UNIVERSIDADE FEDERAL DE SÃO CARLOS
Centro de Ciências Exatas e de Tecnologia
Departamento de Computação

Estágio LaMaV

Estudo de modelos de aprendizado de máquina aplicado à predição de propriedades de vidros

Professor: Murilo Coelho Naldi

Autora
Ana Carolina Castro Rosal

São Carlos, 6 de dezembro de 2023

Sumário

Resumo

Tendo em vista que há uma janela de possíveis combinações químicas que geram vidros ainda não descobertos, desenvolver novos vidros constituídos por múltiplos elementos químicos e com propriedades e aplicações relevantes é um desafio. Dessa forma, a utilização de modelos de aprendizado de máquina para aumentar a acurácia da predição de propriedades vítreas importantes, como a temperatura de transição, Tg, é uma alternativa. Mesmo assim, os modelos de regressão utilizados até o momento não obtiveram sucesso ao predizer valores de Tg altos ou baixos, o que é um problema, porque vidros com temperaturas extremas são continuamente buscados pela indústria. Neste trabalho, o algoritmo HDBSCAN foi utilizado para clusterização de vidros com características similares e, a partir dos grupos encontrados, foram gerados modelos de regressão para predizer a Tg com mais precisão, sobretudo para materiais vítreos com valores de Tg altos ou baixos.

Palavras-chave: vidros, clusterização, HDBSCAN, aprendizado de máquina, temperatura de transição.

Introdução

Os vidros são compostos formados por uma determinada combinação de elementos químicos. Há um grande número de possíveis combinações para formar vidros, em torno de 10^{52} combinações [1]. Mesmo assim, o número de vidros inorgânicos é de somente 10^6, o que revela uma janela de oportunidades de descobrir novos vidros [2].

Tendo em vista essa possibilidade e o fato de que os modelos de Aprendizado de Máquina (AM) têm apresentado grande capacidade de direcionar problemas na área de ciência de materiais [3, 4], a grande quantidade de dados disponíveis sobre vidros permite que inúmeras análises sejam feitas, inclusive o uso de algoritmos de aprendizado não supervisionado com o objetivo de encontrar grupos que tornem a tarefa de predizer as propriedades de vidros de maneira mais acurada.

Em relação à área de desenvolvimento de vidros, a temperatura de transição T_g pode ser definida como a temperatura em que um vidro varia de um material rígido para um material viscoso e é fundamental para o descobrimento de novos vidros. Tendo em vista que é uma propriedade relacionada à estabilidade do vidro contra a cristalização e à estabilidade mecânica do material vítreo, a T_g tem alta dependência da composição química [5]. Além disso, composições que geram vidros com T_g alta ou com T_g baixa são continuamente buscados pela indústria devido à alta resistência dos vidros e à redução de custos, respectivamente

Um estudo prévio [2] utilizou diferentes modelos de aprendizado de máquina para predizer a T_g de vidros, mesmo assim, obteve alta incerteza em predizer altos ($Tg \geq 1150\,K$) e baixos ($Tg \leq 450\,K$) valores de Tg. Assim, melhorar a precisão da predição de vidros com temperaturas extremas é um desafio e uma necessidade industrial.

Por isso, aplicar técnicas de aprendizado de máquina não-supervisionado é uma alternativa para a predição desses tipos de vidros. Um algoritmo do estado da arte é o HDBSCAN [6], que provê uma completa hierarquia de grupos composta por todos os possíveis grupos formados com base em densidade para um infinito intervalo de densidades, além de permitir que os dados sejam visualizados e explorados de diferentes maneiras a partir da hierarquia resultante. Ademais, o algoritmo provê uma solução ótima e global ao problema de maximização da estabilidade dos grupos encontrados, ou seja, os grupos da solução ótima são persistentes ao longo de alguns intervalos de densidade.

As características do HDBSCAN permitem que esse algoritmo seja aplicado aos dados de vidros, como já foi feito em outro estudo [5]. Portanto, nesta pesquisa, o algoritmo

HDBSCAN foi utilizado para a obtenção de grupos relevantes dentro do conjunto de dados de vidros e, a partir desses grupos, regressores específicos foram aplicados aos grupos relevantes, treinados e testados. Depois, a performance dos regressores foi comparada com os resultados obtidos no estudo [5].

Assim, o objetivo desse trabalho é agrupar materiais com composições químicas similares a partir do HDBSCAN e gerar, a partir dos grupos encontrados, modelos de regressão para predizer a Tg com mais precisão do que o modelo de regressão geral já utilizado em um estudo anterior [2]. Portanto, a hipótese é de que os grupos serão formados por vidros com valores de Tg próximos e, consequentemente, os modelos serão mais acurados, principalmente para materiais vítreos com valores de Tg altos ou baixos.

Objetivos

O objetivo geral da pesquisa era aplicar técnicas de aprendizado não supervisionado do estado da arte, como o algoritmo HDBSCAN, para melhorar a performance da predição de T_g de vidros.

Os objetivos específicos foram:

- Estudar os melhores parâmetros para o algoritmo HDBSCAN aplicado ao conjunto de dados;
- Escolher o melhor algoritmo para redução de dimensionalidade do conjunto dados;
- Analisar a composição dos grupos extraídos pelo HDBSCAN;
- Escolher o melhor algoritmo de regressão para os grupos;
- Fazer uma análise comparativa entre os resultados obtidos nas regressões de cada grupo e os resultados do estudo prévio [5].

Metodologia experimental

Nesta seção, são descritos o banco de dados de vidros, as técnicas de pré-processamento aplicadas ao conjunto de dados, o algoritmo de clusterização HDBSCAN, os algoritmos de regressão e as métricas de avaliação utilizados.

Banco de dados

Para a realização deste projeto, foi utilizado um conjunto de dados formado por diversos vidros, representados por tuplas de composições de cada um dos 66 elementos químicos e a respectiva Tg, como pode ser visto na Imagem 1.

Li	Be	B	O	Na	Mg	Al	Si	P	K	...	Hf	Ta	W	Hg	Tl	Pb	Bi	Th	U	Tg
0.000000	0.0	0.000000	0.609091	0.087013	0.019481	0.045455	0.238961	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	899.15
0.333333	0.0	0.000000	0.523810	0.000000	0.000000	0.000000	0.000000	0.142857	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	575.15
0.000000	0.0	0.032441	0.623411	0.032441	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	658.15
0.000000	0.0	0.020248	0.599935	0.126715	0.000000	0.000000	0.253103	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	777.15
0.000000	0.0	0.318182	0.545455	0.109091	0.000000	0.000000	0.000000	0.000000	0.027273	...	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	712.15

Imagem 1 - Estrutura do conjunto de dados utilizado.

