UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA E ENGENHARIA DE MATERIAIS

ESTUDO DA VIABILIDADE DA DETERMINAÇÃO DA DISSOLUÇÃO EM MEIO AQUOSO DE COMPOSTOS ÓXIDOS UTILIZANDO A FERRAMENTA OPEN SOURCE ORANGE DATA MINING

Armando José de Sá Santos

Orientador: Dr. Marcello Rubens Barsi Andreeta

São Carlos-SP
2024

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA E ENGENHARIA DE MATERIAIS

ESTUDO DA VIABILIDADE DA DETERMINAÇÃO DA DISSOLUÇÃO EM MEIO AQUOSO DE COMPOSTOS ÓXIDOS UTILIZANDO A FERRAMENTA OPEN SOURCE ORANGE DATA MINING

Armando José de Sá Santos

Dissertação apresentada ao Programa de Pós-Graduação em Ciência e Engenharia de Materiais como requisito parcial à obtenção do título de MESTRE EM CIÊNCIA E ENGENHARIA DE MATERIAIS

Orientador: Dr. Marcello Rubens Barsi Andreeta

São Carlos-SP
2024

DEDICATÓRIA

Aos meus pais Agnaldo (*in memorian*) e Helga por todo o sacrifício que fizeram e fazem por mim. Não há palavras suficientes para agradecer-lhes.

À minha esposa Fabiely, pelo carinho, compreensão e, principalmente, paciência. Tê-la ao meu lado tornou a caminhada mais suave e me deu forças para seguir em frente.

Ao meu segundo pai, Renato, por ser uma constante fonte de motivação e incentivo ao longo de minha vida.

As minhas irmãs, Andréa e Ana Renata, pelo carinho e alegria mesmo nos momentos mais difíceis.

VITAE DO CANDIDATO

Bacharel em Informática pela UNESA (2010).

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Ciência e Engenharia de Materiais

Folha de Aprovação

Defesa de Dissertação de Mestrado do candidato Armando José de Sá Santos, realizada em 30/07/2024.

Comissão Julgadora:

Prof. Dr. Marcello Rubens Barsi Andreeta (UFSCar)

Prof. Dr. Rodolfo Foster Klein Gunnewiek (UFSCar)

Prof. Dr. Thiago Nicolau Magalhães de Souza Conte (UEPA)

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa de Pós-Graduação em Ciência e Engenharia de Materiais.

AGRADECIMENTOS

Ao professor Marcello R. B. Andreeta, meu orientador, pelas valiosas e incontáveis horas dedicadas ao projeto, sempre com uma presença cheia de otimismo, pela paciência, por seus conselhos e por toda atenção ao longo do projeto. Muito obrigado.

A colega Flávia, que ganhei durante o Programa, pelas valiosas e incontáveis horas dedicadas a me ajudar.

A Universidade Federal de São Carlos, Programa de Pós-Graduação em Ciência e Engenharia de Materiais.

Ao Laboratório de Materiais Vítreos (LaMaV) pela infraestrutura disponibilizada.

Ao “Center for Research, Technology and Education in Vitreous Materials” (CeRTEV).

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001

RESUMO

A solubilidade, ou a capacidade de um material se dissolver em um solvente, é uma propriedade fundamental dos materiais, especialmente para cerâmicas e vidros. A complexidade das interações soluto-solvente torna difícil calcular a solubilidade com precisão, e a falta de métodos precisos de medição pode afetar a confiabilidade dos dados. A taxa de dissolução, que é a velocidade em que um soluto se dissolve, é crucial em áreas como química e farmacologia. A equação de Noyes-Whitney, desenvolvida em 1897, descreve como a taxa de dissolução é influenciada por vários fatores. A composição química de um vidro pode influenciar significativamente suas propriedades, incluindo a solubilidade e a taxa de dissolução. A mineração de dados e o aprendizado de máquina podem ser usados para prever essas propriedades, auxiliando na descoberta e no design de novos materiais. Redes Neurais são particularmente úteis para prever a solubilidade e a taxa de dissolução devido à sua capacidade de modelar relações complexas. O Orange Data Mining é uma ferramenta de mineração de dados e aprendizado de máquina que pode ser usada para prever a solubilidade e a taxa de dissolução de vidros com base em sua composição. A ferramenta é intuitiva e fácil de usar, permitindo aos pesquisadores construir modelos sem necessidade de extensa codificação. Ela também é capaz de lidar com grandes conjuntos de dados, essencial para a mineração de dados, propondo uma abordagem unificada para prever tanto a solubilidade quanto a taxa de dissolução, utilizando apenas a composição do material como entrada. Isso simplifica o processo de previsão, tornando-o mais acessível e menos dependente de dados complexos. Neste trabalho, por meio de software open source, será apresentado um estudo da viabilidade de uso desta ferramenta na predição da solubilidade e taxa de dissolução em meio aquoso aplicada a compostos óxidos

Palavras-chave: Solubilidade; Taxa de dissolução; Materiais cerâmicos; Mineração de dados; Composição

ABSTRACT

STUDY OF THE FEASIBILITY OF DETERMINING THE DISSOLUTION IN AQUEOUS MEDIUM OF OXIDE COMPOUNDS USING THE OPEN SOURCE TOOL ORANGE DATA MINING

Solubility, or the ability of a material to dissolve in a solvent, is a fundamental property of materials, especially for ceramics and glass. The complexity of solute-solvent interactions makes it difficult to accurately calculate solubility, and the lack of precise measurement methods can affect the reliability of the data. The dissolution rate, which is the speed at which a solute dissolve, is crucial in areas such as chemistry and pharmacology. The Noyes-Whitney equation, developed in 1897, describes how the dissolution rate is influenced by various factors. The chemical composition of a glass can significantly influences its properties, including solubility and dissolution rate. Data mining and machine learning can be used to predict these properties, assisting in the discovery and design of new materials. Neural Networks are particularly useful for predicting solubility and dissolution rate due to their ability to model complex relationships. Orange Data Mining is a data mining and machine learning tool that can be used to predict the solubility and dissolution rate of glasses based on their composition. The tool is intuitive and easy to use, allowing researchers to build models without the need for extensive coding. It is also capable of handling large data sets, essential for data mining, proposing a unified approach to predict both solubility and dissolution rate, using only the material composition as input. This simplifies the prediction process, making it more accessible and less dependent on complex data. In this work, through open-source software, a study of the feasibility of using this tool in predicting solubility and dissolution rate in aqueous medium applied to oxide compounds will be presented.

Keywords: Solubility; Dissolution rate; Ceramic materials; Data mining; Composition.

SUMÁRIO

Pág.

ÍNDICE DE TABELAS

ÍNDICE DE FIGURAS

Pág.

SÍMBOLOS E ABREVIATURAS

ADMET: Absorção, Distribuição, Metabolismo, Excreção e Toxicidade

MAE: Mean Absolute Error

MSE: Mean Squared Error

R^2: Coeficiente de Determinação

ReLu: Rectified Linear Unit

RMSE: Root Mean Squared Error

RNA: Redes Neurais Artificiais

SGD: Stochastic Gradient Descent

1 INTRODUÇÃO

A solubilidade é uma propriedade fundamental dos materiais, que se refere à capacidade de um material se dissolver em um solvente específico. Essa propriedade é particularmente importante para cerâmicas e vidros, cuja solubilidade pode ser influenciada por vários fatores, incluindo composição química, temperatura e pressão do ambiente [1].

No entanto, a complexidade das interações soluto-solvente representa um desafio para calcular a solubilidade com precisão. Cada combinação de soluto e solvente pode apresentar características únicas, tornando difícil prever com exatidão a solubilidade. Além disso, a falta de métodos precisos e confiáveis para medir a solubilidade também pode afetar a validade e a confiabilidade dos dados obtidos. Diferentes técnicas de medição podem produzir resultados inconsistentes e imprecisos [2].

A taxa de dissolução, por sua vez, é uma medida da velocidade em que um soluto se dissolve num solvente, sendo crucial em áreas como a química e a farmacologia. Essa taxa pode determinar a eficácia dos medicamentos e o sabor de bebidas, como o café [3].

Os primeiros estudos sobre modelagem da taxa de dissolução foram realizados por Noyes e Whitney, que desenvolveram a equação de Noyes-Whitney em 1897. Essa equação descreve como a taxa de dissolução é influenciada pelo coeficiente de difusão, área de superfície do soluto, concentração de saturação na superfície do sólido e concentração do soluto no volume do solvente [4].

A composição química de um vidro pode influenciar notavelmente suas propriedades, dentre as quais se incluem a solubilidade e a taxa de dissolução. Tal fato se evidencia especialmente relevante em aplicações de setores como a indústria farmacêutica e a gestão de resíduos radioativos. Em paralelo a isso, emerge como um campo de estudo de grande potencial a previsão de propriedades de materiais através da mineração de dados. Esta técnica de análise, quando aplicada conjuntamente com o aprendizado de máquina, possibilita o modelamento e a previsão de propriedades de materiais. Este recurso se mostra de grande valia, pois facilita tanto a descoberta quanto o

design de novos materiais. A mineração de dados, juntamente com o aprendizado de máquina, está sendo cada vez mais adotada para prever e otimizar as propriedades dos materiais, principalmente no que tange aos materiais vítreos [5].

Redes Neurais se tornam uma boa escolha para prever a solubilidade e a taxa de dissolução devido à sua capacidade de modelar relações complexas e não lineares entre variáveis. Além disso, as Redes Neurais são capazes de aprender representações de alto nível dos dados, o que pode ser útil quando se trabalha com dados de composição de materiais. No entanto, muitos estudos na literatura utilizam uma variedade de dados de entrada para prever a solubilidade e a taxa de dissolução, tornando o processo mais complexo e dependente de dados difíceis de obter [6].

Contudo, a capacidade de prever a solubilidade e a taxa de dissolução de vidros baseada em sua composição é certamente um desafio, mas é aqui que entram as ferramentas modernas de mineração de dados e aprendizado de máquina. Uma dessas ferramentas é o Orange Data Mining, uma plataforma de código aberto que oferece uma série de recursos poderosos para a análise e a modelagem de dados [7].

O Orange Data Mining, utilizado neste contexto, apresenta várias vantagens, como sua intuitiva interface e facilidade de uso, permitindo aos pesquisadores construir modelos de aprendizado de máquina sem necessidade de extensa codificação [8]. A ferramenta oferece uma variedade de algoritmos para análise preditiva, adequando-se ao conjunto de dados em questão [7], além de realizar análise de recursos, processo que identifica e seleciona variáveis relevantes em um conjunto de dados, como a composição do vidro e as condições ambientais para a previsão da solubilidade e taxa de dissolução dos vidros [9].

A ferramenta também é capaz de lidar com grandes conjuntos de dados, essencial para a mineração de dados, que é frequentemente desafiada pela "maldição da dimensionalidade", onde a complexidade dos dados aumenta exponencialmente com o número de variáveis. Assim, com todas estas ferramentas e conhecimentos à nossa disposição, a previsão da solubilidade e

taxa de dissolução de vidros a partir da sua composição é não apenas possível, mas também fundamental para a inovação e o progresso em diversas áreas da ciência e da indústria [10].

Por último, mas não menos importante, o Orange Data Mining permite a validação cruzada e o ajuste de parâmetros, garantindo que o modelo produzido seja robusto e preciso [9].

Este trabalho propõe uma abordagem simplificada para prever tanto a solubilidade quanto a taxa de dissolução utilizando exclusivamente a composição química dos materiais como dado de entrada. Esta abordagem simplifica significativamente o processo de previsão, tornando-o mais acessível e menos dependente de dados complexos ou difíceis de obter.

2 OBJETIVOS

2.1 Objetivo geral

O presente projeto de dissertação de mestrado tem como objetivo principal o estudo da viabilidade da predição da solubilidade e taxa de dissolução em meio aquoso de compostos óxidos utilizando a ferramenta *open source Orange Data Mining*.

2.2 Objetivos específicos

Compilar e organizar dados experimentais da literatura sobre composição química, solubilidade e taxa de dissolução de compostos inorgânicos e vidros óxidos.

Preparo e pré-processamento dos dados para a aplicação de técnicas de *data mining* e *machine learning*.

Implementar redes neurais artificiais para estabelecer relações preditivas entre a composição química e as propriedades de solubilidade e taxa de dissolução.

Avaliar a performance dos modelos preditivos através de métricas estatísticas como MSE, R^2 e análise de correlação.

3 REVISÃO BIBLIOGRÁFICA

3.1 Solubilidade

A solubilidade é uma característica crucial em materiais tais como vidros e cerâmicas. Esta propriedade, que pode ser influenciada por fatores como composição química, temperatura, pressão e presença de outras substâncias, impacta diretamente a funcionalidade do material em diversas aplicações. O entendimento aprofundado desses fatores é fundamental na produção e inovação desses materiais [11].

Os estudos pioneiros no que tange aos modelos de solubilidade foram executados por pesquisadores emblemáticos como Samuel H. Yalkowsky e Michael H. Abraham [12], [13]. Em especial, o trabalho de Yalkowsky foi de suma importância para a compreensão da solubilidade de fármacos e compostos químicos. A sua publicação apresenta conceitos basilares de solubilidade e propõe uma metodologia para estimar a solubilidade aquosa de *nonelectrolytes*, tornando-se um marco referencial na área [12].

Michael H. Abraham é conhecido por suas contribuições no desenvolvimento de equações que podem prever a solubilidade em função de várias propriedades físicas e químicas. Abraham apresenta a teoria de interações específicas do soluto-solvente [13].

Um dos principais desafios enfrentados pelos pesquisadores na obtenção da solubilidade é a falta de métodos precisos e confiáveis para medir essa propriedade. Vários estudos têm demonstrado que diferentes técnicas de medição podem produzir resultados inconsistentes e imprecisos, o que pode afetar a validade e a confiabilidade dos dados obtidos [14]

Adicionalmente, é fundamental reconhecer que a solubilidade é sensivelmente afetada pelas condições ambientais, como temperatura, pressão e pH. Mudanças nestes parâmetros ao longo dos experimentos podem modificar de maneira significativa a solubilidade, tornando complexo o desafio de conseguir dados precisos e reprodutíveis. Neste contexto, se discute amplamente a forma como essas variações ambientais podem influenciar a solubilidade e o impacto das técnicas de estado sólido na melhoria dessa propriedade [15].

A seleção do solvente apropriado é outro desafio comum enfrentado pelos pesquisadores. A escolha do solvente tem um impacto considerável na solubilidade, já que a natureza polar ou apolar do solvente pode afetar a interação entre a substância química e o solvente. Fatores como toxicidade, inflamabilidade ou custo elevado de alguns solventes podem limitar a sua utilização. As impurezas também podem influenciar a solubilidade de várias maneiras, seja formando complexos com o solvente ou com a substância solúvel, prejudicando a interação com o solvente. Além disso, as impurezas podem comprometer a reprodutibilidade dos resultados, dificultando a comparação entre diferentes medidas de solubilidade [16].

A disponibilidade de dados de solubilidade representa outro desafio para os pesquisadores. Algumas substâncias têm dados de solubilidade limitados ou inexistentes, principalmente quando se tratam de condições específicas de temperatura, pressão e pH. Esta escassez de dados pode dificultar a previsão da solubilidade [17].

Para superar estes obstáculos, os pesquisadores têm explorado novas técnicas e abordagens para medir e prever a solubilidade de substâncias. Por exemplo, métodos computacionais baseados em modelos termodinâmicos e estatísticos têm demonstrado ser promissores para prever a solubilidade com maior precisão e eficiência. Além disso, técnicas avançadas, como espectroscopia, microscopia e cromatografia, têm sido aplicadas para elucidar os mecanismos de dissolução e fornecer informações mais detalhadas sobre a solubilidade [18]

3.2 Taxa de dissolução

Entender o conceito de "Taxa de Dissolução" é crucial em várias áreas científicas, particularmente na química e farmacologia. "Taxa de Dissolução" se refere à “velocidade” na qual um soluto se dissolve num solvente [19]. Esta taxa é crucial para determinar a velocidade com que uma substância se distribui uniformemente dentro de um solvente, impactando tudo, desde a eficácia dos medicamentos até o sabor do seu café matinal [3].

Os primeiros trabalhos notáveis sobre modelagem da taxa de dissolução foram apresentados por Noyes e Whitney, cujo trabalho culminou na equação de Noyes-Whitney, publicada em 1897. Nesta equação, eles descreveram como a taxa de dissolução é influenciada pelo coeficiente de difusão, área de superfície do soluto, a concentração de saturação na superfície do sólido e a concentração do soluto no volume do solvente [4].

A taxa de dissolução é uma área de estudo ampla e importante na farmacologia e na química física. Ela descreve o quão rápido uma substância sólida se dissolve em um solvente específico sob condições específicas. Nanson Alfred Reinecke foi um dos primeiros cientistas a apresentar trabalhos relevantes sobre a modelagem da taxa de dissolução no início do século XX. Ele propôs a lei de Noyes-Whitney, que descreve a taxa de dissolução. A equação de Noyes-Whitney (1) é dada por[4]:

$$\frac{dC}{dt} = K * A * (Cs - Ct) \qquad (1)$$

onde: dC/dt é a taxa de dissolução, K é o coeficiente de difusão, A é a área de superfície, Cs é a concentração da substância logo após a superfície do sólido, e Ct é a concentração da substância no volume do líquido.

Os principais entraves para calcular a taxa de dissolução incluem o coeficiente de difusão (K), que pode variar com a temperatura, a pressão e a concentração, tornando difícil estimar seu valor preciso. Determinação da área de superfície (A), para sólidos irregulares, é desafiador determinar com precisão a área de superfície. Saturação da superfície (Cs) pode ser difícil medir diretamente e pode mudar com o tempo à medida que o sólido se dissolve.
Concentração no volume do líquido (C): Esta é a concentração da substância no volume do líquido. Ela pode mudar à medida que o sólido se dissolve, afetando a taxa de dissolução. A taxa de dissolução é muito sensível às condições experimentais, incluindo temperatura e agitação. Manter essas condições constantes durante um experimento pode ser difícil, podendo ainda ser afetada pela presença de outras substâncias no solvente [4].

A formulação de Noyes-Whitney é bastante abrangente, porém, vários desafios se apresentam ao tentar calcular a taxa de dissolução em situações práticas: Variação do coeficiente de difusão (K): O coeficiente de difusão é altamente dependente de uma série de fatores, incluindo temperatura, pressão e concentração [20]. Determinação da área de superfície (A): A determinação precisa da área de superfície de um sólido, especialmente um com uma forma irregular, pode ser bastante desafiadora [2].

