UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de Presidente Prudente

LUCAS SANCHES GOMES

ESTATÍSTICA FORENSE: UMA ABORDAGEM DE TÉCNICAS ESTATÍSTICAS PARA FRAGMENTOS DE VIDRO

Revisado pelo Orientador

Assinatura do Orientador

Data ____/ ____ / 2024

PRESIDENTE PRUDENTE

2024

LUCAS SANCHES GOMES

ESTATÍSTICA FORENSE: UMA ABORDAGEM DE TÉCNICAS ESTATÍSTICAS PARA FRAGMENTOS DE VIDRO

Projeto para Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Estatística da FCT/Unesp para aproveitamento na disciplina Trabalho de Conclusão de Curso.
Orientador: Prof. Dr. Guilherme Aparecido Santos Aguilar.

PRESIDENTE PRUDENTE
2024

G633e

Gomes, Lucas Sanches

Estatística forense : uma abordagem de técnicas estatísticas para fragmentos de vidro / Lucas Sanches Gomes. -- Presidente Prudente, 2024

44 p. : tabs., fotos

Trabalho de conclusão de curso (Bacharelado - Estatística) - Universidade Estadual Paulista (UNESP), Faculdade de Ciências e Tecnologia, Presidente Prudente

Orientador: Guilherme Aparecido Santos Aguilar

1. Estatística forense. 2. Classificação. 3. Regressão logística multinomial. 4. Clusterização. 5. Análise descritiva. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

TERMO DE APROVAÇÃO

Lucas Sanches Gomes

ESTATÍSTICA FORENSE: UMA ABORDAGEM DE TÉCNICAS ESTATÍSTICAS PARA FRAGMENTOS DE VIDRO

Relatório Final de Trabalho de Conclusão de Curso aprovado como requisito para obtenção de créditos na disciplina Trabalho de Conclusão do curso de graduação em Estatística da Faculdade de Ciências e Tecnologia da Unesp, pela seguinte banca examinadora:

gov.br Documento assinado digitalmente
GUILHERME APARECIDO SANTOS AGUILAR
Data: 18/12/2024 09:08:52-0300
Verifique em https://validar.iti.gov.br

Orientador: _____

Prof. Dr. Guilherme Aparecido Santos Aguilar

Departamento de Estatística

gov.br Documento assinado digitalmente
JOSE GILBERTO SPASIANI RINALDI
Data: 18/12/2024 15:50:20-0300
Verifique em https://validar.iti.gov.br

_____

Prof. Dr. José Gilberto Spasiani Rinaldi

Departamento de Estatística

Presidente Prudente, 10 de dezembro de 2023.

AGRADECIMENTOS

Dedico primeiramente ao meu pai, um homem que me negligenciou e ignorou por mais de quatorze anos, por ter sido fonte principal de motivação, me ensinando indiretamente desde meus nove anos a transformar tristeza e ódio em aprendizado.

Dedico principalmente a minha amada mãe, que executou excepcionalmente as funções de mãe e pai, e sempre me apoiou em todas as decisões da minha vida, além de focar todas suas energias para que eu pudesse ter um futuro melhor, e atravessou as dificuldades me ensinando o que é se superar a cada dia.

A minha namorada e futura noiva, que esteve comigo a mais de oito anos de relacionamento até agora, e me sustentou emocionalmente quando eu mais precisei, me apoiando nas minhas decisões, secando minhas lágrimas, presenciando minhas vitórias, e me empurrando quando não havia mais forças.

RESUMO

O trabalho visa o estudo da aplicação de técnicas estatísticas na área forense, com ênfase na análise de fragmentos de vidro a fim de clusterizar e classificar as observações utilizando as técnicas de k-médias, criação de um modelo de regressão logística multinomial e análise descritiva, isso permitirá descobrir a origem dos fragmentos em diferentes fontes e futuramente poderia ser utilizado para identificar se um possível suspeito de crime no qual foram encontrados os fragmentos é de fato o responsável pelo ato criminoso. Até o momento a análise descritiva se mostra satisfatória, mostrando uma sumarização das variáveis, correlações entre elas e assimetria nos dados, permitindo interpretarmos que existe sim a diferenciação entre as origens e grupos seletos que irão ser formados posteriormente com base nos índices de refração e elementos químicos presentes em cada tipo de fragmento. A intenção futura é que possamos utilizar dos resultados obtidos para verificar se as técnicas são eficientes do ponto de vista da estatística, ciência de dados e área forense, indicando se o desenvolvimento poderia ser aplicado no futuro para situações reais, ou se seria necessário o uso de outras metodologias e abordagens.

Palavras-chave: Estatística forense; Índice de refração; Regressão logística multinomial, K-médias; Análise descritiva; Clusterização; Classificação; Elementos químicos.

ABSTRACT

The work aims to study the application of statistical techniques in the forensic field, with an emphasis on the analysis of glass fragments in order to cluster and classify observations using k-means techniques, creating a multinomial logistic regression model and descriptive analysis. This will allow us to discover the origin of the fragments from different sources and could potentially be used to identify if a possible crime suspect, where the fragments were found, is indeed responsible for the criminal act. So far, the descriptive analysis has been satisfactory, showing a summarization of variables, correlations between them, and data asymmetry, allowing us to interpret that there is indeed differentiation between the origins and selective groups that will be formed later based on the refractive indices and chemical elements present in each type of fragment. The future intention is to use the obtained results to verify whether the techniques are efficient from the perspective of statistics, data science, and the forensic field, indicating whether the development could be applied in the future to real situations, or if other methodologies and approaches would be necessary.

Keywords: Forensic statistics; Refractive index; Multinomial logistic regression; K-means; Descriptive analysis; Clustering; Classification; Chemical elements.

Sumário

1 INTRODUÇÃO

A estatística é uma área das ciências exatas que surgiu a muitos séculos atrás e se associa a palavra em latim “*Status*”, isto é, estado. Cerca de 3.000 A.C. algumas civilizações necessitavam de um controle numérico de recursos e informações sobre a população, tais como os povos egípcios e babilônicos.

Durante o desenvolvimento da área até os tempos modernos, surge a percepção da versatilidade da estatística no que se diz respeito a uma ramificação de aplicações técnicas em diversos campos da sociedade. Desse modo, nasce a estatística forense, um campo interdisciplinar, que une princípios estatísticos com a investigação criminal. Logo a Estatística Forense se resume em “(...) aplicação da estatística à ciência forense” (ROSADO, F.; NEVES, M. F., 2008, p. 2).

Mesmo sendo de grande importância, o campo ainda é muito pouco conhecido e disseminado no Brasil. E ganha sua força principalmente em outros países como Inglaterra, Estados Unidos e Japão, com fortes abordagens científicas para avaliação de evidências. Por isso, é importante que mais estatísticos e cientistas de dados conheçam tal área de atuação, para que futuramente ela possa se tornar referência na sociedade brasileira e consequentemente trazer mais benefícios e justiça para a população.

Com a finalidade de auxiliar casos policiais e jurídicos, esses cientistas forenses, em suma, investigam evidências, comumente chamadas de provas, e utilizam ferramentas estatísticas que auxiliam a recolha, preservação, análise e interpretação das mesmas, que são encontradas nas cenas de crime e suspeitos (CHADREQUE, M. C., 2012). Para isso deve-se compreender qual a origem da evidência e posteriormente apresentar a significância dos resultados encontrados. Os peritos forenses nem sempre seguem um procedimento linear para tal identificação, porém existem alguns fundamentos principais que são utilizados: transferência, identificação, individualização, associação e reconstrução. A área de atuação pode trabalhar tanto com a estatística frequentista quanto a bayesiana.

Diversas são as evidências que podem ser analisadas, sendo as mais comuns: documentos, balística e cartuchos, áudios, DNA, geolocalização, imagens de segurança e fragmentos de vidro.

1.1 Objetivo geral

O objetivo geral deste trabalho é utilizar algumas técnicas estatísticas para a análise de fragmentos de vidro, uma vez que este tipo de evidência é o mais comum de se encontrar em cenas de crimes e em suspeitos. Geralmente, ao ocorrer um crime como assaltos, agressões ou acidentes de carros, algum vidro de origem é quebrado e repartido em diversos pedaços, então esses fragmentos compõe a cena de origem e parte deles são transferidos ao culpado ou indivíduo que esteve próximo ao local. Quando o indivíduo é tomado como suspeito, esses fragmentos podem ser recolhidos de fibras têxteis, cabelo ou sapatos, e assim podemos recolher tais fragmentos para análise e verificar se suas origens coincidem. Para um entendimento mais consistente, Evett introduziu a estatística na análise de fragmentos de vidro (CURRAN,J. M., 2003).

Existem dois tipos de abordagens mais utilizadas na análise de fragmentos de vidros dentro da estatística forense, a primeira se diz respeito quando os fragmentos são grandes o suficiente para realizarmos análises de cor e densidade, porém, o que se tem de mais comum são fragmentos que variam de 0,1-0,5 mm, assim necessitando de uma análise baseada na composição química elementar e/ou índice de refração (RI) desses fragmentos (CURRAN, J. M., et al, 1997). Os métodos mais famosos atualmente são o *Energy Dispersive X-Ray Spectromer* (SEM-EDX), análise química elementar ou Glass *Refractive Index Mesurement* (GRIM) quando se trata do índice de refração. Vale lembrar que o GRIM é comumente mais utilizado por ser um método mais conservativo para com as provas e se quer preservar a evidência para análises futuras (CURRAN, J. M., 2003). O foco de estudo será utilizar abordagens estatísticas e de *data Science* para comparação de resultados, utilizando o RI e elementos químicos dos fragmentos.

Para o uso das técnicas, foi escolhida uma base de dados de fragmentos de vidro que possui colunas (variáveis) com o índice de refração (RI), colunas de composição química (elementos óxidos) e seis tipos de origem.

1.2 Objetivo específico

O objetivo desse trabalho se divide na aplicação de algumas técnicas dentro da base de dados citada anteriormente no contexto da estatística forense, são elas:

- Análise exploratória e descritiva dos dados: A ideia é resumir e descrever os dados, além de encontrar as medidas de posição e dispersão, tais como as medidas de tendência central, percentis, quantis, amplitude, variância, desvio padrão, coeficiente de variação, matriz de correlação, gráficos e tabelas.
- Técnicas de agrupamento e clusterização: K-médias.
- Técnicas de classificação: Regressão multinomial.
- Teste de hipóteses: Verificar se o resultado de um modelo de classificação seria o mesmo para um de agrupamento, e realizar intervalos de confiança para o teste. Comparar os modelos propostos.
- Interpretação dos resultados: Após os resultados será formulada uma interpretação para uma tomada de decisão, validação dos testes e identificação de outliers, tal interpretação será uma mescla de visões da análise de dados, ciência de dados e forense.

