0336
	boosting	22,14	999,62	30,92	0,9647	0,0543	0,0415
Ensembled CatBoost	bagging	20,84	1025,66	31,04	0,9658	0,0514	0,0380
	boosting	23,59	1139,58	32,90	0,9593	0,0594	0,0452
Blended		18,81	838,71	28,04	0,9710	0,0465	0,0342

Somente regressores base com R^2>95%, validação cruzada com *folders*=10

Tabela 14 – *Ensemble* e *Blend* dos modelos (base de teste)

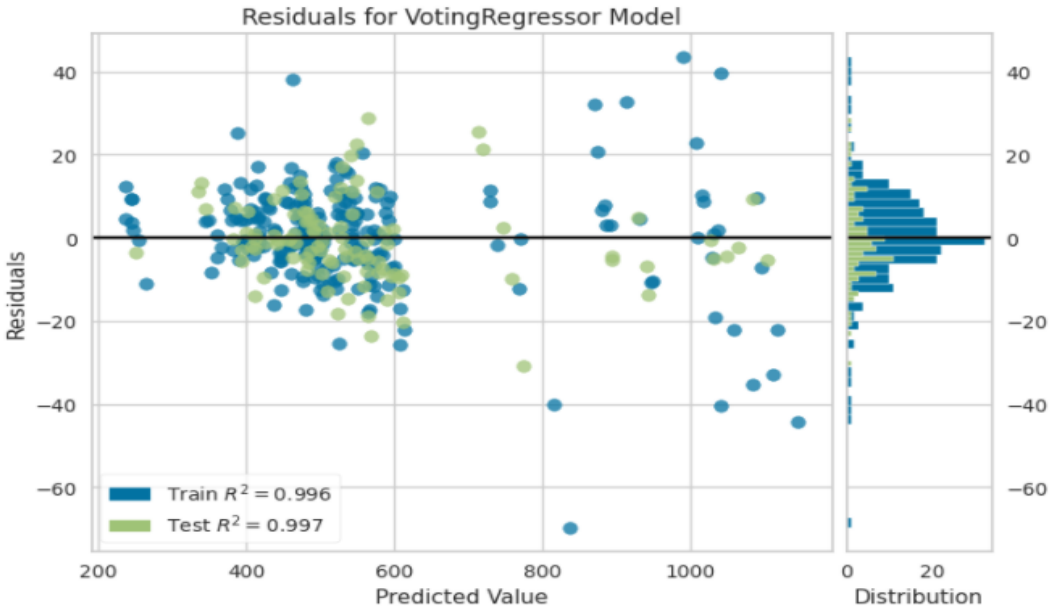
Nome	Tipo	MAE	MSE	RMSE	R^2	RMSLE	MAPE
Ensembled Random Forest	bagging	22,83	907,75	30,13	0,9703	0,0543	0,0418
	boosting	23,30	935,90	30,59	0,9694	0,0532	0,0421
Ensembled Extra Trees	bagging	29,44	1748,46	41,81	0,9428	0,0706	0,0516
	boosting	23,03	868,35	29,47	0,9716	0,0528	0,0423
Ensembled Gradient Boost	bagging	19,45	691,10	26,22	0,9774	0,0444	0,0348
	boosting	21,39	770,99	27,77	0,9748	0,0470	0,0378
Ensembled Extreme Gradient Boosting	bagging	16,74	554,79	23,55	0,9818	0,0406	0,0300
	boosting	20,47	726,05	26,95	0,9762	0,0461	0,0364
Ensembled CatBoost	bagging	17,48	559,56	23,67	0,9817	0,0404	0,0310
	boosting	22,67	885,82	29,76	0,9710	0,0516	0,0411
Blended		15,40	467,85	21,63	0,9847	0,0377	0,0279

Somente regressores base com R^2>95%, validação cruzada com *folders*=10

O regressor baseado em voto (*Blended*) apresentou o melhor desempenho com R^2 = 0,9847, superior aos resultados individuais de cada modelo ajustado bem como superior às propostas de agrupamento *Ensemble*.

