Pré-Processamento de Dados de Identificação de Vidro

Daniel Lemos Simoes
398985
daniellems@alu.ufc.br

Gustavo Filipe do Nascimento
402889
gusnas@alu.ufc.br

Pedro Cercelino Matos
399325
*p*edrocercelino@alu.ufc.br

Clailton Almeida Lopes
400091
*c*lailtonx2@gmail.com

Lucas Martins de Oliveira
398900
*e*ng.lucas@alu.ufc.br

***Resumo*—Com o propósito de obtermos alto rendimento ao trabalhar com diversos dados, é necessário utilizar técnicas que possam deixar estes dados mais limpos o possível, para isso aplicamos técnicas de pré-processamento de dados. Utilizamos a linguagem R para aplicarmos os métodos nos conjuntos de dados fornecidos. Como esperado, as técnicas reduziram a quantidade de dados, fazendo com que o trabalho fosse mais rápido e descomplicado.**

I. Introdução

Com a popularização do uso de inteligência artificial e aumento da geração de dados é importante saber filtrar tais dados. Tendo em vista que o poder computacional cresce num ritmo menor do que a geração de informação.

Por vezes é necessário fazer uma melhor filtragem desses dados já que os mesmos podem apresentar falhas devido a sua natureza, pois o mundo real é complexo e envolve outras variáveis que não podem ser coletadas durante o processo.

Para isto vamos ajustar o dataset, tendo em vista deixar os dados mais limpos, desta forma melhorando a qualidade dos mesmos, facilitando então o processo de predição.

II. Metodologia

Para o trabalho desenvolvido, foi utilizado um dataset que reúne informações a respeito de 214 amostras de vidro com características diversas. O conjunto de dados também apresenta detalhes sobre 9 preditores. São eles: Índice de refração (RI) e a quantidade de elementos químicos Na, Mg, Al, Si, K, Ca, Ba e Fe presentes em cada variação de vidro em forma de porcentagem.

A análise dos dados foi dividida em quatro etapas: unconditional mono-variate analysis, class-conditional mono-variate analysis, unconditional bi-variate analysis e unconditional multi-variate analysis.

A. Unconditional mono-variate analysis

Nesta etapa os dados são analisados sem levar em consideração a divisão de classes proposta pelo próprio dataset. Para visualização dos resultados de média, desvio padrão e assimetria, foi construída a seguinte tabela:

Tabela I
Unconditional mono-variate

	Média	Desvio Padrão	Assimetria
RI	1.51837	0.00304	1.60272
Na	13.40785	0.81660	0.44783
Mg	2.68453	1.44241	-1.13645
Al	1.44491	0.49927	0.89461
Si	72.65093	0.77455	-0.72024
K	0.49706	0.65219	6.46009
Ca	8.95696	1.42315	2.01845
Ba	0.17505	0.49722	3.36868
Fe	0.05701	0.09744	1.72981

Ao observar os valores obtidos, nota-se que o elemento com maior presença na composição das variações de vidro estudadas é o silício (Si), pois possui a média mais alta. Por outro lado, o elemento menos presente é o ferro (Fe). Percebe-se também que o desvio padrão do silício (0.77455) é maior que o do ferro (0.09744), indicando que as quantidades presentes do segundo elemento nas composições estão mais próximas da média, pois o desvio está mais próximo de 0.

A terceira métrica, assimetria (skewness), demonstra a posição geral das amostras no histograma da variável relacionada. O sinal indica o lado em que a ”cauda”da distribuição está posicionada, tendo a média como centro. Como exemplo, segue o histograma do índice de refração (Ri):