1.3 Justificativa

Este trabalho será importante para mostrar como diversos conceitos e técnicas podem auxiliar a área criminal e jurídica do Brasil, propagando a estatística em um campo que não tem o reconhecimento que merece e que poderá ser útil para trazer mais justiça, segurança e assertividade para os brasileiros, isto é, se relacionando com a décima e décima sexta ODS da ONU: reduzir a desigualdade dentro do país e proporcionar o acesso à justiça e paz, respectivamente.

2 REVISÃO BIBLIOGRÁFICA

Diversos materiais foram utilizados para estudo do tema, resumidamente a maioria deles se tratam de técnicas estatísticas e algoritmos de machine learning para programação, clusterização e classificação. O livro Aprendizado de máquina: uma abordagem estatística (IZBICKI, R.; DOS SANTOS, T. M., 2020) nos fornece tais técnicas, como por exemplo k-médias, tanto com a visão teórica quanto aplicada a machine learning.

Uma importante referência é o livro **Forensic Interpretation Of Glass Evidence** (Curran, J. M.; Hicks, T. N.; Buckleton, J. S., 2000), nele podemos encontrar toda a trajetória da análise de fragmentos de vidro como evidência criminal, técnicas comumente utilizadas por peritos, abordagens de evidências, abordagens bayesianas, testes, utilização de softwares entre outros. O livro fornece toda a base teórica para o entendimento da análise forense de fragmentos de vidro e possibilita uma boa visão sobre a comparação destes tipos de dados e como interpretá-los da maneira correta no sentido criminal.

Já os livros **Applied multivariate statistical analysis** (WICHERN, D. W.; JOHNSON, R. A., 2007) e Estatística multivariada aplicada (REIS, E., 2007) serão a base de estudo para o entendimento e aplicação de técnicas multivariadas, tais como a de regressão e o próprio k-médias. Dessa forma será possível compreender toda a parte teórica e realizar juntamente com as demais referências, a aplicação e desenvolvimento de tais modelos utilizando o software R.

Outra grande referência é a dissertação de mestrado Estatística Na Investigação Forense (CHADREQUE, M. C.,2012) que visa abordar a estatística forense de uma forma geral, mostrando a aplicação da estatística de maneira frequencista e bayesiana, citando desde conceitos introdutórios até a identificação de outliers para diversos tipos de evidências, como DNA, fragmentos de vidro, entre outros. Discorrendo também sobre as possibilidades de atuação na área, casos históricos, cálculo de provas e muito mais. É a partir dela, que retiramos o foco do projeto e conceitos iniciais.

Em suma, a ideia é utilizar todas as referências para aplicar as técnicas, realizar os testes estatísticos e algoritmos de machine learning na classificação e clusterização da base de dados encontrada, e assim, avaliar quais obtiveram os melhores resultados.

3 METODOLOGIA

O objeto de estudo será a base de dados sobre fragmentos de vidro obtidos através de *B. German Central Research Establishment Home Office Forensic Science Service Aldermaston*. O banco de dados fornece as variáveis índice de refração (RI),

seis tipos de vidro, e diversas quantidades de composições químicas de óxidos para os mesmos, tais como Na, Fe, K, entre outros. Essa base (chamada de fgl) se encontra também dentro do próprio software R, bastando apenas utilizar o pacote “MASS” e posteriormente o comando “data(fgl)”.

Como já dito, utilizaremos a linguagem e o software estatístico R para auxiliar na aplicação das técnicas, são elas:

- Análise descritiva e exploratória: Se trata da parte inicial de qualquer projeto estatístico, é importante para descrever o banco de dados e as variáveis, utilizarei medidas de posição e dispersão, coeficientes de variação, gráficos e tabelas.

- Sobreamostragem ROSE: é uma técnica de balanceamento dos dados, geralmente utilizada em casos em que as variáveis se encontram com valores muito discrepantes, a mesma proporciona melhores análises, gerando resultados mais significativos. Balanceando igualmente as classes com base na variável de maior comprimento.

- K-médias: É um algoritmo e técnica estatística de clusterização, ele avalia e realiza tais clusteres de acordo com as características dos dados. Para isso iremos definir “k” número de agrupamentos (clusteres), definir aleatoriamente um centróide para cada um deles e para cada ponto, calcular o centróide de menor distância. Depois basta reposicionar o centróide a partir da média calculada e repetir iterativamente os últimos dois passos até encontrar uma posição ideal do centróide.

- KNN: é um algoritmo e técnica de machine learning utilizado em casos de classificação e regressão, se baseando na proximidade de um ponto a outros que temos no conjunto de dados. Geralmente é utilizado em conjunto com uma divisão da base de dados em treino e teste, porém não é obrigatório.

- Regressão multinomial: A regressão é uma técnica que permite verificar se há relação entre a variável resposta e uma ou mais variáveis independentes. Já a multinomial é similar a logística, porém a variável resposta é qualitativa nominal com três ou mais categorias. Escolhemos uma dessas categorias como referência para compará-la as outras. Em suma, serve para classificar se baseando nos valores de um conjunto de variáveis preditoras.

- Teste de hipóteses: Verificar se o resultado de um modelo de classificação seria o mesmo para um de agrupamento, e realizar intervalos de confiança para o teste. Comparar os modelos propostos.

- Interpretação dos resultados: Visando um olhar que mescla a estatística, a ciência forense e a ciência de dados.

A ideia resumidamente é aplicar tais técnicas e compará-las entre si para verificar qual delas seria a mais adequada, simples e eficiente.

4 BALANCEAMENTO ROSE

Quando existe um desequilíbrio de classes em um problema de classificação, podemos utilizar técnicas de balanceamento para melhorar nossas análises e obtermos resultados mais confiáveis e não tendenciosos.

Iremos utilizar então a técnica ROSE (Random OverSampling Examples). Ao invés de apenas duplicar amostras de classes minoritárias, esta técnica de reamostragem gera novos exemplos sintéticos com base na característica original da amostra. Balanceando assim as classes com base na maior amostra, evitando o risco de overfitting.

Primeiro criamos um ponto sintético para as classes minoritárias de acordo com a distribuição de probabilidade original dos dados.

$$X^* = x + e$$

Em qual x é um vetor de características das amostras reais da classe minoritária;

Sendo e um vetor de ruído que é amostrado de uma distribuição de probabilidade.

Essas amostras sintéticas são geradas de acordo com uma distribuição de probabilidade condicional para a classe minoritária, ou seja, $X^* \sim p(x|\, Y = 1)$.

$$X^* \sim N(\mu, \Sigma)$$

Sendo μ o vetor de médias das características da classe minoritária, Σ a matriz de covariância das características da classe minoritária, e $N(\mu, \Sigma)$ uma distribuição normal multivariada.

5 K – MÉDIAS

K – Médias é uma técnica de agrupamento muito utilizada na estatística multivariada (aprendizagem não supervisionada) e na ciência de dados devido a sua simplicidade e poder de lidar com grandes conjuntos de dados.

A mesma tem o intuito de separar de forma significativa os dados em diferentes grupos, onde os indivíduos de cada grupo são semelhantes entre si.

Futuramente esses grupos podem ser utilizados em um modelo de classificação e/ou regressão (no nosso caso, um modelo de regressão multinomial).

A técnica divide o conjunto de dados em k grupos através da soma dos quadrados dentro do grupo (SS_k), isto é, a minimização da soma das distâncias quadráticas de cada observação em relação à média do grupo que foi atribuído.

É importante lembrar que muitas vezes devemos normalizar os dados de variáveis contínuas através do z-score, isto é:

$$Z = \frac{X - \mu}{\sigma}$$

Sendo X as nossas variáveis, μ é nossa média e σ é o desvio-padrão.

Fazendo isso contornamos a possibilidade de variáveis com grande escala dominarem o processo de agrupamento.

Para exemplificar, consideremos dados com tamanho n e duas variáveis x e y. Queremos dividir o nosso conjunto de dados em k grupos, atribuindo cada observação (x_i,y_i) em um grupo k. Dada uma atribuição n_k observações a um grupo k, o centróide ($\overline{x}_k$, $\overline{y}_k$) do grupo será a média dos pontos no grupo, ou seja:

$$\overline{x}_k = \frac{1}{n_k}\sum x_i$$

$$\overline{y}_k = \frac{1}{n_k}\sum y_i$$

No nosso caso, trataremos o termo média do grupo como um vetor das médias das variáveis, e não a um único número.

Sobre a soma dos quadrados dentro de um grupo k (que representa as distâncias quadráticas de cada observação em relação ao centróide do cluster), ou seja, é uma medida de variação total dos dados dentro do grupo, e podemos obtê-la da seguinte maneira:

$$SS_k = \sum (x_i - \overline{x}_k)^2 - (y_i - \overline{y}_k)^2$$

Após isso, a técnica atribui as observações que minimizam a soma dos quadrados dentro do agrupamento, para todos os k grupos ($SS_1 + SS_2 + \cdots + SS_k$):

$$\sum_{i=1}^{k} SS_i$$

O algoritmo é sensível a porção inicial das sementes (centróides), fazendo confusão de clusteres. Por isso devemos rodar o algoritmo diversas vezes para obtermos resultados significativos, associando assim aos clusteres mais frequentes. O processo iterativo garantirá que os centróides não mudem.

Para saber o número de clusteres (k) ideais para o problema, podemos utilizar a distância média (SS_k) para diferentes valores k e plotar o gráfico, assim utilizando o método gráfico do cotovelo, que consiste em observar o ponto k que tivemos uma mudança brusca (esse será o valor k ideal).

Sobre as limitações desta técnica, podemos citar:

É suscetível a problemas quando os cluesteres são de diferentes densidades (uns mais esparsos e outros mais densos).

O mesmo ocorre quando são de tamanhos muito diferentes.

E por fim, quando os clusteres são globulares (de diferentes formatos, ocorrendo devido a topologia do espaço que os dados se encontram).

