Predição da Solubilidade Química: Uma Abordagem Baseada em Modelos de Regressão

1st Adauta Costa Aragão Freire
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brasil
adautacostaa.freire@alu.ufc.br

2nd Byanca Araújo Pinto
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brasil
byancaa@alu.ufc.br

3rd Carlos Eric Alves Ferreira
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brasil
ferreiraeric@alu.ufc.br

4th Lucas Rodrigues de Azevedo
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brasil
lucas.rodrigues@alu.ufc.br

***Abstract*—À medida que a "Era da Informação" avança, cada vez se torna mais perceptível a necessidade de utilizar dados, seja para compreender padrões ou automatizar processos. Eles se fazem presentes em todos os lugares. Atualmente, até mesmo o ato simples de conferir a previsão do tempo, que é algo rotineiro, só é capaz de ser realizado pois existe um enorme processo de análise de dados para que sejam feitas as previsões. Não dificilmente a presença dessas análises passam despercebidas pelo cotidiano. Dessa forma, este artigo busca encontrar maneiras de para realizar boas predições, utilizando modelos como Regressão Ordinária, Regressão Ridge, Partial Least Squares (PLS) e Redes Neurais. O desempenho de cada um destes tipos de regressão será discutida neste artigo, utilizando um dataset que tem o intuito de prever a solubilidade de componentes.**

***Index Terms*—regressão linear, inferência, análise de dados, solubilidade de compostos químicos, modelos, química, regressão ridge, redes neurais.**

I. Introdução

Dado o contexto tecnológico atual, torna-se cada vez mais notório a participação da coleta, análise e interpretação de dados no cotidiano de milhares de pessoas. Com o aumento exponencial da geração de informação que a internet possibilita, como coautora, faz-se necessário o bom uso de dados. Segundo [1], entender o mundo por meio de dados é como tentar juntar as peças da realidade usando um quebra-cabeça com peças a mais. A análise de dados é essencial para compreender comportamentos, padrões e tomar decisões, baseando-se nos modelos de predição.

Para entender a real importância dos dados, é necessário entender primeiramente o contexto em que eles estão inseridos. O simples ato de checar a previsão do tempo no celular já faz parte de um sistema complexo que carrega inúmeras informações para fornecer uma expectativa adequada sobre o tempo e ajudar a evitar imprevistos. Dado este exemplo, percebe-se então que, mesmo que muitas coisas passem despercebidas, a análise de dados é fundamental em diversos âmbitos.

É necessário ressaltar que, quando há coleta de dados, erros são frequentes. Se existe um dataset coletado para a temperatura de uma cidade, é de suma importância entender se os valores são reais. Se, por exemplo, houvesse um dado que indica uma temperatura negativa em uma cidade do Nordeste do Brasil, seria algo que chamaria atenção por ser um *outlier*, um dado que destoa dos outros. Pode ser um fenômeno raro, mas a maior probabilidade é de que isso seja um valor errado. *Outliers* podem interferir bastante no resultado de uma predição, dependendo do método utilizado. A regressão linear ordinária é um dos métodos mais sensíveis a essa discrepância no conjunto. Por outro lado, modelos como regressão baseada em árvores lida melhor com tais valores. O processo de desenvolvimento desses tipos de ferramentas evoluiu em vários campos, como química, ciência da computação, física e estatística. [2]

Assim, técnicas de pré-processamento de dados são necessárias para evitar certas situações que podem afetar a precisão do modelo, principalmente ao trabalhar com os mais sensíveis a *outliers*, por exemplo. De acordo com [3], para muitas tarefas, é difícil saber quais recursos devem ser extraídos. Algumas estratégias para melhorar o conjunto de amostras envolvem substituir dados ausentes pela média de valores da variável no conjunto, detecção e remoção de *outliers*, normalização, centralização, dentre outras. É crucial interpretar o conjunto de dados e o modelo utilizado para entender qual a melhor forma de realizar o pré-processamento, levando em conta suas necessidades.

