Pontifícia Universidade Católica de Campinas

Ciência de dados e Inteligência Artificial

Projeto 3 - Algoritmo KNN
Relatório de Análise do *Dataset* "Glass Identification"

Giovane Bruno Nardari - 21000244

João Roberto Crespi Junior - 21003801

Introdução

Para este relatório, teremos como objeto de estudo o *Dataset "Glass Identification"*, que pode ser encontrado no website https://archive.ics.uci.edu/ml/datasets/glass+identification. O *Dataset* foi criado com o intuito de auxiliar investigações criminais, uma vez que a identificação correta do tipo de vidro coletado, pode transformar essa prova em uma evidência concreta para casos diversos.

O *Dataset* contém dados sobre 214 amostras de vidros, que são classificadas em sete diferentes tipos, nomeados em inglês como: ***Building windows float-processed (Tipo 1), Building windows non-float-processed (Tipo 2), Vehicle windows float-processed (Tipo 3), Vehicle windows non-float-processed (Tipo 4), Containers (Tipo 5), Tablewares (Tipo 6), Headlamps (Tipo 7)*** (entretanto, nesse *Dataset,* existem apenas exemplares de seis dos sete diferentes tipos de vidro, uma vez que não foi contabilizado nenhum vidro do tipo *"**Vehicle windows non-float-processed**"*), além de conter informações sobre a composição química, como as suas concentrações de **Sódio (Na), Magnésio (Mg), Alumínio(Al), Silício (Si), Potássio (K), Cálcio (Ca), Bário (Ba) e Ferro (Fe)**, e também uma de suas propriedades físicas, o **Índice de Refração (RI)**.

Como base do estudo iremos aplicar o algoritmo KNN, apresentado em aula, para a tarefa de classificação dos tipos de vidros. O resultado será capaz de indicar a precisão do algoritmo em analisar corretamente as similaridades entre as características de cada exemplar do *Dataset*. Contudo, é necessário compreender que o *Dataset* possuí apenas 214 amostras,

distribuídas de forma desigual entre as classes, ou seja, apresenta dois fatores que podem afetar negativamente os resultados do algoritmo KNN.

Visualização e Análise

Inicialmente, buscamos constatar quaisquer tipos de ruídos que poderiam existir no *Dataset* original, concluímos que não existiam dados faltantes para quaisquer atributos dos exemplares. Em seguida, foram verificadas algumas medidas estatísticas - as primeiras análises feitas foram acerca de: a quantidade de exemplares por tipo de vidro (**Figura 1**), a extração da média dos atributos de cada tipo (**Figura 2**), e a visualização dos outliers (**Figura 3**).