Para burlar o último problema, podemos utilizar diferentes métricas de distância, como Pearson, Cosseno ou Minkowsk. No nosso caso, usaremos a convencional, a Euclidiana:

$D(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$, onde $[0, \infty)$

Já sobre as vantagens da técnica, podemos citar:

É de fácil implementação e interpretação.
Pode ser usada em grandes conjuntos de dados.
É de fácil adaptação a novos problemas e exemplos.
Já existem algoritmos e versões aprimoradas com intuito de resolver grande parte das limitações.

6 KNN

O algoritmo de K-vizinhos próximos, é uma técnica de machine learning amplamente utilizada na ciência de dados, principalmente em questões envolvendo classificação e regressão. O mesmo se destaca por sua fácil implementação e praticidade, se baseando na proposta de que elementos semelhantes tendem estar próximos uns dos outros no espaço trabalhado, e será utilizado neste projeto caso haja um ruim desempenho do modelo de K-médias.

A utilização do método visa abranger o lado de machine learning e da ciência de dados, por isso iremos dividir nosso conjunto de dados em treino e teste (na proporção de 70% - 30%, respectivamente e aleatoriamente). Os dados de treino serão importantes para treinar o nosso modelo no ambiente desejado, e depois disso o modelo é testado nos dados “novos” do conjunto de teste, que visa simular como o modelo criado irá reagir a dados novos futuramente.

Ele irá classificar uma nova instância com base de uma classe predominante entre as k vizinhas mais próximas nos nossos dados de treinamento. Primeiramente é realizado o cálculo da distância para cada nova amostra:

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

No qual p e q são pontos no espaço;

Sendo $p_i \; e \; q_i$ valores das n dimensões para os pontos p e q.

Após calcular as distâncias, as ordenamos de forma crescente e é selecionado k valores menores.

Como iremos classificar as variáveis, a classe da nova amostra irá ser definida pela maioria entre os vizinhos selecionados:

$$C(x) = argmax_c \sum_{i=1}^{k} \delta(c_i, c)$$

O qual c_i é a classe do i-ésimo vizinho e c é cada classe possível;

$\delta(c_i, c)$ é função indicadora, que irá retornar 1 caso $c_i = c$, ou 0 caso contrário.

7 REGRESSÃO LOGÍSTICA MULTINOMIAL

A regressão logística é uma técnica estatística conhecida e muito utilizada para construir um modelo a partir de um conjunto de observações, com intuito de realizar predição ou classificação de valores tendo como base variáveis categóricas em função de uma ou mais variáveis independentes contínuas ou binárias.

São três tipos principais, sendo elas binomial, ordinal e multinomial (que é o foco desta pesquisa).

A multinomial é um modelo que visa classificar objetos em três ou mais categorias sem ordem específica, no nosso caso, será utilizada em função dos elementos

químicos presente em cada observação de fragmento de vidro, classificando o fragmento em grupos semelhantes.

Em qualquer tipo de regressão, a variável resposta Y dado o valor da variável independente X possui o valor médio chamado de valor médio condicional expresso pela equação linear:

$$E[Y\,|X = x] = \beta_0 + \beta_1 X$$

Para $-\infty < E[Y|X = x] < +\infty$.

Desta maneira, a regressão logística é representada pela quantidade:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

Logo, o modelo logístico se fundamenta na função logística que descreve a forma matemática. E para linearizar o modelo aplicamos o logaritmo, tal que:

$$g(x) = \left(\frac{\pi(x)}{1-\pi(x)}\right) = ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = ln\left(\frac{\frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}}{1 - \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}}\right) = ln\left(e^{\beta_0+\beta_1 X}\right)$$

$$= \beta_0 + \beta_1 X$$

Temos então a transformação logit para a quantidade $\pi(x)$.

Para o caso múltiplo do estudo, possuindo então nossas variáveis independentes de quantidade de elementos químicos presentes em cada observação de fragmento de vidro, podemos denotá-las por um vetor dado por $x^t = \left(x_1, x_2, \dots, x_p\right)$.

Logo, a quantidade será expressa por:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p}}$$

Respectivamente, terá a função logit:

$$g(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

O modelo múltiplo possui um ajustamento bastante semelhante ao simples, por isso vamos abordá-lo de forma direta. Tal ajustamento se faz calculando o EMV (estimador de máxima verossimilhança), que parte do uso de um vetor de parâmetros $\beta^t = (\beta_0, \beta_1, \beta_{2,} \ldots, \beta_p)$. Supondo que as observações dos nossos dados possuem independência entre si, podemos calcular o estimador da seguinte maneira:

$$L(\beta) = ln[L(\beta)] = \sum_{i=1}^{n} \{y\ ln[\pi(x)](1-y)\ ln[1-\pi(x)]\} = \cdots$$

$$= \sum_{i=1}^{n} \left[y\ \beta_0 + y\ \beta_1 X_1 + y\ \beta_2 X_2 + \cdots + y\ \beta_p X_p - \ln\left(1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p}\right)\right]$$

Após ajustar o modelo é obrigatória checarmos sua significância, e para isso utilizaremos o teste da razão de verossimilhança. O teste indica que o modelo obteve êxito se os coeficientes de regressão associados a β forem nulos (com exceção de β_0) quando testados simultaneamente.

O teste é calculado da seguinte maneira:

$$D = -2ln\left(\frac{Função\ de\ máxima\ verossimilhança\ do\ modelo\ corrente}{Função\ de\ máxima\ verossimilhança\ do\ modelo\ saturado}\right)$$

Sendo que, o modelo corrente possui a variáveis desejadas para o estudo, e o saturado contém todas as variáveis e interações possíveis.

O teste então pode ser apresentado por:

$$D = -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\hat{\pi}_l}{y_i}\right) + (1-y_i)\ ln\left(\frac{1-\hat{\pi}_l}{1-y_i}\right)\right]$$

O teste possui respectivamente, as hipóteses e estatística do teste:

$$\begin{cases} H_0: \beta_1 = \cdots = \beta_p = 0 \\ H_1: Existe\ algum\ \beta \neq 0 \end{cases}$$

$$G = -2\ ln\left(\frac{modelo\ corrente}{modelo\ saturado}\right) \sim sob\ H_0\ X^2(p-1)$$

Se rejeitarmos a hipótese nula H_0, temos que existe ao menos um dos coeficientes diferente de zero.

8 RESULTADOS

8.1 Apresentação dos Dados

Os dados presentes nesse estudo foram obtidos do próprio software estatístico R, proveniente da biblioteca MASS, denominado por fgl, as variáveis analisadas e presentes em cada observação são os índices de refração e elementos químicos, são eles:

- RI (Índice de refração)
- Na (Sódio)
- Mg (Magnésio)
- Al (Alumínio)
- Si (Silício)
- K (Potássio)
- Ca (Cálcio)
- Ba (Bário)
- Fe (Ferro)

8.2 Análise Descritiva

Ao realizar a sumarização das variáveis, obtemos os seguintes resultados:

Tabela 1 – Sumarização das variáveis

	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe
Mín	-6.8500	10.73	0.000	0.290	69.81	0.0000	5.430	0.000	0.0000
1º Qu.	-1.4775	12.91	2.115	1.190	72.28	0.1225	8.240	0.000	0.0000
Mediana	-0.3200	13.30	3.480	1.360	72.79	0.5550	8.600	0.000	0.0000
Média	0.3654	13.41	2.685	1.445	72.65	0.4971	8.957	0.175	0.0570
3º Qu.	1.1575	13.82	3.600	1.630	73.09	0.6100	9.172	0.000	0.1000
Máx	15.9300	17.38	4.490	3.500	75.41	6.2100	16.190	3.150	0.5100

- RI (Índice de refração): média de 0.3654, a maioria das amostras possui refratividade entre -1.4775 e 1.1575 e um intervalo de -6.85 a 15.93, indicando grande variabilidade, o que faz sentido, uma vez que o tamanho do fragmento também varia muito.
- Na (Sódio): Média de 13.41, a maioria tem concentração entre 12.91 e 13.82 com intervalo de 10.73 a 17.38.
- Mg (Magnésio): Média de 2.685, a maioria tem concentração entre 2.115 e 3.6 com intervalo de 0 a 4.49.
- Al (Alumínio): Média de 1.445, a maioria tem concentração entre 1.19 e 1.63 e um intervalo de 0.29 a 3.5.
- Si (Silício): Média de 72.65, a maioria tem concentração entre 72.28 e 73.09 e intervalo de 69.81 a 75.41.
- K (Potássio): Média de 0.4971, a maioria tem concentração entre 0.1225 e 0.61 com intervalo de 0 a 6.21.
- Ca (Cálcio): Média de 8.957, a maioria tem concentração entre 8.24 e 9.172 com intervalo de 5.43 a 16.19.
- Ba (Bário): Média de 0.175, a maioria tem concentração igual a 0 e intervalo de 0 a 3.15.
- Fe (Ferro): Média de 0.05701, a maioria tem concentração igual a 0 e um intervalo de 0 a 0.51.

A sumarização serve para termos uma ideia inicial e geral de cada variável, e auxilia ao surgimento de insights.

A próxima ideia foi partir para as medidas de tendência central, foram elas:

Tabela 2 – Medidas de tendência central das variáveis

	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe
Média	0.3654	13.4078	2.6845	1.4449	72.6509	0.4970	8.9569	0.1750	0.0570
Desvio-padrão	3.0368	0.8166	1.4424	0.4992	0.7745	0.6521	1.4231	0.4972	0.0974
Variância	9.2225	0.6668	2.0805	0.2492	0.5999	0.4253	2.0253	0.2472	0.0094

Coeficiente de varia-ção	831.0598	6.0904	53.7303	34.5537	1.0661	131.2109	15.8887	284.0494	170.9170

Tais resultados permitem ter uma noção e entende melhor sobre a distribuição de cada variável que temos, nos mostrando por exemplo como cada uma delas varia em torno da média, e verificamos se ela possui uma dispersão fraca, moderada, forte, entre outros.