Os dados se limitaram a vidros formados por óxidos e foram coletados de revistas científicas, livros e patentes obtidas do *SciGlass* [7]. Foram considerados apenas vidros com composição acima de 30% de fração atômica de oxigênio. Foram excluídos todos os compostos que poderiam modificar o balanço do oxigênio, como os que contém enxofre, hidrogênio, carbono, flúor, cloro, nitrogênio, bromo, iodo e os metais nobres, assim como foi feito em um trabalho prévio [2].

Foram coletadas aproximadamente 51.000 composições de vidros, cada composição com de 2 a 32 elementos químicos diferentes de um total de 65 elementos. Além disso, as composições duplicadas foram removidas e substituídas pela mediana da Tg. No fim, foram obtidos 43.641 vidros e os valores de Tg variam de 342 K até 1495 K. A distribuição dos vidros de acordo com a Tg pode ser vista na Imagem 2.

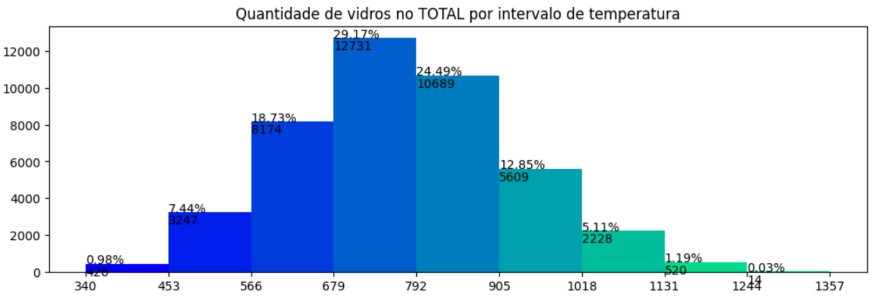

Imagem 2 - Gráfico da distribuição dos vidros de acordo com a Tg.

Pré-processamento

Antes do pré-processamento, o conjunto de dados foi separado em matriz de atributos, contendo a composição dos vidros com 43.641 linhas e 65 colunas, e em vetor de alvos, contendo 43.641 valores correspondentes às Tg de cada uma das composições.

Depois, foi feita a aplicação do *MinMaxScaler* na matriz de atributos, que considera o valor máximo e valor mínimo de cada uma das colunas da matriz de atributos e os transforma em 1 e 0, respectivamente. Assim, todos os valores da matriz são normalizados e variam de 0 a 1. Isso permite que todos os pontos dentro do espaço da matriz de atributos tenham o valor do módulo do vetor até a origem de, no máximo, igual a 1.

Posteriormente, foi aplicado um algoritmo de redução de dimensionalidade na matriz de atributos, tendo em vista que o espaço em questão é muito esparso, porque possui 65 dimensões. Dessa forma, foi utilizado o algoritmo redutor de dimensionalidade UMAP (Uniform Manifold Approximation and Projection) [8] que oferece uma vantagem de estabilidade ao transformar os dados, o que não acontece, por exemplo com o algoritmo t-SNE que obtém um resultado diferente a cada transformação feita nos dados, não sendo interessante para este trabalho.

Algoritmo de agrupamento

Com o objetivo de encontrar grupos significativos dentro do conjunto de dados de vidros, o algoritmo do estado da arte HDBSCAN [6] foi utilizado. Esse algoritmo é uma extensão do DBSCAN, cuja abordagem de agrupamento se baseia no conceito de alcançabilidade.

A alcançabilidade diz que um ponto ***q*** dito diretamente alcançado por outro ponto ***p***, se a distância entre os pontos é menor do que um valor limite ϵ e ***p*** é cercado por um número suficiente de pontos. Então, ***q*** é considerado alcançável por ***p***, se existe uma sequência p_1, p_2, p_3, ..., p_n tais que $p_1 = p$ e p_{i+1} é diretamentamente alcançável por p_i.

Dessa forma, um *cluster* é um conjunto de pontos que satisfaz duas propriedades:

1. Todos os pontos de um cluster são mutuamente conectados, o que significa que para quaisquer dois pontos distintos *p* e *q* em um *cluster*, existe um ponto *o*, tal que *p* e *q* são alcançáveis por *o*;
2. Se um ponto é conectado a qualquer ponto em um *cluster*, esse ponto também faz parte do *cluster*.

O algoritmo DBSCAN, portanto, possui dois parâmetros: o raio da vizinhança de um ponto, *eps*, e o número mínimo de pontos da vizinhança de um ponto para que ele seja considerado um ponto de centro, *minpts*.

A abordagem do algoritmo HDBSCAN é complementar a ideia do DBSCAN, porque utiliza o mesmo conceito de alcançabilidade, porém, adiciona o conceito de hierarquia para contemplar todos os possíveis grupos de um conjunto de dados a partir de um dendrograma, por meio da variação do *eps*, como pode ser visto na Imagem 3. Dessa forma, o HDBSCAN se trata da extensão do DBSCAN, porque converte o último em um algoritmo de agrupamento hierárquico e utiliza uma técnica de extração de um agrupamento com base na estabilidade dos grupos encontrados.

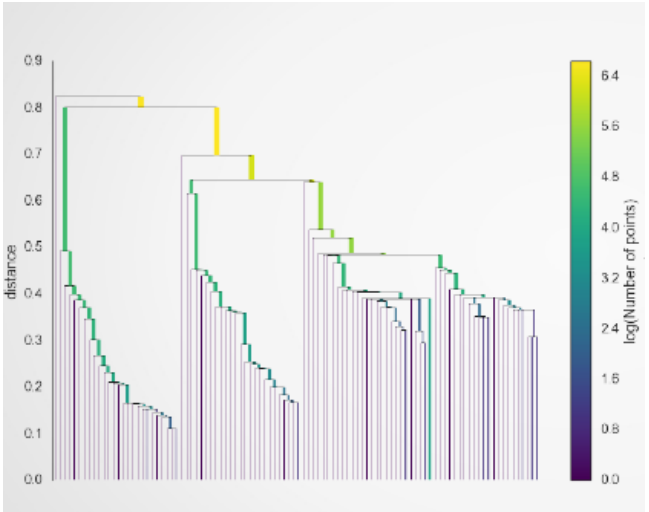

Imagem 3 - Exemplo de dendrograma gerado pela biblioteca do sklearn, 2016.

Além disso, o HDBSCAN se utiliza do conceito da distância de alcançabilidade mútua, que pode ser definida matematicamente pela fórmula a seguir,

$$d_{\text{mreach}-k}(a,b) = \max\{\text{core}_k(a), \text{core}_k(b), d(a,b)\}$$

em que *d(a,b)* é a distância entre os pontos *a* e *b*, e *core* é a distância necessária para que um ponto seja considerado ponto central.