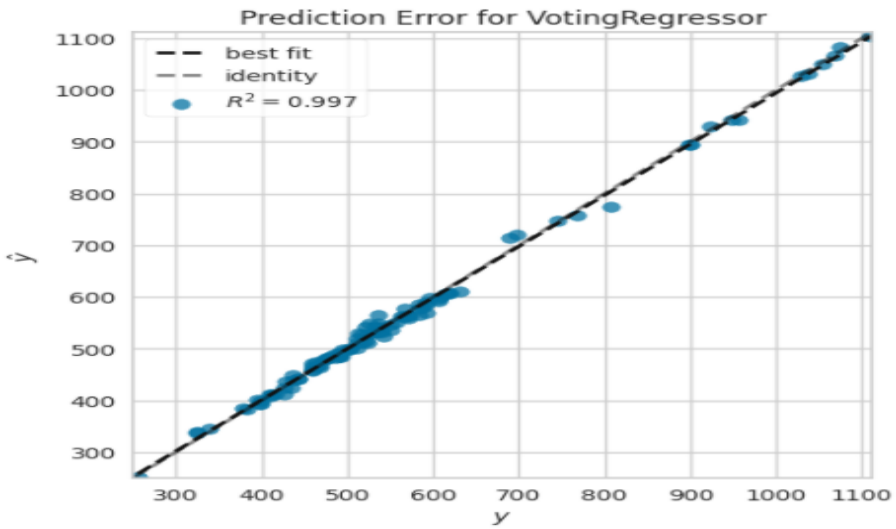
A finalização do modelo é a última etapa recomendada ao usar o *framework* do Pycaret. Um fluxo usual de tarefas iniciaria com setup do PyCaret, seguido pela comparação inicial de todos os modelos de regressão disponíveis e identificando alguns potenciais modelos candidatos com base nas métricas disponíveis.

Figura 29 – **Visualização dos resíduos do regressor baseado em voto (*Blended*)**



Fonte: Elaborada pelo autor.

Figura 30 – **Predição do regressor baseado em voto (*Blended*)**

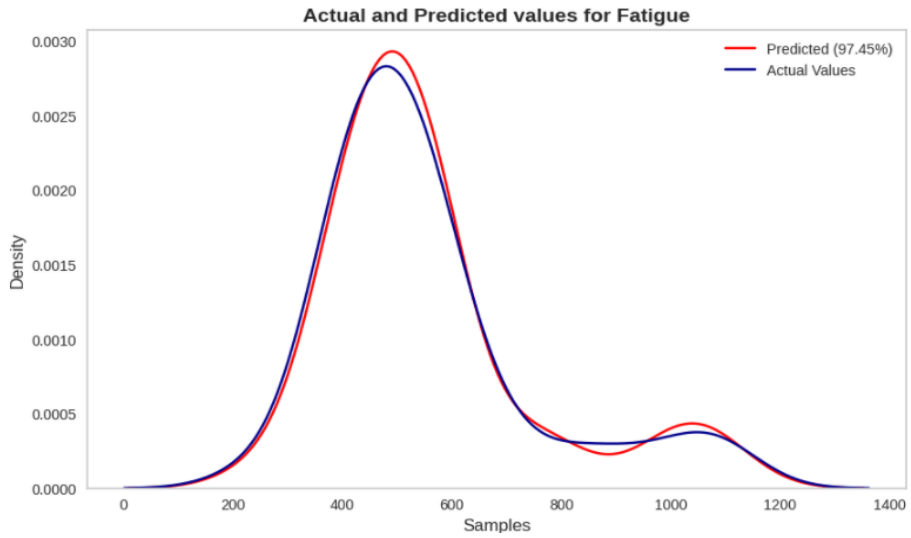


Fonte: Elaborada pelo autor.

A função *finalize_model()* do PyCaret tem o objetivo de ajustar o modelo ao conjunto completo de dados (treino e teste). Para o regressor escolhido (*Blended*), o resultado pode ser visto nas Figuras 29 e 30.