Já citando este tópico, obtivemos os gráficos de dispersão:

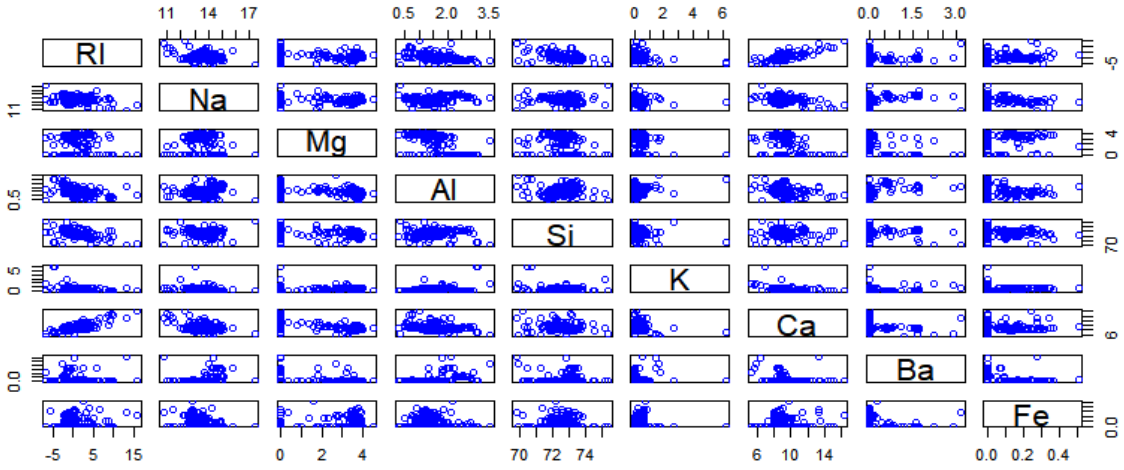

Figura 1 – Gráfico de dispersão

Visualmente podemos notar uma assimetria das distribuições, e a variável Cálcio e RI, aparentam ter uma linearidade positiva.

Verificando as correlações, começamos calculando a matriz de correlação entre variáveis e depois plotando o gráfico para melhor visualização. Correlação essa que varia de -1 a 1.

Tabela 3 – Correlações entre variáveis

	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe
RI	1.0000	-0.1918	-0.1222	-0.4073	-0.5420	-0.2898	0.8104	-0.0003	0.1430
Na	-0.1918	1.0000	-0.2737	0.1567	-0.0698	-0.2660	-0.2754	0.3266	-0.2413
Mg	-0.1222	-0.2737	1.0000	-0.4817	-0.1659	0.0053	-0.4437	-0.4922	0.0830
Al	-0.4073	0.1567	-0.4817	1.0000	-0.0055	0.3259	-0.2595	0.4794	-0.0744
Si	-0.5420	-0.0698	-0.1659	-0.0055	1.0000	-0.1933	-0.2087	-0.1021	-0.0942
K	-0.2898	-0.2660	0.0053	0.3259	-0.1933	1.0000	-0.3178	-0.0426	-0.0077

Ca	0.8104	-0.2754	-0.4437	-0.2595	-0.2087	-0.3178	1.0000	-0.1128	0.1249
Ba	-0.0003	0.3266	-0.4922	0.4794	-0.1021	-0.0426	-0.1128	1.0000	-0.0586
Fe	0.1430	-0.2413	0.0830	-0.0744	-0.0942	-0.0077	0.1249	-0.0586	1.0000

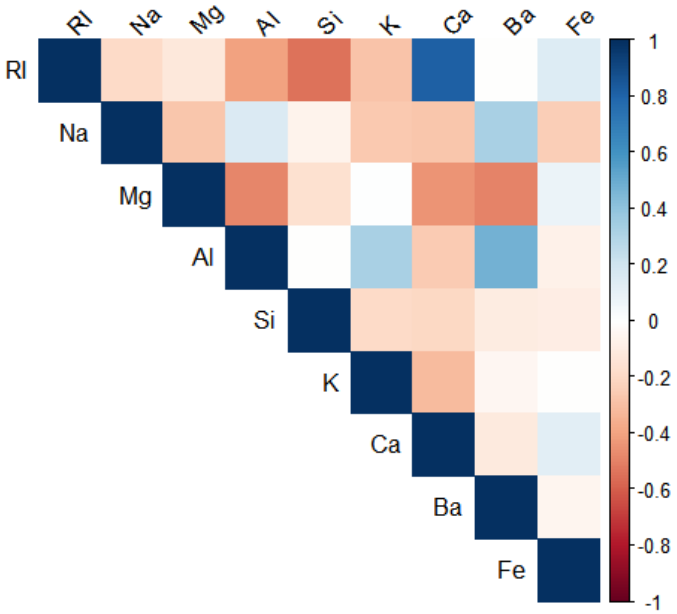

Figura 2 – Gráfico de correlação

Analisando as correlações positivas mais fortes, temos:

- Ri e Ca: Com correlação de 0.8104, muito forte.

  Não obtemos correlações negativamente fortes.

  E outras correlações relevantes (moderadas) são:

- RI e Al: Correlação negativa de -0.4073260.
- RI e Si: Correlação negativa de -0.5420522.
- RI e K: Correlação negativa de -0.2898327.
- Na e Ba: Correlação positiva de 0.3266029.

Como podemos observar, a maioria das correlações importantes, se dizem respeito ao índice de refração dos fragmentos (RI), e vale lembrar que correlação não implica em causalidade, isto é, não é porque o uma variável aumenta, que a outra também aumentará, nem mesmo para o caso de correlação negativa.

Ao realizar uma tabela com os tipos de vidro existentes, obtivemos:

Tabela 4 – Contagem dos tipos de vidro

WinF	WinNf	Veh	Con	Tabl	Head
70	76	17	13	9	29

Sendo eles:

- WinF: Fragmentos de vidros flutuantes de janelas.
- WinNF: Fragmentos de vidros não flutuantes de janelas.
- Veh: Fragmentos vidros de janelas de veículos.
- Con: Fragmentos vidros de recipientes.
- Tabl: Fragmentos vidros de louças.
- Head: Fragmentos vidros de faróis de veículos.

Logo em seguida realizamos os cálculos dos percentis, que nos auxiliam sobre como está a distribuição dos dados de cada variável, e dividem os dados de maneira igual (cada uma representando 25%).

Tabela 5 – Percentis das variáveis

	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe
25%	-1.4775	12.9075	2.115	1.19	72.2800	0.1225	8.2400	0	0
50%	-0.3200	13.3000	3.480	1.36	72.7900	0.5550	8.6000	0	0
75%	1.1575	13.8250	3.600	1.63	73.0875	0.6100	9.1725	0	0.1

Após isso, calculamos a amplitude através da diferença entre o maior valor com o menor de cada variável. Lembrando que a amplitude pode estar fortemente afetada caso haja outliers, e amplitudes grandes indicam grande variabilidade dos dados.

Iremos interpretá-las juntamente com os percentis acima.

- RI (Índice de refração): Amplitude de 22.78, com base no seu percentil, eles variam significativamente.
- Na (Sódio): Amplitude de 6.65, com base no seu percentil, tem variação não tão significativa e mais tranquila em relação a outras variáveis.
- Mg (Magnésio): Amplitude de 4.49, com base no seu percentil, possui uma variação moderada.
- Al (Alumínio): Amplitude de 3.21, com base no seu percentil, temos uma variação moderada.
- Si (Silício): Amplitude de 5.60, com base no seu percentil, obtivemos variação significativa.
- K (Potássio): Amplitude de 6.21, com base no seu percentil, também obtivemos uma variação relativamente grande.
- Ca (Cálcio): Amplitude de 10.76, com base no seu percentil, mais uma variável com variabilidade significativa.
- Ba (Bário): Amplitude de 3.15, com base no seu percentil, não existe variação.
- Fe (Ferro): Amplitude de 0.51, com base no seu percentil, obtemos uma variação extremamente pequena.

Por fim, para uma análise visual, realizamos o histograma de cada variável.

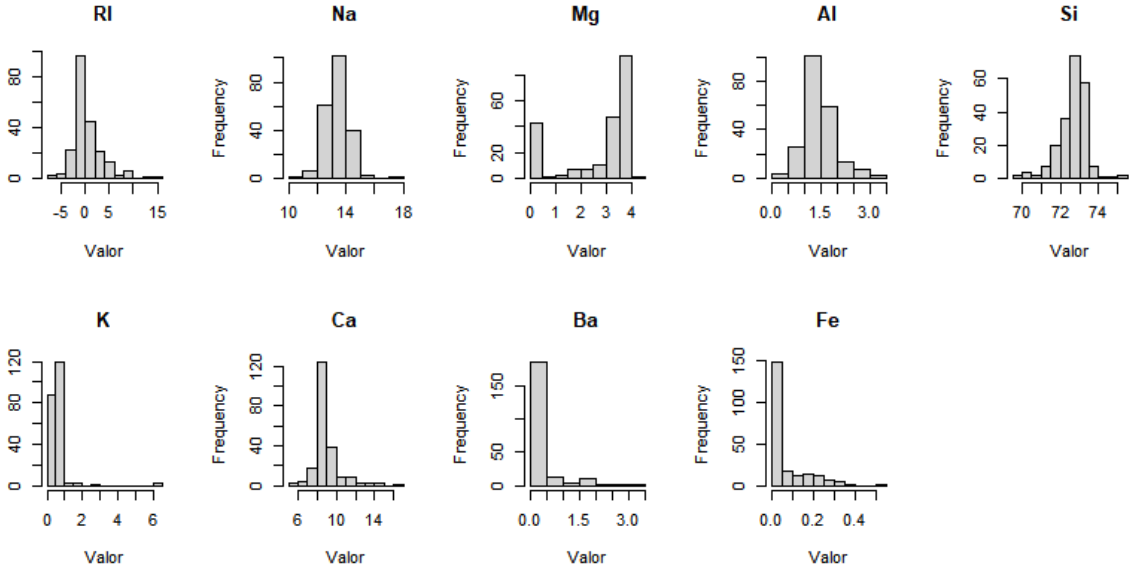

Figura 3 – Histograma de cada variável

Os que mais se aproximam de uma distribuição normal (visualmente), são o Na e Al. Enquanto RI, Mg e Si possuem uma assimetria pequena.

Todavia, K, Ca, Ba e Fe possuem uma assimetria muito grande.

8.3 Balanceamento dos dados

Após a aplicação do balanceamento utilizando a técnica ROSE, obtivemos o seguinte resultado:

Tabela 6 - Dados balanceados

WinF	WinNF	Veh	Con	Tabl	Head
76	76	76	76	76	76

8.4 K-médias

Para sabermos o número ideal k de clusteres, iremos avaliar o gráfico da silhueta, pois o gráfico do cotovelo não indicou um “cotovelo”.