Depois que as informações são coletadas e faz-se o pré-processamento, é preciso escolher um modelo para realizar as predições. Uma alternativa para usar esses dados e fazer previsões é a regressão linear. Quando o resultado esperado na saída do modelo é um valor numérico, a inferência pode ser feita de maneira precisa, tendo as variáveis certas para isso. Embora possa parecer um pouco simples em comparação com algumas das abordagens de aprendizado estatístico mais mod-

ernas, a regressão linear ainda é um método de aprendizado estatístico útil e amplamente utilizado [1].

Um exemplo de regressão linear simples:

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{1}$$

β_0 é o intercepto ou constante, β_1 é um coeficiente de regressão e ϵ é o erro do modelo. Essa é uma abordagem linear simples para prever uma resposta quantitativa Y com base em X, uma variável preditora. Ela assume que existe aproximadamente uma relação linear entre X e Y[1].

Um dos objetivos principais da regressão é minimizar o erro, que é calculado subtraindo o valor real da predição y e o valor calculado pelo modelo $\hat{y}$.

$$e_i = y_i - \hat{y} \tag{2}$$

Utilizando o mesmo princípio da fórmula vista anteriormente, a *RMSE (Root Mean-Square Error)*, também conhecida como Raiz do Erro Quadrado Médio é uma métrica importante para avaliar o desempenho de um modelo de regressão. Ele utiliza o princípio da fórmula vista anteriormente para estimar o quão bem o modelo consegue aproximar os dados reais.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

Um exemplo das inúmeras aplicações da regressão linear é a previsão da solubilidade de um componente. Na indústria farmacêutica, por exemplo, pode-se combinar química, biologia e análises estatísticas para desenvolver medicamentos e produtos de beleza utilizando de previsões sobre a solubilidade de componentes, buscando produtos com melhor custo-benefício, mas que mantenham a eficácia, levando em consideração suas especificidades, o que pode alavancar a indústria e tecnologia nesse meio.

II. Métodos

A. Descrição do Dataset

O *dataset* [2] utilizado neste artigo fornece uma visão geral de observações sobre compostos químicos, contém 1267 observações de compostos químicos, com o objetivo principal de investigar e prever seus valores de solubilidade, obtidos por meio de medições experimentais. Cada observação no dataset é caracterizada por um conjunto de 228 variáveis preditoras, que podem ser divididas em três principais grupos distintos:

1) 208 variáveis binárias (impressões digitais químicas) que indicam a presença ou ausência de subestruturas químicas específicas nos compostos.
2) 16 descritores de contagem, que incluem informações como o número de ligações químicas e a quantidade de átomos de determinados elementos, como o oxigênio.
3) 4 descritores contínuos, que fornecem informações como pesos moleculares e áreas de superfície.

Das 1267 observações, o *dataset* será subdivido em um *dataset* de treino contendo 951 observações e um *dataset* de teste, que possui 316 observações. Ademais, os valores de solubilidade, que são o resultado dos modelos de regressão construídos a partir desses datasets.

Os preditores originais do *dataset* de treino e teste, são representados por solTrainX e solTestX, foram mantidos em suas unidades naturais, fornecendo uma base de comparação direta com os dados transformados. Para melhorar a qualidade da análise, foram aplicadas técnicas de pré-processamento como tratamento de assimetrias, centralização e escalonamento aos preditores, gerando os conjuntos solTrainXtrans e solTestXtrans. Essas transformações são essenciais para assegurar a normalização dos dados e a redução de possíveis vieses durante a modelagem. Outrossim, os valores de solubilidade são armazenados nos vetores solTrainY para treinamendo dos modelos e solTestY para validação dos modelos.

B. Análise Exploratória

Inicialmente, é notório a importância de primeiro entender o conjunto de dados a ser trabalhado. Portanto, o primeiro passo a ser feito é a análise exploratória dos dados, um passo crucial para entender os padrões e comportamentos do conjunto utilizado, antes de iniciar processos mais complexos. Durante essa etapa, os valores ausentes, discrepantes e distribuições enviesadas são percebidos, o que vai ajudar a entender os próximos passos a serem trabalhados, selecionando as abordagens mais adequadas.