Saturação da superfície (Cs): Medir diretamente a concentração de saturação na superfície do sólido pode ser difícil, e a concentração pode mudar com o tempo à medida que o sólido se dissolve [21] Concentração no volume do líquido (Ct): A concentração da substância no volume do líquido pode mudar à medida que o sólido se dissolve, afetando a taxa de dissolução [22]

A equação de Noyes-Whitney desempenha um papel fundamental na quantificação da taxa de dissolução. Esta relação matemática leva em conta os fatores mencionados e fornece uma maneira de calcular a Taxa de Dissolução [4]. Na ciência farmacêutica, entender a taxa de dissolução é fundamental para projetar sistemas eficazes de liberação de medicamentos. Os medicamentos devem se dissolver nos fluidos corporais para serem absorvidos e exercerem seus efeitos terapêuticos. Assim, a Taxa de Dissolução pode influenciar significativamente a eficácia do medicamento [23].

3.3 Predição de propriedades dos materiais através de mineração de dados

A predição de propriedades dos materiais através de mineração de dados tem se mostrado um campo de estudo extremamente promissor. A mineração de dados, combinada com aprendizado de máquina, tem sido aplicada para modelar e prever propriedades dos materiais, auxiliando na descoberta e no design de novos materiais [24]. Os bancos de dados de materiais, que contêm informações detalhadas sobre a composição, estrutura e propriedades dos materiais, fornecem um terreno fértil para a aplicação de técnicas de mineração de dados. A aplicação dessas técnicas pode revelar relações não lineares complexas e tendências que são difíceis de identificar com métodos tradicionais

[25]. O aprendizado de máquina, um método amplamente usado na mineração de dados, tem sido aplicado para prever propriedades dos materiais, como a energia de formação de compostos, propriedades mecânicas, propriedades termelétricas e muito mais [26].

Por exemplo, algoritmos de aprendizado de máquina foram usados para prever a energia de formação de compostos binários com uma precisão que rivaliza com os métodos teóricos mais avançados [27]. Além disso, a mineração de dados foi usada para prever a ductilidade dos metais, que é uma propriedade mecânica crucial na indústria de manufatura [28].

Apesar desses avanços, ainda existem desafios na predição de propriedades dos materiais através da mineração de dados. A qualidade e a quantidade dos dados disponíveis são limitações significativas, pois os modelos de aprendizado de máquina dependem de grandes quantidades de dados de alta qualidade para treinamento [24].

3.4 Análise da composição de vidros e como influencia na solubilidade e na taxa de dissolução

A composição química de um vidro pode influenciar significativamente suas propriedades, incluindo a solubilidade e a taxa de dissolução. Essas propriedades são de particular interesse em várias aplicações de vidro, como na indústria farmacêutica e na área de resíduos radioativos [30].

A solubilidade dos vidros depende fortemente dos óxidos de rede, como o dióxido de silício (SiO_2), e dos óxidos modificadores, como o óxido de sódio (Na_2O) ou o óxido de cálcio (CaO). Em geral, quanto maior a concentração de óxidos de rede e menor a concentração de óxidos modificadores, menor a solubilidade do vidro [31].

A taxa de dissolução dos vidros também é uma função de sua composição. Por exemplo, a adição de alumínio (Al_2O_3) pode retardar a taxa de dissolução do vidro devido à formação de uma camada de gel resistente na superfície do vidro durante a dissolução. Além disso, a adição de elementos como ferro (Fe_2O_3) ou titânio (TiO_2) também pode reduzir a taxa de dissolução do vidro [32].

Na indústria farmacêutica, a solubilidade do vidro é de particular importância na fabricação de fármacos. Vidros solúveis, como os vidros bioativos, são utilizados para a entrega de fármacos no corpo humano, onde a taxa de dissolução do vidro pode ser ajustada para controlar a liberação do fármaco [33]. Na área de resíduos radioativos, a composição do vidro é ajustada para minimizar a solubilidade e a taxa de dissolução, a fim de garantir a segurança a longo prazo do armazenamento de resíduos [34].

3.5 Redes Neurais Artificiais (RNA)

As Redes Neurais Artificiais (RNA) são uma metodologia computacional inspirada na estrutura e no funcionamento do cérebro humano, sendo um dos principais pilares da Inteligência Artificial (IA). Elas visam a reproduzir as sinapses neurais humanas, onde a comunicação e a transferência de conhecimento ocorrem [35]

Uma rede neural é constituída por uma coleção de unidades de processamento, denominadas neurônios, organizadas em camadas. A informação flui da camada de entrada, passa pelas camadas ocultas e chega à camada de saída [36].

Aprendizado em Redes Neurais acontece por um processo chamado de retropropagação [37]. Na retropropagação, o algoritmo compara a saída prevista com a saída real e, a partir da diferença (erro), ajusta os pesos associados a cada neurônio na tentativa de minimizar o erro nas próximas previsões [37].

Redes Neurais são bem aplicadas em problemas de classificação, reconhecimento de padrões e previsões. Elas são particularmente eficazes em lidar com dados de alta dimensão e não-lineares [38]. Na medicina, as redes neurais têm sido usadas para prever o diagnóstico e o prognóstico de diversas doenças, como câncer e doenças cardiovasculares. Na finança, são utilizadas para prever a movimentação do mercado de ações [39]

Os principais desafios no uso de Redes Neurais incluem a escolha dos parâmetros, como o número de camadas ocultas e neurônios, a complexidade computacional, o risco de sobreajuste e a interpretabilidade dos modelos [40]

3.6 AdaBoost

O algoritmo AdaBoost, ou *Adaptive Boosting*, é um algoritmo de aprendizado de máquina que é usado para melhorar o desempenho de outros algoritmos de aprendizado. Ele faz isso criando um "comitê" de aprendizes fracos e combinando suas previsões para criar uma previsão final mais forte. Um aprendiz fraco é um modelo que é apenas um pouco melhor do que adivinhar aleatoriamente. AdaBoost é frequentemente usado em conjunto com árvores de decisão, mas pode ser usado com qualquer tipo de algoritmo de aprendizado de máquina[41], [42].

AdaBoost tem sido aplicado com sucesso em uma variedade de problemas. Por exemplo, foi usado para melhorar a precisão dos sistemas de detecção de intrusão de rede, reduzindo a taxa de falsos positivos. Também foi usado para melhorar a detecção de câncer de mama, aumentando a precisão, sensibilidade e especificidade do diagnóstico [43].

O AdaBoost funciona atribuindo pesos a cada exemplo de treinamento e ajustando esses pesos após cada aprendiz fraco ser treinado. Os exemplos que são classificados incorretamente recebem pesos maiores, incentivando o próximo aprendiz fraco a se concentrar neles. Isso permite que o AdaBoost se adapte a exemplos difíceis de classificar [44].

No entanto, o AdaBoost pode sofrer de *overfitting* em situações de alto ruído. Isso ocorre porque o algoritmo tende a se concentrar em exemplos difíceis de classificar, que podem ser simplesmente ruído. Várias técnicas de regularização foram propostas para mitigar esse problema, incluindo a introdução de uma "desconfiança" nos dados e a realização de uma descida de gradiente com relação à margem suave [45].

Em comparação com outros algoritmos de aprendizado de máquina, o AdaBoost tem uma complexidade computacional relativamente baixa e pode produzir resultados comparáveis ou melhores. No entanto, é importante notar que o desempenho do AdaBoost depende fortemente da qualidade dos aprendizes fracos e da quantidade de ruído nos dados [42].

3.7 AdaBoost x Redes Neurais Artificiais

O AdaBoost é um algoritmo de aprendizagem de conjunto que combina vários aprendizes fracos para criar um modelo forte. Ele é conhecido por sua simplicidade e eficácia, especialmente em tarefas de classificação. O AdaBoost tem a vantagem de ser menos propenso a *overfitting* em comparação com alguns outros algoritmos, especialmente em ambientes de baixo ruído [41]. No entanto, o AdaBoost pode sofrer de *overfitting* em situações de alto ruído e pode ser sensível a outliers [45].

As Redes Neurais são modelos de aprendizagem de máquina que se inspiram no funcionamento do cérebro humano. Elas são especialmente eficazes para tarefas que envolvem dados não estruturados, como reconhecimento de imagem e processamento de linguagem natural [46]. As Redes Neurais são capazes de aprender representações complexas e não lineares, o que as torna poderosas para muitas tarefas. No entanto, elas também têm desvantagens. Por exemplo, elas podem ser propensas a *overfitting*, especialmente se a rede for muito complexa em relação à quantidade de dados disponíveis. Além disso, as Redes Neurais podem ser computacionalmente intensivas e requerem mais tempo para treinar do que outros algoritmos [6].

Em termos de desempenho, tanto o AdaBoost quanto as Redes Neurais podem alcançar alta precisão em várias tarefas. No entanto, o melhor algoritmo a ser usado depende do problema específico em questão. Por exemplo, em tarefas de classificação de imagens, as Redes Neurais, especialmente as Redes Neurais Convolucionais, tendem a superar outros algoritmos, incluindo o AdaBoost [6]. No entanto, em tarefas onde os dados são estruturados e o ruído é baixo, o AdaBoost pode ser uma escolha eficaz [47].

3.8 Orange Data Mining

O Orange é uma plataforma de mineração de dados de código aberto, fornecendo técnicas de análise de dados como visualização, pré-processamento de dados, classificação e agrupamento, com uma interface intuitiva de arrastar e soltar que permite aos usuários implementar fluxos de trabalho de mineração de dados complexos, independentemente de sua experiência em programação.

Destaca-se a modularidade do Orange, onde os usuários podem combinar diferentes *widgets* para criar um fluxo de trabalho personalizado, abrangendo funções desde a leitura e limpeza de dados até a implementação e avaliação de modelos de aprendizado de máquina, como Redes Neurais e Máquinas de Vetores de Suporte [7].

Outra característica notável é a sua capacidade de visualização de dados, oferecendo vários *widgets* de visualização que permitem aos usuários explorar seus dados de várias maneiras, incluindo gráficos de dispersão, gráficos de barras, mapas de calor e muito mais [9].

Além disso, o Orange também permite a integração com a linguagem de programação Python. Isso significa que os usuários podem expandir suas análises além dos *widgets* pré-existentes, escrevendo seus próprios scripts Python dentro do ambiente Orange [8].

A ferramenta é uma plataforma de aprendizado de máquina e mineração de dados de código aberto que permite a análise visual de dados e a construção interativa de modelos de aprendizado de máquina. Na literatura existente, muitos estudos utilizam linguagens de programação como Python ou R para construir e treinar seus modelos de aprendizado de máquina [48], [49], [50]. Embora essas linguagens sejam poderosas e flexíveis, elas também têm uma curva de aprendizado íngreme e podem ser inacessíveis para aqueles sem formação em programação.

Em contraste, o Orange oferece uma interface gráfica de usuário intuitiva que permite a construção de modelos de aprendizado de máquina através de um processo de arrastar e soltar. Isso torna o aprendizado de máquina e a mineração de dados acessíveis a um público mais amplo, incluindo aqueles sem formação em programação.

Além disso, o Orange inclui uma variedade de recursos de visualização de dados que podem ajudar a entender melhor os dados e os resultados do modelo. Isso pode ser particularmente útil, onde a relação entre a composição do material e a solubilidade e a taxa de dissolução pode ser complexa e não linear.

Há suporte a uma variedade de algoritmos de aprendizado de máquina, permitindo experimentar diferentes abordagens para prever a solubilidade e a taxa de dissolução. Isso pode permitir que você encontre o melhor modelo para seus dados, melhorando ainda mais a precisão de suas previsões.

O Orange é uma poderosa ferramenta de análise de dados que se destaca pela sua interface visual intuitiva, ideal para usuários sem experiência em programação. Em um estudo de caso publicado [51], foi utilizado para analisar dados de oxidação de materiais, permitindo identificar condições que aumentam a resistência à oxidação.

Através da visualização de padrões e desenvolvimento de modelos preditivos, descobriu-se que certas ligas metálicas e tratamentos térmicos específicos resultam em maior resistência à oxidação, contribuindo para o desenvolvimento de novos materiais resistentes à corrosão. Utilizando dados históricos e algoritmos de regressão, a resistência de novos materiais foi prevista com alta precisão, reduzindo a necessidade de testes extensivos [51].

Apesar das muitas vantagens do Orange, ele também tem algumas limitações. Por exemplo, a quantidade de dados que pode lidar é limitada pela memória do computador. Além disso, embora ofereça uma ampla gama de algoritmos de aprendizado de máquina, pode não incluir alguns dos algoritmos mais recentes ou mais especializados [7].

3.9 Predição de propriedades dos materiais através de mineração de dados

A predição de propriedades dos materiais por meio da mineração de dados tem se tornado uma área de estudo promissora, com a combinação de mineração de dados e aprendizado de máquina sendo utilizada para modelar e prever propriedades dos materiais, auxiliando na descoberta e no design de novos materiais [24]. Bancos de dados de materiais, repletos de informações detalhadas sobre composição, estrutura e propriedades, têm fornecido um terreno fértil para aplicação dessas técnicas, permitindo revelar complexas relações não lineares e tendências difíceis de identificar com métodos tradicionais [25].

Além disso, o aprendizado de máquina, um método bastante usado na mineração de dados, tem sido aplicado na previsão de diversas propriedades dos materiais, como a energia de formação de compostos, propriedades mecânicas, propriedades termelétricas, entre outras [26]. Exemplos notáveis incluem o uso de algoritmos de aprendizado de máquina para prever a energia de formação de compostos binários com alta precisão [27] e a utilização da mineração de dados para prever a ductilidade dos metais, uma propriedade mecânica crucial na indústria de manufatura [28].

Entretanto, apesar desses avanços, existem ainda desafios na predição de propriedades dos materiais através da mineração de dados. As limitações significativas são a qualidade e quantidade de dados disponíveis, pois os modelos de aprendizado de máquina requerem grandes quantidades de dados de alta qualidade para treinamento [24].

3.10 Pesquisas sobre Solubilidade e Taxa de Dissolução

A previsão da solubilidade e da taxa de dissolução são tópicos de pesquisas importantes, e várias abordagens foram exploradas na literatura. No entanto, a aplicação de técnicas de aprendizado de máquina e mineração de dados para esses problemas específicos ainda é uma área de pesquisa emergente.

O estudo "*Impacts of glass composition, pH, and temperature on glass forward dissolution rate*" [52] aborda a taxa de dissolução de vidros nucleares em condições aquosas diluídas, destacando a importância da composição do vidro, pH e temperatura nesse processo. O trabalho analisa 19 vidros de resíduos nucleares, utilizando o modelo $rf=k_0 \cdot 10^{-\eta \cdot pH} \cdot \exp(-Ea/RT)$ para avaliar a taxa de dissolução. Embora pH e temperatura sejam reconhecidos como fatores significativos, a influência da composição do vidro permaneceu incerta. A pesquisa revelou que 90% da variação na taxa de dissolução pode ser explicada apenas pelos efeitos do pH e da temperatura, sugerindo que os efeitos da composição são relativamente pequenos.

Os resultados mostraram que, ao normalizar as diferenças de pH e temperatura, a única diferença notável na taxa de dissolução entre os vidros

estava correlacionada com a fração de tetraedros de vidro formados pelo boro tetraédrico (f([4]B)). Observou-se um limiar abrupto em um valor alto de f([4]B) (~0.22), onde taxas de dissolução mais altas são previstas sem efeitos discerníveis da composição abaixo desse limiar. Essa descoberta indica que, acima de um certo ponto, a presença de boro tetraédrico tem um impacto significativo na taxa de dissolução, independentemente de outras variáveis de composição [52].

Ao longo do estudo, os parâmetros do modelo (k_0, η, Ea) foram ajustados individualmente para cada vidro, e uma análise estatística detalhada foi conduzida para determinar as correlações entre esses parâmetros e a composição do vidro. No entanto, devido à forte correlação positiva entre log[k0] e Ea, não foi possível estabelecer uma correlação preditiva clara entre os parâmetros individuais do modelo e a composição do vidro. Assim, a abordagem de ajustar diretamente a taxa de dissolução combinada de todos os 19 vidros mostrou ser mais eficaz, confirmando que os efeitos da composição são mínimos em comparação com os impactos do pH e da temperatura [52].

Por exemplo, Sun e colaboradores [48] utilizaram uma rede neural profunda para classificar compostos com base em dados experimentais de difração de raios-X, permitindo a identificação rápida de materiais com *band gaps* de interesse para aplicações de colheita de energia.

Jónsdóttir e colaboradores [53] discutem a importância de métodos preditivos para classificar a adequação de compostos químicos como potenciais medicamentos, bem como para prever suas propriedades físico-químicas e ADMET.

Horev-Azaria e colaboradores [49] aplicaram técnicas de aprendizado de máquina para modelar a toxicidade de nanopartículas de ferrita de cobalto em diferentes modelos celulares. Apesar do trabalho não se concentrar diretamente na solubilidade e na taxa de dissolução, ele demonstra o potencial do aprendizado de máquina para prever propriedades complexas de materiais.

Boobier e colaboradores [54] utilizaram *machine learning* para prever a solubilidade em solventes orgânicos e água. A pesquisa aplicou algoritmos como regressão linear, árvores de decisão, redes neurais, entre outros, usando dados

estruturais e informações físico-químicas dos compostos. Os resultados mostraram que os modelos de *machine learning* foram capazes de prever a solubilidade com alta precisão, destacando-se a eficácia das redes neurais e dos métodos de *ensemble learning* como o *gradient boosting*.

Zhuyifan Ye e Defang Ouyang [55] focaram na previsão da solubilidade de compostos de pequenas moléculas em solventes orgânicos usando algoritmos de *machine learning*. O estudo utilizou uma grande base de dados com 5081 entradas de solubilidade experimental e aplicou algoritmos como DNN, SVM, lightGBM, entre outros. O modelo lightGBM se destacou com desempenho superior, demonstrando uma forte correlação entre os valores preditos e os observados. A abordagem ajudou a identificar os melhores solventes para dissolução de compostos específicos, reduzindo a necessidade de experimentação extensiva.

Jin e colaboradores [56] desenvolveram modelos baseados em *machine learning* para prever a solubilidade de ingredientes farmacêuticos ativos. A metodologia envolveu o uso de *fingerprints* moleculares e dados de solubilidade padronizados. Modelos como regressão linear, árvores de decisão e redes neurais foram avaliados, com as redes neurais apresentando os melhores resultados em termos de precisão preditiva. Esse estudo destacou a importância de considerar interações específicas da rede cristalina e a temperatura na modelagem da solubilidade.