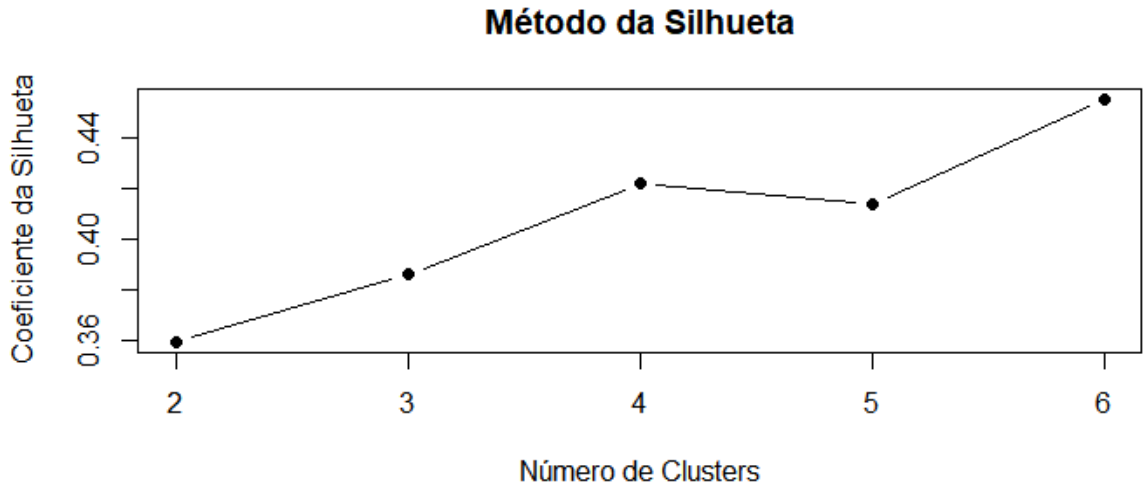

Figura 4 - Gráfico de silhueta

Podemos notar então, que o número ideal de clusteres é k = 6. Como parte final desta análise, precisamos verificar o quão bem ele está agrupando os dados, e podemos verificar através do valor da Inércia e da plotagem do gráfico.

Tabela 7 - Eficiência do método

Método	Inércia
K-médias	1668.128

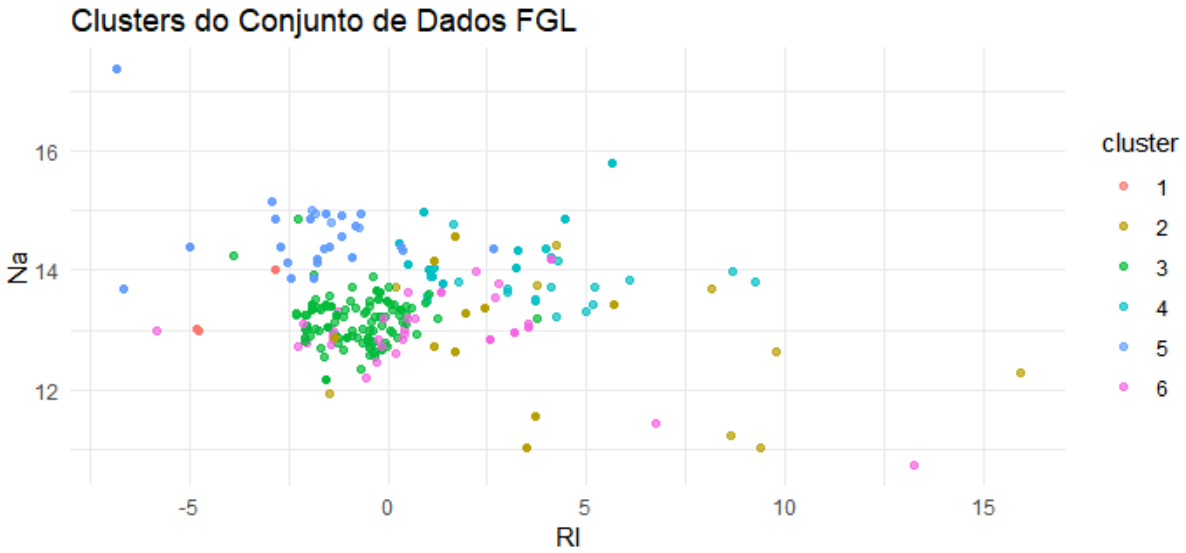

Figura 5 - Gráfico de agrupamento K-médias

É possível verificar então, que a inércia está extremamente alta e o algoritmo não agrupa os dados de forma eficiente (para qualquer variável), mesmo com os dados balanceados. Indicando que o método não é ideal para o nosso problema em questão, dessa forma, partiremos para o uso do KNN como alternativa ao K-médias.

8.5 KNN

Após separado os dados em treino e teste, aleatoriamente escolhemos um valor para k (no nosso caso o valor escolhido foi 5), e treinamos nosso modelo. Retornando os seguintes resultados no conjunto de teste.

Tabela 8 - Matriz confusão (Observado X Predito)

	WinF	WinNF	Veh	Con	Tabl	Head
WinF	15	2	2	0	0	1
WinNF	4	15	0	0	0	1
Veh	2	3	20	0	0	2
Con	0	2	0	20	0	0
Tabl	1	0	0	2	22	2
Head	0	0	0	0	0	16

Tabela 9 - Acurácia do modelo

Modelo	Acurácia
KNN	81.82%

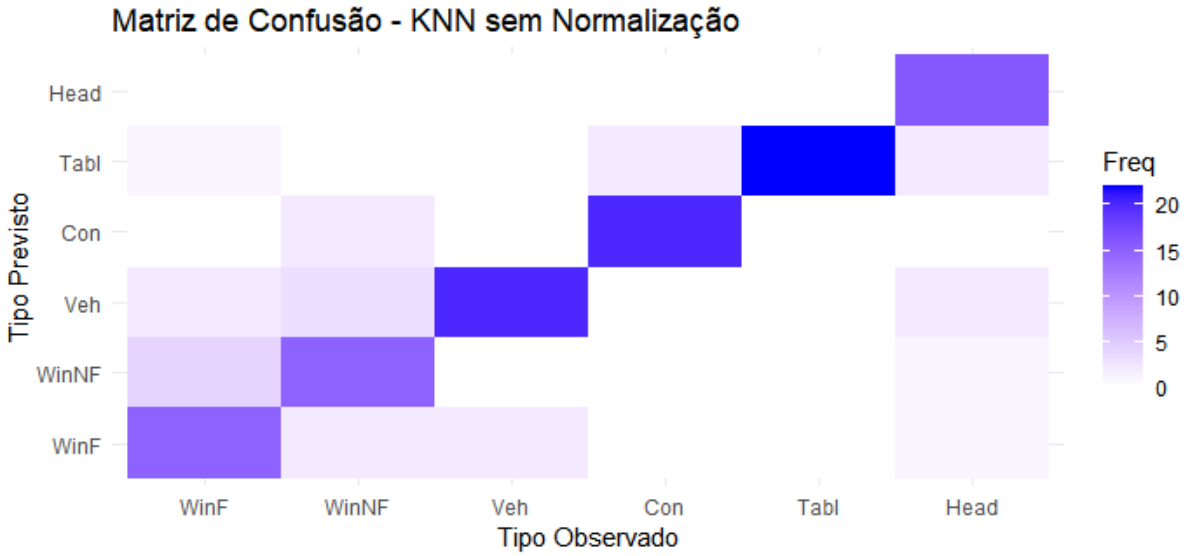

Figura 6 - Gráfico da matriz de confusão

Os resultados obtidos se mostram satisfatórios e muito melhores do que o uso do K-médias, o valor da acurácia nos mostra que o modelo acerta muito bem ao predizer o tipo de vidro com base nas características químicas, enquanto a matriz e gráfico de confusão nos detalha quais são as classes em que ele acerta e erra mais ao tentar realizar esse agrupamento.

8.6 Regressão multinomial

Foi escolhido utilizar o modelo completo, com todas as variáveis, uma vez que é de tamanha importância para o contexto analisado, e como referência, um modelo nulo que leva em conta apenas o intercepto, assim podendo assumir que a probabilidade de cada tipo de vidro não depende de variáveis preditoras, ou seja, é constante. São eles:

$$Completo: Tipo\ de\ vidro \sim RI + Na + Mg + Al + Si + K + Ca + Ba + Fe$$

$$Nulo: Tipo\ de\ vidro \sim 1$$

No qual tipo de vidro é a variável dependente;

RI(Índice de refração), Na, Mg, Al, Si, K, Ca, Ba e Fe são as variáveis independentes.

A fim de comparar as verossimilhanças, realizamos a seguinte ANOVA.

Tabela 10 - ANOVA

Modelo	Graus de liberdade	Desvio residual	Teste	Df	Estatística LR	P-valor
Nulo	2275	1634.085				
Completo	2230	381.313	1 vs 2	45	1252.772	p<0.001

Podemos notar que o desvio do modelo completo é muito menor, indicando um melhor ajuste aos dados, e um p-valor muito baixo mostrando que há evidências significativas estatisticamente sobre o modelo completo ser superior ao nulo, isto é, as variáveis preditoras são importantes e contribuem para explicar o tipo de vidro.

Tabela 11 - Pseudo- R2

Métrica	Valor
Pseudo-R^2	0.9626

A métrica do Pseudo-R^2 nos mostra que o modelo está bem ajustado aos dados e tem uma excelente capacidade explicativa.

Tabela 12 - Análise geral

Variável	X^2	Graus de liberdade	P-valor	Significante
RI	121.842	5	< 2.2e-16	Sim
Na	-0.868	5	1.0000000	Não
Mg	43.008	5	3.682e-08	Sim
AL	27.513	5	4.531e-05	Sim
Si	25.930	5	9.208e-05	Sim
K	52.244	5	4.808e-10	Sim
Ca	30.866	5	9.954e-06	Sim
Ba	39.280	5	2.086e-07	Sim
Fe	21.764	5	0.0005806	Sim

Através dessa análise podemos notar a significância de cada variável para nosso modelo completo e concluímos que a única não significativa é o sódio presente no fragmento de vidro.