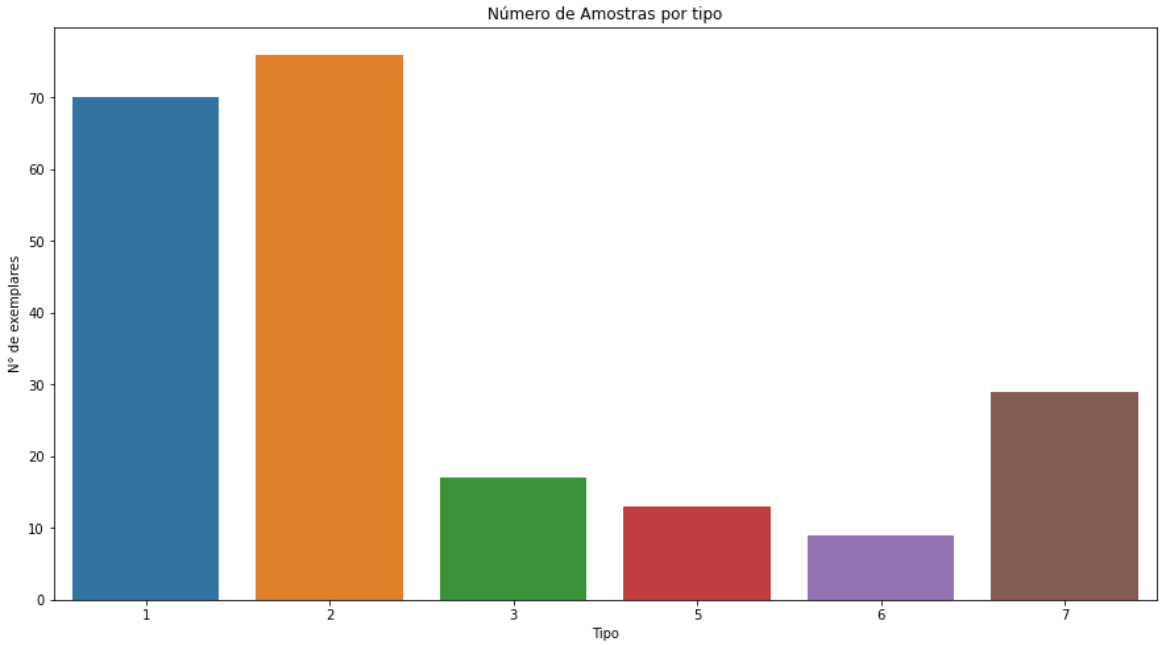

Figura 1 - Gráfico de barras contabilizando o número total de exemplares para cada tipo.

Type	Ri	Na	Mg	Al	Si	K	Ca	Ba	Fe
1	1.518718	13.242286	3.552429	1.163857	72.619143	0.447429	8.797286	0.012714	0.057000
2	1.518619	13.111711	3.002105	1.408158	72.598026	0.521053	9.073684	0.050263	0.079737
3	1.517964	13.437059	3.543529	1.201176	72.404706	0.406471	8.782941	0.008824	0.057059
5	1.518928	12.827692	0.773846	2.033846	72.366154	1.470000	10.123846	0.187692	0.060769
6	1.517456	14.646667	1.305556	1.366667	73.206667	0.000000	9.356667	0.000000	0.000000
7	1.517116	14.442069	0.538276	2.122759	72.965862	0.325172	8.491379	1.040000	0.013448

Figura 2 - Média dos valores dos atributos para cada tipo de vidro.

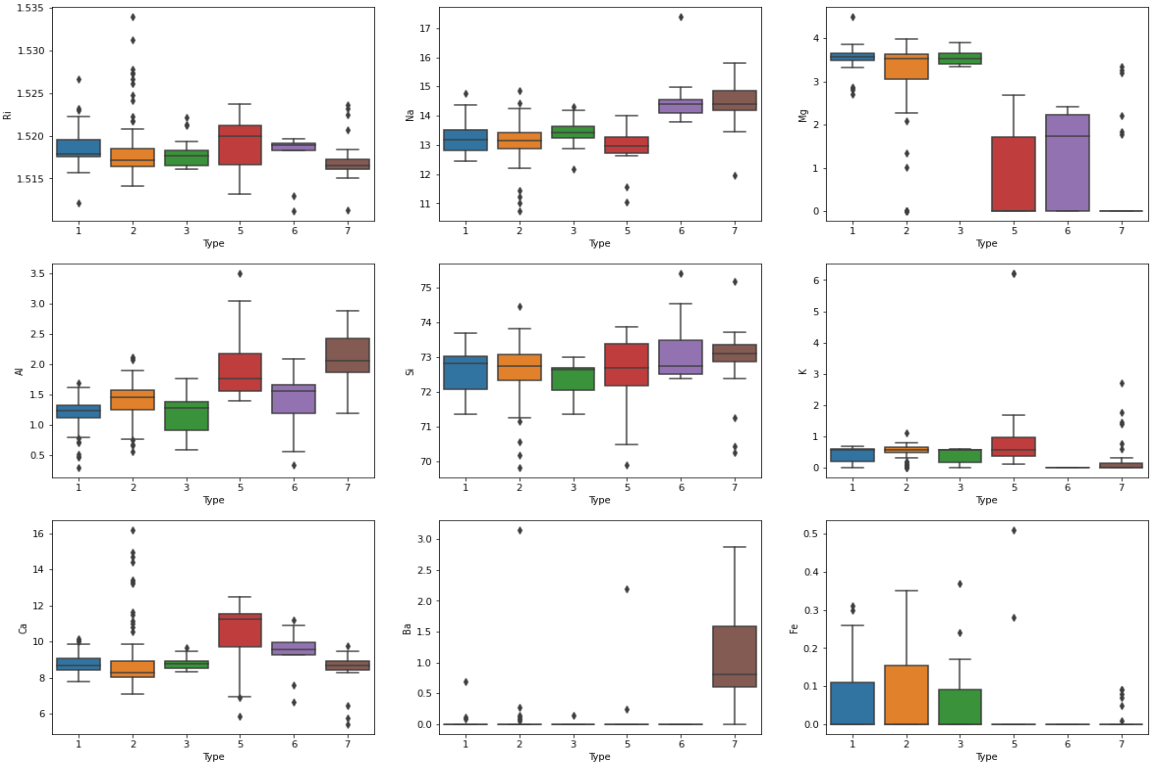

Figura 3 – *Boxplots* comparando os tipos de vidro para cada um de seus atributos.