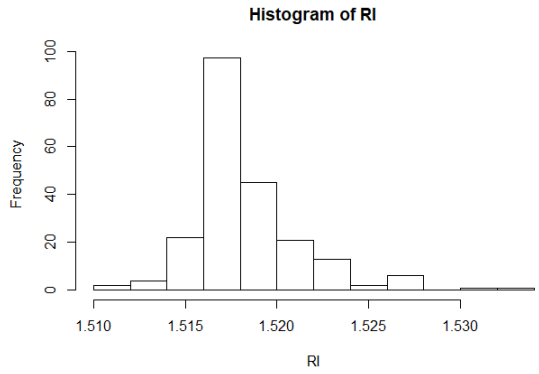

Figura 1. Índice de refração do dataset

A assimetria do Ri é de 1.60272. Por ser um valor positivo, a cauda encontra-se à direita do gráfico. Portanto, o oposto da cauda, a maior parte das distribuições, encontram-se à esquerda do histograma. Isso indica que a maioria das amostras apresentam valores abaixo da média para o índice de refração.

B. Class-conditional mono-variate analysis

A próxima análise a ser feita leva em consideração as classes presentes no dataset. Essas classes variam de 1 até 7, mas com nenhuma amostra na classe 4.

A seguir são apresentadas as tabelas contendo a média, desvio padrão e assimetria das classes 1, 3 e 7.

Tabela II
CLASSE 1

	Média	Desvio Padrão	Assimetria
RI	1.51872	0.00227	0.74373
Na	13.24229	0.49930	0.75382
Mg	3.55243	0.24704	-0.67673
Al	1.16386	0.27316	-1.08004
Si	72.61914	0.56948	-0.55427
K	0.44743	0.21488	-0.89977
Ca	8.79729	0.57481	0.68638
Ba	0.01271	0.08384	7.56199
Fe	0.05700	0.08907	1.30412

Tabela III
CLASSE 3

	Média	Desvio Padrão	Assimetria
RI	1.51796	0.00192	0.97080
Na	13.4371	0.50689	-0.46194
Mg	3.54353	0.16279	0.60216
Al	1.20118	0.34749	-0.33255
Si	72.4047	0.51228	-0.69140
K	0.40647	0.22989	-0.63772
Ca	8.78294	0.38011	0.78536
Ba	0.00882	0.03638	3.42403
Fe	0.05706	0.10786	1.69194

Tabela IV
CLASSE 7

	Média	Desvio Padrão	Assimetria
RI	1.51712	0.00255	0.97652
Na	14.44206	0.68636	-1.44730
Mg	0.53828	1.11768	1.63207
Al	2.12276	0.44273	-0.29284
Si	72.96586	0.94023	-1.21153
K	0.32517	0.66849	2.13951
Ca	8.49138	0.97351	-1.93554
Ba	1.04000	0.66534	0.45058
Fe	0.01345	0.02979	1.78060

Assim como feito na etapa anterior, pode-se observar características nos resultados que refletem nos histogramas referentes. Toma-se como exemplo a quantidade de sódio (Na) nas amostras pertencentes a classe 7. O seu valor de assimetria é -1.44730. Por ser um valor negativo, sabe-se que a maioria das amostras estará na direita da média, que por sua vez é de 14.44206. Note-se também que, nesse caso, o módulo da skewness é muito próximo de 1, indicando que o deslocamento é perceptível. Essas informações podem ser comprovadas com a observação do histograma referente ao preditor mencionado.

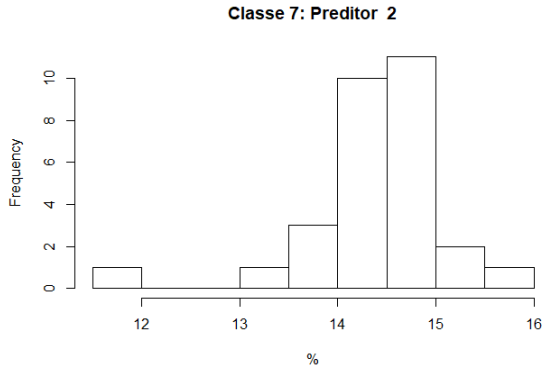

Figura 2. Histograma referente a quantidade de sódio (Na) nas amostras da classe 7