Tabela 13 - Parâmetros estimados e teste de Wald

Classe	Intercepto	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe
WinF	56.49	0.28	-0.06	-2.84	3.52	-0.52	-0.82	-1.48	-2.27	3.00
Veh	146.23	-2.47	1.02	3.13	1.27	-3.05	-2.80	5.69	0.37	2.58
Con	13.90	0.64	-3.63	-12.06	21.40	0.95	2.59	-4.60	-8.35	-20.57
Tabl	-15.64	4.87	12.32	-18.11	16.48	2.09	-93.94	-29.00	-72.95	-445.58
Head	-21.63	9.64	1.55	-34.06	15.08	5.57	-12.13	-38.32	-22.42	-229.17

Os valores nos indicam quanto cada variável influencia positivamente ou negativamente em cada tipo de vidro, não devemos dar tanta atenção ao Na, uma vez que ele não é estatisticamente significante.

Tabela 14 - Razões de chances

Classe	Intercepto	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe
WinF	3.43e+24	1.32	0.94	0.06	33.74	0.60	0.44	0.23	0.10	20.05
Veh	3.20e+63	0.08	2.78	22.88	3.55	0.05	0.06	296.62	1.44	13.15
Con	1.08e+06	1.90	0.03	0.00	1.96e+09	2.60	13.36	0.01	0.00	0.00
Tabl	1.61e-07	129.71	2.24e+05	0.00	1.44e+07	8.09	0.00	0.00	0.00	0.00
Head	4.03e-10	1.53e+04	4.70	0.00	3.53e+06	259.51	0.00	0.00	0.00	0.00

O Odds Ratios ou razão de chances, nos mostra o impacto de cada variável no quesito de probabilidade para o modelo, por exemplo RI e Al tem alto impacto no tipo WinF de vidro.

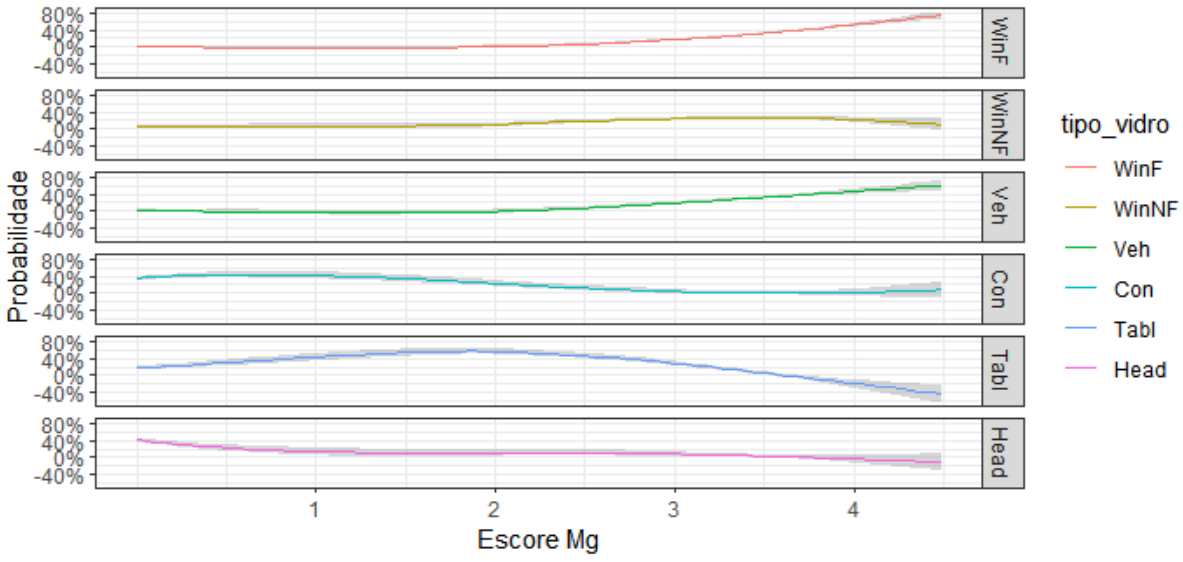

Figura 7 - Gráfico de probabilidades por escore

Este tipo de gráfico nos permite analisar como a probabilidade de uma variável muda de acordo com seu escore, nesse caso, a variável Mg.

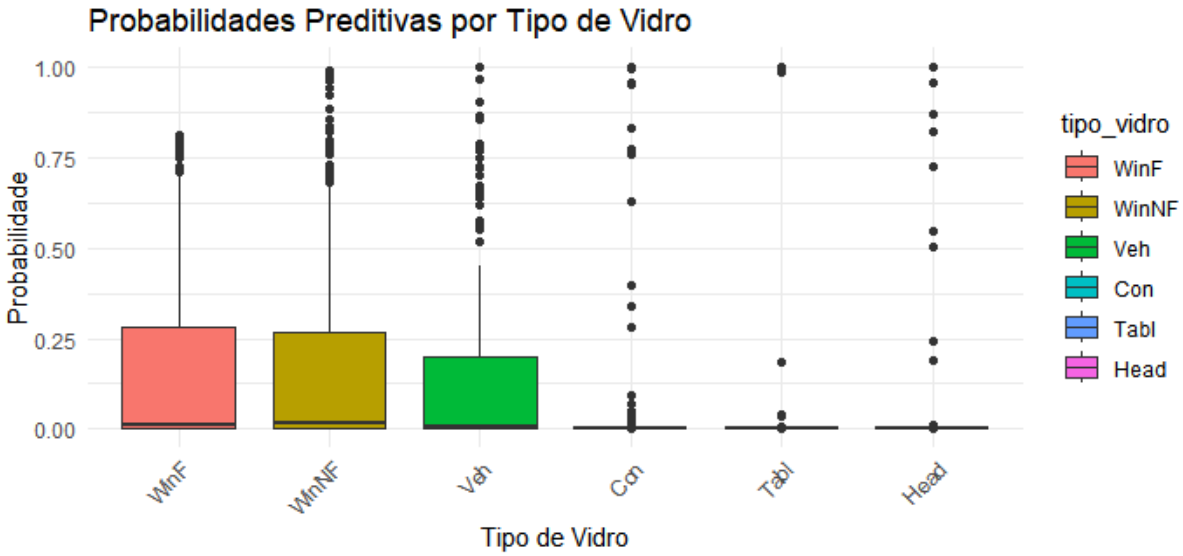

Figura 8 - Probabilidades preditas por tipo de vidro

Com o boxplot, podemos avaliar a dispersão das probabilidades em relação ao tipo de vidro, verificando diversos outliers, o que faz sentido no problema em questão.

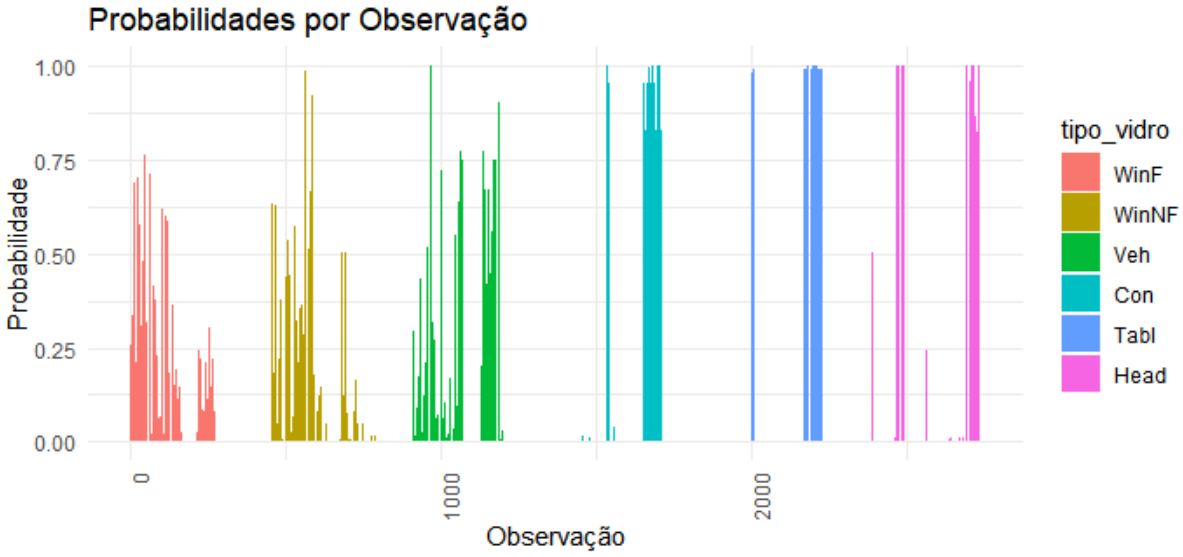

Figura 9 - Gráfico de probabilidades por observação

Aqui podemos avaliar a probabilidade predita para cada observação de forma individual, separado por tipo de vidro. Neste caso, podemos notar probabilidades altas.

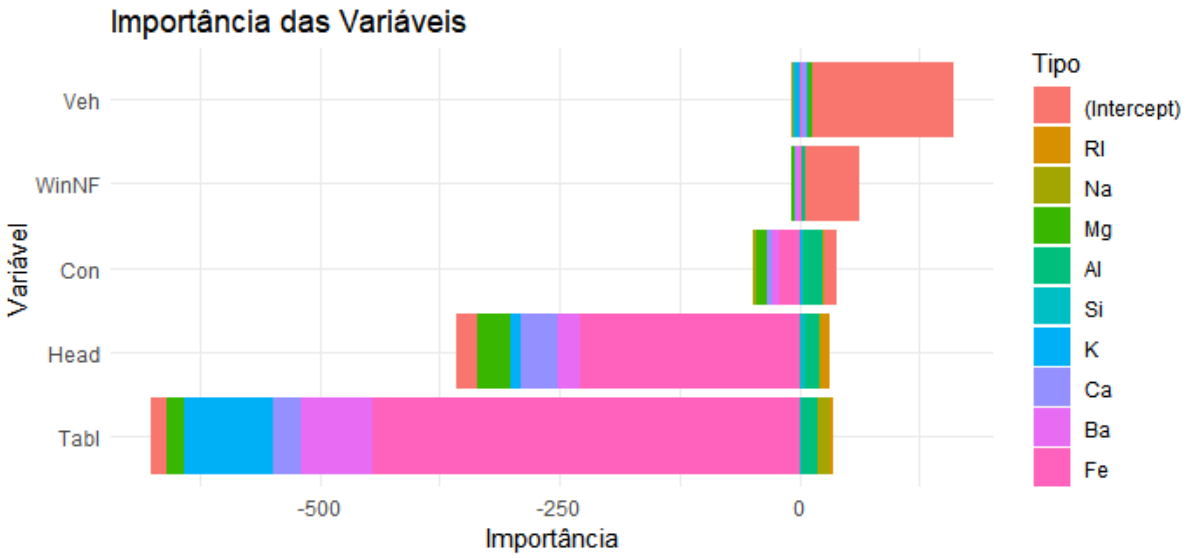

Figura 10 - Gráfico de importância por variável

Este gráfico é um dos mais importantes, representa a importância das variáveis para a classificação de cada tipo de vidro, seja ela negativa ou positiva, permitindo que avaliemos qual elemento químico ou índice de refração mais importa na composição do fragmento.