Essa métrica faz com que pontos com baixa distância central permaneçam à mesma distância um do outro, mas os pontos mais esparsos são afastados com objetivo de que eles estejam separados de, pelo menos, a distância central de qualquer outro ponto, o que torna o algoritmo bastante poderoso para lidar com diferentes densidades e com *outliers*, pontos que foram gerados por mecanismos diferentes da maioria dos pontos, como pode ser visto na partição da Imagem 4.

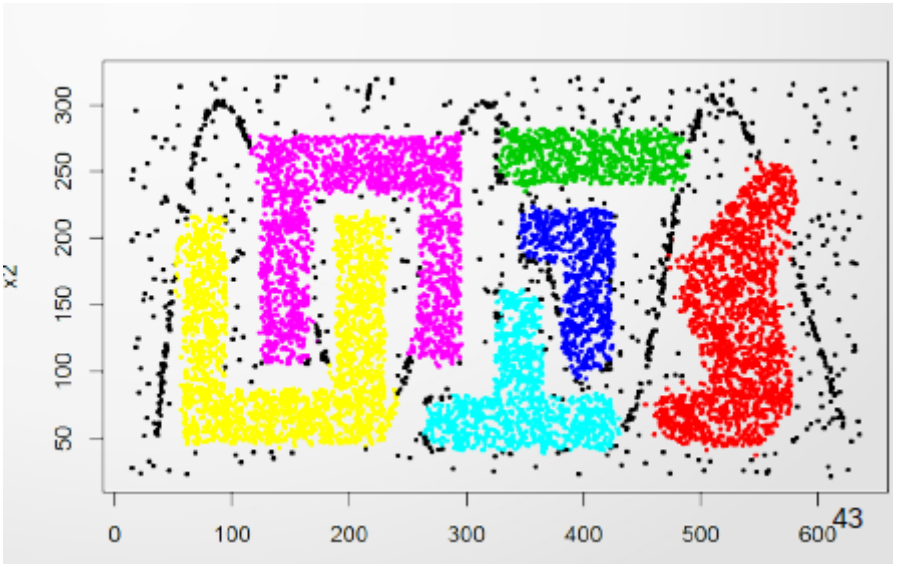

Imagem 4 - Partição gerada pelo algoritmo HDBSCAN.

Os parâmetros do algoritmo hierárquico são *min_cluster_size* e *min_samples*. O parâmetro *min_cluster_size* indica o menor grupo que será considerado *cluster*. Mesmo assim, esse parâmetro não é suficiente para que o algoritmo funcione como é esperado, então é necessário definir *min_samples*, que indica a quantidade de amostras que a vizinhança de um ponto deve ter para que ele seja considerado um ponto central, o que vai definir a quantidade de pontos que serão considerados *outliers*.

Diante disso, para saber quais valores dos parâmetros do HDBSCAN adequados para o conjunto de dados, é necessário analisar que parâmetros explicam melhor a natureza dos dados. Neste estudo, considerando os 43.641 vidros, concluiu-se que o valor mais adequado de *min_cluster_size* era de 160, ou seja, é necessário que um grupo tenha mais de 160 pontos para que ele seja considerado *cluster*.

A partir desse valor fixado, para saber o valor mais adequado de *min_samples*, foram analisados os dendrogramas gerados a partir do algoritmo, a distribuição de probabilidade dos pontos serem *outliers*, a porcentagem de *outliers* e a quantidade de pontos por *cluster*, como pode ser visto na Imagem 5.

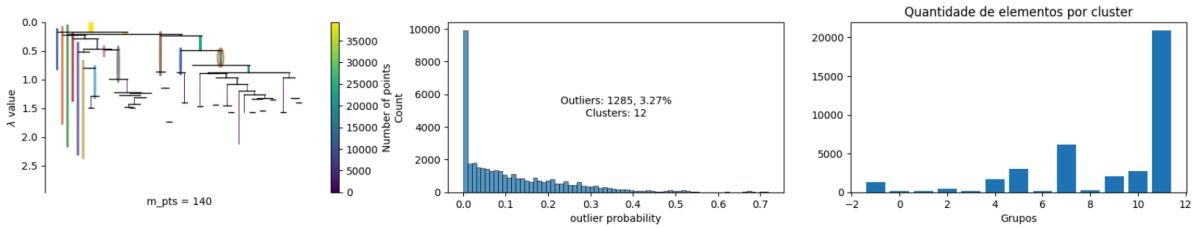

Imagem 5 - Gráficos norteadores da escolha de *min_samples*, para o caso em que *min_cluster_size* = 160 e *min_samples* = 140.

Tendo em vista os gráficos acima, é possível ressaltar que é importante selecionar grupos que contenham mais pontos do que a quantidade de *outliers*, além de que é importante considerar uma quantidade de *clusters* com adequada, para que depois seja possível interpretar o significado dos grupos encontrados à luz do conjunto de dados.

Assim, para o conjunto de dados dos vidros, considerando os conjuntos separados pelo *k-fold*, os valores adequados de *min_samples* ficavam entre 80 e 140, o que gerava uma quantidade adequada de grupos, além de grupos com tamanhos maiores em relação à quantidade de outliers e de uma partição com uma porcentagem de outliers menor que 20%.

Algoritmos de regressão

Em um estudo prévio [2], diferentes algoritmos de regressão foram utilizados para realizar a predição de diferentes propriedades de vidros. Os experimentos realizados no estudo foram reproduzidos neste estudo com o objetivo de escolher o melhor regressor para predizer a Tg do conjunto de dados de vidros.

Nesse sentido, o algoritmo que resultou nos melhores valores da métricas de validação, que serão detalhadas em uma seção posterior, foi o *Random Forest* (RF).

O *Random Decision Forest* [9] ou *Random Forest* [10] é um algoritmo baseado na ideia de modelos múltiplos (*ensemble*), em que modelos de aprendizado poderosos podem ser formados a partir de modelos chamados *weak learners*. Então, o RF se baseia no *bagging* [11], no qual os dados são primeiramente separados em diferentes amostras e, para cada uma das amostras, é criado um modelo de aprendizado e depois esses modelos são agregados gerando o classificador/regressor final.

No caso particular do RF, os modelos são árvores de decisão induzidas durante a fase de treino e, quando aplicadas ao conjunto de treino, a saída do modelo é a média das árvores individuais. Cada árvore de decisão é induzida por meio do *bagging*. Assim, o RF tende a ter menos *overfitting* do que as árvores individuais.