Dessa forma, análises como a monovariada e bivariada são realizadas. A análise monovariada é importante por observar como os preditores são individualmente, avaliando suas métricas, como média, desvio padrão e assimetria. A assimetria pode ser observada por um histograma, por exemplo, o que também ajuda a entender os valores que os preditores assumem. Já a análise bivariada examina o comportamento entre duas variáveis para entender possíveis relações entre as duas. Uma forma interessante de buscar essa correlação entre os preditores é plotando gráficos de dispersão, o que ajuda a identificar tendências. Como os *datasets* são divididos entre ”treinamento” e ”teste”, os procedimentos são realizados entre os dois.

Além disso, dada a sua necessidade, o pré-processamento foi realizado. Inicialmente, os preditores possuíam forte assimetria à direita, com uma média de 1,6. Dessa maneira, foi necessária a aplicação de Box-Cox nos preditores contínuos, utilizando λ como parâmetro de transformação, conforme a fórmula abaixo:

$$x^* = \begin{cases} \frac{x^\lambda - 1}{\lambda}, & \text{se } \lambda \neq 0, \\ \log(x), & \text{se } \lambda = 0. \end{cases} \tag{4}$$

Agora, o que se pode dizer sobre a relação entre os preditores com os dados transformados? Algumas observações podem indicar alguma relação linear entre o preditor e o resultado, como é o caso do peso molecular. Ademais, também existem relações não lineares, como o cloro.

C. Modelos de Regressão Linear Ordinária

A regressão linear é uma das ferramentas mais amplamente utilizadas em aprendizagem estatística e de máquina. No meio desta ferramenta, é possível encontrar a Regressão Linear Ordinária, que utiliza de métodos para estimar os parâmetros de um modelo de regressão linear. Ela fornece abordagens matemáticas que buscam encontrar a melhor maneira de ajustar uma linha aos dados. Ou seja, uma reta que representa relações entre uma ou mais variáveis.

$$y_i = \beta_0 + \sum_{j=1}^{D} x_{ij}\beta_j \tag{5}$$

O y_i representa o valor da resposta para a i-ésima observação. Ele é o valor previsto ou estimado, conforme a fórmula. β_0 é o intercepto ou constante, o valor quando x é igual a zero. Já o somatório de $x_{ij}\beta_j$ é uma combinação linear das variáveis preditoras x_{ij} e seus respectivos coeficiente β_j. Esse coeficiente determina o tamanho da influência que a variável tem sobre o resultado final. O somatório vai depender do tamanho de D, que aqui é o número total de variáveis independentes no modelo.

Essa regressão tornou-se popular devido à sua eficácia e simplicidade. Contudo, ao utilizar o seu método, ele supõe que as variáveis não são perfeitamente correlacionadas (ou seja, não há uma multicolinearidade muito forte). Sendo assim, em casos em que essa suposição não é correta, o modelo pode ter problemas. Ou seja, mesmo sendo uma abordagem direta e simples, a regressão linear ordinária pode ser sensível a multicolinearidade e *outliers*. Essa é uma das razões que levam à utilização de regressões penalizadas, como Ridge e Lasso, que podem tornar o modelo mais adaptado a dados complexos.

O primeiro passo para construir o modelo de regressão linear ordinário é utilizar os preditores transformados do conjunto de treinamento de X para aprender o modelo. Dessa maneira, os coeficientes β são calculados utilizando a fórmula da regressão ordinária e vão sendo ajustados de maneira que a soma dos resíduos quadrados seja a menor possível.