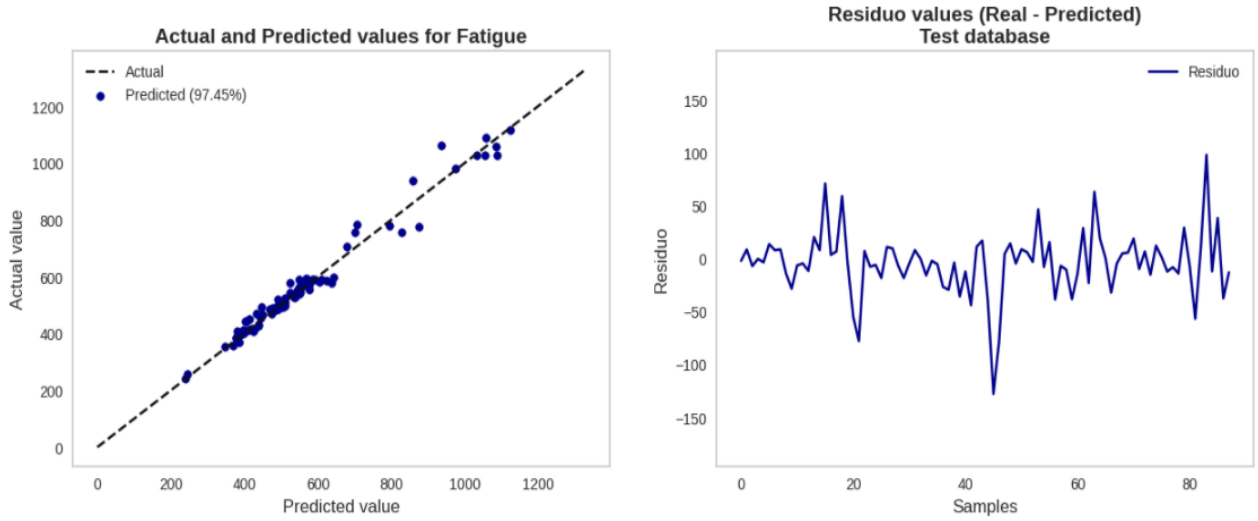
Uma vez definido qual o melhor modelo a ser adotado, vamos utilizar 20% da base de dados que foi originalmente separada de forma aleatória para ser destinada somente à dados desconhecidos ao modelo. O regressoe baseado em voto apresentou $R^2 = 0{,}9750$, resultado muito bom considerando que são dados completamente desconhecidos ao modelo. Uma visualização mais gráfica desse resultado é apresentada nas Figuras 31 e 32.

Figura 31 – **Análise do melhor modelo escolhido (*Blended*)**



Fonte: Elaborada pelo autor.

Figura 32 – **Análise do melhor modelo escolhido (*Blended*)**



Fonte: Elaborada pelo autor.

5 CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

Durante a revisão bibliográfica desse trabalho pode-se notar que é praticamente um consenso entre os pesquisadores de mineração de dados sobre modelagem preditiva: é mais útil saber sobre um conjunto de técnicas com bom desempenho de predição para um determinado problema, em vez de do que identificar um única técnica vencedora. Usando essa premissa examinou-se mais de 12 diferentes técnicas para modelagem preditiva de resistência à fadiga em aços especiais.

Usualmente o processo de desenvolvimento de novas ligas de aços especiais é demorado e bastante caro, além do fato que diversos fatores externos pode afetar à correlação entre composição química, variáveis de processo (analisando sob a ótica da escala industrial de produção) e as propriedades mecânicas como fadiga.

Nesse sentido, utilizar aprendizado de máquina para gerar possíveis correlações entre química, processamento, estrutura e propriedades será de grande interesse para os engenheiros de materias de forma a aumentar a velocidade do processo de design, acelerando o *time-to-market* no desenvolvimento de novas ligas de aços especiais.

Nesse estudo, por questões de confidencialidade e necessidade de autorização por parte da alta administração da Aperam, foi utilizado a base de dados pública NMIS para estudar a correlação entre composição química, processamento industrial e resistência à fadiga de diferentes de tipos de aços, onde obteve-se um nível de confiança acima de 95% em diversos modelos testados, em concordância com as diversas referências bibliográficas estudadas como (GAUTHAM et al., 2011), (DESHPANDE et al., 2013), (DOBRZANSKI; KOWALSKI; MADEJSKI, 2005) e (AGRAWAL et al., 2014). É muito encorajador ver que apesar a quantidade limitada de dados disponíveis neste conjunto de dados, os modelos analíticos baseados em dados foram capazes de atingir um grau alto de precisão.