Durante as observações feitas sobre a análise inicial, constatamos que existe uma disparidade na quantidade de exemplares para cada tipo, os exemplares dos tipos um e dois apresentam uma quantidade maior de amostras - a causa disso provavelmente se deve a fonte dos dados de estudo, o Serviço de Ciência Forense dos Estados Unidos, que, possivelmente, contabilizou em maior quantidade amostras provindas de invasões domiciliares.

A partir da análise realizada sobre os atributos, foi possível observar algumas tendências para os tipos de vidro em relação aos seus atributos. Notou-se que o índice de refração é similar para todos os tipos, entretanto, o tipo 6 contém exemplares com valores mais concentrados para os índices de refração, e, em contrapartida, o tipo 5 possuí uma variação elevada dos valores para esse mesmo atributo entre seus exemplares. Para o sódio, os tipos 6 e 7 conseguem se destacar em relação aos outros, por terem seus exemplares com concentrações mais elevadas desse elemento. No Magnésio os tipos 1, 2 e 3 apresentam uma média acima de 3% em sua composição. A concentração de Alumínio nas amostras é maior para os tipos 5 e 7. O Potássio aparece em baixa quantidade, em especial no tipo 6 em que nenhum de seus exemplares possuí K em sua composição. O Cálcio é mais presente no tipo 5 e o Bário no tipo 7. O Ferro apresenta baixas concentrações em todos os tipos, entretanto, dentre eles, 1, 2 e 3 apresentam valores maiores. Não foi possível realizar conclusões relevantes sobre a Sílica, uma vez que todos os

tipos de vidro, nesse caso, apresentam quantidades semelhantes do elemento em sua composição, com exceção de alguns *outliers*.

Também foi realizado um gráfico de *Heatmap* (**Figura 4**) para encontrar multicolinearidade entre os atributos, que indicaria irrelevância na análise de um dos atributos dos pares de variáveis que apresentam índice de correlação alto, e também um gráfico de distribuição normal (**Figura 5**), para encontrar quais seriam os melhores atributos para distinguir, por exemplo, as classes 1, 2 e 3, que, como foi visto anteriormente, trazem maior dificuldade. Quando observados, nota-se, no *Heatmap*, que existe apenas um par de atributos com correlação significativa, que seriam o RI e o Ca, portanto, entende-se que não seria necessário analisar ambos para impactar a tarefa de classificação, justificando uma possível redução de dimensionalidade. Por sua vez, nos gráficos da **Figura 5,** foi possível identificar uma sobreposição das curvas entre os tipos 1 e 3, o que indica sua similaridade, e ao mesmo tempo, pode diferenciá-los do tipo 2, nos gráficos de **Ri**, **Mg**, **Ba**, **Ca** e **Ba.**