Treino e validação

Com o objetivo de treinar e validar os algoritmos de regressão, os dados foram divididos em conjuntos de treino e conjuntos de teste seguindo a estratégia de k-fold *cross validation*, com k=10. Essa estratégia divide aleatoriamente os dados em 10 conjuntos de tamanho similar, na qual 9 conjuntos são utilizados para treinar o modelo de regressão e o conjunto restante é usado para testar o modelo. Assim, obtém-se 10 conjuntos diferentes de treino e de teste, e a performance do regressor pode ser aferida por meio da média da performance de predição dos conjuntos de teste.

A partir dessa divisão, um modelo RF foi utilizado para treinar todos os dados de treino, chamado de regressor geral, e outros modelos de RF, chamados de regressores específicos, foram gerados a partir dos *clusters* da partição definida pelo algoritmo HDBSCAN. Os grupos gerados pela partição foram considerados significativos somente se o grupo tivesse uma quantidade maior de pontos do que a quantidade de *outliers*.

Com o objetivo de distribuir os pontos de treino nos grupos definidos pelo HDBSCAN, o algoritmo k-Nearest Neighbors (k-NN) [12] foi utilizado, tendo em vista que ele utiliza os k vizinhos mais próximos para classificar um ponto no conjunto de dados, o que é aplicável aos dados de vidros, porque vidros semelhantes possuem valores de Tg próximos. Dessa forma, se um ponto de teste fosse definido

como pertencente a um grupo significativo, ele faria parte do conjunto de teste do regressor correspondente àquele grupo. Caso contrário, o ponto seria considerado um *outlier* e faria parte do conjunto de teste do regressor geral.

Assim, as métricas de validação da performance dos regressores para o caso do método dos regressores específicos foram agregadas por meio de uma média ponderada, considerando a porcentagem de pontos de cada grupo significativo e dos *outliers*.

Os parâmetros do regressor Random Forest utilizados foram definidos conforme a Imagem 6. Esses parâmetros foram os que geraram melhores resultados no estudo prévio [2].

```
reg_geral = ensemble.RandomForestRegressor(n_estimators=933,
                                           random_state = 0,
                                           n_jobs=-1,
                                           max_features="sqrt")
```

Imagem 6 - Trecho de código contendo os parâmetros do regressor RF.

Métricas de validação

Para validar e comparar as regressões geral e específica, as métricas escolhidas foram:

- **Coeficiente de determinação (R^2):** É uma métrica estatística que representa o quão bem o modelo se adapta ao conjunto de dados. O valor dela varia entre 0 e 1, em que 0 diz que o modelo não representa em nada o conjunto de dados e 1 diz que o modelo está completamente adaptado ao conjunto de dados. A fórmula matemática que representa essa métrica pode ser vista logo abaixo:

$$R^2 = 1 - \frac{\text{sum squared regression (SSR)}}{\text{total sum of squares (SST)}},$$
$$= 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}.$$

- **Erro médio quadrático (RMSE):** Se trata do desvio padrão dos erros de predição para cada um dos pontos do conjunto de dados. Em outras palavras, essa métrica indica o quão concentrados estão os dados ao redor da linha que representa o modelo de regressão. A fórmula matemática que representa essa métrica pode ser vista logo abaixo.

$$\mathrm{RMSE}_{fo} = [\sum_{i=1}^{N} (z_{f_i} - z_{o_i})^2 / N]^{1/2}$$

Where:

- Σ = summation (“add up”)
- $(z_{f_i} - Z_{oi})^2$ = differences, squared
- N = sample size.

- **Desvio Absoluto Médio (MAD):** É uma medida estatística que quantifica a dispersão dos dados em relação à média. É calculada encontrando a diferença absoluta entre cada valor dos dados e a média desses valores, e então calculando a média dessas diferenças. O MAD é útil para entender a variabilidade dos dados em um conjunto, sendo menos sensível a valores extremos do que outras medidas de dispersão, como o desvio padrão. Isso faz com que o MAD seja uma escolha adequada quando se deseja uma medida robusta da variabilidade dos dados.

$$MAD = \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}|$$

Onde:

- MAD é o Desvio Absoluto Médio,
- n é o número total de observações,
- x_i são os valores individuais das observações,
- $\bar{x}$ é a média dos valores das observações.

Resultados e discussões

Para fazer a análise dos resultados, um dos conjuntos de dados de treino será analisado de maneira a exemplificar os resultados obtidos com o objetivo de facilitar o entendimento do experimento.

Nesse conjunto, o algoritmo HDBSCAN identificou 6 grupos significativos, ou seja, cujo número de elementos é maior do que a quantidade de *outliers*. A quantidade de elementos de cada grupo e as respectivas porcentagem em relação ao conjunto de treinamento são descritas abaixo:

- Grupo 1: 1662 (4,23%);
- Grupo 2: 2064 (5,26%);
- Grupo 3: 2686 (6,84%);
- Grupo 4: 3039 (7,74%);
- Grupo 5: 6139 (15,64%);
- Grupo 6: 20921 (53,29%).

A composição de cada um dos grupos quanto a Tg pode ser vista no conjunto de gráficos das imagens de 7 a 12.

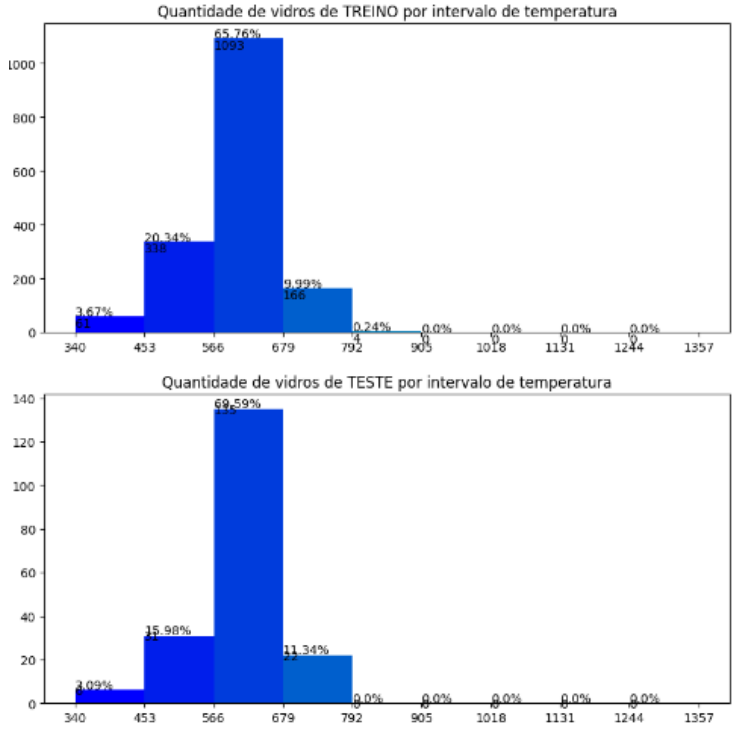

Imagem 7 - Distribuição dos vidros do grupo 1 em relação a Tg.