O estudo também buscou compreender a influência dos vários atributos na resistência à fadiga, utilizando diversas técnicas de determinação de importância de atributos. O resultado obtido demostra que o uso dessas técnicas podem dar uma orientação inicial aos engenheiros de materiais no projeto de novas ligas e aplicações, podendo ser validado com a abordagem fisico-metalurgica clássica. Ao ser buscar fazer uma seleção de atributos deve-se levar em consideração a opinião dos engenheiros metalurgistas especializados. Como *insights* gerados tems-se:

- Os atributos de parâmetros de processo relacionadas às temperaturas de fase (NT, THT, CT, QmT e TT) são mais importantes para o comportamento da resistência à Fadiga do que os atributos relacionados à tempo (THQCr, Ct e DT).

- A composição química do aço é um fator fundamental na determinação da resistência à Fadiga. devendo ser consideradas como variáveis importantes durante a seleção dos atributos.

- Diversas simulações de cenários de seleção de atributos devem ser realizadas, a partir da análise de correlação simples, de forma a determinar, juntamente com as técnicas de machine learning de seleção de atributos, qual deverá ser o melhor conjunto a ser considerado.

Para os modelos baseados em boosting estudandos pode-se notar a frequência com que cada uma das 7 variáveis mais importantes aparecem: Cr (7), NT (5), Tt (4), TT (3), QmT (3), Dt (3) e C (2), o que demostra que existem duas variáveis relacionadas à composição química, três relacionadas à temperaturas de processo e duas relacioadas à tempos de processo, o que pode ser explicado pelas famílias de aços adotados na composição da base de dados NMIS. Por hora utilizou-se a base de dados quase completa (atributos DT, THt, TCr foram desconsiderados por apresentar correlação acima de 95% com outros atributos).

Importante ressaltar que, apesar dos bons resultados obtidos, O banco de dados NIMS referente à fadiga de aços especiaos não fornece informações complementares importantes nos parâmetros estruturais, como fração de fase, tamanho de grão, fração de precipitado, etc. Vamos discutir a seguir algumas das limitações do conjunto de dados NIMS

Uma vez que o volume de dados usado neste estudo é pequeno se comparado com as quantidades típicas de dados usados em estudos de mineração de dados, a busca por modelos cada vez mais precisos pode ser vista como um incentivo para usar mais dados (possivelmente combinar mais dados de diferentes fontes) para melhor validar os resultados e eventualmente tornar o modelo ainda mais robusto.

Outra limitação da base NMIS é o número significativamente diferente de dados correspondentes aos diferentes tipos de aços. Isso faz com que os modelos preditivos, que por sua vez foram desenvolvidos sobre todos os dados podem não ser altamente precisos para todos os tipos de aço considerados. Recomenda-se aumentar o volume de dados considerados para cada tipo de aço da base de dados.

Como recomendações para trabalhos futuros pode-se considerar:

- Por se tratar de natureza não linear inerente dos dados da base de dados de resistência a fadiga recomenda-se explorar outros métodos não lineares de *machine learning*.

- Os modelos estudados podem ser usado para construir previsões para outras propriedades mecânicas importates como % alongamento, resistência à tensão aplicada,

resistência à fratura ou dureza, cujos dados também estão disponível no conjunto de dados NIMS.

- Aprofundar nas análises relacionadas à seleção de atributos, em especial se a base de dados NMIS pude ser extendida para mais registros por família de aço bem como incluir novas famílias de forma a ter um modelo de predição mais generalizado.
- Explorar as diversas técnicas de *feature importance* disponíveis na literatura e já implementadas nos diversos pacotes Python, simulando com o conjunto mais restrito de atributos e comparando com os resultados obtidos nesse trabalho.

REFERÊNCIAS

ABDI, H.; WILLIAMS, L. J. Principal component analysis. **WIREs Computational Statistics**, v. 2, n. 4, p. 433–459, 2010. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.101>.

AGRAWAL, A.; CHOUDHARY, A. A fatigue strength predictor for steels using ensemble data mining: Steel fatigue strength predictor. p. 2497–2500, 2016. Disponível em: <https://doi.org/10.1145/2983323.2983343>.

AGRAWAL, A. et al. Exploration of data science techniques to predict fatigue strength of steel from composition and processing parameters. **Integrating Materials and Manufacturing Innovation**, Springer, v. 3, n. 1, p. 90–108, 2014. ISSN 2193-9772. Disponível em: <https://doi.org/10.1186/2193-9772-3-8>.