Depois de ajustar o modelo a partir do conjunto de treinamento, ele é testado utilizando o conjunto de teste de X depois do pré-processamento, comparando com os valores de Y. Agora, para calcular a precisão do modelo, utiliza-se a RMSE, Raiz do Erro Quadrado Médio e o R^2, Coeficiente de Determinação para ver o quanto o modelo está ajustado aos dados de teste. O R^2 é calculado por:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{6}$$

Em seguida, os métodos de validação cruzada são utilizados. A *cross-validation*, ou validação cruzada é amplamente utilizada para avaliar a perfomance de modelos. Ela divide os dados em conjuntos de treinamento e teste e valida diferentes combinações. Sendo assim, ela busca reduzir o viés do modelo, o que poderia acontecer se ele fosse avaliado em uma única divisão de dados, o que poderia ser ruim para um modelo que é sensível à distribuição de dados.

No caso, a validação cruzada utilizada vai ser feita utilizando a técnica do *K-fold*, na qual o conjunto de dados é treinado K vezes, cada vez utilizando K-1 dobras para treinamento e a restante para o teste. O processo é repetido para cada uma das K dobras e as métricas de desempenho (RMSE e R^2) são calculadas a cada iteração. No caso desse dataset, serão utilizadas 5 e 10 dobras. Essa técnica também é importante, pois o modelo é avaliado em diferentes subconjuntos de dados.

Sobre os resultados da RMSE, quando comparados os valores sem validação cruzada e com validação cruzada de 5 dobras, eles são bem próximos: 0.7456 e 0.7338. Mas, utilizando 10 dobras, ele já reduz para 0.7005. Agora, sobre o R^2, nos três casos ele manteve valores próximos, assumindo um valor próximo de 8.700.

Fig. 1. Valores originais vs valores previstos - Regressão Linear Ordiária sem validação cruzada.

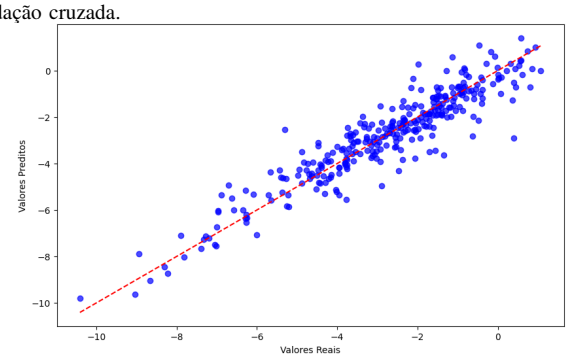


Fig. 2. Valores originais vs valores previstos - Regressão Linear Ordiária com validação cruzada de 5 dobras.

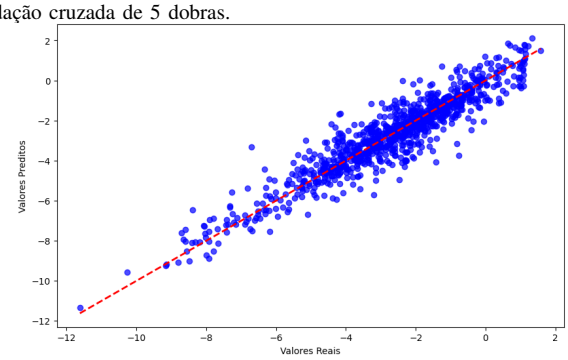
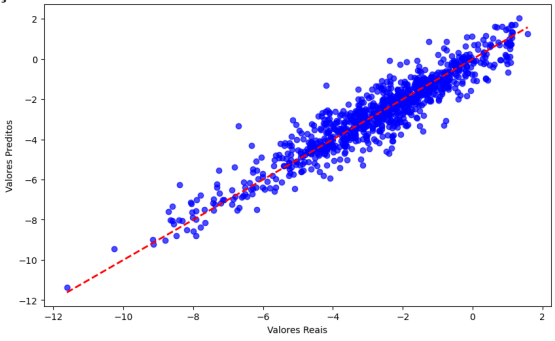


Fig. 3. Valores originais vs valores previstos - Regressão Linear Ordiária com validação cruzada de 10 dobras.