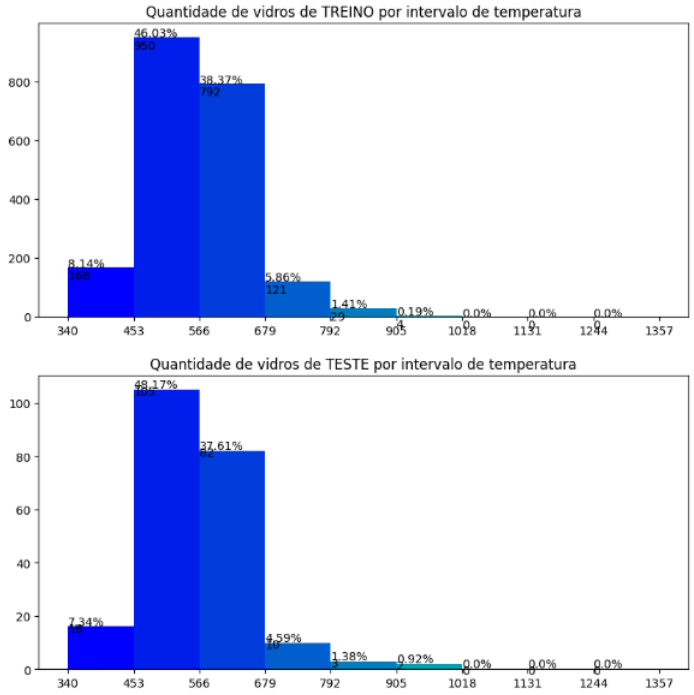

Imagem 8 - Distribuição dos vidros do grupo 2 em relação a Tg.

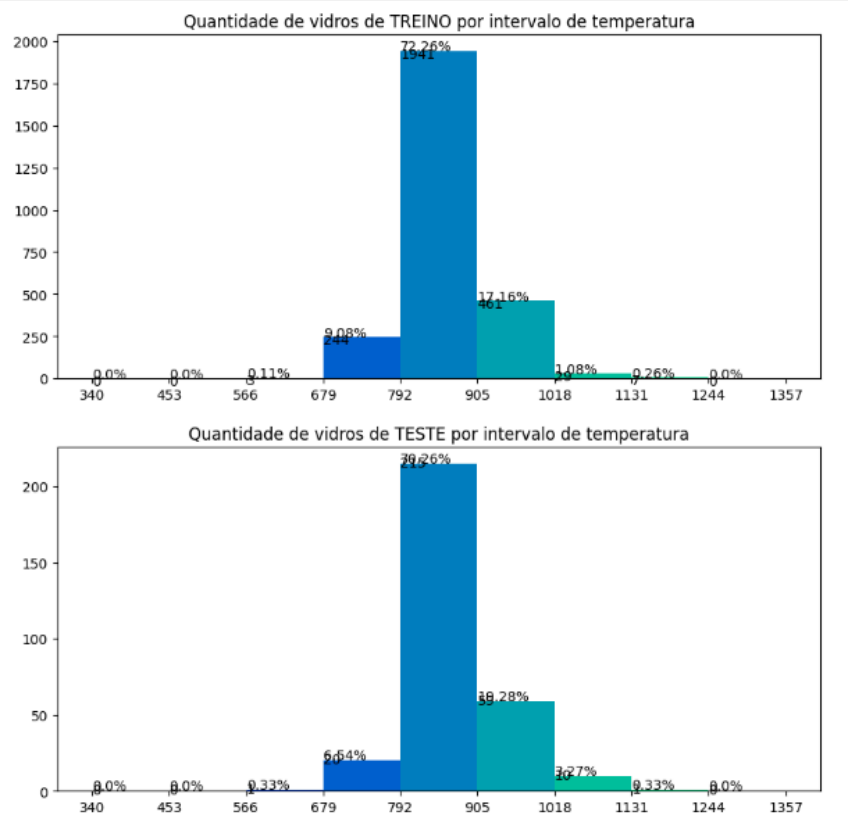

Imagem 9 - Distribuição dos vidros do grupo 3 em relação a Tg.

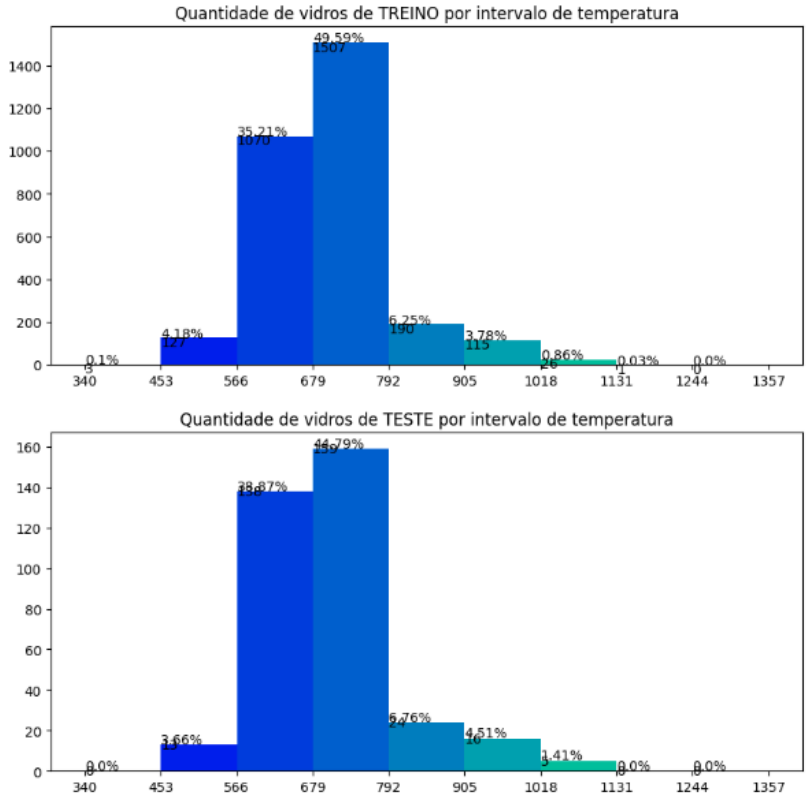

Imagem 10 - Distribuição dos vidros do grupo 4 em relação a Tg.