Em outro estudo, Yin e colaboradores [57] aplicaram *machine learning* para prever a solubilidade de CO_2 em diferentes solventes, utilizando modelos como regressão de vetor suporte, *gradient boosting* e redes neurais *multilayer perceptron*. Entre os modelos, o XGBoost apresentou a melhor performance com um coeficiente de determinação (R^2) de 0.9838, demonstrando alta precisão nas previsões. A abordagem auxiliou na otimização de processos industriais envolvendo captura e armazenamento de carbono.

Em um estudo semelhante, Horev-Azaria e colaboradores [58] compararam a toxicidade de nanopartículas de cobalto e íons de cobalto usando técnicas de aprendizado de máquina. Finalmente, Taheri e colaboradores [50]

aplicaram algoritmos de aprendizado de máquina baseados em Bayes para mapear a suscetibilidade a sumidouros em uma província do Irã.

Tayyebi e colaboradores [59] compararam modelos baseados em descritores químicos e *fingerprints* moleculares (ECFPs). Utilizou algoritmos de Random Forest (RF) e Regressão Linear Múltipla (MLR) para prever a solubilidade aquosa de compostos orgânicos. O modelo RF mostrou melhor desempenho com valores de R^2 mais altos (0.88 para RF vs. 0.80 para MLR) e menores erros médios quadráticos (RMSE) e absolutos (MAE). Os valores de RMSE e MAE foram 0.64 e 0.41, respectivamente, para o conjunto de teste, indicando uma previsão mais precisa da solubilidade.

Vansh Ramani e Tarak Karmakar [60] desenvolveram um modelo utilizando Redes Neurais de Grafos (GNNs) para prever a solubilidade de compostos em diferentes solventes. A abordagem incorpora interações soluto-solvente utilizando descritores estruturais e cargas parciais atômicas. O modelo proposto, chamado MolMerger, mostrou-se eficiente na previsão da solubilidade sem a necessidade de cálculos químicos complexos ou dados experimentais caros. A inclusão de interações soluto-solvente melhorou significativamente a precisão das previsões.

Embora esses trabalhos não se concentrem especificamente na previsão da solubilidade e da taxa de dissolução com a mesma abordagem proposta neste trabalho, eles demonstram a aplicabilidade das técnicas de aprendizado de máquina e mineração de dados para prever propriedades complexas de materiais e compostos. Esses estudos podem fornecer uma base para o desenvolvimento de novas abordagens para a previsão da solubilidade e da taxa de dissolução.

3.11 A escolha de Redes Neurais

A escolha de Redes Neurais (RN) para prever a solubilidade e a taxa de dissolução pode ser justificada por várias razões.

Primeiramente, as Redes Neurais são conhecidas por sua capacidade de modelar relações complexas e não lineares entre variáveis. Isso é particularmente útil em seu trabalho, onde a relação entre a composição do

material e a solubilidade e a taxa de dissolução pode ser altamente complexa. As Redes Neurais podem capturar essas relações complexas de uma maneira que outros algoritmos, como o AdaBoost, podem não ser capazes [61].

Em segundo lugar, as Redes Neurais são capazes de aprender representações de alto nível dos dados, o que pode ser útil quando se trabalha com dados de composição de materiais. Essas representações de alto nível podem revelar padrões e estruturas nos dados que podem ser úteis para prever a solubilidade e a taxa de dissolução [62].

Além disso, as Redes Neurais são altamente flexíveis e podem ser ajustadas para se adequar a uma variedade de problemas. Uma possibilidade é ajustar a arquitetura da rede, a função de ativação e o algoritmo de otimização para adequá-los ao problema específico. Isso pode permitir que você obtenha o melhor desempenho possível de seu modelo [63]. A escolha das RNAs é justificada pela sua capacidade de modelar relações complexas e não lineares entre variáveis, como demonstrado por Agatonovic-Kustrin e Beresford [64] em aplicações de química farmacêutica.

3.12 Machine Learning e a Engenharia de Materiais

A aplicação de técnicas de aprendizado de máquina na engenharia de materiais é de grande importância devido ao seu potencial para acelerar a descoberta e o desenvolvimento de novos materiais. As técnicas de aprendizado de máquina podem ser usadas para identificar correlações importantes entre as propriedades dos materiais e seus processos de fabricação, permitindo a criação de modelos de ordem reduzida que podem ser usados para prever o comportamento de novos materiais [65].

Por exemplo, as técnicas de aprendizado de máquina têm sido usadas para desenvolver novas regras químicas para engenharia de estruturas eletrônicas em materiais termoelétricos de meio-Heusler. Essas regras podem ser usadas para projetar estruturas de banda com convergência de banda e alta degeneração de vale, melhorando a eficiência dos materiais termoelétricos [66].

Além disso, as técnicas de aprendizado de máquina também têm sido usadas para desenvolver estratégias para sistemas com propriedades de

invariância. Isso é particularmente relevante na engenharia de materiais, onde muitos sistemas possuem propriedades de simetria ou invariância. As técnicas de aprendizado de máquina podem ser usadas para ensinar um modelo de aprendizado de máquina uma propriedade de invariância, melhorando a precisão e a eficiência do modelo [67].

As técnicas de aprendizado de máquina também têm sido usadas para mineração de ligações de estrutura-propriedade em compósitos de alto contraste. Essas ligações podem ser usados para prever as propriedades de novos compósitos, acelerando o processo de descoberta e desenvolvimento de materiais [68].

Finalmente, as técnicas de aprendizado de máquina também têm sido usadas para prever as propriedades dos materiais a partir de apenas a composição elementar. Isso é particularmente útil na engenharia de materiais, onde a composição elementar de um material pode ter um impacto significativo em suas propriedades [69].

3.13 Parâmetros do algoritmo de redes neurais

A análise da combinação de diferentes parâmetros em redes neurais artificiais (RNAs) é crucial para otimizar o desempenho dos modelos preditivos, avaliando as inúmeras possibilidades de combinação entre o número de neurônios (NEURONS), a função de ativação (ACTIVATION), o número máximo de iterações, o erro quadrático médio (MSE) e o coeficiente de determinação (R^2). Estas variáveis são fundamentais para ajustar a complexidade e a capacidade de aprendizado das RNAs, influenciando diretamente na precisão e na robustez das previsões.

Primeiramente, o número de neurônios em uma camada escondida (NEURONS) desempenha um papel crucial na modelagem da capacidade da rede de capturar padrões complexos nos dados. Um número muito baixo de neurônios pode resultar em um modelo subajustado, enquanto um número excessivo pode levar ao sobreajuste, onde o modelo se adapta muito bem aos dados de treinamento, mas falha em generalizar para novos dados. Estudos

anteriores, como os de Heaton [70], destacam a importância de um equilíbrio adequado no número de neurônios para otimização de desempenho.

Além disso, a escolha da função de ativação (ACTIVATION) é essencial para introduzir não-linearidades na rede e permitir que ela aprenda funções complexas. Entre as diversas funções de ativação, a ReLU (*Rectified Linear Unit*) tem se mostrado particularmente eficaz, conforme relatado por Nair e Hinton[71], devido à sua simplicidade computacional e capacidade de mitigar o problema do desaparecimento do gradiente, comum em funções de ativação sigmoide e tangente hiperbólica. A função ReLU, com sua definição matemática simples e vantagens computacionais, contribui significativamente para o desempenho do modelo, ajudando a capturar padrões complexos nos dados sem sofrer os problemas associados ao desaparecimento do gradiente [71]. Esses achados corroboram com a literatura existente e oferecem uma base sólida para aplicações práticas em modelagem preditiva com RNAs [72].

Comparando a ReLU com outras funções de ativação como *Identity*, *Logistic* e *Tanh*, podemos observar diferenças significativas em termos de comportamento e desempenho. A função de ativação *Identity* é simplesmente a função linear f(x)=x. Não há introdução de não linearidade na rede, o que implica que, mesmo com múltiplas camadas, a combinação das ativações permanecerá linear. Isso limita a capacidade da rede de aprender representações complexas dos dados [71].

A função *Logistic*, também conhecida como *Sigmoide*, é definida como $f(x) = \frac{1}{1+e^{-x}}$. Ela mapeia qualquer valor de entrada para um intervalo entre 0 e 1, introduzindo uma não-linearidade suave. No entanto, a função *Sigmoide* tende a sofrer com o problema do desaparecimento do gradiente, especialmente em redes profundas, onde os gradientes podem se tornar extremamente pequenos durante a retropropagação, dificultando o treinamento eficaz dos pesos [73].

A função Tanh (tangente hiperbólica) é definida como f(x)=tanh(x), que mapeia os valores de entrada para um intervalo entre -1 e 1. Embora a Tanh centralize os dados em torno de zero, o que pode ajudar na convergência durante o treinamento, ela ainda sofre com o problema do desaparecimento do gradiente, semelhante à função Sigmoide [74].

Por outro lado, a ReLU é definida como f(x)=max(0,x), retornando zero para entradas negativas e a própria entrada para entradas positivas. Essa característica simples e não-linear permite que a ReLU evite o problema do desaparecimento do gradiente, facilitando o treinamento de redes profundas e melhorando a convergência [71]. Além disso, a ReLU pode introduzir esparsidade na rede, ativando apenas um subconjunto de neurônios, o que pode levar a modelos mais eficientes e com maior capacidade de generalização [75].

No que diz respeito ao número máximo de iterações, este parâmetro define o limite de épocas para o treinamento da rede. Um valor muito baixo pode impedir que a rede alcance um mínimo global do erro, enquanto um valor muito alto pode resultar em um tempo de treinamento desnecessariamente longo e possível sobreajuste. A escolha adequada deste parâmetro é frequentemente feita com base em experimentações e validações cruzadas, como sugerido por Goodfellow e colaboradores [76].

O erro quadrático médio (MSE) é uma métrica fundamental para a avaliação de modelos preditivos, especialmente em trabalhos que envolvem redes neurais artificiais (RNAs) [76], [77]. Ele quantifica a discrepância média entre os valores preditos pelo modelo e os valores reais observados, elevando ao quadrado essa diferença para enfatizar e penalizar erros maiores [76]. Podemos calcular através da fórmula: $MSE = (1/n) * \Sigma(y_i - \hat{y}_i)^2$, onde n é o número total de amostras no conjunto de dados, y_i é o valor real da i-ésima amostra e $\hat{y}_i$ é o valor predito pelo modelo para a i-ésima amostra. O MSE é amplamente utilizado devido à sua sensibilidade a grandes erros [77], uma característica que o torna particularmente útil para identificar e penalizar previsões imprecisas, auxiliando na otimização do modelo e na busca por melhores resultados [76], [77].

No contexto de redes neurais, o MSE desempenha um papel crucial na função de perda, que é otimizada durante o processo de treinamento do modelo. Um MSE baixo indica que o modelo está fornecendo previsões que estão, em média, próximas aos valores reais, sugerindo uma boa capacidade de generalização. Em contrapartida, um MSE elevado pode sinalizar que o modelo está subajustado ou que há necessidade de ajustes nos hiperparâmetros, como

o número de neurônios, a função de ativação ou o número de iterações de treinamento.

A escolha do MSE como métrica de avaliação é corroborada por diversos estudos recentes. Por exemplo, Zhang e colaboradores [78] destacam a utilização do MSE em aplicações de aprendizado profundo devido à sua eficiência em lidar com problemas de regressão. Da mesma forma, Chollet [79] enfatiza a relevância do MSE na avaliação da performance de modelos preditivos em sua obra sobre aprendizado profundo com Python e Keras, um *framework* popular para a construção e treinamento de RNAs.

Além disso, o MSE facilita a comparação direta entre diferentes modelos e configurações de hiperparâmetros. Ao testar várias combinações de neurônios na camada oculta e iterações máximas, como descrito anteriormente, o MSE serve como um indicador objetivo para determinar qual configuração oferece a melhor performance. Essa análise detalhada permite que pesquisadores e praticantes de ciência de dados identifiquem rapidamente as configurações que minimizam o erro, otimizando assim a precisão das previsões.

O MSE é uma métrica indispensável para trabalhos que envolvem a análise e otimização de redes neurais artificiais. Sua aplicação fornece *insights* valiosos sobre a performance do modelo, orientando ajustes necessários para alcançar uma maior precisão preditiva. A relevância do MSE é amplamente reconhecida e apoiada por referências contemporâneas na literatura de aprendizado de máquina.

4 MATERIAIS E MÉTODOS

4.1 Etapas do desenvolvimento

O desenvolvimento do trabalho consiste em várias etapas, começando com a Coleta de Dados, onde são reunidas todas as informações necessárias para a análise. Em seguida, esses dados passam pela Modelagem dos Dados, etapa em que são organizados e estruturados de maneira adequada para a análise subsequente. A Verificação de Padrões com Orange é realizada para identificar padrões e tendências significativas nos dados usando a ferramenta de mineração de dados Orange. Após essa verificação, ocorre o Ajuste do Algoritmo, que visa otimizar o desempenho do modelo de análise.

A próxima etapa é a Separação das Bases de Dados (Aprendizado e Teste), onde os dados são divididos em conjuntos de treinamento e teste para validação do modelo. Posteriormente, a Predição: Base de Aprendizado vs. Teste é realizada para avaliar a precisão do modelo com os dados de teste. A título de inferir sobre os dados e a precisão da determinação da solubilidade, foram realizados testes de predição para solubilidades altas, médias e baixas em separado. O mesmo foi realizado para as taxas de solubilidade.

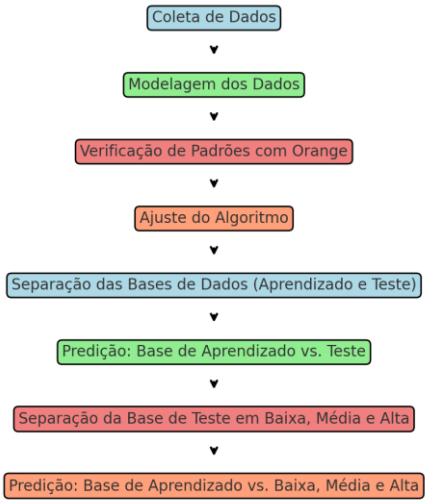

Figura 1: Fluxograma das etapas do desenvolvimento do trabalho.

4.2 Coleta e Modelagem dos Dados

Para o desenvolvimento deste trabalho, a bases de dados para a solubilidade foi obtida a partir de dados coletados em Handbook [80] e para a taxa de dissolução, utilizamos a coleta de dados realizada por Vienna et. al. [52] para dados de vidros de sílica.

A sistematização dos dados para a predição da solubilidade e taxa de dissolução foi realizada em duas etapas principais. Primeiramente, os dados de solubilidade e taxa de dissolução foram extraídos da literatura e organizados em uma planilha, com cada composto representado em uma coluna e a quantidade de cada elemento químico em colunas subsequentes.

Os dados utilizados neste trabalho foram modelados através de decomposição de sua fórmula química, por exemplo $Yb_2(SO_4)_3$, todos os foram colocados em uma planilha, com cada coluna da tabela representando os elementos químicos da tabela periódica. Na coluna correspondente ao Yb, foi atribuído o valor “2”, para S, 3 e para O, 12. Nas demais colunas que não há átomos, foi atribuído “0”. Desta forma temos a quantidade de átomos de cada composto, sendo essa a nossa entrada. Utilizando as ferramentas do Orange, aplicamos um pré-processamento para que valores nulos sejam automaticamente ignorados e não entrem no processo de aprendizado-predição, conforme Tabela 1 – Solubilidade e Tabela 2 – Taxa de Dissolução.

Tabela 1: Amostra dos dados de solubilidade apresentando a distribuição da quantidade de átomos em cada coluna, seguida por uma coluna adicional que contém a média da solubilidade de cada composto.

Formula	C	Cr	H	Li	N	Na	Nd	O	S	Zn	SOL
(NH4)2C2O4	4	0	8	0	2	0	0	4	0	0	5,2
BeSO4	0	0	0	0	0	0	0	4	1	0	41,3
Ca(C2H3O2)2	2	0	6	0	0	0	0	4	0	0	34,2
CoSO4	0	0	0	0	0	0	0	4	1	0	38,3
Cr(ClO4)3	0	1	0	0	0	0	0	12	0	0	58
CrO3	0	1	0	0	0	0	0	3	0	0	169,39
Li2SO4	0	0	0	2	0	0	0	4	1	0	34,2
Na2Cr2O7	0	2	0	0	0	2	0	7	0	0	187
Na2CrO4	0	1	0	0	0	2	0	4	0	0	87,6
NaClO4	0	0	0	0	0	1	0	4	0	0	205
NaCN	1	0	0	0	1	1	0	0	0	0	64,2
Nd(NO3)3	0	0	0	0	3	0	1	9	0	0	152
NdCl3	0	0	0	0	0	0	1	0	0	0	100,26
TlIO3	0	0	0	0	0	0	0	3	0	0	0,0667
TlNO3	0	0	0	0	1	0	0	3	0	0	12,11
ZnSO4	0	0	0	0	0	0	0	4	1	1	57,7

Tabela 2: Amostra dos dados da Taxa de Dissolução e solubilidade apresentando a distribuição da quantidade de átomos em cada coluna, seguida por uma coluna com a Temperatura, pH e o valor da Taxa de cada composto.