Para entendermos como o modelo classifica os fragmentos em cada tipo de gráfico analisaremos a matriz de confusão e em seguida verificamos a acurácia do modelo, verificando o quão bem ele está realizando tal classificação dessa matriz.

Tabela 15 - Matriz de confusão (predito X Observado)

	WinF	WinNF	Veh	Con	Tabl	Head
WinF	52	15	9	0	0	0
WinNF	15	48	9	2	0	2
Veh	5	4	67	0	0	0
Com	0	0	0	76	0	0
Tabl	0	0	0	0	76	0
Head	0	0	0	0	0	76

Aqui podemos ver claramente o modelo classificando cada tipo de vidro para uma classe específica, por exemplo, para a classe WinF, nosso modelo classifica

corretamente 52 observações, e erroneamente 24 observações, 15 em WinNF e 9 em Veh.

Tabela 16 - Acurácia

Modelo	Acurácia
Completo	0.8662

Podemos notar então, que o modelo tem uma acurácia de 86.62%, ou seja, é a quantidade que ele prevê corretamente cada fragmento na classe correta, uma boa acurácia.

Ainda com intuito de avaliar o desempenho do modelo, podemos realizar o gráfico de curvas ROC multiclasse, que irá medir a capacidade do mesmo em distinguir as classes.

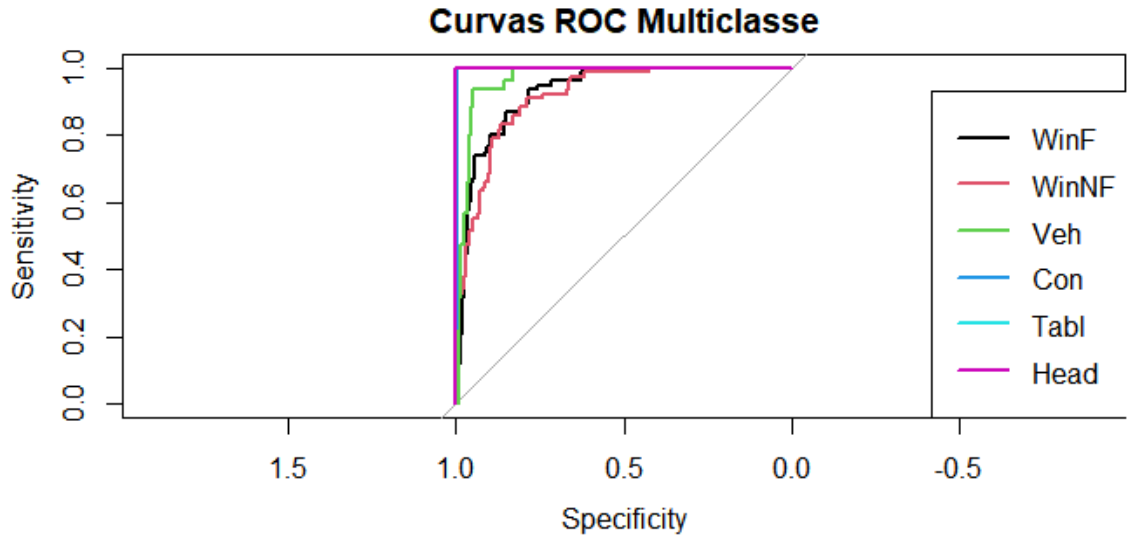

Figura 11 - Curvas ROC

A linha diagonal nos mostra como referência um modelo aleatório sem poder de distinção, enquanto o eixo x representa a taxa de falsos positivos (especificidade) e o eixo y a taxa de verdadeiros positivos (sensibilidade), logo quanto mais a curva tende a esquerda superior, melhor. Podemos notar então que as curvas ROC de multiclasse indicam que o modelo distingue de forma muito boa as classes.

Tabela 17 - Validação final do ajuste

Métrica	Valor
Deviância do modelo completo	381.313
Deviância do modelo nulo	1634.085
Diferença das deviâncias	1252.772
P-valor	0

Quanto menor a deviância, melhor é o ajuste do modelo, realizando a diferença podemos estimar o ganho no ajuste e em conjunto com o p-valor de 0, calculado através do teste de razão de verossimilhanças, podemos concluir que o modelo completo é extremamente superior e efetivo.

Partindo agora para a análise gráfica dos resíduos de Pearson e deviance.