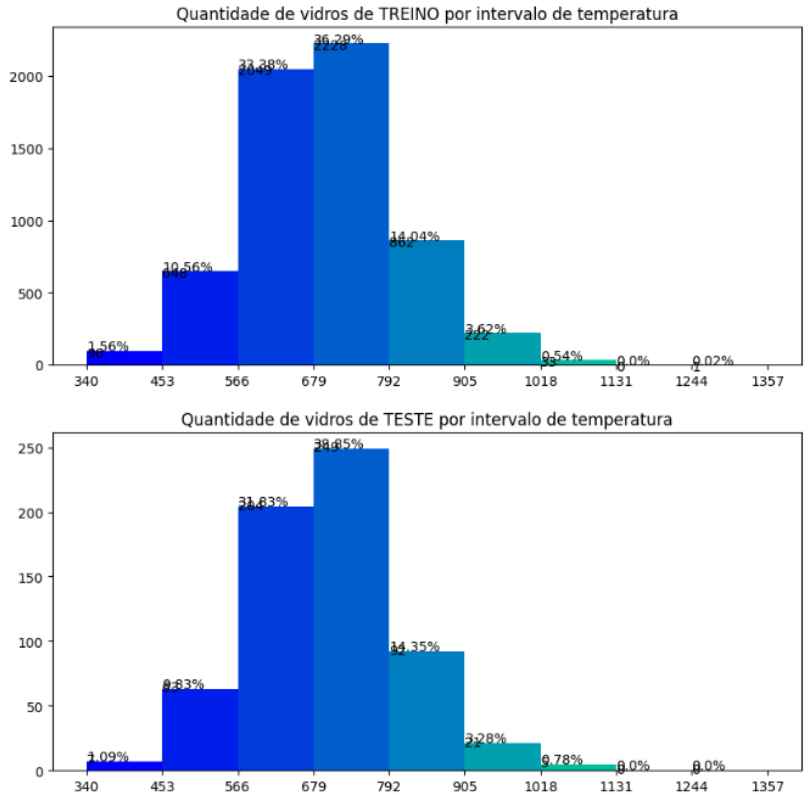

Imagem 11 - Distribuição dos vidros do grupo 5 em relação a Tg.

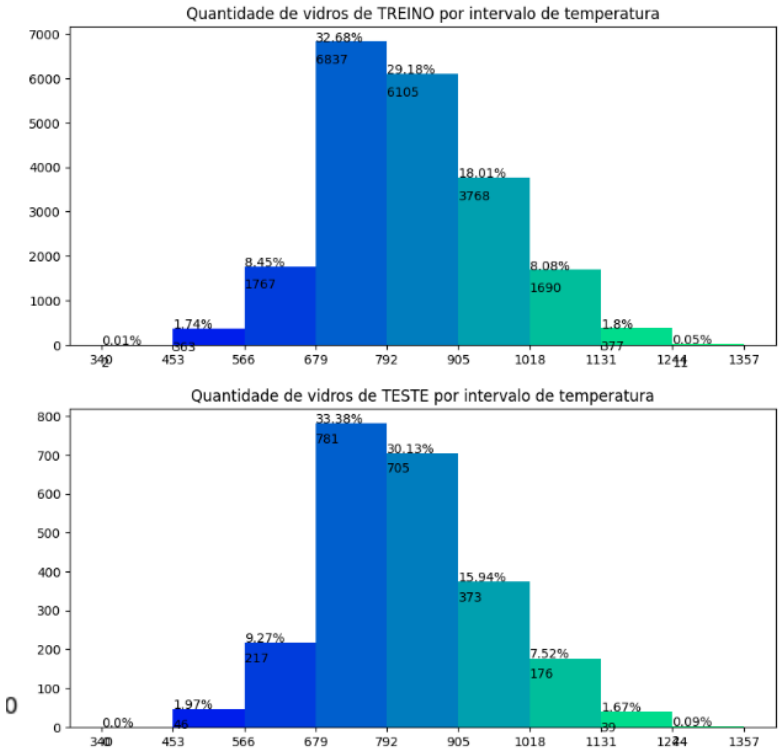

Imagem 12 - Distribuição dos vidros do grupo 6 em relação a Tg.

Ademais, o erro absoluto da predição da Tg para cada vidro dos grupos pode ser visualizada nas imagens de 13 a 18.

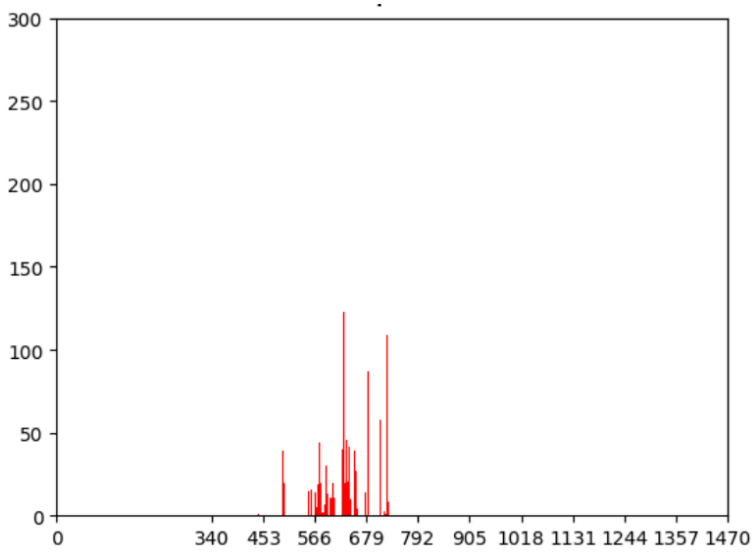

Imagem 13 - Erro absoluto da predição da Tg para cada vidro no grupo 1.

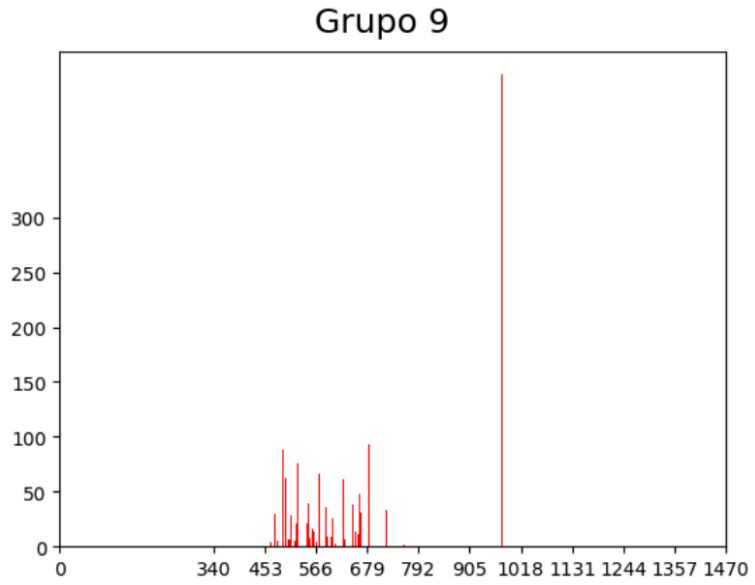

Imagem 14 - Erro absoluto da predição da Tg para cada vidro no grupo 2.