C. Unconditional bi-variate analysis

Foi então realizada uma análise bi-variada dos preditores,para que desta forma, fosse possível avaliar o nível de correlação entre as variáveis e verificar a existência de informação redundante. Esta análise é importante, pois através dela é possível conceber a possibilidade de reduzir a complexidade de nosso modelo, desprezando dados que não trazem informações relevantes para a nossa análise. Na *Figura 6* (em anexo), temos uma matriz de gráficos de dispersão, que detêm variáveis de dois a dois , sendo indicado o elemento respectivo em sua diagonal principal. Inspecionando-a , é possível conjecturar uma relação de linearidade positiva entre os preditores *Ri e Ca*, e uma sutil linearidade negativa entre os preditores *Ri e Si*. Através da matriz de correlação (*Figura 3*), é possível constatar que de fato existe uma linearidade

positiva significante entre *Ri e Ca* (próxima de 1) e uma sutil linearidade negativa entre os preditores *Ri e Si*.

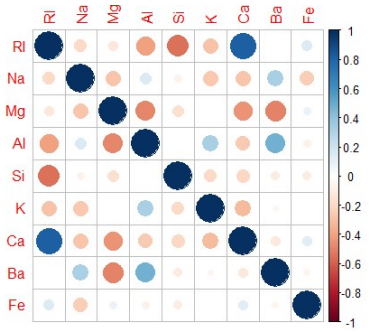

Figura 3. *Matriz de Correlação das Variáveis*

D. Unconditional multi-variate analysis

Nesta última etapa faremos a análise multi-variada dos preditores que implica em realizar uma análise dos componentes principais dos preditores (do inglês Principal Component Analysis), que visa extrair as informações importantes de um conjunto de dados multivariado e para expressar essas informações como um conjunto reduzido de preditores chamados componentes principais.

O passo a passo resumido para o cálculo do PCA é o seguinte:

- Obtenção dos dados.
- Subtrair os dados pela sua média e dividir pelo desvio padrão.
- Calcular a matriz de covariância da matriz de dados do item anterior.
- Calcular os autovetores e autovalores da matriz de covariância.
- Escolher os preditores com maiores autovalores associados para formar os componentes principais.
- Obtenção do novo conjunto de dados a partir dos PCA's.

A primeira etapa após a obtenção dos dados é a normalização, que consiste em subtrair os dados originais pela média em cada preditor e dividir pelo seu desvio padrão. A normalização na análise de um conjunto de dados é importante principalmente para agrupar os dados em uma certa faixa de valores e assim facilitar a visualização por meio de gráficos. Este método também é conhecido como z-score.

O próximo passo consiste em realizar o cálculo da matriz de covariância dos dados normalizados pela média e o desvio padrão. Portanto, convém definirmos o conceito de covariância.

Covariância: A covariância é uma medida estatística que mede o grau de inter-relação entre duas ou mais variáveis, é feita essencialmente em duas ou mais dimensões, dado que a covariância em uma dimensão resulta na variância.

A covariância para três dimensões é dada por:

$$cov(x,y,z) = \begin{vmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{vmatrix}$$

Ao calcularmos a variação conjunta dos preditores, é importante que possamos realizar a redução dos dados em componentes principais, para isso, utilizamos a matriz já calculada no método de Decomposição de valor singular (do inglês, Singular Value Decomposition - SVD), que consistem em decompor a matriz de covariância em autovetores e autovalores. Definição de autovetores e autovalores:

$$(A - \lambda I)u = 0$$

Onde 'A' é uma matriz quadrada n x n; 'I' é a matriz identidade nas dimensões de A; 'Lambda' são o(s) autovalore(s) de A; 'u' são o(s) autovetore(s) de A;

Em posse do vetor de autovalores da matriz de covariância dos dados normalizados, escolhemos os preditores com os dois maiores autovalores para se tornarem componentes principais da distribuição, isso se deve ao fato de que estes preditores detêm uma maior variabilidade dos dados do que os demais preditores, ao fim dessa análise a validação desta escolha é verificada através do gráfico de porcentagem de variância versus número de componentes no PCA.