GLASS	Al	B	Ca	F	Fe	K	La	Li	Mg	Na	O	Si	Zn	Zr	T (°C)	pH(*T*)	log(r_i) M
LD6-5412	15	9	5	1	0	2	0	0	0	41	183	60	0	0	20,00	5,90	-3,85
IDF21-EC14	13	20	2	0	0	1	0	0	2	53	182	44	3	3	23,00	8,00	-3,38
LD6-5412	15	9	5	1	0	2	0	0	0	41	183	60	0	0	20,00	8,10	-3,28
SRL-202	6	13	1	0	10	3	2	20	2	17	189	57	0	1	23,00	9,00	-3,17
IDF1-B2	13	14	1	2	1	0	0	0	2	55	175	45	3	3	23,00	8,30	-3,10
EWG-C	12	19	6	1	1	1	1	9	2	40	176	43	2	2	23,00	8,00	-2,95
LAW-ABP1	14	18	0	0	2	3	1	0	2	45	187	48	2	3	23,00	7,10	-2,94
IDF21-EC14	13	20	2	0	0	1	0	0	2	53	182	44	3	3	23,00	9,00	-2,89
EWG-C	12	19	6	1	1	1	1	9	2	40	176	43	2	2	23,00	9,00	-2,83
AFCI	12	18	6	0	0	0	2	20	0	15	191	58	0	1	22,70	9,00	-2,79
IDF1-B2	13	14	1	2	1	0	0	0	2	55	175	45	3	3	23,00	9,00	-2,79
ORPLG9	9	17	3	0	0	8	0	0	2	46	176	46	3	3	22,70	9,00	-2,78
LD6-5412	15	9	5	1	0	2	0	0	0	41	183	60	0	0	20,00	9,10	-2,77
LAW-B45	8	22	7	0	4	0	0	19	5	13	186	50	2	2	23,00	9,00	-2,70
SON68	7	28	5	0	3	0	3	9	0	23	195	53	2	2	23,00	9,00	-2,67
AFCI	12	18	6	0	0	0	2	20	0	15	191	58	0	1	22,70	10,00	-2,66
IDF7-E12	10	18	11	1	0	1	0	11	2	33	174	44	3	2	23,00	8,00	-2,64
ORLEC33	12	21	3	0	0	1	0	0	2	50	182	46	2	3	22,00	9,00	-2,58
ORLEC28	14	20	2	0	1	5	0	0	2	49	182	43	3	3	22,00	9,00	-2,54
IDF7-E12	10	18	11	1	0	1	0	11	2	33	174	44	3	2	23,00	9,00	-2,49
SRL-202	6	13	1	0	10	3	2	20	2	17	189	57	0	1	40,00	8,70	-2,49

Essa abordagem de representação molecular em vetores numéricos é consistente com técnicas de aprendizado de máquina para química, como descritas por Rogers e Hahn[81].

Na segunda fase, a modelagem dos dados foi refinada com a adição de uma nova coluna, onde incluímos os valores de solubilidade para 25°, conforme mostrado na Tabela 3. Definimos 25°C como faixa de temperatura padrão para a modelagem por se tratar de uma condição ambiente relevante e por concentrar o maior volume de dados experimentalmente validados na literatura.

Essa abordagem foi adotada para preservar a estequiometria dos compostos e garantir que as variações de temperatura, um fator crucial na solubilidade, fossem devidamente consideradas. Ao adicionar essa nova

variável, buscamos melhorar a capacidade do software de capturar padrões ocultos nos dados e testar seu desempenho no reconhecimento desses padrões.

Tabela 3: Dados de amostra após a adição de uma nova coluna com os valores calculados em 25° (SOL) para cada composto.

Formula	**0°**	**10°**	**20°**	**30°**	**40°**	**50°**	**60°**	**70°**	**80°**	**90°**	**100°**	**SOL**
Ag2SO4	0,56	0,67	0,78	0,88	0,97		1,13		1,26	1,32	1,39	0,84
AgC2H3O2	0,73	0,89	1,04	1,23	1,43		1,93		2,59			1,12
AgNO3	122	167	216	265	311		440		585	652	733	234,00
Rb2SO4	37,5	42,4	48,1	53,6	58,5		67,5		75,1	78,6	81,8	50,80
Sm(BrO3)3	34,2	47,6	62,54	79	98							70,65
Sr(ClO4)2	233,8	258,7	291,7	327,5	363,9							306,00
Sr(HCO2)2	9,1	9,54	12,7	13,25	14,68	17,83	25		31,9	32,9	34,4	10,82
Sr(IO3)2			0,19								0,35	0,17
Tl2SeO4		2,17	2,8						8,5		10,8	3,27
Tl2SO4	2,73	3,7	4,87	6,16	7,53		11		14,6	16,5	18,4	5,47
TlBr	0,0238	0,032	0,059	0,068	0,097		0,204					0,06
YBr3	63,9		75,1	83,3	87,3	96,1	101		116	123	129,6(95)	79,10
YCl3	77,3	78,1	75,1	79,6	80,8							75,28
Zn(C2H3O2)2	44,5	37,36	30	25,63	20,48						44,6	28,20
Zn(ClO3)2	145	152	200	209	223							204,20
Zn(HCO2)2	3,7	4,3	5,2	6,1	7,4		11,8		21,2	28,8	38	5,61
ZnSO4	41,6	47,2	53,8	61,3	70,5		75,4		71,1		60,5	57,70

A decisão de alterar a modelagem da solubilidade foi motivada pelos resultados insatisfatórios obtidos no primeiro cenário, onde o coeficiente de correlação R foi de apenas 0,09, conforme ilustrado na Figura 2. Este valor indicou uma baixa capacidade preditiva do modelo inicial, sugerindo que a abordagem utilizada não capturava adequadamente as complexas interações entre os elementos químicos e suas contribuições para a solubilidade dos compostos. Diante disso, tornou-se evidente a necessidade de uma revisão na estratégia de modelagem para melhorar a precisão e a confiabilidade das previsões. A partir dos resultados obtidos, optou-se por modificar a abordagem de modelagem, ajustando os métodos de representação dos dados e considerando outras variáveis e características físico-químicas que poderiam influenciar a solubilidade. Essa mudança foi essencial para desenvolver um

modelo mais robusto e capaz de fornecer previsões mais acuradas, refletindo melhor a realidade dos sistemas estudados.

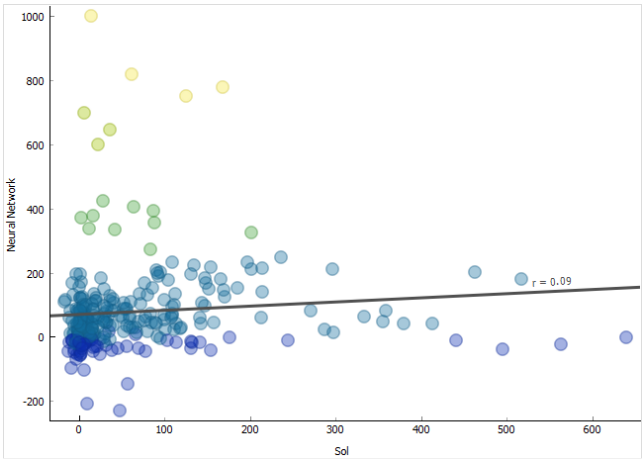

Figura 2: Dados da solubilidade predita em função dos dados experimentais, considerando a modelagem inicial de cada composto representado em uma coluna e a quantidade de cada elemento químico em colunas subsequentes.

4.3 Verificação de padrões e ajuste do algoritmo

O algoritmo de RNA foi otimizado ajustando-se parâmetros como o número de iterações e a quantidade de neurônios na camada oculta. A otimização de hiperparâmetros em RNAs é uma prática comum para melhorar o desempenho do modelo, conforme discutido por Bergstra e Bengio[82]. O processo de otimização foi iterativo, testando diferentes configurações até que o desempenho do modelo não apresentasse mais melhorias significativas.

A etapa de Verificação de Padrões e Ajuste do Algoritmo foi uma parte fundamental da nossa metodologia. Nesta fase, realizamos ajustes minuciosos para otimizar o desempenho dos modelos preditivos, incluindo a combinação e calibração de diversos parâmetros nas redes neurais artificiais (RNAs). Esse processo envolveu a escolha criteriosa do número de neurônios (NEURONS) na camada escondida, a seleção da função de ativação (ACTIVATION) mais adequada, além do ajuste do número máximo de iterações. Essas variáveis foram ajustadas para garantir a máxima precisão e robustez nas previsões.

Durante o processo de ajuste, também monitoramos o erro quadrático médio (MSE) e o coeficiente de determinação (R^2) para avaliar a performance do modelo. Realizamos ajustes iterativos para refinar continuamente a configuração dos parâmetros, assegurando que a melhor configuração fosse utilizada. Essa abordagem permitiu prever separadamente tanto a solubilidade quanto a taxa de dissolução, utilizando a mesma modelagem básica para ambos os tipos de predição, resultando em alta acurácia e robustez nas previsões finais.

Após a escolha da função de ativação, ReLU, procedeu-se à realização de uma nova série de testes com o objetivo de otimizar a performance da rede neural ao analisar o erro quadrático médio (MSE) em diferentes configurações. O MSE é uma métrica fundamental neste tipo de análise, pois mede a média dos quadrados das diferenças entre os valores previstos e os valores reais, permitindo quantificar a precisão do modelo. Quanto menor o MSE, mais preciso é o modelo [76]. A metodologia envolveu a execução de testes com números variados de iterações e neurônios, permitindo uma avaliação abrangente das combinações possíveis. Foram realizadas simulações para cada configuração de parâmetros, e o MSE foi calculado para cada uma delas. Esta abordagem permitiu identificar quais configurações ofereciam o menor MSE, indicando um ajuste mais preciso do modelo aos dados [83].

Esses testes são essenciais para entender como o modelo se comporta sob diferentes circunstâncias e garantir que a rede neural esteja bem ajustada. Através da análise dos valores de MSE obtidos, foi possível determinar a combinação ideal de neurônios e iterações que proporcionou o melhor equilíbrio entre a complexidade do modelo e sua capacidade de generalização, garantindo a eficácia da rede neural em capturar os padrões subjacentes dos dados sem superestimar o ruído [84].

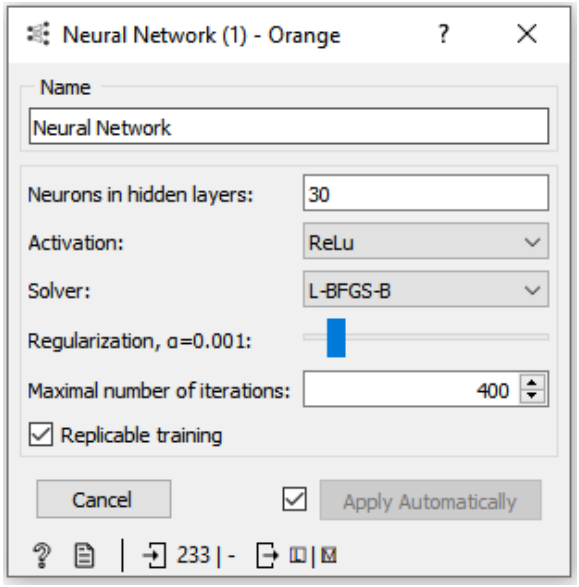

Figura 3: Exemplo de "janela" do *software* Orange, demonstrando alguns dos parâmetros ajustados no processo de refino das predições.

Utilizamos diversas funcionalidades da interface para personalizar a configuração da rede neural, visando otimizar seu desempenho e adaptabilidade. Nomeamos as configurações para facilitar a organização dos experimentos. Ajustamos o número de neurônios nas camadas ocultas para equilibrar a capacidade de aprendizado com a prevenção de *overfitting*. Escolhemos a função de ativação apropriada para introduzir não-linearidade, fundamental para modelar relações complexas nos dados [36], [71].

Selecionamos o algoritmo de otimização (*solver*) mais adequado, o que impactou diretamente a velocidade e eficácia do treinamento. Controlamos a força da regularização através do parâmetro alfa (Figura 3) para evitar o *overfitting*, incentivando a criação de modelos mais simples e generalizáveis. Definimos o número máximo de iterações para determinar o tempo de treinamento e utilizamos a opção de treinamento replicável para garantir a consistência dos resultados em múltiplas execuções [85], [86], [87].

A otimização dos hiperparâmetros foi realizada iterativamente até que o desempenho do modelo se estabilizasse. Para avaliar o desempenho do modelo, utilizamos o *widget Test and Score*. Métricas como o erro quadrático médio (MSE) e o coeficiente de determinação (R^2) foram monitoradas para garantir a precisão das predições.

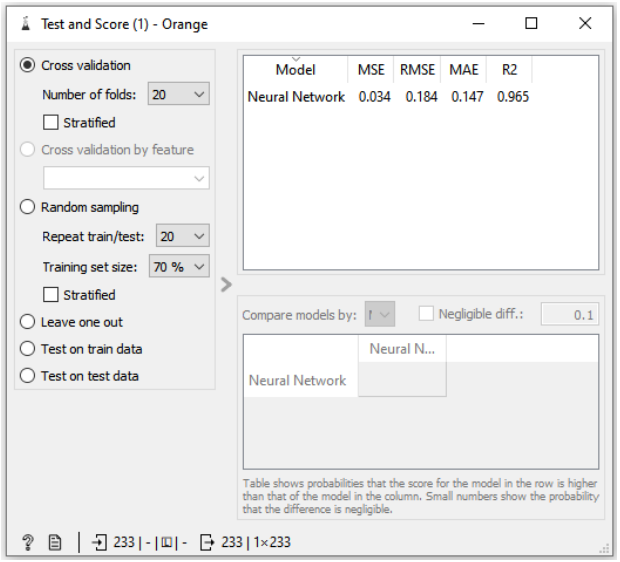

Figura 4: Janela *Test and Score* no Orange para Avaliação de Modelos de *Machine Learning* e Otimização de Hiperparâmetros.

4.4 Separação das Bases de Dados e Predição

Os dados foram separados em Dados de Aprendizado e Dados de Testes (70% para aprendizado e 30% para teste). Como o Orange utiliza o conceito de Canvas, podemos criar, visualizar e analisar modelos de negócio de maneira interativa.

O Canvas é dividido em diferentes componentes que representam os elementos essenciais do modelo de negócio. Esses componentes são representados por blocos que podem ser arrastados e conectados entre si, facilitando a criação de um modelo visual e estruturado, como visto na figura 4:

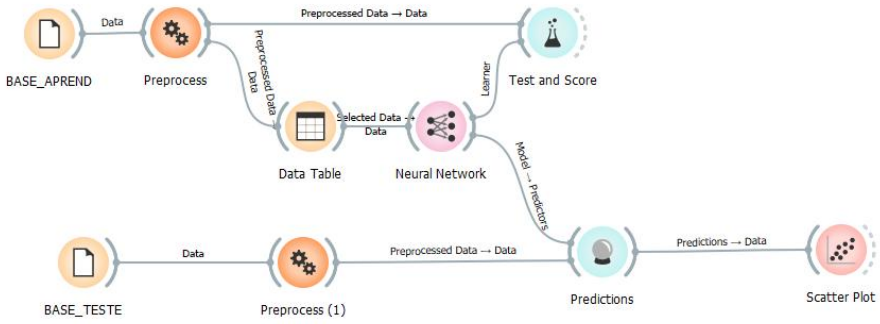

Figura 5: Desenho esquemático da interface do processo de predição da solubilidade.

Na seção apresentada na Figura 5 é possível observar as duas entradas, a de aprendizagem e a de teste. A base para aprendizagem passa por um pré-processamento para ser aplicado o algoritmo de Redes Neurais. Os elementos de *Data Table* (mostra a visualização dos dados após o pré-processamento) e *Test and Score* (permite a visualização das referências de MSE, RMSE, MAE e R^2) possibilitam uma compreensão melhor dos dados de entrada e do processamento. Da mesma forma que os dados de aprendizagem passam por um pré-processamento, os de teste também, se comunicando com o algoritmo (que já aprendeu com os dados de entrada) através do elemento *Predictions*, que mostra, através de uma tabela, um comparativo dos dados “imputados” com os dados preditos.

Os mesmos critérios adotados para a predição da Solubilidade, foram adotados na predição da Taxa de Dissolução, conforme Figura 6.

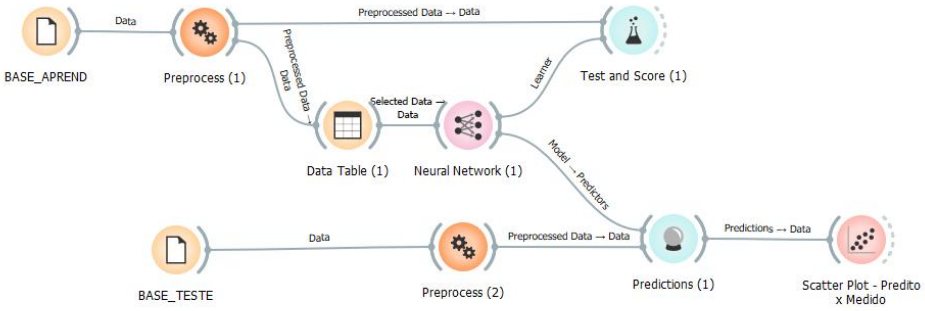

Figura 6: desenho esquemático da interface do processo de predição da Taxa de Dissolução.

Por fim, para entendermos como é feito a predição, no momento do *input* dos dados (aprendizagem e teste), é preciso apontar para a ferramenta qual coluna deverá se comportar como nosso alvo (*target*) para que o Orange possa prever qual será o valor com base no que foi aprendido.

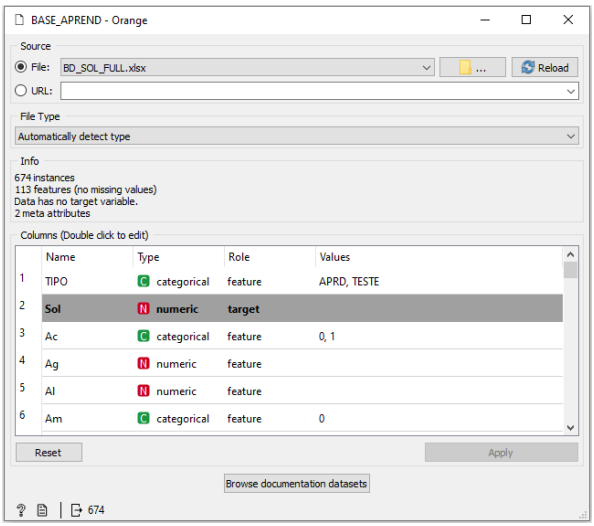

Figura 7: Desenho esquemático da interface do processo de definição de um alvo (*target*).

4.5 Separação e Predição da Base Teste

A análise dos dados de solubilidade e da taxa de dissolução foi aprimorada pela divisão da base em categorias distintas, facilitando a compreensão das nuances e particularidades de cada grupo. Essa estratégia permitiu uma abordagem mais direcionada e eficiente na aplicação das redes neurais, otimizando o aprendizado e a capacidade de predição do modelo para cada faixa de solubilidade. Adicionalmente, a análise individualizada de cada categoria possibilitou a identificação de padrões e relações específicas entre a estrutura química e a solubilidade, informações valiosas para o desenvolvimento de modelos preditivos mais precisos e robustos [84].

A base de dados utilizada para os testes foi cuidadosamente selecionada e dividida em solubilidade baixa (< 25) inferior a 25g/100 ml, média (< 100) com solubilidade entre 25 e 100g/100 ml e alta (≥ 100) com solubilidade igual ou superior a 100g/100ml. Para a taxa de dissolução, foi adotada como dissolução

baixa os valores inferiores a -2 (log(f)), dissolução média com valores entre -2 e 0 e dissolução alta, os valores acima de 0.