Agora, observando os gráficos das Figuras 1, 2 e 3, percebe-se uma grande diferença quanto à utilização da validação cruzada. Os dados brutos não apresentaram um bom ajuste entre os valores originais e os previstos em comparação aos dados com a validação cruzada, que possuem uma melhor capacidade de generalização.

D. Modelos de Regressão Linear L2-penalizado

Conforme mencionado anteriormente, os modelos de regressão ordinária não conseguem lidar muito bem com dados que possuem multicolinearidade ou outliers. Dessa maneira, segundo [4], os desafios em aprender com os dados levaram a uma revolução nas ciências estatísticas. Portanto, como uma alternativa mais sofisticada, existem os modelos de regressão penalizados. Aqui o assunto que será tratado é sobre a Regressão Ridge, um modelo que aplica uma penalidade à soma dos erros quadrados, dependendo de um parâmetro λ, conforme a fórmula abaixo:

$$SSE_{L2} = \sum_{i=1}^{n} e_i^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{7}$$

O método que calcula o erro quadrático com penalização L2 vai combinar o erro quadrático dos dados com a regularização dos coeficientes. O primeiro termo $\sum_{i=1}^{n} e_i^2$ representa a soma dos erros quadrados (os erros são calculados a partir da diferença entre os valores reais (y_i) e os valores preditos ($\hat{y_i}$)). O segundo termo $\lambda \sum_{j=1}^{p} \beta_j^2$ é justamente a penalização L2, que vai adicionar uma penalidade proporcional ao quadrado dos coeficientes β_j.

O parâmetro de regularização, representado por λ vai indicar o tamanho do impacto do termo no modelo. Ou seja, ajuda a mitigar problemas como *overfitting* (ou superajuste) e multicolinearidade. Sendo assim, percebe-se que a Regressão Ridge é uma excelente alternativa para ajudar a manter a precisão do modelo e conseguir generalizar bem para novos dados.

Para implementar a Regressão Ridge, após ter feito o pré-processamento (muito importante nessa etapa, pois a penalização depende da escala dos coeficientes), é necessário a divisão dos dados entre os conjuntos de treinamento e teste. Como o *dataset* utilizado já estava nesse formato, o próximo passo é utilizar o conjunto de treinamento, pré-processado da mesma maneira do de testes, para ajustar o modelo, variando o hiperparâmetro λ em um intervalo de escala logarítmica. Além disso, para cada valor de λ, as validações cruzadas com 5 e 10 dobras são utilizadas para calcular RMSE e R^2.

Com a validação cruzada de 5 dobras, a melhor performance do λ foi quando ele assumiu o valor de 17, conseguindo uma RMSE de 0.7207 e R^2 de 0.8742. Já com a validação cruzada com 10 dobras, ele atingiu o melhor desempenho quando ele assumiu o valor de 18, com a RMSE em 0.7104 e R^2 em 0.8731.

E. Modelos de Regressão PLS OU PCR

A PLS *(Partial Least Squares)*, ou Mínimos Quadrados Parciais, combina regressão linear com redução de dimensionalidade e o seu objetivo principal é encontrar componentes latentes que maximizem a covariância entre as variáveis preditoras e o resultado, garantindo que os componentes extraídos sejam relevantes para prever o resultado do modelo.

Sua vantagem é que ela lida bem com colinearidade entre as variáveis preditoras e identifica componentes relevantes diretamente relacionados à variável resposta. Porém, tem uma interpretação mais difícil em comparação à PCR, devido à mistura de informação das variáveis preditoras e a variável resposta nos componentes latentes e é suscetível a overfitting se não houver uma seleção criteriosa do número de componentes.

A PCR *(Principal Component Regression)*, ou Regressão por Componentes Principais, utiliza a PCA *(Principal Component Analysis)* para transformar as variáveis preditoras em componentes não correlacionados e posteriormente, a regressão linear é ajustada sobre os principais componentes selecionados, priorizando aqueles que capturam maior variância entre as variáveis preditoras, mas sem considerar diretamente a relação com a variável resposta.