ALPAYDIN, E. **Introduction to Machine Learning (Adaptive Computation and Machine Learning)**. [S.l.]: The MIT Press, 2004. ISBN 0262012111.

BHATTAD, C. R. **Building a fatigue strength predictor for steel using an ensemble deep learning model**. 2019. 1–65 p. Dissertação (Mestrado) — DEPARTMENT OF METALLURGICAL AND MATERIALS ENGINEERING, INDIAN INSTITUTE OF TECHNOLOGY MADRAS, 2019.

BIGRIVERSTEEL. **BigRiverSteel**. 2021. Disponível em: <https://bigriversteel.com/>. Acesso em: 17 julho 2021.

BIOENERGIA, A. **Aperam Bioenergia - Fornos FAP 2000**. 2022. Disponível em: <https://aperambioenergia.com.br/aperam-bioenergia-investe-em-inovacao-e-sustentabilidade/>. Acesso em: 05 janeiro 2022.

BISHOP, C. M. **Pattern Recognition and Machine Learning (Information Science and Statistics)**. Berlin, Heidelberg: Springer-Verlag, 2006. ISBN 0387310738.

BREIMAN, L. Machine learning, volume 45, number 1 - springerlink. **Machine Learning**, v. 45, p. 5–32, 2001. Disponível em: <https://doi.org/10.1023/A:1010933404324>.

_____. Random forests. **Machine Learning**, v. 45, p. 5–32, 2001. Disponível em: <https://doi.org/10.1023/A:1010950718922>.

BREWKA, G. Artificial intelligence—a modern approach by stuart russell and peter norvig. **The Knowledge Engineering Review**, Prentice Hall. Series in Artificial Intelligence, Englewood Cliffs, NJ., Cambridge University Press, v. 11, n. 1, p. 78–79, 1996.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. Association for Computing Machinery, 2016. (KDD '16), p. 785–794. ISBN 9781450342322. Disponível em: <https://doi.org/10.1145/2939672.2939785>.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, v. 20, p. 273–297, 1995. ISSN 1573-0565. Disponível em: <https://doi.org/10.1007/BF00994018>.

Deepmind Technologies. **AlphaGo - The story so far**. 2017. Disponível em: <https://deepmind.com/research/case-studies/alphago-the-story-so-far>. Acesso em: 17 julho 2021.

DESHPANDE, P. et al. Application of statistical and machine learning techniques for correlating properties to composition and manufacturing processes of steels. In: _____. [S.l.: s.n.], 2013. p. 155–160. ISBN 978-3-319-48585-0.

DESU, R. K. et al. Mechanical properties of austenitic stainless steel 304l and 316l at elevated temperatures. **Journal of Materials Research and Technology**, v. 5, n. 1, p. 13–20, 2016. ISSN 2238-7854. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2238785415000605>.

DIEZ-OLIVAN, A. et al. Data fusion and machine learning for industrial prognosis: Trends and perspectives towards industry 4.0. **Information Fusion**, Elsevier, v. 50, p. 92–111, 2019.

DOBRZANSKI, L.; KOWALSKI, M.; MADEJSKI, J. Methodology of the mechanical properties prediction for the metallurgical products from the engineering steels using the artificial intelligence methods. **Journal of Materials Processing Technology**, v. 164-165, p. 1500–1509, 2005.

FENG, W.; YANG, S. Thermomechanical processing optimization for 304 austenitic stainless steel using artificial neural network and genetic algorithm. **Applied Physics A**, Springer Berlin Heidelberg, Berlin/Heidelberg, v. 122, n. 12, p. 1–10, 2016. ISSN 0947-8396.

FRAGASSA, C. et al. Predicting the tensile behaviour of cast alloys by a pattern recognition analysis on experimental data. **Metals**, v. 9, n. 5, 2019. ISSN 2075-4701. Disponível em: <https://www.mdpi.com/2075-4701/9/5/557>.

FREUND, Y.; SCHAPIRE, R. E. A desicion-theoretic generalization of on-line learning and an application to boosting. In: **Computational Learning Theory**. Berlin, Heidelberg: Springer Berlin Heidelberg, 1995. p. 23–37. ISBN 978-3-540-49195-8.

FUJII, H.; MACKAY, D. J. C.; BHADESHIA, H. K. D. H. Bayesian neural network analysis of fatigue crack growth rate in nickel base superalloys. **ISIJ International**, v. 36, n. 11, p. 1373–1382, 1996.