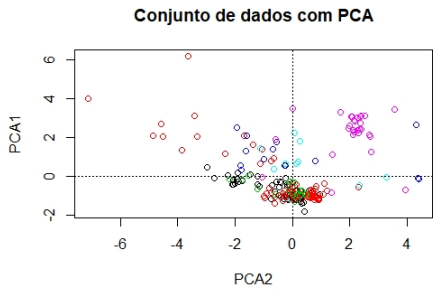

Figura 4. *Novo conjunto de dados com uso dos componentes principais*

Analisando o gráfico de dispersão, observamos que com o uso dos componentes principais podemos obter uma melhor visualização da distribuição dos dados, embora pelo menos 3 classes abaixo do eixo y tenham ficado sobrepostas, dificultando a utilização de aproximações para esta região do gráfico.

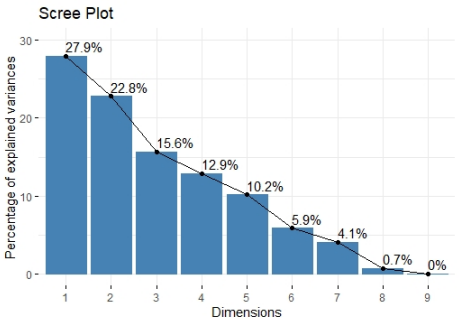

Figura 5. *Gráfico porcentagem de variância ao adicionar componentes na análise*

A análise do gráfico da figura 5 indica que a escolha de dois PCA's para representar os dados, possui apenas 50.7 por cento da variância dos dados. Mas que nesse caso já possibilitou uma melhor visualização.

III. Resultados

O pré-processamento de dados é fundamental para análise de modelos preditivos, pois reduz a redundância com o uso do método SVD (Singular Value Decomposition), principalmente quando há muitos preditores correlacionados entre si, também ameniza o impacto computacional de outliers no conjunto de dados, além de melhorar a visualização do conjunto de dados através da utilização de componentes principais, gerando um ganho computacional por utilizar apenas alguns preditores, todos estes aspectos contribuem bastante para uma melhor compreensão e predição dos dados.

Referências

[1] **Kuhn, M., Johnson, K. (2013).** ***Applied Predictive Modeling*** **New York: Springer.**

[2] **Michela Mulas. (2019).** ***Data pre-processing*****. Slides de Aula.**

ANEXO

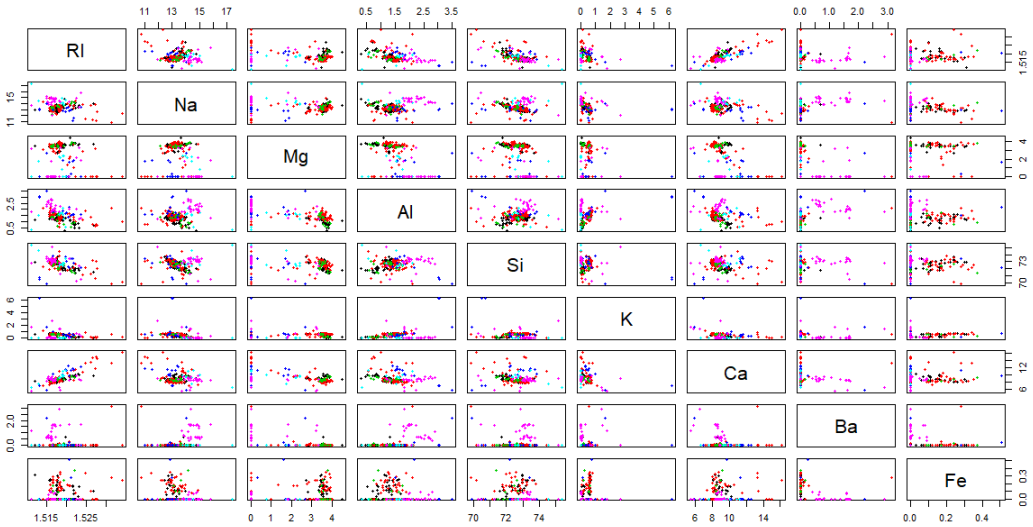

Figura 6. *Matriz de Gráficos de Dispersão*