A PLS popularizou-se por reduzir a dimensionalidade ao eliminar redundâncias entre os preditores e ser simples de interpretar e ter um menor risco de multicolinearidade. Contudo, ela não garante que os componentes selecionados sejam relevantes para a variável resposta e também pode descartar informações úteis para a previsão, caso os componentes úteis para a variável resposta capturem pouca variação nos preditores.

Para determinar o modelo mais adequado e a quantidade ideal de componentes M para a regressão, foram realizadas validações cruzadas utilizando 5 e 10 dobras. Durante este processo, foi considerado que as variáveis preditoras apresentam forte correlação entre si, uma característica que os métodos PLS e PCR lidam de maneira eficiente devido à sua robustez em cenários de multicolinearidade.

O principal objetivo da validação cruzada foi identificar o valor de M que proporcionasse o melhor desempenho do modelo, medido através do erro quadrático médio da raiz e do coeficiente de determinação. Para isso, foram testados valores de M variando de 1 até o ponto de melhor ajuste.

Fig. 4. Métricas da validação cruzada do PLS

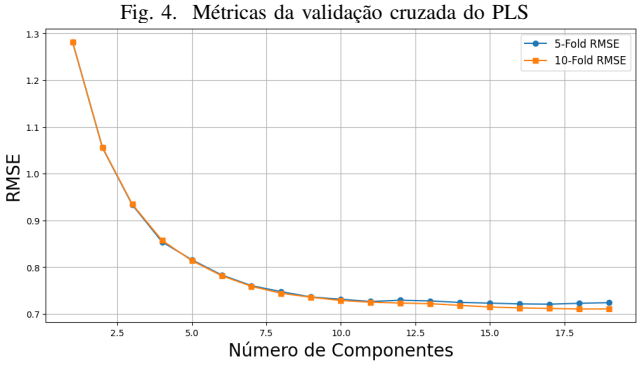
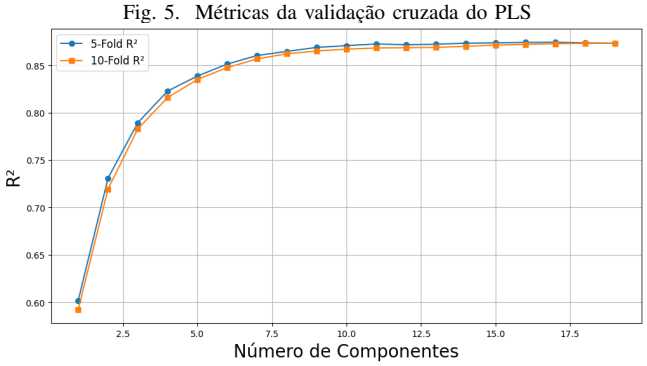


Fig. 5. Métricas da validação cruzada do PLS



Como pode ser observado nas Figuras 4 e 5, mesmo com o aumento no número de dobras, o desempenho do modelo não sofreu grandes alterações. As diferenças observadas nos valores de M são mínimas, mostrando uma boa consistência dos modelos a partir da utilização de dez componentes.

Portanto, entre os dois métodos de regressão, o PLS se destacou ligeiramente, apresentando um desempenho superior em relação ao PCR. Apesar de ambos os modelos terem alcançado resultados semelhantes em termos de M, o PLS foi mais eficiente na previsão, com um erro ligeiramente menor e uma capacidade de explicação ligeiramente superior.

F. Redes Neurais

As redes neurais são modelos inspirados no cérebro humano, por isso são compostas por camadas de neurônios interconectados, o que permite que elas aprendam relações complexas entre os dados. Para tarefas de regressão, elas podem modelar relações não lineares entre as variáveis preditoras e os resultados, superando algumas limitações dos modelos lineares.