GAUTHAM, B. et al. More efficient icme through materials informatics and process modeling. **Proceedings of the 1st World Congress on Integrated Computational Materials Engineering (ICME)**, p. 35–42, 09 2011. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118147726.ch5>.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. [S.l.]: MIT Press, 2016. <http://www.deeplearningbook.org>.

GORNI, A. **A Siderurgia e a Indústria 4.0**. 2021. Disponível em: <https://www.industrialheating.com/articles/94718-a-siderurgia-e-a-indústria-40>. Acesso em: 17 julho 2021.

GUO, Z.; SHA, W. Modelling the correlation between processing parameters and properties of maraging steels using artificial neural network. **Computational Materials Science**, v. 29, n. 1, p. 12–28, 2004. ISSN 0927-0256. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0927025603000922>.

GUYON, I. et al. Gene selection for cancer classification using support vector machines. **Mach. Learn.**, Kluwer Academic Publishers, USA, v. 46, n. 1–3, p. 389–422, 2002. ISSN 0885-6125. Disponível em: <https://doi.org/10.1023/A:1012487302797>.

HACKERS, D. **Data Science para Executivos**. 2021. Disponível em: <https://medium.com/data-hackers/data-science-para-executivos-c2bbbbeed333>. Acesso em: 17 julho 2021.

HODGSON, P.; KONG, L.; DAVIES, C. The prediction of the hot strength in steels with an integrated phenomenological and artificial neural network model. **Journal of Materials Processing Technology**, v. 87, n. 1, p. 131–138, 1999. ISSN 0924-0136. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0924013698003446>.

JIAO, P.; ALAVI, A. H. Artificial intelligence-enabled smart mechanical meta-materials: advent and future trends. **International Materials Reviews**, Taylor & Francis, v. 66, n. 6, p. 365–393, 2021. Disponível em: <https://doi.org/10.1080/09506608.2020.1815394>.

JONES, D. M.; WATTON, J.; BROWN, K. J. Comparison of hot rolled steel mechanical property prediction models using linear multiple regression, non-linear multiple regression and non-linear artificial neural networks. **Ironmaking & Steelmaking**, Taylor & Francis, v. 32, n. 5, p. 435–442, 2005. Disponível em: <https://doi.org/10.1179/174328105X48151>.

KAELBLING, L.; LITTMAN, M.; MOORE, A. Reinforcement learning: A survey. **Journal of Artificial Intelligence Research**, p. 237–285, 1996. Disponível em: <https://arxiv.org/abs/cs/9605103>.

KE, G. et al. Lightgbm: A highly efficient gradient boosting decision tree. In: GUYON, I. et al. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2017. v. 30. Disponível em: <https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf>.

LINDSTRöM, J. et al. Towards intelligent and sustainable production systems with a zero-defect manufacturing approach in an industry4.0 context. **Procedia CIRP**, v. 81, p. 880–885, 01 2019.

MANDAL, S. et al. Artificial neural network modeling to evaluate and predict the deformation behavior of stainless steel type aisi 304l during hot torsion. **Applied Soft Computing**, v. 9, n. 1, p. 237–244, 2009. ISSN 1568-4946. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1568494608000604>.

MCKINSEY. **How a steel plant in India tapped the value of data—and won global acclaim**. 2021. Disponível em: <https://www.mckinsey.com/industries/metals-and-mining/how-we-help-clients/how-a-steel-plant-in-india-tapped-the-value-of-data-and-won-global-acclaim>. Acesso em: 17 julho 2021.

MEHTA, M.; AGRAWAL, R.; RISSANEN, J. Sliq: A fast scalable classifier for data mining. In: **EDBT**. IBM Almaden Research Center, 650 Harry Road, San Jose, CA: [s.n.], 1996. p. 18–32.

MYLLYKOSKI, P.; LARKIOLA, J.; NYLANDER, J. Development of prediction model for mechanical properties of batch annealed thin steel strip by using artificial neural network modelling. **Journal of Materials Processing Technology**, v. 60, n. 1, p. 399–404, 1996. ISSN 0924-0136. Proceedings of the 6th International Conference on Metal Forming. Disponível em: <https://www.sciencedirect.com/science/article/pii/0924013696023618>.