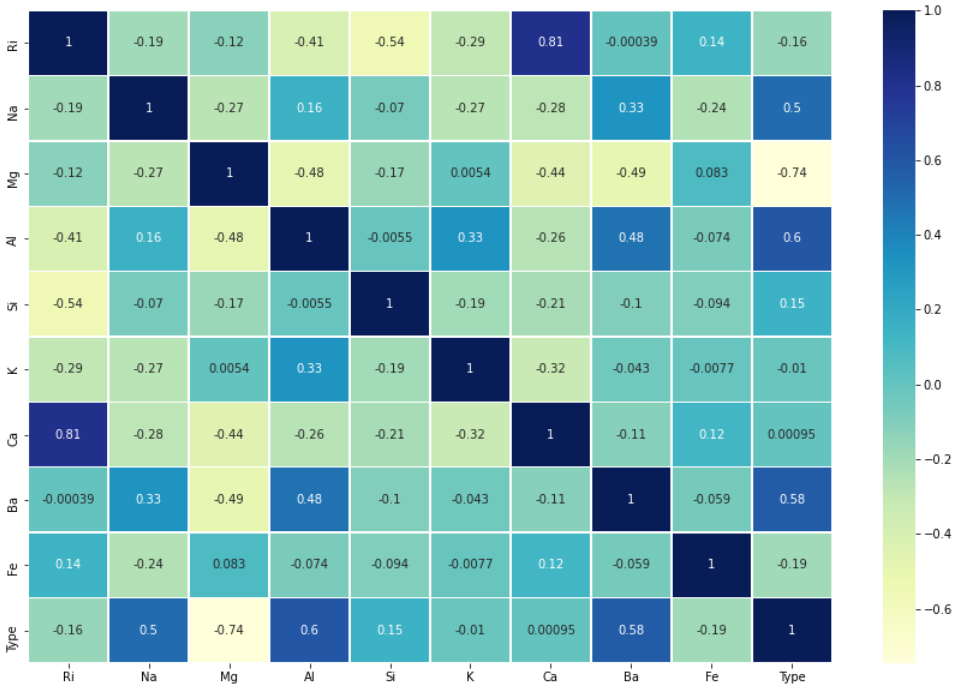

Figura 4 – *Heatmap* indicando multicolinearidade entre os atributos.

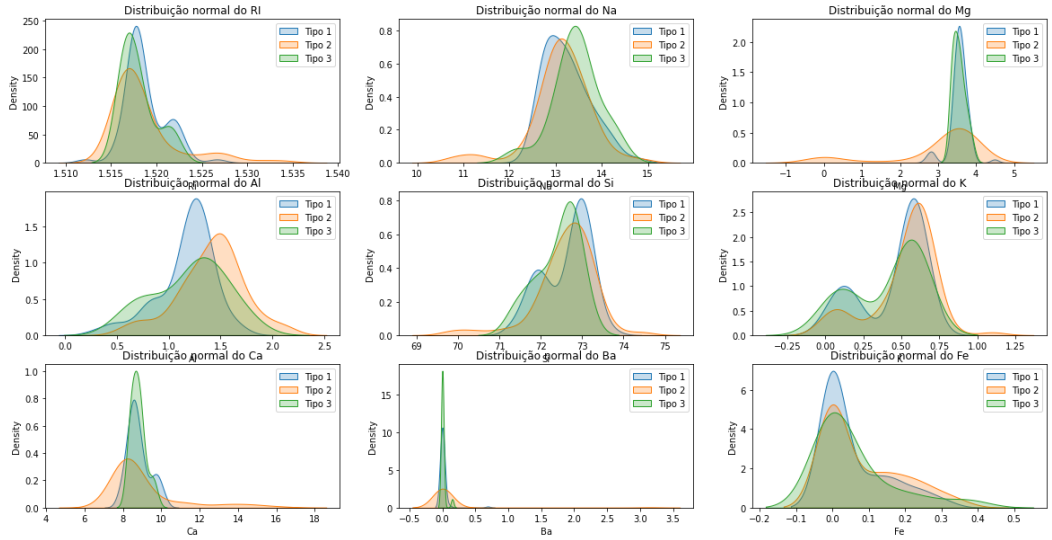

Figura 5 – Distribuição normal dos atributos para os tipos 1, 2 e 3.

Algoritmo K-Neareset Neighbor

Para a tarefa de classificação dos exemplares, aplicamos o algoritmo KNN sobre o *Dataset.* A princípio, os resultados para o *score*, desse algoritmo, variavam, aproximadamente, entre 44,2% e 93%, como exibido no gráfico com a distribuição desses resultados (**Figura 6**), e com média de 65,6%, dependendo do número "k" de vizinhos que seriam comparados com cada exemplar, e das amostras que eram selecionadas, para treino e teste. Para a visualização da classificação feita pelo algoritmo, também plotamos uma Matriz de Confusão para constatar os principais erros do algoritmo, quando ele obteve seu *score* mais alto (**Figura 7**).