A divisão em categorias de solubilidade e taxa de dissolução permite uma análise mais detalhada dos resultados, facilitando a identificação de padrões e tendências relacionados à solubilidade dos compostos. Adicionalmente, a inclusão de compostos com elementos incomuns enriquece a base de dados e contribui para uma compreensão mais abrangente da solubilidade em diferentes classes de substâncias.

Foi adotado o mesmo processo para a predição tanto da solubilidade quanto da taxa de dissolução. Os dados de teste também passaram por um pré-processamento antes de serem aplicados ao algoritmo. Esses dados foram então comunicados com o modelo treinado. Este processo completo assegurou a consistência e precisão das previsões realizadas, seguindo o modelo esquemático apresentado na Figura 5.

5 RESULTADOS E DISCUSSÃO

5.1 Ajuste do algoritmo

Concluídas as etapas de preparação e predição, passamos agora a apresentar e discutir os resultados obtidos com o modelo ajustado. Com base nos experimentos realizados, o melhor conjunto de parâmetros encontrado foi NEURONS (30), ACTIVATION (RELU), Número máximo de iterações (400), MSE (0,034) e R^2 (0,83). Esses resultados indicam um bom equilíbrio entre a complexidade do modelo e a precisão preditiva, demonstrando que a combinação desses parâmetros permite à rede neural capturar de forma eficaz os padrões subjacentes nos dados sem superestimar o ruído. Esses achados corroboram com a literatura existente e oferecem uma base sólida para aplicações práticas em modelagem preditiva com RNAs [88].

Uma amostra da tabela que embasa a explicação do efeito do número de neurônios e da função de ativação ReLU no erro (MSE) e no coeficiente de determinação (R^2). Essa amostra ilustra como a escolha de 30 neurônios, função de ativação ReLU e 400 iterações apresentou o melhor desempenho geral do modelo de rede neural.

Tabela 4: Amostra com resultados obtidos com as simulações dos dados de Solubilidade.

NEURONS	ACTIVATION	MAX_NUM_IT	MSE	R^2
10	IDENTITY	10	3×10^3	0,79
20	IDENTITY	10	5×10^1	0,99
30	IDENTITY	10	1×10^2	0,99
40	IDENTITY	10	3×10^1	0,99
50	IDENTITY	10	1×10^2	0,99
10	LOGISTIC	10	1×10^4	0,01
20	LOGISTIC	10	1×10^4	0,16
30	LOGISTIC	10	1×10^4	0,25
40	LOGISTIC	10	9×10^3	0,35
50	LOGISTIC	10	9×10^3	0,38
10	TANH	10	1×10^4	0,02
20	TANH	2000	2×10^3	0,84
30	TANH	1000	2×10^3	0,86
50	TANH	2000	2×10^3	0,87
100	TANH	2000	2×10^3	0,88
10	RELU	10	1×10^3	0,98
20	RELU	100	2	1,00
30	RELU	335	0,5	1,00
30	RELU	400	0,5	1,00
30	RELU	600	0,5	1,00

Tabela 5: Amostra com resultados obtidos com as simulações dos dados de Taxa de Dissolução.

NEURONS	ACTIVATION	MAX_NUM_IT	MSE	R^2
10	IDENTITY	10	0,05	0,95
20	IDENTITY	10	0,05	0,95
30	IDENTITY	10	0,04	0,96
40	IDENTITY	10	0,04	0,95
50	IDENTITY	10	0,06	0,94
10	LOGISTIC	10	0,08	0,92
20	LOGISTIC	10	0,06	0,94
30	LOGISTIC	10	0,05	0,95
40	LOGISTIC	10	0,06	0,94
50	LOGISTIC	10	0,06	0,94
10	TANH	10	0,1	0,90
20	TANH	2000	0,06	0,94
30	TANH	1000	0,09	0,91
50	TANH	2000	0,07	0,93
100	TANH	2000	0,04	0,96
10	RELU	10	0,2	0,77
20	RELU	100	0,03	0,97
30	RELU	335	0,03	0,97
30	RELU	400	0,03	0,97

A amostra demonstra que, com 30 neurônios e a função de ativação ReLU, os resultados obtidos foram significativamente melhores em termos de MSE e R^2, especialmente quando o número máximo de iterações foi de 400. Para os dados de solubilidade, o modelo alcançou um MSE de 0,5 e um R^2 de 1,00, indicando um excelente ajuste. Observa-se também que o MSE de 0,034 e o R^2 de 0,97 foram os melhores resultados encontrados para os dados de taxa de dissolução, indicando um equilíbrio ideal entre a complexidade do modelo e a precisão preditiva. O valor tão elevado destes valores de R^2 iguais ou próximos a 1, não são ainda completamente compreendidos. Esses valores, em especial dos MSE, suportam a conclusão de que a combinação desses parâmetros permite à rede neural capturar de forma eficaz os padrões subjacentes nos dados sem superestimar o ruído.

A análise dos resultados permite observar o efeito do número de neurônios na camada oculta em relação ao erro (MSE) e ao coeficiente de determinação (R^2) ao utilizar a função de ativação ReLU (Rectified Linear Unit).

A quantidade de neurônios na camada oculta influencia diretamente o desempenho do modelo. Observa-se que, inicialmente, o aumento do número de neurônios tende a reduzir o MSE, indicando uma melhor capacidade do modelo em capturar os padrões dos dados. Por exemplo, com 10 neurônios, o MSE é de 0,022 e o R^2 é de 0,77. Com 20 neurônios, o MSE varia de 0,030 e o R^2 de 0,97. Com 30 neurônios, o MSE varia de 0,031 a 0,038 e o R^2 de 0,97 a 0,96. Esse comportamento mostra que adicionar neurônios melhora a capacidade de aprendizagem até um ponto ótimo, após o qual os ganhos adicionais se tornam marginais ou até prejudiciais devido ao *overfitting*.

Além disso, ao observar o desempenho com 40 ou mais neurônios, há uma estabilização ou aumento no erro (MSE). Por exemplo, com 45 neurônios, o MSE pode aumentar para 0,036 e o R^2 cair para 0,96, sugerindo que a rede neural pode estar sobreajustada aos dados de treinamento, perdendo sua capacidade de generalização para novos dados.

A escolha de 30 neurônios com a função de ativação ReLU e 400 iterações foi justificada pelo balanço ótimo entre a redução do erro e a melhoria na capacidade preditiva do modelo, conforme evidenciado pelos valores de MSE e R^2 na planilha.

Tabela 6: Representação do melhor ajuste para cada ativação (Solubilidade).

NEURONS	ACTIVATION	MAX_NUM_IT	MSE	R^2
20	IDENTITY	10	5 x10	0,99
30	LOGISTIC	10	$1x10^4$	0,25
100	TANH	2000	1×10^3	0,88
30	**RELU**	**400**	**0,5**	**1,00**

Tabela 7: Representação do melhor ajuste para cada ativação (Taxa de Dissolução).

NEURONS	ACTIVATION	MAX_NUM_IT	MSE	R^2
30	IDENTITY	10	0,04	0,96
30	LOGISTIC	10	0,05	0,95
100	TANH	2000	0,04	0,96
30	**RELU**	**400**	**0,03**	**0,97**

Inicialmente, a quantidade máxima de iterações foi testada em sete níveis distintos: 10, 20, 30, 40, 50, 100 e 200. Para cada uma dessas configurações de iteração, a quantidade de neurônios na camada oculta foi variada, começando com valores de 1 a 30 em incrementos de 1. Após esta faixa inicial, os testes continuaram com incrementos de 5, abrangendo valores de 35 até 150 neurônios. Este procedimento sistemático garantiu uma cobertura ampla das possíveis combinações, permitindo uma análise detalhada do impacto de cada configuração no MSE.

Os resultados desses experimentos foram fundamentais para a identificação das configurações que proporcionaram o menor MSE. A análise foi conduzida por meio da geração de gráficos que ilustravam a performance de diferentes combinações de neurônios e iterações, permitindo visualizar de forma clara as tendências de desempenho e, assim, facilitar a identificação das configurações mais adequadas. Observou-se que, à medida que o número de iterações aumentava, a performance da rede neural melhorava até certo ponto, após o qual os ganhos adicionais tornavam-se marginais.

Além disso, a variação no número de neurônios da camada oculta revelou um papel crucial no equilíbrio entre subajuste (*underfitting*) e sobreajuste (*overfitting*). Um número muito reduzido de neurônios resultava em modelos incapazes de capturar a complexidade dos dados, enquanto valores excessivos conduziam à superparametrização da rede, elevando o MSE. Essa relação se expressou no comportamento em formato de curva U, o que possibilitou identificar claramente o número ótimo de neurônios para a condição avaliada.

De maneira semelhante, os experimentos também demonstraram que o aumento indiscriminado do número de camadas ocultas não implica, necessariamente, em melhor desempenho. Embora múltiplas camadas possam ampliar a capacidade de representação da rede, observou-se um comportamento mais nuançado, em que o aumento da complexidade levava a estados de superparametrização. Nesse contexto, foi evidenciado o chamado regime moderno de desempenho, em que a rede escapa da curva clássica em “U” e mantém o MSE em níveis baixos e estáveis, coerente com o fenômeno do *double descent* descrito na literatura.

Essa abordagem metódica e detalhada está alinhada às práticas recomendadas em aprendizado de máquina. Estudos de Bengio [82] e LeCun e colaboradores [75] ressaltam a importância de explorar sistematicamente diferentes configurações de hiperparâmetros para otimizar o desempenho das redes neurais. Da mesma forma, a utilização de visualizações gráficas para avaliar a performance experimental, conforme sugerido por Brownlee [89], constitui uma prática consolidada e eficaz para a análise e compreensão dos resultados.

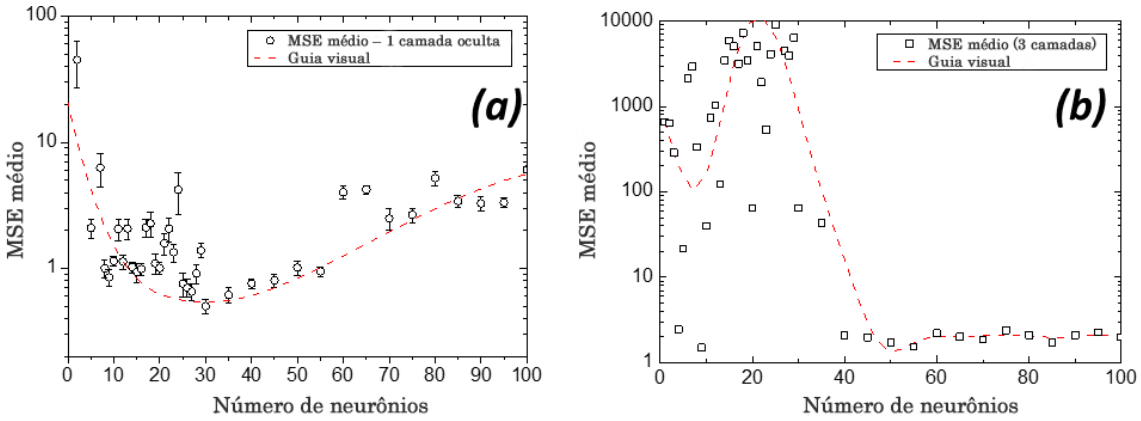

Figura 8: Efeito no MSE da RNA para os dados de solubilidade em função do número de neurônios na camada oculta. (a) Regime clássico com a curva em formato de “U” esperada para 1 camada oculta. (b) Com 3 camadas ocultas, observa-se o comportamento do regime moderno, ilustrando a transição do estado subparametrizado para o superparametrizado, em conformidade com o fenômeno do *double descente* [89].

Embora esse tipo de análise não possa ser realizada diretamente no software Orange, os experimentos podem ser produzidos e os resultados extraídos de forma simples, podendo ser concluídos em uma única aula, sob a supervisão dos professores. Cada execução leva apenas alguns segundos ou minutos para ser concluída em um computador convencional. Dessa forma, nossos resultados mostram que a escolha de 30 neurônios e da função de ativação ReLU resulta em uma configuração significativamente melhor para as RNAs de solubilidade e dissolução, com 1 camada oculta, em termos de MSE e R^2.

A função ReLU, nesses casos, com sua definição matemática simples e vantagens computacionais, contribui de forma significativa para o desempenho do modelo, auxiliando na captura de padrões complexos nos dados sem sofrer com problemas associados ao desaparecimento do gradiente (gradient vanishing) [90]. Esses achados são consistentes com a literatura existente e fornecem uma base sólida para aplicações práticas em modelagem preditiva com RNAs [91].

Outro aspecto interessante que pode ser verificado na RNA é o número de iterações. Como mencionado anteriormente, esse parâmetro está intimamente relacionado à escolha do solucionador (solver). O solucionador L-BFGS-B pode ser considerado como um otimizador em batch, o que significa que ele avalia todo o conjunto de dados de treinamento para calcular a função de perda e o gradiente em cada iteração (ou passo). Dessa forma, para RNAs simples (1 camada oculta), geralmente converge rapidamente para uma boa condição de treinamento. Esse pode ser o motivo da baixa sensibilidade observada em relação ao parâmetro de iteração, dentro das configurações utilizadas.

Por outro lado, o solucionador ADAM trabalha utilizando pequenos subconjuntos aleatórios de dados (mini-batches) em cada iteração. Por esse motivo, o ADAM pode ser mais sensível ao número de iterações em RNAs simples, geralmente seguindo o número de iterações definido pelo usuário.

Uma forma simples de verificar a influência do parâmetro de iteração é utilizar o banco de dados de teste para a previsão de seus valores. A Figura 9 ilustra a influência do número de iterações para os dois solucionadores diferentes (L-BFGS-B e ADAM), em uma RNA com 1 camada oculta e 30 neurônios, mas com valores de iteração iguais a 10 e 100. As Figuras 9a e 9b ilustram o efeito para o solucionador L-BFGS-B. É possível observar que a diferença entre 10 e 100 iterações é quase imperceptível para os valores previstos, exceto por uma pequena dispersão nos valores de baixa solubilidade, indicando uma possível convergência após poucas iterações.

Por outro lado, para o solucionador ADAM, ambas as condições apresentaram resultados bastante insatisfatórios em termos de valores preditos.

Embora com 100 iterações seja perceptível uma melhora na previsão, a RNA com solver ADAM exigiria um número ainda maior de passos de iteração para alcançar a convergência, nesse caso.

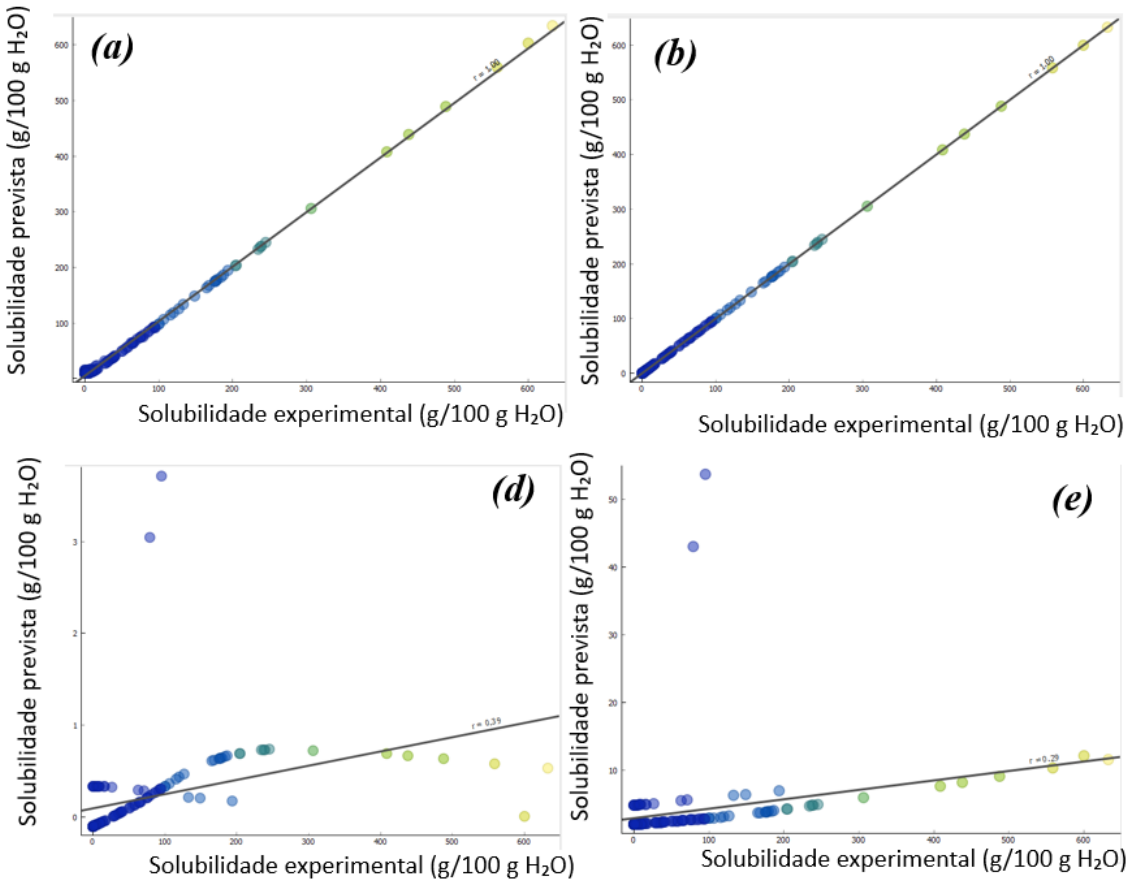

Figura 9: Saída dos valores previstos versus valores experimentais para a rede neural de solubilidade, considerando 30 neurônios em uma única camada oculta. (a) Solucionador L-BFGS-B com 10 iterações, (b) Solucionador L-BFGS-B com 100 iterações, (c) Solucionador ADAM com 10 iterações e (d) Solucionador ADAM com 100 iterações.

A modificação desses parâmetros em tempo real permite observar diretamente os efeitos dessas mudanças nos resultados da predição. Esta abordagem interativa não só facilita a compreensão dos conceitos teóricos, mas também destaca a importância de ajustar corretamente os parâmetros para obter modelos eficazes.

A utilização do Orange permite uma visualização imediata dos impactos, reforçando a aprendizagem ativa e promovendo um ambiente de exploração e

experimentação. É possível ver como diferentes configurações levam a diferentes resultados, compreendendo melhor a sensibilidade dos modelos de machine learning às suas configurações internas.