NARAYANA, P. et al. Modeling high-temperature mechanical properties of austenitic stainless steels by neural networks. **Computational Materials Science**, v. 179, p. 109617, 2020. ISSN 0927-0256. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0927025620301087>.

NARAYANA, P. L. et al. Modeling mechanical properties of 25cr-20ni-0.4c steels over a wide range of temperatures by neural networks. **Metals**, v. 10, n. 2, 2020. ISSN 2075-4701. Disponível em: <https://www.mdpi.com/2075-4701/10/2/256>.

PALLA, S. et al. Artificial neural network modeling of the tensile properties of indigenously developed 15cr-15ni-2.2mo-ti modified austenitic stainless steel. **Transactions of the Indian Institute of Metals**, v. 59, 08 2006.

PATEL, S. V.; JOKHAKAR, V. N. A random forest based machine learning approach for mild steel defect diagnosis. In: **2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)**. [S.l.: s.n.], 2016.

QUINLAN, J. R. Induction of decision trees. **MACH. LEARN**, v. 1, p. 81–106, 1986.

RAJAN, K.; SUH, C.; MENDEZ, P. Principal component analysis and dimensional analysis as materials informatics tools to reduce dimensionality in materials science and engineering. **Stat. Anal. Data Min.**, v. 1, p. 361–371, 2009. Disponível em: <https://doi.org/10.1002/sam.10031>.

SANTOS, B. P. et al. Industry 4.0: Challenges and opportunities. **Revista Produção e Desenvolvimento**, v. 4, n. 1, p. 111–124, Mar. 2018. Disponível em: <https://revistas.cefet-rj.br/index.php/producaoedesenvolvimento/article/view/e316>.

SINGH, S. B. et al. Neural network analysis of steel plate processing. **Ironmaking & Steelmaking**, v. 25, p. 355–365, 1998.

SOURMAIL, T.; BHADESHIA, H. K. D. H.; MACKAY, D. J. C. Neural network model of creep strength of austenitic stainless steels. **Materials Science and Technology**, Taylor Francis, v. 18, n. 6, p. 655–663, 2002. Disponível em: <https://doi.org/10.1179/026708302225002065>.

STOLL, A.; BENNER, P. Machine learning for material characterization with an application for predicting mechanical properties. **GAMM-Mitteilungen**, v. 44, 03 2021. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/gamm.202100003>.

SUTTON, R. S.; BARTO, A. G. **Reinforcement learning: An introduction**. [S.l.]: Cambridge, Massachusetts, London, England : The MIT Press, 2018. 552 p. ISBN 9780262039246.

TAKAHASHI H. J. E TEIXEIRA, R. Aplicação de técnicas de inteligência computacional para predição de propriedades mecânicas de aços de alta resistência microligados. **Tecnologia em metalurgia e materiais, Sao Paulo**, v. 5, n. 2, p. 100–104, 2008. Disponível em: <https://doi.org/10.4322/tmm.00502007>.

TOLLE, K.; TANSLEY, S.; HEY, T. The fourth paradigm: Data-intensive scientific discovery [point of view]. **Proceedings of the IEEE**, v. 99, p. 1334–1337, 08 2011.

WANG, Y. et al. Prediction and analysis of tensile properties of austenitic stainless steel using artificial neural network. **Metals**, v. 10, n. 2, 2020. ISSN 2075-4701. Disponível em: <https://www.mdpi.com/2075-4701/10/2/234>.

WEN, Y. et al. Corrosion rate prediction of 3c steel under different seawater environment by using support vector regression. **Corrosion Science**, v. 51, n. 2, p. 349–355, 2009. ISSN 0010-938X. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0010938X08004721>.

WU, Y.; YAN, Y.; LV, Z. Novel prediction model for steel mechanical properties with msvr based on mic and complex network clustering. **Metals**, v. 11, n. 5, 2021. ISSN 2075-4701. Disponível em: <https://www.mdpi.com/2075-4701/11/5/747>.

YAN, X.; YAN, X.; SU, X. G. **Linear regression analysis: theory and computing.** [S.l.]: World Scientific Publishing Co. Pte. Ltd, 2009. ISBN 9789812834102.

ZOU, H. et al. Multi-class adaboost. **Statistics and its Interface**, International Press of Boston, Inc., v. 2, p. 349–360, 2009. ISSN 1938-7989.

5