Em seguida, em busca de melhorar os resultados obtidos pelo algoritmo aplicado, foi adicionado ao código dois processos de remoção de *outliers*, ruídos que já haviam sido identificados visualmente durante a análise dos *boxplots*, foi testada, principalmente, a remoção dos *outliers* de todos os tipos de vidro. O critério utilizado para identificar os *outliers* foi baseado no método de ajuste dos limites superior e inferior dos próprios *boxplots*, ou seja, foram calculados o primeiro e terceiro quartil (Q1 e Q3, respectivamente) e o Alcance Interquartil (AIQ), para os dados de cada tipo de vidro sobre cada um dos seus atributos. Então, calculou-se o ajuste dos limites superior e inferior de cada *boxplot*, dados pelas fórmulas: Limite inferior = Q1 – AIQ * 1,5; e Limite superior = Q3 + AIQ * 1,5. Após, os dados que superavam esses limites foram considerados *outliers*, e então removidos do *Dataset* para novos testes do algoritmo KNN.

Os novos testes do algoritmo, realizados sobre o *Dataset* após a remoção dos dados que foram considerados ruidosos para todas as classes, obtiveram *scores* melhores. Os *scores* para esse novo *Dataset* variavam entre **55,2%** e **93,1%**, porém, agora, com média de **73,3%**. Entretanto, esses resultados foram considerados não suficientemente satisfatórios para justificar a remoção dos *outliers*, pois essa remoção reduziu a quantidade de exemplares para **144**, ou seja, **70** exemplares (ou, aproximadamente, **32,7%** dos dados originais) foram removidos para uma melhora de **7,7** pontos percentuais na média dos *scores*, que se traduzem em, apenas, **3** classificações corretas a mais, em média.

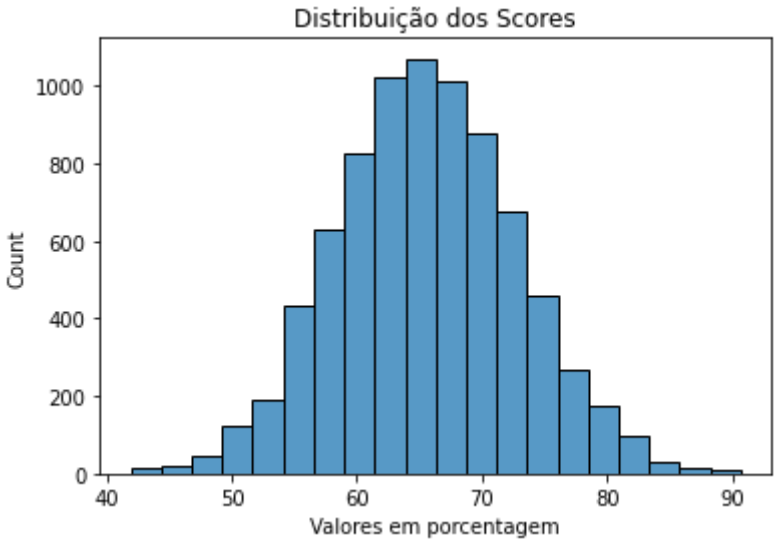

Figura 6 - Distribuição dos resultados (*scores)* obtidos para diferentes "k" dentro do algoritmo KNN, quando aplicado sobre o *Dataset* original.

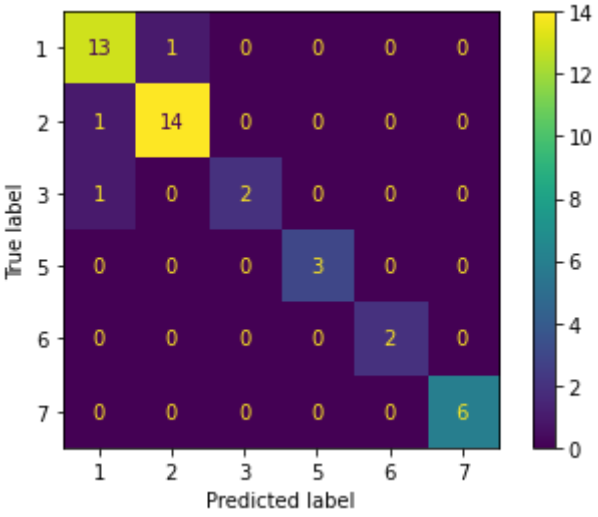

Figura 7 - Matriz de confusão pós algoritmo KNN, aplicado sobre o *Dataset* original.