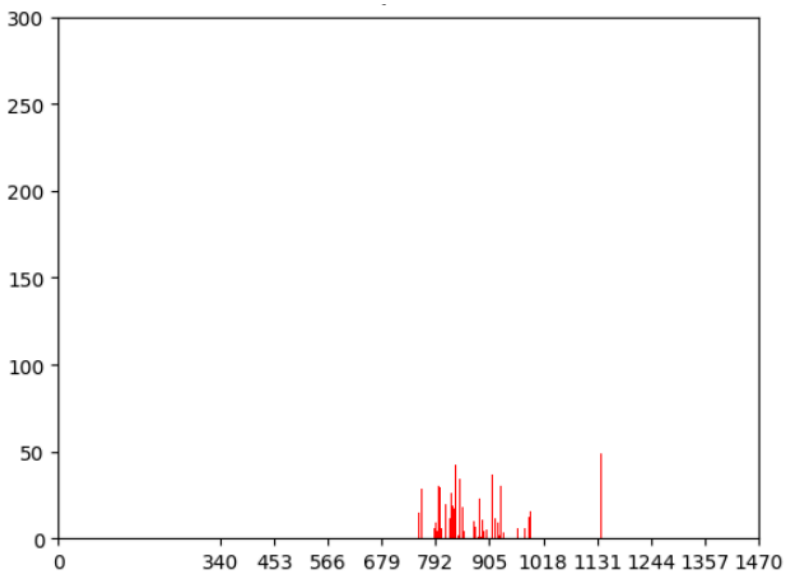

Imagem 15 - Erro absoluto da predição da Tg para cada vidro no grupo 3.

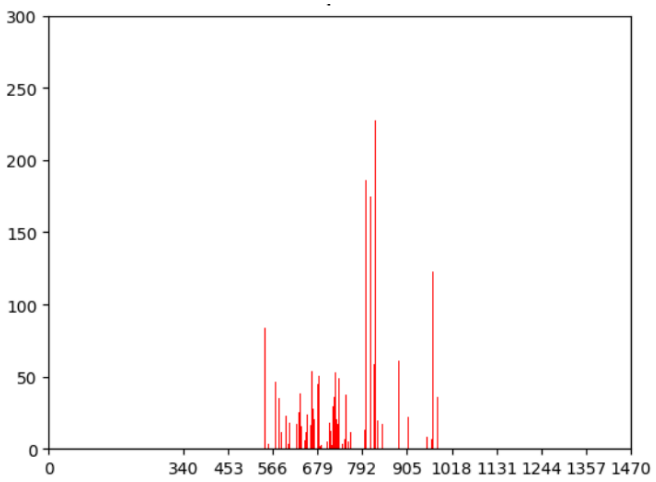

Imagem 16 - Erro absoluto da predição da Tg para cada vidro no grupo 4.

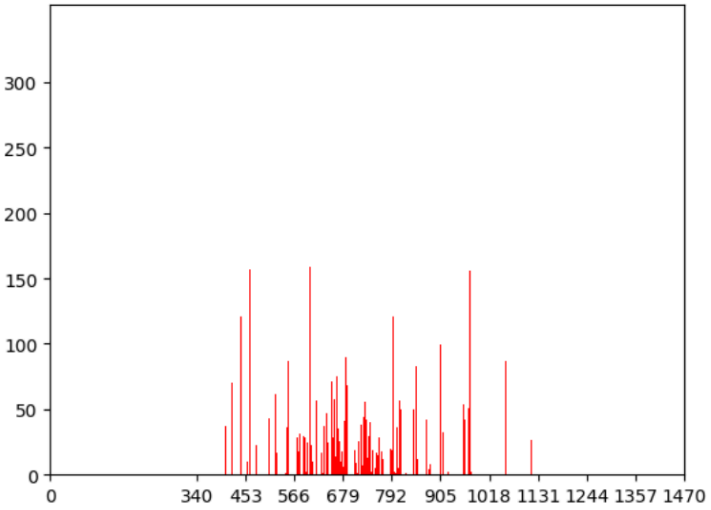

Imagem 17 - Erro absoluto da predição da Tg para cada vidro no grupo 5.

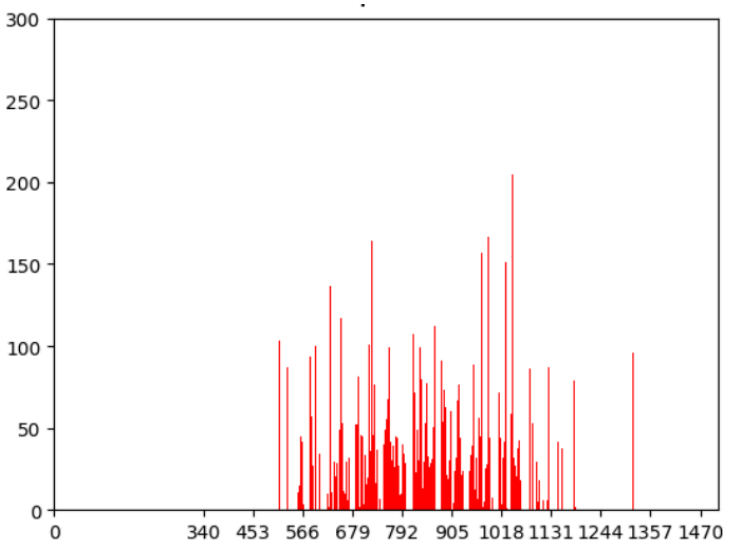

Imagem 18 - Erro absoluto da predição da Tg para cada vidro no grupo 6.

Por fim, os resultados das métricas de avaliação para cada um dos grupos podem ser visualizados na Tabela 1.

	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6
RMSE	28,47	58,94	20,58	37,95	43,47	35,35
R^2	0,79	0,55	0,9	0,85	0,84	0,93

Tabela 1 - Resultados das métricas avaliativas dos regressores específicos de cada um dos grupos.