Ensinar machine learning através de uma abordagem prática e interativa, como a oferecida pelo Orange, torna a aprendizagem mais eficaz e envolvente. Alterar parâmetros de redes neurais e observar seus efeitos em tempo real proporciona uma compreensão profunda de como esses modelos funcionam e como diferentes configurações podem impactar os resultados. Este método não só melhora a assimilação dos conceitos teóricos, mas também prepara para enfrentar desafios reais no campo da ciência de dados.

A experiência de aprendizado proporcionada pelo Orange vai além da teoria e se traduz em aplicações práticas na construção de modelos preditivos. A interface intuitiva do software e suas ferramentas de visualização interativa, que facilitam a compreensão dos modelos de aprendizado de máquina, foram essenciais para o desenvolvimento e avaliação dos modelos de predição de taxa de dissolução e solubilidade. A capacidade de manipular parâmetros em tempo real e observar seus efeitos, como demonstrado no treinamento da rede neural para a taxa de dissolução e na divisão dos dados para a solubilidade, reflete a abordagem prática e interativa que o Orange oferece.

5.2 Predição dos Dados

O software permite a visualização interativa e a análise de dados, facilitando a compreensão dos modelos de aprendizado de máquina. Tanto no caso da taxa de dissolução como Solubilidade, os dados foram divididos em dois conjuntos: um para treinamento e outro para teste. Este procedimento permitiu avaliar a capacidade do modelo de generalizar para novos dados.

A plataforma busca simplificar a análise de dados, permitindo que usuários explorem suas informações sem a necessidade de configurações complicadas ou conhecimentos avançados de programação. Como algoritmo de predição, foi utilizado *Neural Network*, tanto para predição de solubilidade, como para a taxa de dissolução.

A Figura 10 ilustra a dispersão comparando os valores preditos pela rede neural com os valores observados da taxa de dissolução. O eixo vertical (*Neural Network*) mostra os valores preditos pelo modelo de rede neural, enquanto o eixo horizontal (Taxa Dissolução) exibe os valores observados[36].

Na Figura 10, cada ponto representa uma amostra do conjunto de dados utilizado para a validação do modelo. A linha diagonal representa a linha de melhor ajuste (y = x), onde as predições seriam perfeitas. A proximidade dos pontos a esta linha indica a precisão do modelo: quanto mais próximos os pontos estiverem da linha, melhor é o desempenho preditivo do modelo[71].

O valor de R^2 obtido utilizando o modelo de rede neural no Orange foi 0.99, conforme ilustrado na Figura 10. Este valor é um indicador de quão bem os dados preditos correspondem aos dados observados. Um R^2 de 0.99 sugere que 99% da variação nos dados observados pode ser explicada pelo modelo preditivo, indicando um excelente ajuste[85].

Os resultados obtidos utilizando a metodologia adotada nesta dissertação estão compatíveis com resultados obtidos por Vienna e colaboradores[52], com um R^2 superior a 0.9. Embora este também seja um valor alto, o resultado obtido com o Orange demonstra uma ligeira melhoria na precisão da predição.

O estudo comparativo com os resultados de Vienna e colaboradores [52] destaca a eficácia do uso de técnicas avançadas de aprendizado de máquina, como redes neurais, na modelagem de propriedades complexas de materiais. O coeficiente de determinação (R^2 = 0.99) ao usar o Orange demonstra a capacidade das redes neurais em capturar relações não lineares nos dados, superando métodos tradicionais[92].

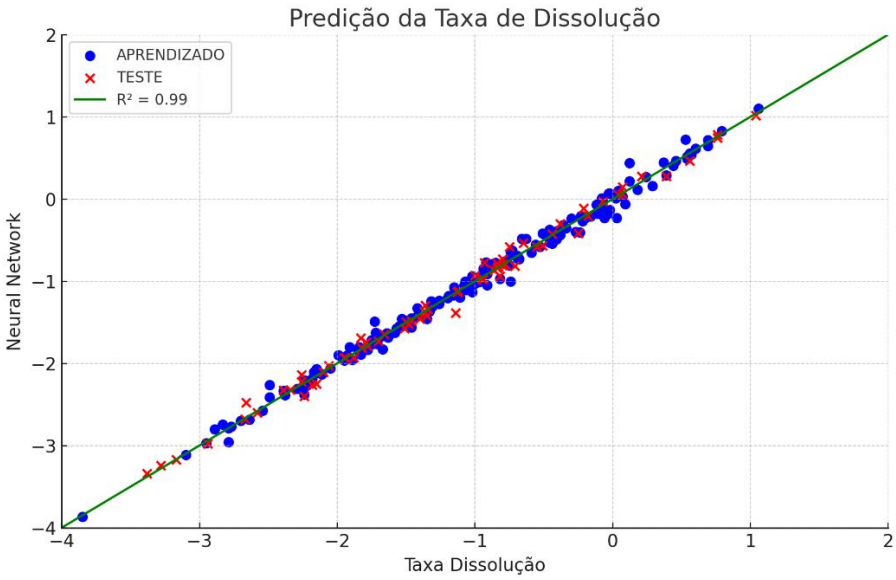

Figura 10: Dados da taxa de dissolução predita em função dos dados experimentais, ilustrando em vermelho os dados referentes a predição da base de teste e em azul os dados da base de aprendizado.

A figura 10 representa um gráfico de dispersão que compara os valores preditos pela rede neural com os valores observados da taxa de dissolução. O eixo vertical (*Neural Network*) exibe os valores preditos pelo modelo de rede neural, enquanto o eixo vertical (Taxa Dissolução) mostra os valores observados da taxa de dissolução.

Na Figura 11, cada ponto representa uma amostra do conjunto de dados utilizado para a validação do modelo. A linha diagonal representa a linha de melhor ajuste (y = x), onde as predições seriam perfeitas. A proximidade dos pontos a esta linha indica a precisão do modelo: quanto mais próximos os pontos estiverem da linha, melhor é o desempenho preditivo do modelo[36], [71].

A solubilidade é uma propriedade fundamental em diversas aplicações, como na indústria farmacêutica, onde a solubilidade de um composto pode influenciar sua biodisponibilidade e eficácia. A capacidade de prever com precisão a solubilidade a partir da composição química dos materiais pode acelerar o desenvolvimento de novos fármacos e materiais, reduzindo a necessidade de testes experimentais extensivos[85], [93].

A utilização de redes neurais para a predição da solubilidade demonstra o potencial destas técnicas avançadas de aprendizado de máquina em modelar

propriedades complexas de materiais. O coeficiente de determinação ($R^2 = 0.89$) indica que o modelo de rede neural é capaz de capturar de maneira eficiente as relações entre a composição dos materiais e sua solubilidade[92].

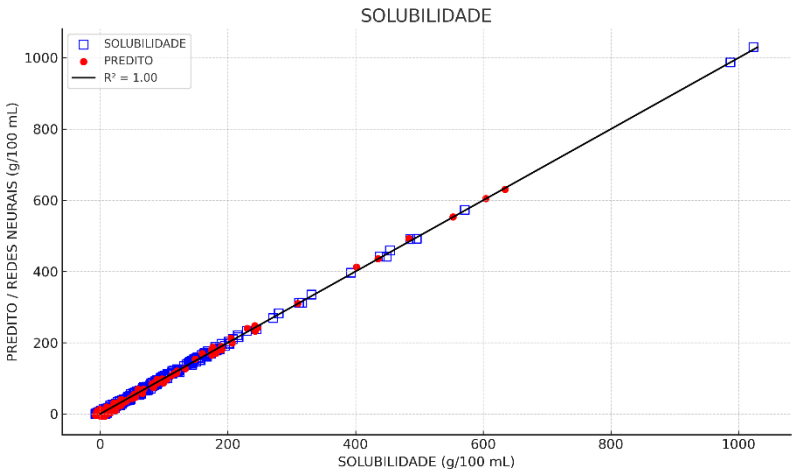

Figura 11: Dados da solubilidade predita em função dos dados experimentais, ilustrando em vermelho os dados referentes a predição da base de teste e em azul os dados da base de aprendizado.

Os resultados obtidos através da implementação das Redes Neurais Artificiais (RNAs) e do uso do Orange para a predição da solubilidade e taxa de dissolução dos compostos óxidos em meio aquoso demonstram a eficácia da abordagem proposta neste trabalho. A aplicação dessas técnicas avançadas de mineração de dados e aprendizado de máquina permitiu não apenas a modelagem precisa dessas propriedades, mas também ofereceu *insights* valiosos sobre a relação entre a composição química dos materiais e suas propriedades de dissolução.

A escolha de utilizar a composição do material como único dado de entrada para a predição da solubilidade e taxa de dissolução se mostrou eficiente, simplificando significativamente o processo de previsão e tornando-o mais acessível. Essa abordagem inovadora difere das metodologias tradicionais que frequentemente requerem uma variedade de dados de entrada, como propriedades físico-químicas e dados espectroscópicos, destacando a capacidade das RNAs de capturar relações complexas a partir de um conjunto limitado de dados.

A utilização do Orange como ferramenta de mineração de dados e aprendizado de máquina facilitou significativamente a análise de dados e a construção de modelos preditivos. Sua interface gráfica de usuário intuitiva e o processo simplificado de arrastar e soltar permitiram a implementação de fluxos de trabalho complexos de mineração de dados sem a necessidade de conhecimentos avançados de programação. Além disso, os recursos de visualização de dados do Orange proporcionaram uma compreensão mais profunda dos modelos de aprendizado de máquina e dos resultados obtidos. Outra diferença fundamental vista na literatura está em utilizar linguagens de programação como Python ou R, que, apesar de poderosas e flexíveis, apresentam uma curva de aprendizado mais acentuada e podem ser menos acessíveis para pesquisadores sem formação em programação.

A comparação entre os métodos de aquisição de dados utilizados neste trabalho e outros métodos existentes na literatura revela várias diferenças significativas e implicações para a pesquisa e aplicação prática em ciência e engenharia de materiais. O enfoque deste estudo na utilização da composição do material como único dado de entrada para a predição de solubilidade e taxa de dissolução representa uma simplificação significativa em comparação com abordagens mais tradicionais, que frequentemente recorrem a uma ampla gama de dados de entrada, incluindo propriedades físico-químicas, dados espectroscópicos e até mesmo dados de difração de raios-X.

O estudo conduzido por Vienna e colaboradores [52] e o presente trabalho compartilham o objetivo de compreender a influência da composição do vidro na taxa de dissolução. No entanto, as abordagens metodológicas adotadas em cada pesquisa diferem significativamente, resultando em perspectivas complementares sobre o problema.

Estamos propondo um enfoque metodológico distinto ao utilizar exclusivamente a composição do vidro como dado de entrada para a análise das taxas de dissolução. Ao compilar e reproduzir dados de diversas composições de vidros nucleares, utilizando a mesma base de dados compilada como referência, esta pesquisa visa isolar e quantificar a contribuição específica de cada componente da composição do vidro na taxa de dissolução.

Diferentemente dos autores, que ajustaram os dados de dissolução combinados a um modelo geral que inclui pH e temperatura, este trabalho se concentra em desenvolver um modelo preditivo baseado apenas na composição, desconsiderando variáveis ambientais.

Os resultados obtidos a partir desta metodologia simples revelaram que, mesmo ao utilizar exclusivamente a composição como variável de entrada, os resultados foram compatíveis com os obtidos por Vienna e colaboradores [52]. Isso demonstra que a composição do vidro, quando analisada isoladamente, pode fornecer *insights* significativos sobre a taxa de dissolução. A simplicidade da abordagem facilita sua aplicação em sala de aula, por exemplo, onde os alunos podem replicar os experimentos e compreender a importância de cada componente da composição do vidro.

A fim de manter a consistência e facilitar a comparação com os dados, a mesma nomenclatura foi mantida no presente estudo. Essa escolha permite que os resultados sejam facilmente interpretados e comparados, destacando como metodologias diferentes podem levar a resultados semelhantes. A manutenção da nomenclatura, além de simplificar a comunicação dos resultados, reforça a relevância dos dados compilados e a robustez dos modelos utilizados.

Assim, ao comparar as duas abordagens, é possível observar que, enquanto Vienna e colaboradores [52] destacaram a dominância do pH e da temperatura sobre a composição, este trabalho enfatiza a potencial capacidade da composição de ser um fator preditivo autônomo.

Tabela 8: Comparação dos Valores de Solubilidade Experimental e Previsões da Rede Neural com Seus Respectivos Erros e Desvios.

Composto	**Solubilidade**	**Neural Network**	**Desvio**	**Referência**
$Zn(IO_3)_2$	0,64	0,859142	34%	[94]
$Tl_2C_2O_4$	1,90	2,0635	8%	[95]
$RbClO_3$	6,63	6,6359	1%	[95]
$Sr(HCO_2)_2$	10,82	10,6863	-1%	[94], [95]
$Yb_2(SO_4)_3$	26,72	26,7244	0%	[95]
Na_2SeO_4	58,5	58,2392	0%	[94]
YBr_3	79,1	78,1695	-1%	[94], [95]
AgF	179,5	180,023	0%	[95]

$Zn(NO_3)_2$	120	120,306	0%	[94]
$Na_2Cr_2O_7$	187	187,53	0%	[94]
$Sr(ClO_4)_2$	306	306,432	0%	[94]
$RbNO_3$	65	64,8174	0%	[94]

Tabela 9: Comparação dos Valores da Taxa de Dissolução Experimental e Previsões da Rede Neural com Seus Respectivos Erros e Desvios.

Composto	**Taxa Dissolução**	**Neural Network**	**Desvio**	**Referência**
IDF21-EC14	-3,38	-3,34	-1%	[52], [96]
LD6-5412	-3,28	-3,25	-1%	[52], [97]
SRL-202	-3,17	-3,17	0%	[52], [98]
ORLEC33	-2,58	-2,60	1%	[52], [99]
LD6-5412	-1,50	-1,50	0%	[52], [97]
IDF7-E12	-1,13	-1,13	0%	[52], [100]
IDF21-EC14	-1,12	-1,12	0%	[52], [96]
SON68	-0,84	-0,84	0%	[52], [101]
IDF7-E12	-0,80	-0,81	1%	[52], [100]
IDF1-B2	0,76	0,78	3%	[52], [100]
SRL-202	0,76	0,74	-2%	[52], [98]
ORLEC33	1,04	1,02	-2%	[52], [99]

Tabela 10: Composição dos vidros – Taxa de Dissolução (mol% de óxidos e halogênios).

Vidros	IDF1-B2	IDF21-EC14	IDF7-E12	LD6-5412	ORLEC33	SON68	SRL-202
Al_2O_3	6.65	6.68	4.75	7.54	6.08	3.39	3.09
B_2O_3	7.11	9.78	9.00	4.60	10.61	14.02	6.62
CaO	1.33	2.37	11.40	4.57	2.99	5.01	1.31
Cl	0.21	0.38	0.04	0.63	0.38	0.00	0.00
F	1.75	0.29	0.67	0.98	0.28	0.00	0.00
Fe_2O_3	0.47	0.14	0.10	0.00	0.08	1.31	5.18
K_2O	0.09	0.36	0.37	0.99	0.36	0.00	1.41
LN_2O_3[a]	0.15	0.07	0.05	0.09	0.07	1.45	1.00
Li_2O	0.00	0.00	5.32	0.00	0.00	4.60	10.00
MgO	1.85	1.69	1.65	0.00	1.67	0.00	2.47
MnO	0.06	0.00	0.00	0.00	0.00	0.38	3.04
MoO_3	0.00	0.00	0.00	0.07	0.00	0.85	0.00
Na_2O	27.35	26.37	16.48	20.69	24.93	11.39	8.42
SO_3	0.44	0.51	1.20	0.17	0.84	0.00	0.00
SiO_2	45.00	43.87	43.76	59.65	45.66	52.69	56.56
SnO_2	0.49	1.05	0.00	0.00	0.00	0.01	0.00
TiO_2	0.00	0.29	0.00	0.00	0.00	0.00	0.20

ZnO	3.04	2.51	2.52	0.00	2.48	2.15	0.07
ZrO_2	2.99	3.33	1.83	0.00	2.74	1.54	0.64
Outros[b]	1.03	0.30	0.87	0.02	0.83	1.22	0.00

[a] LN_2O_3: é a soma dos óxidos de terras raras, incluindo Ce_2O_3, Eu_2O_3, Gd_2O_3, La_2O_3, Nd_2O_3, Pr_2O_3, Sm_2O_3 e Y_2O_3.

[b] Outros: é a soma dos componentes com <1,0% em massa em qualquer vidro, que inclui Ag_2O, BaO, CdO, Cr_2O_3, Cs_2O, NiO, P_2O_5, PbO, PdO, Rb_2O, Re_2O7, Rh_2O_3, RuO_2, SeO_2, SrO, TeO_2 e V_2O_5.

A solubilidade de compostos químicos é uma propriedade fundamental em diversas áreas, como química, farmácia e ciência dos materiais. No entanto, a vasta gama de compostos existentes e a diversidade de condições experimentais dificultam a obtenção de dados de solubilidade abrangentes e confiáveis. Adicionalmente, muitos estudos se concentram em compostos comuns, negligenciando aqueles com elementos incomuns ou propriedades únicas.

Diante desse cenário, este trabalho buscou investigar a solubilidade de uma variedade de compostos químicos, incluindo aqueles com elementos incomuns, a fim de ampliar o conhecimento sobre essa propriedade e fornecer dados relevantes para futuras pesquisas e aplicações. Para tanto, foram realizados testes de solubilidade em condições controladas, utilizando uma metodologia rigorosa e consistente.

Para analisar e comparar os dados das diferentes bases de solubilidade, empregamos redes neurais artificiais (RNAs) como ferramenta de modelagem e predição [76]. Essa abordagem computacional nos permitiu explorar as relações entre a estrutura química dos compostos e suas respectivas solubilidades, buscando identificar padrões e tendências que auxiliassem na compreensão e previsão dessa propriedade. Para avaliar a força e a direção da relação linear entre as características químicas dos compostos e suas solubilidades, utilizamos o coeficiente de correlação (r) [84]. Essa métrica nos forneceu informações valiosas sobre a qualidade do ajuste dos modelos de RNA, indicando se as características químicas dos compostos são boas preditoras de sua solubilidade.

Um coeficiente de correlação próximo de +1 indica uma forte correlação positiva, sugerindo que as características químicas analisadas influenciam

significativamente a solubilidade. Por outro lado, um valor próximo de -1 indica uma forte correlação negativa, enquanto um valor próximo de 0 sugere uma correlação fraca ou inexistente [102]. Ao analisar os valores de r obtidos, pudemos identificar quais características químicas são mais relevantes para explicar a solubilidade dos compostos, permitindo-nos construir modelos preditivos mais precisos e aprimorar nossa compreensão dos mecanismos que governam essa propriedade [76].