Sua principal vantagem é a capacidade de capturar padrões complexos, sendo eficaz em problemas em que existem relações não lineares. No entanto, as redes neurais tendem ao*overfitting* devido ao grande número de coeficientes de regressão. Para combater esse problema, várias abordagens diferentes podem ser utilizadas.

$$y = f\left(\sum_{i=1}^{n} w_i x_i + b\right) \tag{8}$$

A fórmula de uma rede neural descreve o processo de cálculo da saída (previsão) a partir das entradas. Cada entrada x_i é multiplicada por um peso w_i, que vai determinar a sua importância. A soma ponderada das entradas é ajustada pelo viés b e a função de ativação f é aplicada para introduzir a não linearidade no modelo. O resultado final, y, é a previsão gerada pela rede. Durante o treinamento, os pesos e o viés são ajustados para que a rede faça previsões cada vez mais precisas.

Com os dados separados em treinamento e teste, a construção da rede neural pode ser iniciada. Após isso, é necessário o treinamento do modelo, onde há o ajuste dos pesos das conexões entre os neurônios. Depois, a precisão do modelo pode ser calculada a partir da RMSE e R². O resultado obtido após nove iterações gerou um RMSE de 0.7678, enquanto o R² foi de 0.8652. Com base nos resultados apresentados, o modelo não linear não superou o modelo linear, uma vez que o RMSE e o R² não apresentaram melhorias significativas. Isso sugere que, para os dados em questão, a relação entre os preditores e a variável resposta é suficientemente bem modelada por uma abordagem linear. A falta de melhora no modelo não linear indica que a relação entre as variáveis não é complexa o suficiente para justificar a introdução de não linearidade. Portanto, o modelo linear pode ser mais eficiente e interpretável para esse dataset.

III. Resultados

A. Resultados da Análise Exploratória

Na análise dos dados do conjunto utilizado, é possível perceber as principais características do *dataset*, plotando valores como a média, assimetria e desvio padrão. A partir dos dados plotam-se os histogramas para observar comportamentos que indiquem necessidade de atenção. Conforme se pode observar na Figura 6 e Figura 8, existe uma grande assimetria, assim como alguns *outliers*. Após realizado o pré-processamento, na Figura 7 e Figura 9, nota se a diferença, já que a assimetria foi removida, e agora existe a centralização e escalonamento dos dados. Portanto, nota-se a importância dessa análise inicial, para deixar o conjunto de dados suscetíveis a bons resultados nos modelos.

Fig. 6. Histograma Fator Hidrofílico - Sem pré-processamento

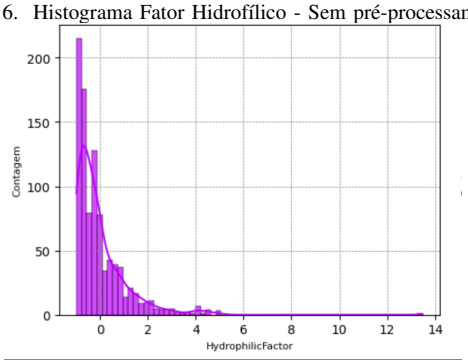


Fig. 7. Histograma Fator Hidrofílico - Com pré-processamento

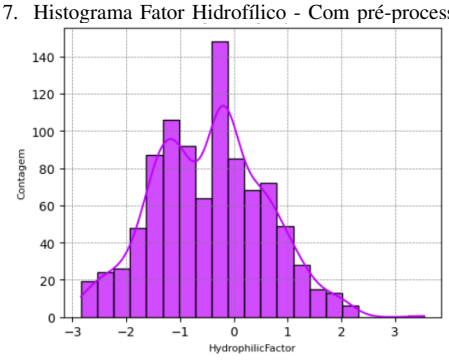


Fig. 8. Histograma Número de Hidrogênio - Sem pré-processamento

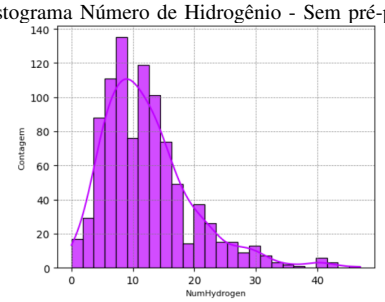
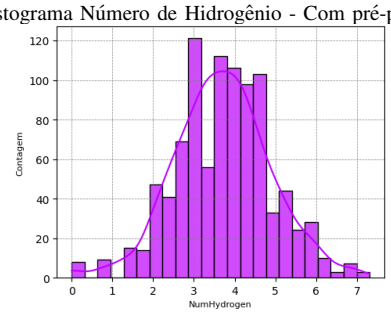