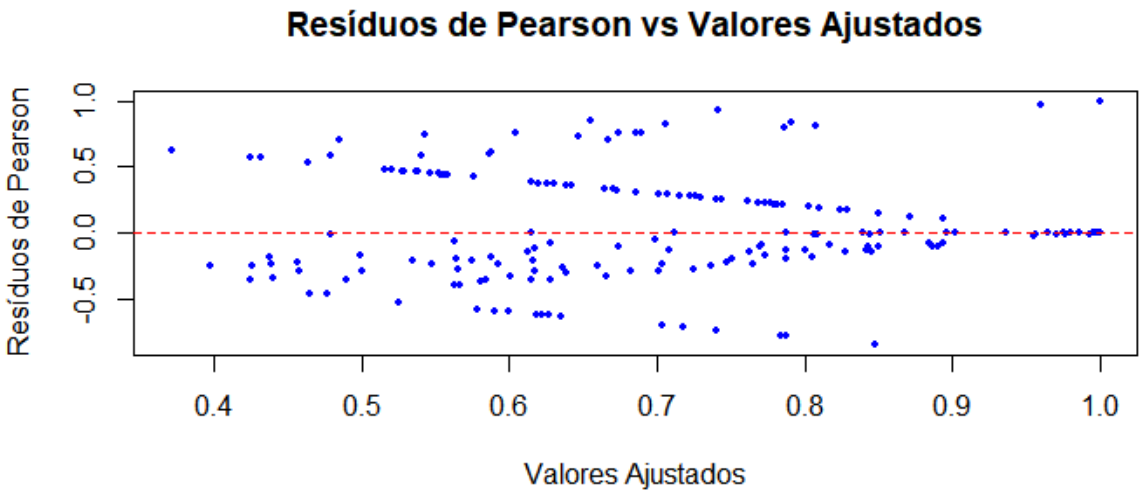

Figura 12 - Dispersão dos resíduos de pearson

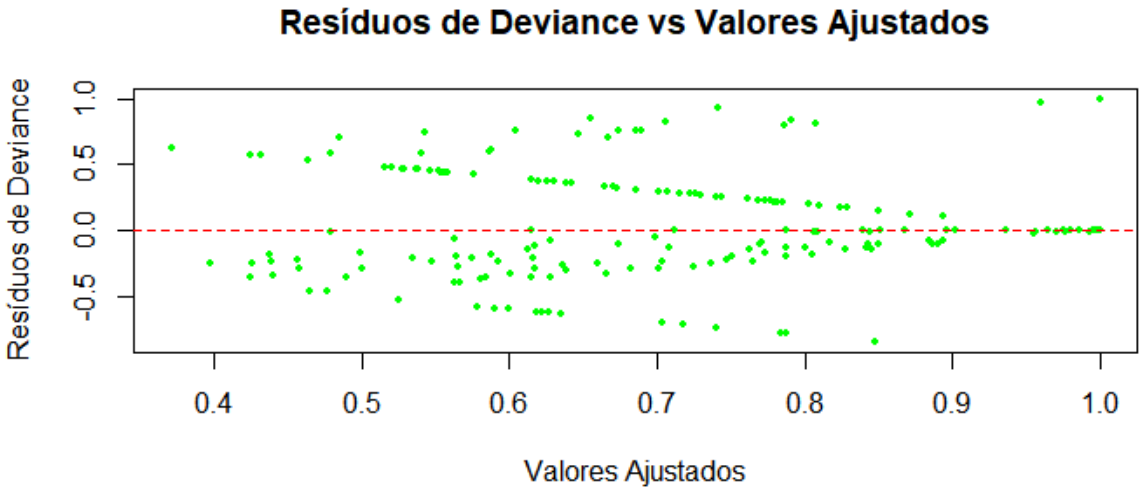

Figura 13 - Dispersão dos resíduos da deviância

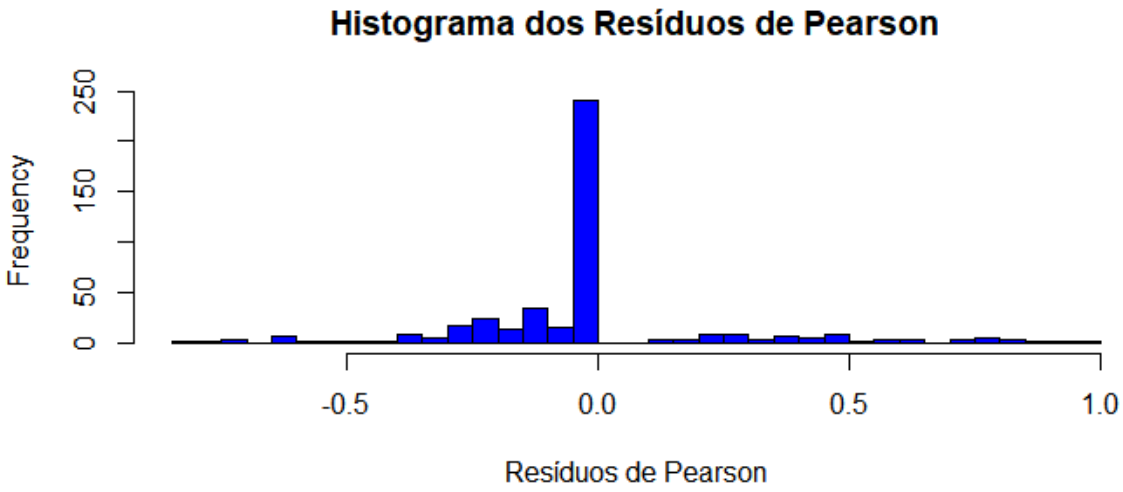

Figura 14 - Histograma dos resíduos de pearson

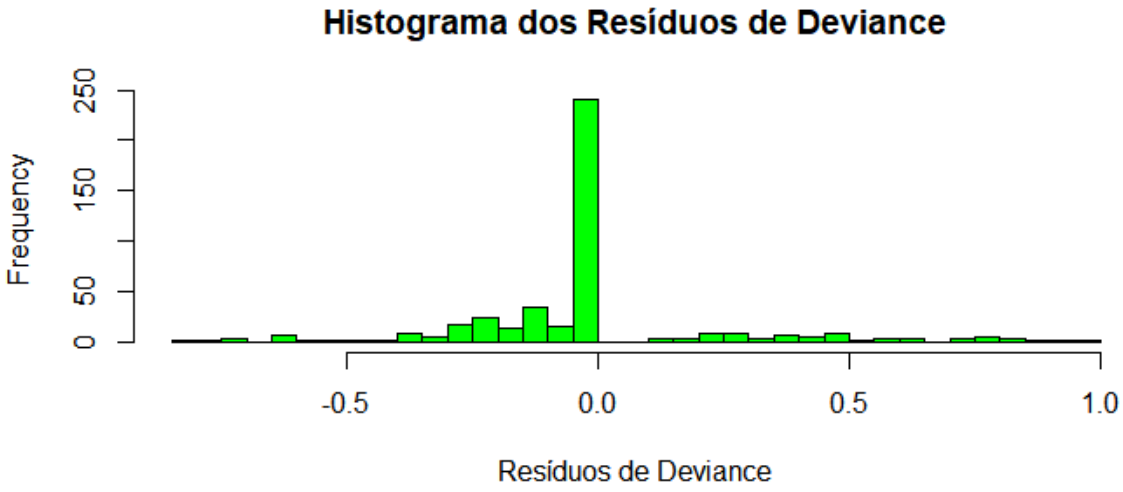

Figura 15 - Histograma dos resíduos da deviância

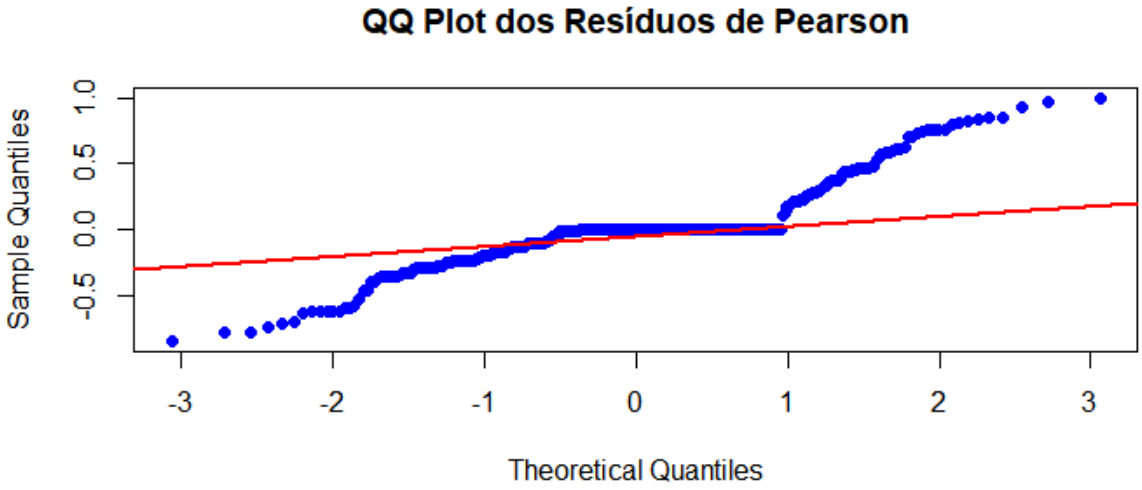

Figura 16 - Grafico QQ dos resíduos

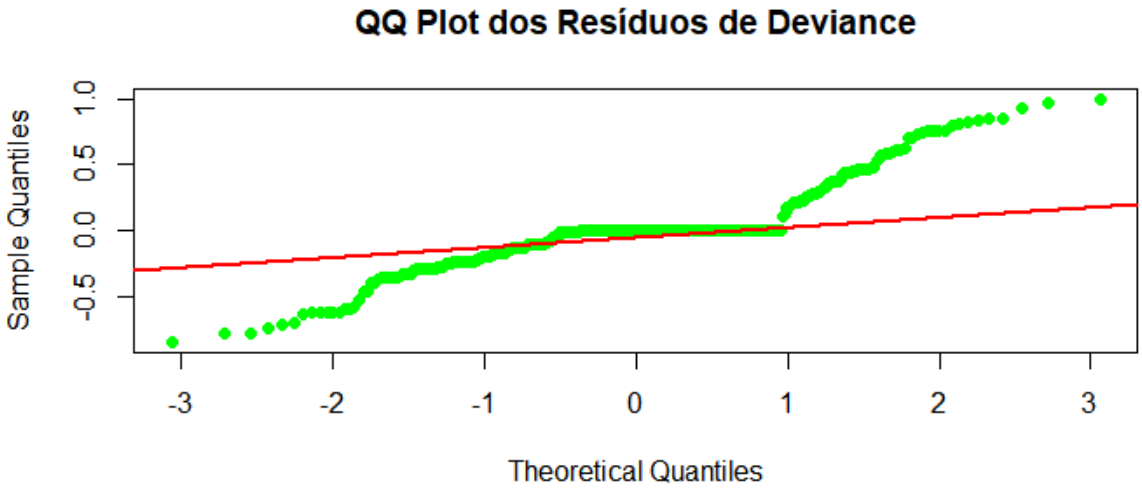

Figura 17 - Gráfico QQ dos resíduos da deviância

Tabela 18 - Medidas de resíduos

Medida	Valor
Resíduos Pearson (média)	-4.669299e-04
Resíduos deviância (média)	-4.669299e-04

Para que os resíduos sigam normalidade, os gráficos de dispersão deveriam ter dispersão aleatória sem tendência, os histogramas deveriam seguir uma distribuição normal, e os QQ-Plots deveriam ter pontos mais próximos a linha vermelha. Podemos concluir então que os resíduos não seguem uma distribuição normal, porém como nosso modelo é de classificação e não regressão, a normalidade dos resíduos não é uma premissa, uma vez que o modelo está bem ajustado e classificando as observações de forma satisfatória, isso se comprova com a tabela das medidas dos resíduos, onde a média se encontra próxima a zero.

Caso futuramente quisermos lidar com regressão, poderíamos pensar na alternativa de transformação de variáveis, uso de outros algoritmos ou modelos, como por exemplo o de árvore de decisão ou floresta aleatória, ou mesmo métodos de reamostragem.

A fim de finalizarmos a análise, iremos realizar o teste de Wald para teste de hipóteses:

$$H_0 \ vs \ H_1$$

No qual H_0 representa que as variáveis não são significativas para o modelo; H_1 que as variáveis são significativas.

Para cada variável, obtivemos os seguintes resultados:

Tabela 19 - Teste de wald (hipóteses)

Variável	X^2	Grau de liberdade	P-valor
RI	140.625	5	< 2.2e-16
Na	8.551	5	0.12838
Mg	49.407	5	1.833e-09
Al	29.378	5	1.955e-05
Si	37.086	5	5.757e-07
K	62.364	5	3.943e-12
Ca	28.465	5	2.952e-05
Ba	51.623	5	6.447e-10
Fe	11.299	5	0.04577

Para a variável ser significativa (rejeitando H_0), o seu p-valor deve ser menor que 0.05, temos então que todas as variáveis são significativas, com exceção da variável Na. Podendo futuramente então criarmos e testarmos um modelo sem esta variável e verificarmos se os resultados serão melhores.

9 CONCLUSÕES

Este trabalho teve como objetivo mesclar visões da ciência de dados, estatística e análise forense, a fim de usar métodos para classificar fragmentos de vidro a uma classe de vidro específica, assim podendo serem utilizados por profissionais da área criminal em investigações onde tais fragmentos são encontrados em suspeitos e nas cenas de crime, auxiliando a tomarem decisões mais justas e certeiras de quem é o criminoso.

Podemos então concluir que balancear os dados com sobreamostragem exerce um papel melhor nas análises, melhorando o ajuste dos modelos e suas acurácias e classificações.

O algoritmo de K-médias não se mostrou efetivo para este tipo de problema, classificando de forma errônea os fragmentos, a saída então foi optar por outro método. Escolhemos o KNN, com uma visão da ciência de dados, utilizando a técnica conjuntamente de dividir a base de dados em treino e teste, conseguimos obter resultados satisfatórios ao classificar os fragmentos em grupos distintos de vidro, isto é, a técnica pode ser utilizada com êxito futuramente.

O uso de regressão multinomial teve como intuito a visão estatística, e se mostrou excelente para classificar os tipos de vidro, gerando resultados interessantes de quais elementos químicos e índice de refração presentes no vidro são interessantes para tal. Além de obtermos também a proporção da predição correta de cada classe, oferecendo probabilidades que podem ser utilizadas para subjugar um suspeito do crime em questão.

Na visão forense, podemos concluir de forma geral que as técnicas aqui executadas podem auxiliar e muito os casos criminais, ainda mais no Brasil, onde a área da estatística forense não é amplamente usada e disseminada, podendo tornar futuramente o país mais justo e seguro, incentivando também mais profissionais estatísticos a ingressarem nesta área.

REFERÊNCIAS

CURRAN, J. M.; CHAMPOD, T. N. H.; BUCKLETON, J. S. (Ed.). **Forensic interpretation of glass evidence**. CRC Press, 2000.

IZBICKI, R.; DOS SANTOS, T. M. **Aprendizado de máquina: uma abordagem estatística**. Rafael Izbicki, 2020.

CURRAN, J. M. **Introduction to data analysis with R for forensic scientists**. CRC Press, 2010.

VENABLES, W. N.; RIPLEY, B. D. **Modern applied statistics with S-PLUS**. Springer Science & Business Media, 2013.

CHADREQUE, M. C. **Estatística na investigação forense**. 2012. Dissertação de Mestrado.

ZELTERMAN, D. **Estatística multivariada aplicada com R** . Basileia, Suíça: Springer International Publishing, 2015.

WICHERN, D. W.; JOHNSON, R. A. **Applied multivariate statistical analysis**. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 6th ed., 2007.

REIS, E. **Estatística multivariada aplicada.** Lisboa: Silabo, 1997.