5.3 Predição da Base Teste: baixa, média e alta

Para contextualizar e exemplificar o impacto das redes neurais na predição de solubilidade de compostos, após a predição inicial, realizamos uma nova predição, desta vez separando a base de teste em categorias de baixa, média e alta tanto para solubilidade quanto para taxa de dissolução. A seguir, apresentamos gráficos com dados de teste que não foram utilizados durante o treinamento da rede neural, permitindo avaliar a capacidade do modelo em generalizar para novos dados. Os gráficos incluem os valores experimentais de solubilidade, as previsões da rede neural e o desvio observado para compostos classificados como de baixa, média e alta solubilidade.

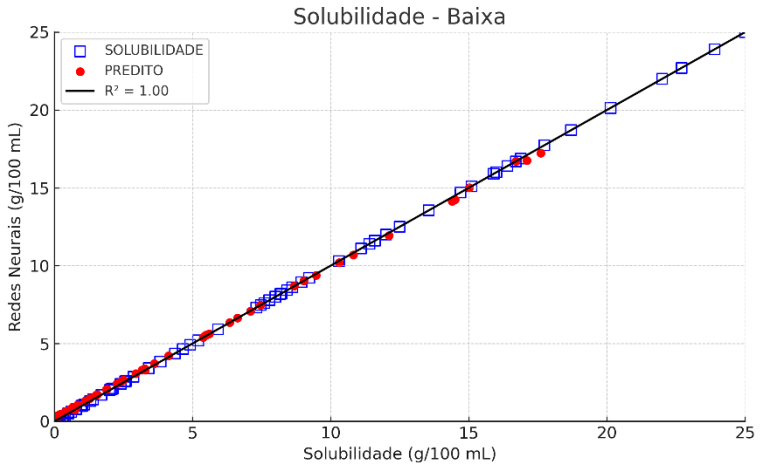

Figura 12: Gráfico da predição da Solubilidade baixa.

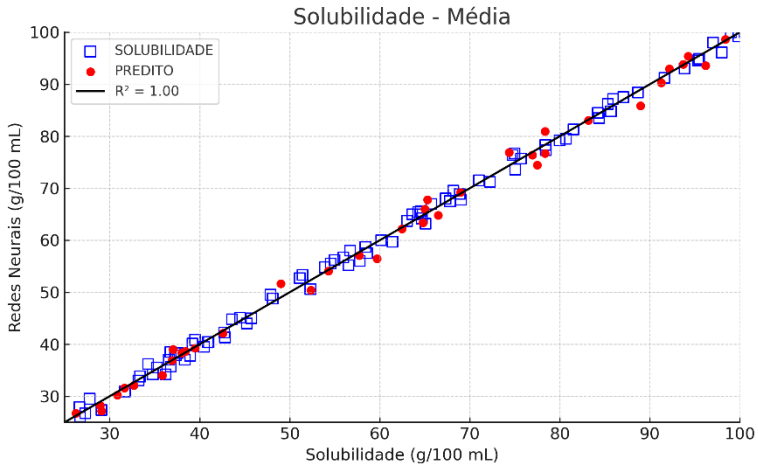

Figura 13: Gráfico da predição da Solubilidade média.

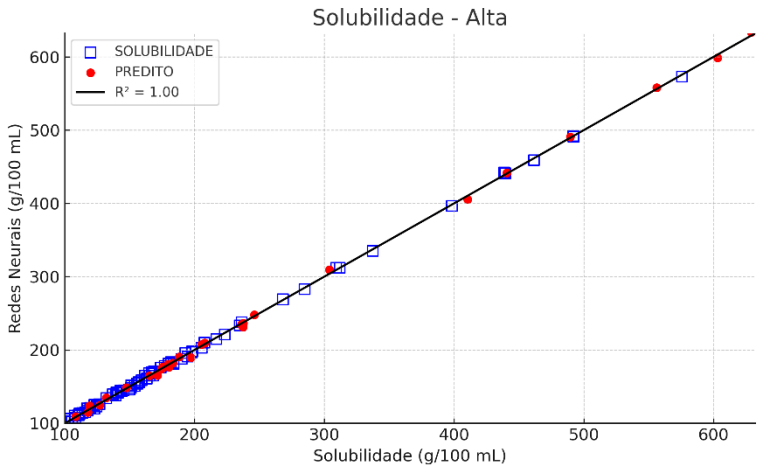

Figura 14: Gráfico da predição da Solubilidade alta.

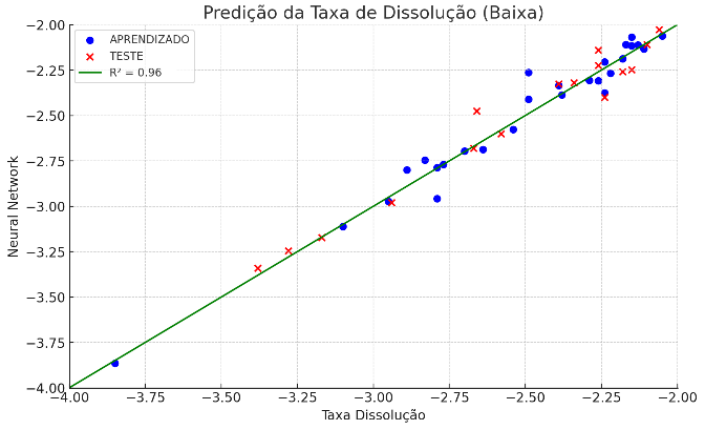

Figura 15: Gráfico da predição da Taxa de Dissolução baixa.

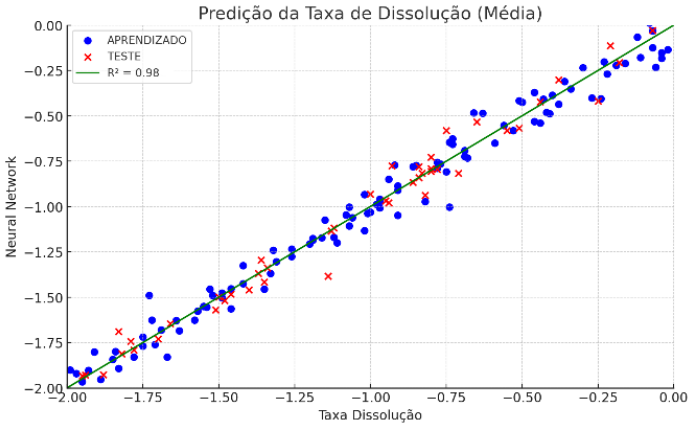

Figura 16: Gráfico da predição da Taxa de Dissolução média.

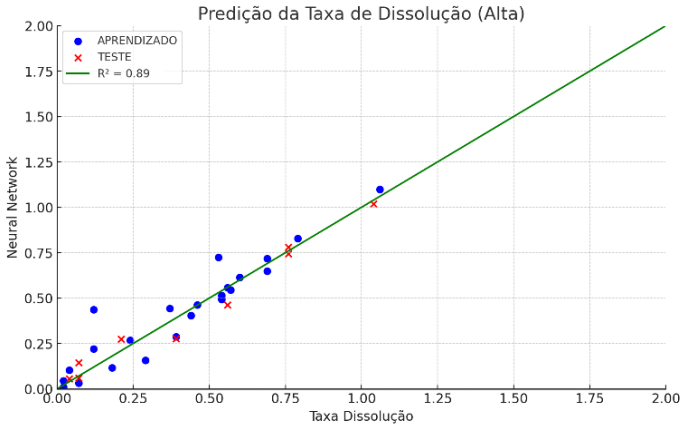

Figura 17: Gráfico da predição da Taxa de Dissolução alta.

Os testes realizados mostraram que a predição das propriedades dos compostos, utilizando redes neurais no Orange, permaneceu consistente, mesmo com diferentes formas estequiométricas, como KCl e K_2Cl_2, demonstrando que essas variações não influenciaram significativamente os resultados.

Tabela 11: Predição para diferentes estequiometrias.

Composto	Neural Network
K_2Cl_2	34,9609
K_4Cl_4	34,9609
K_6Cl_6	34,9609
$K_{12}Cl_{12}$	34,9609
$K_{40}Cl_{40}$	34,9609

Essa robustez do modelo preditivo é uma indicação da capacidade das redes neurais de generalizar bem a partir dos dados fornecidos, capturando as características fundamentais que influenciam as propriedades dos compostos. Isso é particularmente importante na ciência dos materiais e na química, onde pequenas variações na composição podem levar a mudanças significativas no comportamento do material [85], [93].

A capacidade de prever com precisão as propriedades dos compostos sem que a estequiometria específica afete os resultados tem várias implicações práticas. Isso simplifica o processo de modelagem e análise, permitindo que os pesquisadores usem representações estequiométricas equivalentes sem comprometer a precisão das predições. Essa abordagem pode ser aplicada em diversas áreas, incluindo o desenvolvimento de novos materiais, a formulação de medicamentos e a engenharia química [92].

A atribuição exata de compostos no material vítreo apresenta desafios significativos devido à sua composição complexa e multifacetada. Por exemplo, a soma dos óxidos de terras raras, incluindo Ce_2O_3, Eu_2O_3, Gd_2O_3, La_2O_3, Nd_2O_3, Pr_2O_3, Sm_2O_3 e Y_2O_3, é representada como LN_2O_3. Além disso, há componentes que, embora presentes em quantidades menores que 1,0% em massa, somam-se para formar um grupo conhecido como 'Others', que inclui Ag_2O, BaO, CdO, Cr_2O_3, Cs_2O, NiO, P_2O_5, PbO, PdO, Rb_2O, Re_2O_7, Rh_2O_3, RuO_2, SeO_2, SrO, TeO_2 e V_2O_5. Esta complexidade na composição química do vidro torna desafiador determinar a contribuição exata de cada elemento ou composto individual, complicando a análise e a caracterização do material.

6 CONCLUSÃO

O trabalho demonstrou a viabilidade da utilização da ferramenta *open source* Orange Data Mining para a predição da solubilidade e da taxa de dissolução de compostos óxidos em meio aquoso. A proposta do trabalho, que considera apenas a composição química dos materiais como entrada, simplificou significativamente o processo de previsão e mostrou-se eficaz na captura das relações complexas entre a composição dos materiais e suas propriedades.

A aplicação de Redes Neurais Artificiais (RNA) se mostrou particularmente eficaz, devido à sua capacidade de modelar relações não lineares e complexas entre as variáveis. Os modelos desenvolvidos apresentaram um bom desempenho preditivo, evidenciado por métricas como o coeficiente de determinação (R^2) e o erro quadrático médio (MSE). Em especial, a configuração com 30 neurônios na camada oculta e a função de ativação ReLU apresentou os melhores resultados, equilibrando a precisão preditiva e evitando o sobreajuste.

Os resultados indicam que manter a fórmula estrutura, preservando a estequiometria, e a adição da média da solubilidade melhorou a capacidade preditiva do modelo. Além disso, a robustez do modelo preditivo, independentemente das variações estequiométricas dos compostos, demonstra que as RNAs conseguem generalizar eficientemente a partir dos dados fornecidos.

A utilização do Orange Data Mining facilitou a implementação e a análise dos modelos preditivos, graças à sua interface intuitiva e às ferramentas de visualização de dados. A possibilidade de ajuste em tempo real dos parâmetros das redes neurais e a visualização dos efeitos dessas mudanças nos resultados da predição proporcionaram uma compreensão aprofundada dos modelos e de seus comportamentos.

Este trabalho contribui para a literatura ao apresentar uma metodologia simplificada e eficaz para a predição de propriedades de materiais, utilizando técnicas de mineração de dados e aprendizado de máquina. Os achados têm implicações práticas significativas, especialmente na indústria farmacêutica e na

gestão de resíduos radioativos, onde a solubilidade e a taxa de dissolução dos materiais são de extrema importância.

Os objetivos propostos foram atingidos. A viabilidade da determinação da dissolução em meio aquoso de compostos óxidos utilizando a ferramenta Orange Data Mining foi comprovada, e os resultados obtidos mostraram que a metodologia é eficiente e pode ser aplicada em diferentes contextos de análise de materiais. A simplificação do processo de previsão e a eficácia dos modelos preditivos desenvolvidos confirmam a utilidade e a aplicabilidade da abordagem proposta.

7 SUGESTÕES PARA FUTUROS TRABALHOS

As possibilidades futuras para o uso da ferramenta Orange Data Mining em estudos de predição são vastas e promissoras, abrangendo diversas áreas do conhecimento e aplicações práticas. A seguir, destacamos algumas das direções potenciais para a exploração e desenvolvimento futuro desta poderosa ferramenta de mineração de dados e aprendizado de máquina.

O Orange Data Mining pode ser ampliado para incluir uma gama ainda maior de algoritmos de aprendizado de máquina, tanto clássicos quanto modernos, como redes neurais convolucionais (CNNs), redes neurais recorrentes (RNNs), e modelos de aprendizado profundo (*deep learning*). A incorporação desses algoritmos permitirá a análise e predição de dados complexos, como imagens, séries temporais e dados textuais.

Para tirar o máximo proveito do Orange, é recomendado seguir algumas melhores práticas, como manter os dados organizados, testar diferentes algoritmos e ajustar parâmetros para otimizar os resultados. A experimentação com diferentes combinações de *widgets* e a análise dos resultados podem ajudar a identificar as melhores abordagens para cada tipo de problema. Além disso, a documentação detalhada e os tutoriais disponíveis na comunidade do Orange são recursos valiosos para aprender e implementar novos métodos de análise [72].

Apesar das muitas vantagens, o Orange tem suas limitações. Alguns usuários podem encontrar restrições na customização de algoritmos ou na escala dos projetos que podem ser executados. É importante avaliar se a ferramenta atende às necessidades específicas de cada projeto. Por exemplo, para projetos que requerem a customização avançada de algoritmos ou que lidam com grandes volumes de dados, ferramentas como TensorFlow[103] ou PyTorch[104] podem ser mais adequadas [72].

Como os dados de solubilidade e taxa de dissolução podem variar em escalas diferentes, aplicar normalização ou padronização dos dados poderia melhorar a convergência dos modelos de aprendizado de máquina, como mencionado por Rogel *e colaboradores* [81].

O tratamento adequado de dados ausentes, seja através de imputação por médias ou algoritmos mais avançados como K-Nearest Neighbors, pode ser um importante passo no pré-processamento para garantir que a análise seja precisa e não introduza vieses indesejados [78].

Técnicas como Análise de Componentes Principais (PCA) poderiam ser aplicadas para reduzir a dimensionalidade dos dados, o que pode ser particularmente útil ao lidar com grandes bases de dados de compostos químicos. Isso também ajudaria a combater a "maldição da dimensionalidade", como discutido por Bengio e colaboradores [82].

A presença de valores discrepantes pode influenciar negativamente a precisão dos modelos preditivos. O uso de métodos como a detecção de outliers baseada em IQR (Intervalo Interquartil) ou métodos baseados em aprendizado de máquina, como *Support Vector Machines* (SVM) para detecção de anomalias, poderia melhorar a qualidade dos dados de entrada [25].

A ampliação das capacidades do Orange e a exploração de novas áreas de aplicação podem consolidar ainda mais a sua utilidade como ferramenta de análise preditiva. Estudos futuros poderiam focar na integração do Orange com outras plataformas de aprendizado de máquina, na adaptação de novos algoritmos e na validação da metodologia proposta em diferentes contextos industriais e acadêmicos. Dessa forma, espera-se que a ferramenta continue a evoluir e a contribuir significativamente para o campo da ciência de dados e da engenharia de materiais.

8 REFERÊNCIAS BIBLIOGRÁFICAS

[1] E. M. Pierce *et al.*, “Experimental determination of the effect of the ratio of B/Al on glass dissolution along the nepheline (NaAlSiO4)-malinkoite (NaBSiO4) join”, *Geochim Cosmochim Acta*, vol. 74, nº 9, p. 2634–2654, maio 2010, doi: 10.1016/j.gca.2009.09.006.

[2] N. Jain e S. H. Yalkowsky, “Estimation of the aqueous solubility I: Application to organic nonelectrolytes”, *J Pharm Sci*, vol. 90, nº 2, p. 234–252, 2001, doi: 10.1002/1520-6017(200102)90:2<234::AID-JPS14>3.0.CO;2-V.

[3] H. Lennernäs e B. Abrahamsson, “The use of biopharmaceutic classification of drugs in drug discovery and development: current status and future extension”, *Journal of Pharmacy and Pharmacology*, vol. 57, nº 3, p. 273–285, fev. 2010, doi: 10.1211/0022357055263.

[4] A. A. Noyes e W. R. Whitney, “The rate of solution of solid substances in their own solutions”, *J Am Chem Soc*, 1897.

[5] K. Rajan, “Materials Informatics: The Materials ‘gene’ and Big Data”, *Annu Rev Mater Res*, vol. 45, p. 153–169, jul. 2015, doi: 10.1146/annurev-matsci-070214-021132.

[6] A. Fuentes, S. Yoon, S. C. Kim, e D. S. Park, “A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition”, *Sensors (Switzerland)*, vol. 17, nº 9, set. 2017, doi: 10.3390/s17092022.

[7] J. Demšar *et al.*, “Orange: Data Mining Toolbox in Python Tomaž Curk Matija Polajnar Laň Zagar”, 2013.

[8] T. Curk *et al.*, “Microarray data mining with visual programming”, *Bioinformatics*, vol. 21, nº 3, p. 396–398, fev. 2005, doi: 10.1093/bioinformatics/bth474.

[9] M. Toplak *et al.*, “Infrared Orange: Connecting Hyperspectral Data with Machine Learning”, *Synchrotron Radiat News*, vol. 30, nº 4, p. 40–45, jul. 2017, doi: 10.1080/08940886.2017.1338424.

[10] M. Zitnik, F. Nguyen, B. Wang, J. Leskovec, A. Goldenberg, e M. M. Hoffman, “Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities”, *Information Fusion*, vol. 50, p. 71–91, out. 2019, doi: 10.1016/j.inffus.2018.09.012.

[11] J. E. Shelby, *Introduction to Glass Science and Technology*. The Royal Society of Chemistry, 2005.

[12] S. H. Yalkowsky e S. C. Valvani, “Solubility and partitioning I: Solubility of nonelectrolytes in water.”, 1980.