Fig. 9. Histograma Número de Hidrogênio - Com pré-processamento



Também foram separados dois preditores para realizar essa comparação, conforme as Tabelas I e II. Dessa forma, analisando em termos numéricos, é possível perceber uma diferença considerável em termos de assimetria.

TABLE I
ASSIMETRIA ENTRE PREDITORES ANTES DO PRÉ-PROCESSAMENTO

Preditor	**Assimetria**
Número de Halogênio	2.6914
Fator Hidrofílico	3.403785

TABLE II
ASSIMETRIA ENTRE PREDITORES ANTES DO PRÉ-PROCESSAMENTO

Preditor	**Assimetria**
Número de Halogênio	1.034808
Fator Hidrofílico	0.100462

B. Resultados das Regressões

Conforme as tabelas de comparação de desempenho abaixo, é notório que a utilização da validação cruzada melhora consideravelmente os resultados em termos de RMSE e R². Por exemplo, utilizando o modelo simples de Regressão Linear Ordinária, visto na Tabela III, o RMSE diminui de 7.338 para 0.7005, e o R² aumenta de 0.8698 para 0.8761. Isso se dá pela divisão dos dados em mais grupos de treinos e validação, o que ajudou a tornar o modelo mais estável e preciso.

TABLE III
DESEMPENHO DO MODELO OLS COM DIFERENTES ESTRATÉGIAS DE VALIDAÇÃO

Modelo	**RMSE**	**R²**
Ordinária sem validação cruzada	0.7456	0.8709
Ordinária com validação cruzada de 5 dobras	0.7338	0.8698
Ordinária com validação cruzada de 10 dobras	0.7005	0.8761

É notório observar que o melhor resultado para o RMSE ocorre quando há a utilização de um modelo que possui penalização, como visto na Tabela IV, ocorre com λ com o valor de 2.154 e 10 dobras utilizando a Regressão Ridge.

TABLE IV
MELHOR λ NO MODELO RIDGE PARA CADA QUANTIDADE DE DOBRAS USADAS E ANÁLISE DO DESEMPENHO.

Nº de dobras	**Melhor λ**	**RMSE**	**R^2**
5	10.00	0.7299	0.8711
10	2.1544	0.6849	0.8816

TABLE V
MELHOR M NO PLS PARA CADA QUANTIDADE DE DOBRAS USADAS E ANÁLISE DO DESEMPENHO.

Nº de dobras	**Melhor M**	**RMSE**	**R^2**
5	17.00	0.7207	0.8742
10	18.00	0.7104	0.8731

Já o resultado que mostrou o pior desempenho foi o da Regressão Linear Ordinária sem validação cruzada. Outrossim, quando se compara a PLS com os demais modelos, ela mostra-se melhor que a Regressão Ridge com 5 dobras, mas foi inferior a ela quando utilizadas 10 dobras. Ademais, quando comparamos os modelos lineares com o modelo não linear feito pela rede neural, esses obtiveram desempenhos superiores, sendo o modelo de penalização Ridge o que apresenta destaque pelo melhor resultado.

REFERENCES

[1] Zheng Alice and Casari Amanda. 2018. Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists. O'Reilly Media, Inc, 99
[2] M. Kuhn e K. Johnson, Applied Predictive Modeling. New York, NY, USA: Springer, 2013.
[3] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.
[4] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning. New York, NY: Springer, 2013.
[5] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning. 2nd ed., New York, NY, USA: Springer, 2009.