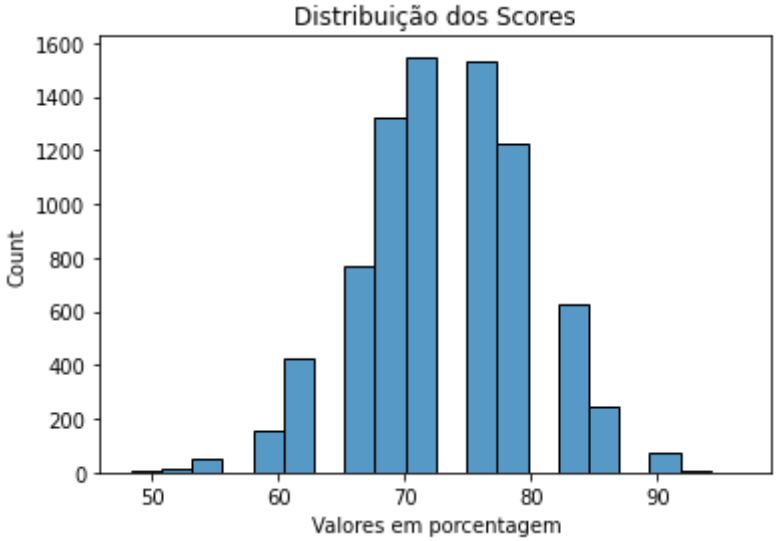

Figura 8 - Distribuição dos resultados (*scores*) obtidos para diferentes "k" dentro do algoritmo KNN, quando aplicado sobre o *Dataset* modificado.

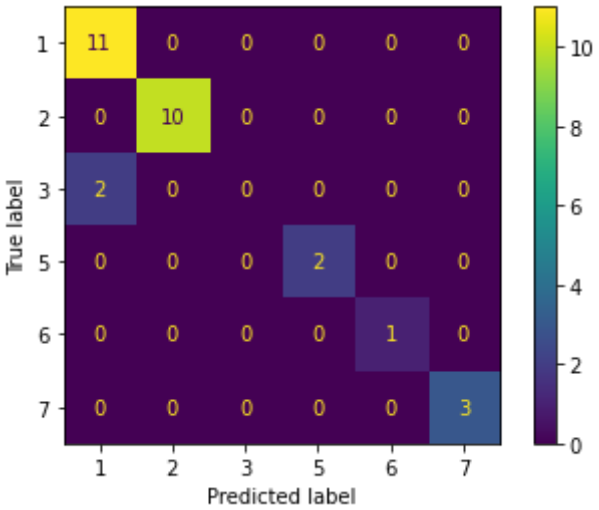

Figura 9 - Matriz de confusão pós algoritmo KNN, aplicado sobre o *Dataset* modificado.

Conclusão sobre a aplicação do algoritmo KNN

Durante os testes realizados, com múltiplas combinações, a melhor configuração encontrada no algoritmo, levando em conta o objetivo da classificação e o contexto do *Dataset*, foi de k = 1 para o *random_state* 9917, com o *Dataset* original, obtendo o *score* de 93,02%, já que o modificado apresenta uma remoção significativa dos dados e obteve 0% de precisão sobre a classe 3.

Tratando-se novamente da remoção dos *outliers*, constatou-se que o *Dataset* modificado apresentou melhores *scores*, contudo, percebemos que a redução de 32,7% dos dados, que já estavam sob o problema de possuir poucos exemplares, poderia refletir um resultado irreal, uma

vez que, estatisticamente, é sempre importante ressaltar que são as amostragens maiores que melhor representam a realidade estudada, então, por isso, foi decidido que o *Dataset* original deveria ser mantido.

Os erros de classificação ocorridos, envolviam, majoritariamente, os tipos 1, 2 e 3, e dependiam da amostra selecionada e do "k" utilizado. Esses erros foram, provavelmente, causados pela alta semelhança entre os exemplares dessas três classes, aliado a menor quantidade de exemplares do tipo 3. Entretanto, entende-se que, para aplicação prática, os erros de classificação entre as três classes mencionadas podem ser considerados menos significativos, pois olhares profissionais do ramo forense são capazes de distinguir os vidros que se encaixam nas categorias "*float-processed*" e "*non-float-processed*".

Os testes realizados nos deixam a conclusão de que, seria possível classificar corretamente todas as classes, com poucos erros, caso existissem mais dados para serem oferecidos no treino do algoritmo.