A partir dos resultados apresentados, é possível perceber que o grupo que segue uma distribuição de vidros com relação a Tg semelhante a distribuição geral da Imagem 2, como o grupo 6 com distribuição ilustrado na Imagem 12, apresenta valor maior de R^2, uma métrica relacionada ao quanto um modelo de regressão é explicativo em relação ao conjunto de dados.

Dessa forma, quando a distribuição muda, como é o caso dos grupos 1 e 2, que apresentam uma maior concentração de vidros no intervalo de temperatura de 453 K a 679 K, do que no intervalo de 795 K a 905K, o valor de R^2 diminui.

Outro ponto importante é que é possível perceber que o HDBSCAN conseguiu separar os dados em grupos de vidros que possuem Tgs próximas, o que indica que o algoritmo funcionou como o esperado.

Por fim, é possível comparar o regressor geral e os regressores específicos por meio das métricas de validação de regressão quantitativas, conforme mostrado na Tabela 2.

	MAD	RMSE	R^2
Regressor geral	**23,36**	37,44	**0,94**
Regressores específicos	23,95	**36,8**	0,88

Tabela 2 - Resultados das métricas avaliativas dos regressores específicos e do regressor geral.

A comparação evidencia que o regressor geral obtém um melhor resultado de predição do que o conjunto de regressores específicos, o que refuta a hipótese inicial de que separar o conjunto de dados em grupos menores e fazer regressores específicos para cada um

deles seria uma maneira de tornar a predição mais precisa.

Esse resultado pode ser explicado por diferentes perspectivas. A primeira é o fato de que o algoritmo RF utiliza o *bagging* que gera um modelo que generaliza bem o conjunto de dados e evita o *overfitting*. Além disso, diferentemente de outros algoritmos de regressão, o RF é mais acurado com o aumento do número de árvores de decisão, tendo em vista que o resultado final é uma média dos resultados de cada uma das árvores individuais. Por isso que o regressor geral é mais acurado do que os regressores específicos e a métrica de R^2 resultou em um valor maior para o modelo geral.

A segunda perspectiva está relacionada ao fato de que o HDBSCAN, por se tratar de um algoritmo baseado em densidade, cria grupos a partir da proximidade dos pontos no espaço do conjunto de dados. Porém, é possível que o "efeito corrente" [6], um efeito que faz com que grupos contenham elementos distintos, porque estão em pontas opostas do grupo, tenha deixado os grupos com vidros que podem até possuir temperaturas próximas, mas possuem composições químicas distintas e não podem ser considerados vidros semelhantes.

Por fim, a terceira perspectiva está relacionada ao processo de redução de dimensionalidade feita na fase de pré-processamento dos dados. Sabendo que a dimensão original do conjunto de dados é de 65 e que ela foi reduzida para 10, isso tornou o espaço de pontos menos esparso, mesmo assim, pode ter criados pontos que não necessariamente correspondem aos vidros iniciais. Essa perspectiva combinada à perspectiva do "efeito corrente" mostra que o pré-processamento pode não ter sido adequado para o conjunto de dados e fez com que os grupos encontrados não fossem formados por vidros parecidos, o que foi verificado ao final do estudo.

Conclusões

O projeto de pesquisa permitiu o aprendizado de diversos tópicos da teoria de aprendizado de máquina supervisionado e não-supervisionado aplicados ao estudo de dados de vidros, além do ganho de conhecimento sobre redução de dimensionalidade do conjunto de dados.

Conclui-se, portanto, que os objetivos específicos deste trabalho foram alcançados, tendo em vista que o estudo e aplicação dos conhecimentos da teoria de aprendizado de máquina supervisionado e não-supervisionado, da teoria de algoritmos de redução de dimensionalidade e da teoria da avaliação do desempenho dos algoritmos de regressão foram aplicados aos dados de vidros.

Apesar do objetivo geral de melhoramento da performance da predição de Tg de vidros não ter sido alcançado, foi possível avançar na pesquisa do uso de técnicas avançadas de aprendizado de máquina visando encontrar novas composições de vidros.

Assim, a partir do estudo feito neste trabalho, é possível aplicar outros modelos de regressão, como algoritmos que utilizam *boosting*, uma técnica muito poderosa de predição. Além disso, é extensivo a este trabalho utilizar outras técnicas de pré-processamento, como redução de dimensionalidade, com o objetivo de avaliar quais as melhores técnicas para o conjunto de dados de vidros.

Ademais, é possível estudar a predição de outras propriedades de vidros, como o índice de refração, tendo em vista que se trata de uma propriedade cuja medição é muito mais precisa do que a medição da Tg, por exemplo, o que pode melhorar a performance dos algoritmos de predição em relação ao que foi feito nesta pesquisa.

Referências

[1] ZANOTTO, E., COUTINHO, F. How many non-crystalline solids can be made from all the elements of the periodic table? Journal of Non-Crystalline Solids, 347, 285–288 (2004). URL http://www.sciencedirect.com/ science/article/pii/S0022309304005101.

[2] ALCOBAÇA, E., et al. Explainable Machine Learning Algorithms for Predicting Glass Transition Temperatures. Acta Materialia, v. 188, p. 92–100, abr. 2020, https://doi.org/10.1016/j.actamat.2020.01.047.

[3] TSHITOYAN, V. et al. Unsupervised word embeddings capture latent knowledge from materials science literature. Nature 571, 95 (2019).

[4] HUO, H. et al. Semi-supervised machine-learning classification of materials synthesis procedures. npj Computational Materials 5, 62 (2019).

[5] MICCIO, L. A.; SCHWARTZ, G. A. Mapping Chemical Structure-Glass Transition Temperature Relationship through Artificial Intelligence. Macromolecules, v. 54, n. 4, p. 1811-1817, fev. 2021.

[6] CAMPELLO, R. J. G. B., et al. Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection. ACM Transactions on Knowledge Discovery from Data, v. 10, n. 5, jul. 2015. Disponível em: https://doiorg.ez31.periodicos.capes.gov.br/10.1145/2733381.

[7] MAZURIN, O. V. & Priven, A. I. SciGlass - Glass Information System - Glass Database - Glass Properties (ITC, Inc., 2017). URL http://www.sciglass.info/.

[8] UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2018. Disponível em: https://umap-learn.readthedocs.io/en/latest/#umap-uniform-manifold-approximation-and-projection-for-dimension-reduction .

[9] HO, T. K. Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition, vol. 1, 278–282 (IEEE, 1995).

[10] BREIMAN, L. Random forests. Machine learning 45, 5–32 (2001).

[11] BREIMAN, L. Bagging predictors. Machine learning 24, 123–140 (1996).

[12] WEINBERGER, K. Q. & Saul, L. K. Distance metric learning for large margin nearest neighbor classification. Journal of Machine Learning Research 10, 207–244 (2009).