[13] M. H. Abraham’~, H. S. Chadha’, G. S. Whi, T. I. Ng, e R. C. Mitchell$, “Hydrogen Bonding. 32. An Analysis of Water-Octanol and Water-Alkane Partitioning and the AIog P Parameter of Seiler”, 1085.

[14] S. H. Yalkowsky e Y. He, *Handbook of Aqueous Solubility Data*, vol. 46, nº 19. CRC Press, 2003. doi: 10.1021/jm0303251.

[15] J. R. Hughey, S. Huang, e R. O. Williams, “Solid-state techniques for improving solubility”, em *AAPS Advances in the Pharmaceutical Sciences Series*, vol. 22, Springer Verlag, 2016, p. 121–163. doi: 10.1007/978-3-319-42609-9_3.

[16] O. A. El Seoud, “Solvation Simplified”, *Quim Nova*, vol. 33, nº 10, 2010.

[17] N. Jain, G. Yang, S. G. Machatha, e S. H. Yalkowsky, “Estimation of the aqueous solubility of weak electrolytes”, *Int J Pharm*, vol. 319, nº 1–2, p. 169–171, ago. 2006, doi: 10.1016/j.ijpharm.2006.04.022.

[18] V. Krishnamoorthy e V. Priya Ranjan Prasad, “Physicochemical characterization and in vitro dissolution behavior of olanzapine-mannitol solid dispersions”, 2012.

[19] A. N. Martin, P. J. Sinko, e Yashveer. Singh, *Martin’s physical pharmacy and pharmaceutical sciences : physical chemical and biopharmaceutical principles in the pharmaceutical sciences.* Lippincott Williams & Wilkins, 2011.

[20] W. L. Jorgensen, D. S. Maxwell, e J. Tirado-Rives, “Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids”, 1996.

[21] V. J. Stella, V. M. Rao, E. A. Zannou, e V. Zia, “Mechanisms of drug release from cyclodextrin complexes”, 1999.

[22] P. Costa, “An alternative method to the evaluation of similarity factor in dissolution testing”, 2001. [Online]. Disponível em: www.elsevier.com/locate/ijpharm

[23] R. Löbenberg e G. L. Amidon, “Modern bioavailability, bioequivalence and biopharmaceutics classification system: New scientific approaches to international regulatory standards”, *European Journal of Pharmaceutics and Biopharmaceutics*, vol. 50, 2000, [Online]. Disponível em: www.elsevier.com/locate/ejphabio

[24] K. Rajan, “Materials informatics”, 2005.

[25] Y. Liu, T. Zhao, W. Ju, e S. Shi, “Materials discovery and design using machine learning”, 2017, *Chinese Ceramic Society*. doi: 10.1016/j.jmat.2017.08.002.

[26] L. Ward, A. Agrawal, A. Choudhary, e C. Wolverton, “A general-purpose machine learning framework for predicting properties of inorganic materials”, *NPJ Comput Mater*, vol. 2, ago. 2016, doi: 10.1038/npjcompumats.2016.28.

[27] A. Seko, T. Maekawa, K. Tsuda, e I. Tanaka, “Machine learning with systematic density-functional theory calculations: Application to melting temperatures of single- and binary-component solids”, *Phys Rev B Condens Matter Mater Phys*, vol. 89, nº 5, fev. 2014, doi: 10.1103/PhysRevB.89.054303.

[28] D. Xue, P. V. Balachandran, J. Hogden, J. Theiler, D. Xue, e T. Lookman, “Accelerated search for materials with targeted properties by adaptive design”, *Nat Commun*, vol. 7, abr. 2016, doi: 10.1038/ncomms11241.

[29] K. Konstantinou, “Computational modelling of structural, dynamical and electronic properties of multicomponent silicate glasses”.

[Online]. Disponível em: https://www.researchgate.net/publication/316558284

[30] B. C. Bunker, “Molecular mechanisms for corrosion of silica and silicate glasses”, 1994.

[31] E. H. Oelkers e S. R. Gislason, “The mechanism, rates and consequences of basaltic glass dissolution: I. An experimental study of the dissolution rates of basaltic glass as a function of aqueous Al, Si and oxalic acid concentration at 25°C and pH 3 and 11”, 2001.

[32] I. Izquierdo-Barba, A. J. Salinas, e M. Vallet-Regí, “Bioactive Glasses: From Macro to Nano”, *Int J Appl Glass Sci*, vol. 4, nº 2, p. 149–161, jun. 2013, doi: 10.1111/ijag.12028.

[33] M. I. Ojovan e O. G. Batyukhnova, “Glasses for Nuclear Waste Immobilization”.

[34] W. S. Mcculloch e W. Pitts, “A LOGICAL CALCULUS OF THE IDEAS IMMANENT IN NERVOUS ACTIVITY* n”, 1990.

[35] S. S. Haykin e S. S. Haykin, *Neural networks and learning machines*. Prentice Hall/Pearson, 2009.

[36] “rumelhart1986”.

[37] M. Jordan, J. Kleinberg, e B. Schölkopf, “Pattern Recognition and Machine Learning”.

[38] G. S. Atsalakis e K. P. Valavanis, “Surveying stock market forecasting techniques - Part II: Soft computing methods”, *Expert Syst Appl*, vol. 36, nº 3 PART 2, p. 5932–5941, 2009, doi: 10.1016/j.eswa.2008.07.006.

[39] A. Géron, *HANDS-ON MACHINE LEARNING WITH SCIKIT-LEARN, KERAS, AND TENSORFLOW concepts, tools, and techniques to build intelligent systems*. O’Reilly Media, Inc, 2022.

[40] M. Mazini, B. Shirazi, e I. Mahdavi, “Anomaly network-based intrusion detection system using a reliable hybrid artificial bee colony and AdaBoost algorithms”, *Journal of King Saud University - Computer and Information Sciences*, vol. 31, nº 4, p. 541–553, out. 2019, doi: 10.1016/j.jksuci.2018.03.011.

[41] W. Hu, W. Hu, e S. Maybank, “AdaBoost-based algorithm for network intrusion detection”, *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 38, nº 2, p. 577–583, abr. 2008, doi: 10.1109/TSMCB.2007.914695.

[42] J. Zheng, D. Lin, Z. Gao, S. Wang, M. He, e J. Fan, “Deep Learning Assisted Efficient AdaBoost Algorithm for Breast Cancer Detection and Early Diagnosis”, *IEEE Access*, vol. 8, p. 96946–96954, 2020, doi: 10.1109/ACCESS.2020.2993536.

[43] L. Breiman, “Random Forests”, 2001.

[44] G. R. ¨ Atsch e K.-R. M. ¨ Uller, “Soft Margins for AdaBoost”, 2001. [Online]. Disponível em: www.first.gmd.de

[45] B. Zhang *et al.*, “Ensemble learners of multiple deep cnns for pulmonary nodules classification using ct images”, *IEEE Access*, vol. 7, p. 110358–110371, 2019, doi: 10.1109/ACCESS.2019.2933670.

[46] J. Hu, J. Lu, e Y. P. Tan, “Discriminative deep metric learning for face verification in the wild”, em *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, set. 2014, p. 1875–1882. doi: 10.1109/CVPR.2014.242.

[47] S. Sun *et al.*, “Accelerated Development of Perovskite-Inspired Materials via High-Throughput Synthesis and Machine-Learning Diagnosis”, *Joule*, vol. 3, nº 6, p. 1437–1451, jun. 2019, doi: 10.1016/j.joule.2019.05.014.

[48] L. Horev-Azaria *et al.*, “Predictive Toxicology of cobalt ferrite nanoparticles: comparative in-vitro study of different cellular models using methods of knowledge discovery from data”, 2013. [Online]. Disponível em: http://www.particleandfibretoxicology.com/content/10/1/32

[49] K. Taheri, H. Shahabi, K. Chapi, A. Shirzadi, F. Gutiérrez, e K. Khosravi, “Sinkhole susceptibility mapping: A comparison between Bayes-based machine learning algorithms”, *Land Degrad Dev*, vol. 30, nº 7, p. 730–745, abr. 2019, doi: 10.1002/ldr.3255.

[50] J. F. Rodrigues, L. Florea, M. C. F. de Oliveira, D. Diamond, e O. N. Oliveira, “Big data and machine learning for materials science”, 1º de dezembro de 2021, *Springer Nature*. doi: 10.1007/s43939-021-00012-0.

[51] J. D. Vienna, J. J. Neeway, J. V. Ryan, e S. N. Kerisit, “Impacts of glass composition, pH, and temperature on glass forward dissolution rate”, *Npj Mater Degrad*, vol. 2, nº 1, dez. 2018, doi: 10.1038/s41529-018-0042-5.

[52] S. Ósk Jónsdóttir, F. S. Jørgensen, e S. Brunak, “Prediction methods and databases within chemoinformatics: Emphasis on drugs and drug candidates”, 15 de maio de 2005, *Oxford University Press*. doi: 10.1093/bioinformatics/bti314.

[53] S. Boobier, D. R. J. Hose, A. J. Blacker, e B. N. Nguyen, “Machine learning with physicochemical relationships: solubility prediction in organic solvents and water”, *Nat Commun*, vol. 11, nº 1, dez. 2020, doi: 10.1038/s41467-020-19594-z.

[54] Z. Ye e D. Ouyang, “Prediction of small-molecule compound solubility in organic solvents by machine learning algorithms”, *J Cheminform*, vol. 13, nº 1, dez. 2021, doi: 10.1186/s13321-021-00575-3.

[55] H. Jin, Z. Jin, Y. G. Kim, e C. Fan, “Development of machine learning-based solubility models for estimation of Hydrogen solubility in oil: Models assessment and validation”, *Case Studies in Thermal Engineering*, vol. 51, nov. 2023, doi: 10.1016/j.csite.2023.103622.

[56] G. Yin *et al.*, “Multiple machine learning models for prediction of CO2 solubility in potassium and sodium based amino acid salt solutions”, *Arabian Journal of Chemistry*, vol. 15, nº 3, mar. 2022, doi: 10.1016/j.arabjc.2021.103608.

[57] L. Horev-Azaria *et al.*, “Predictive toxicology of cobalt nanoparticles and ions: Comparative in vitro study of different cellular models using methods of knowledge discovery from data”, *Toxicological Sciences*, vol. 122, nº 2, p. 489–501, ago. 2011, doi: 10.1093/toxsci/kfr124.

[58] A. Tayyebi *et al.*, “Prediction of organic compound aqueous solubility using machine learning: a comparison study of descriptor-based and fingerprints-based models”, *J Cheminform*, vol. 15, nº 1, dez. 2023, doi: 10.1186/s13321-023-00752-6.

[59] V. Ramani e T. Karmakar, “Graph Neural Networks for Predicting Solubility in Diverse Solvents using MolMerger incorporating Solute-solvent Interactions”, fev. 2024, [Online]. Disponível em: http://arxiv.org/abs/2402.11340

[60] Y. H. Zhao *et al.*, “Evaluation of human intestinal absorption data and subsequent derivation of a quantitative structure - Activity relationship (QSAR) with the Abraham descriptors”, *J Pharm Sci*, vol. 90, nº 6, p. 749–784, 2001, doi: 10.1002/jps.1031.

[61] A. Jouyban, “Review of the cosolvency models for predicting solubility of drugs in water-cosolvent mixtures”, 2008.

[62] S. L. Diamond, “Systems biology of coagulation”, junho de 2013. doi: 10.1111/jth.12220.

[63] S. Agatonovic-Kustrin e R. Beresford, “Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research”, 2000. [Online]. Disponível em: www.elsevier.com/locate/jpba

[64] D. M. Dimiduk, E. A. Holm, e S. R. Niezgoda, “Perspectives on the Impact of Machine Learning, Deep Learning, and Artificial Intelligence on Materials, Processes, and Structures Engineering”, 1º de setembro de 2018, *Springer Science and Business Media Deutschland GmbH*. doi: 10.1007/s40192-018-0117-8.

[65] M. T. Dylla, A. Dunn, S. Anand, A. Jain, e G. J. Snyder, “Machine Learning Chemical Guidelines for Engineering Electronic Structures in Half-Heusler Thermoelectric Materials”, *Research*, vol. 2020, jan. 2020, doi: 10.34133/2020/6375171.

[66] J. Ling, R. Jones, e J. Templeton, “Machine learning strategies for systems with invariance properties”, *J Comput Phys*, vol. 318, p. 22–35, ago. 2016, doi: 10.1016/j.jcp.2016.05.003.

[67] Z. Yang *et al.*, “Deep learning approaches for mining structure-property linkages in high contrast composites from simulation datasets”, *Comput Mater Sci*, vol. 151, p. 278–287, ago. 2018, doi: 10.1016/j.commatsci.2018.05.014.

[68] D. Jha *et al.*, “ElemNet: Deep Learning the Chemistry of Materials From Only Elemental Composition”, *Sci Rep*, vol. 8, nº 1, dez. 2018, doi: 10.1038/s41598-018-35934-y.

[69] Jeff. Heaton, *Introduction to neural networks with Java*. Heaton Research, 2005.

[70] V. Nair e G. E. Hinton, “Rectified Linear Units Improve Restricted Boltzmann Machines”.

[71] J. Demšar *et al.*, “Orange: Data Mining Toolbox in Python Tomaž Curk Matija Polajnar Laň Zagar”, 2013.

[72] X. Glorot, A. Bordes, e Y. Bengio, “Deep Sparse Rectifier Neural Networks”, 2011.

[73] A. L. Maas, A. Y. Hannun, e A. Y. Ng, “Rectifier Nonlinearities Improve Neural Network Acoustic Models”, 2013.

[74] Y. Lecun, Y. Bengio, e G. Hinton, “Deep learning”, 27 de maio de 2015, *Nature Publishing Group*. doi: 10.1038/nature14539.

[75] I. Goodfellow, Y. Bengio, e A. Courville, “Deep Learning”.

[76] Z. Wang e A. C. Bovik, “Mean squared error: Lot it or leave it? A new look at signal fidelity measures”, *IEEE Signal Process Mag*, vol. 26, nº 1, p. 98–117, 2009, doi: 10.1109/MSP.2008.930649.

[77] Z. Zhang, P. Cui, e W. Zhu, “Deep Learning on Graphs: A Survey”, dez. 2018, [Online]. Disponível em: http://arxiv.org/abs/1812.04202

[78] F. Chollet, *DEEP LEARNING with Python*. MANNING, 2021.

[79] Pradyot. Patnaik, *Handbook of inorganic chemicals*. McGraw-Hill, 2003.

[80] D. Rogers e M. Hahn, “Extended-connectivity fingerprints”, *J Chem Inf Model*, vol. 50, nº 5, p. 742–754, maio 2010, doi: 10.1021/ci100050t.

[81] J. Bergstra, J. B. Ca, e Y. B. Ca, “Random Search for Hyper-Parameter Optimization Yoshua Bengio”, 2012. [Online]. Disponível em: http://scikit-learn.sourceforge.net.

[82] A. Géron, “Hands-On Machine Learning with Scikit-Learn and TensorFlow”.

[83] C. M. Bishop, “Neural Networks for Pattern Recognition CLARENDON PRESS • OXFORD 1995”.

[84] R. Kohavi, “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection”. [Online]. Disponível em: http//roboticsStanfordedu/"ronnyk

[85] D. C. Liu e J. Nocedal, “ON THE LIMITED M E M O R Y BFGS M E T H O D FOR LARGE SCALE O P T I M I Z A T I O N”, 1989.

[86] A. Y. Ng, “Feature selection, L 1 vs. L 2 regularization, and rotational invariance”.

[87] J. Heaton, “Artificial Intelligence for Humans, Volume 3: Deep Learning and Neural Networks”.

[88] M. Belkin, D. Hsu, S. Ma, e S. Mandal, “Reconciling modern machine learning practice and the bias-variance trade-off”, *Proc Natl Acad Sci U S A*, vol. 116, nº 32, p. 15849–15854, dez. 2018, doi: 10.1073/pnas.1903070116.

[89] A. L. Maas, A. Y. Hannun, e A. Y. Ng, “Rectifier Nonlinearities Improve Neural Network Acoustic Models”, 2013.

[90] J. Demšar *et al.*, “Orange: Data Mining Toolbox in Python”, *Journal of Machine Learning Research*, vol. 14, p. 2349–2353, 2013.

[91] T. Chai e R. R. Draxler, “Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature”, *Geosci Model Dev*, vol. 7, nº 3, p. 1247–1250, jun. 2014, doi: 10.5194/gmd-7-1247-2014.

[92] B. Efron e R. Tibshirani, *An introduction to the bootstrap*. Chapman & Hall, 1994.

[93] W. M. (Ed.) HAYNES, *CRC Handbook of Chemistry and Physics*, 95. ed. Boca Raton: CRC Press, 2014.

[94] Atherton. SEIDELL, “Solubilities of Inorganic and Metal Organic Compounds: a compilation of quantitative solubility data from the periodical literature.”, *New York: D. Van Nostrand Company*, 1940.

[95] I. S. Muller, F. Perez-Cardenas, I. Joseph, I. L. Pegg, e L. M. Ayers, “ORP-63487 Revision 0”.

[96] X. Feng e I. L. Pegg, “A glass dissolution model for the effects of S/V on leachate pH”, 1994.

[97] C. M. Jantzen, K. G. Brown, e J. B. Pickett, “Durable Glass for Thousands of Years”, *Int J Appl Glass Sci*, vol. 1, nº 1, p. 38–62, mar. 2010, doi: 10.1111/j.2041-1294.2010.00007.x.

[98] A. A. Kruger *et al.*, “ORP-56502 Revision 0 Final Report-ILAW PCT, VHT, Viscosity, and Electrical Conductivity Model Development, VSL-07R1230-1”, 2007.

[99] J. J. Neeway *et al.*, “FY2016 ILAW Glass Corrosion Testing with the Single-Pass Flow-Through Method”, 2017.

[100] P. Frugier *et al.*, “SON68 nuclear glass dissolution kinetics: Current state of knowledge and basis of the new GRAAL model”, *Journal of Nuclear Materials*, vol. 380, nº 1–3, p. 8–21, out. 2008, doi: 10.1016/j.jnucmat.2008.06.044.

[101] J. L. Rodgers e ; W Alan Nicewander, “Thirteen Ways to Look at the Correlation Coefficient”, 1988.

[102] M. Abadi *et al.*, “TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems”, mar. 2016, [Online]. Disponível em: http://arxiv.org/abs/1603.04467

[103] A. Paszke *et al.*, “PyTorch: An Imperative Style, High-Performance Deep Learning Library”, dez. 2019, [Online]. Disponível em: http://arxiv.org/abs/1912.01703