Estudo Dirigido - Glass Identification

Ítalo Gonçalves e José Antonio

06/10/2024

Sobre a base de dados

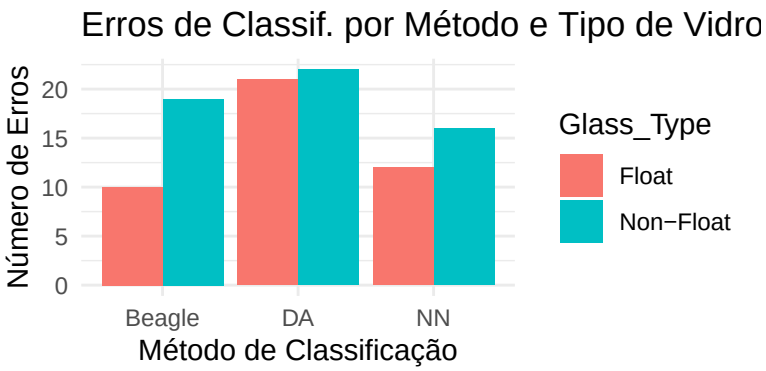
A base de dados **Glass Identification** é usada para classificar diferentes tipos de vidro com base em suas propriedades químicas, o que é útil em investigações criminais. Ao analisar fragmentos de vidro encontrados em cenas de crime, é possível identificar sua origem e ligá-los a um local ou suspeito. A base de dados contém variáveis que medem a quantidade de diferentes óxidos químicos presentes nas amostras de vidro, ajudando a distinguir entre os tipos, sendo eles:

- Sódio (Na)
- Magnésio (Mg)
- Alumínio (Al)
- Silício (Si)
- Cálcio (Ca)
- Bário (Ba)
- Ferro (Fe)

O objetivo principal desse estudo é determinar se o vidro pertence a janelas produzidas pelo método de "float glass" ou a outro tipo. Essa classificação é feita com base em padrões químicos, usando três métodos de análise para melhorar a precisão na identificação. Para isso, foi realizada uma análise com três métodos de classificação:

- **Beagle**: Resultados incorretos ao classificar vidros de janelas "float" foram 10, e 19 para vidros que não eram "float".
- **NN (Nearest Neighbors)**: Houve 12 respostas incorretas para vidros "float" e 16 para vidros que não eram "float".
- **DA (Discriminant Analysis)**: Este método apresentou 21 erros para vidros "float" e 22 para outros vidros.

A seguir, o gráfico mostra esses resultados:



Dimensionalidade dos Dados

```
dim(Glass)
```

```
## [1] 214  10
```

Essa base de dados possui 214 amostras e 10 variáveis. E cada variável representa uma característica química ou física dos fragmentos de vidro, sendo a última variável uma classe, sendo o tipo do vidro.

Análise Estrutural das Variáveis

Utilizando a função `str()` nos dá uma visão geral da base de dados, como o tipo de cada variável e uma amostra de seus valores:

```
str(Glass)
```

```
## 'data.frame':    214 obs. of  10 variables:
##  $ RI  : num  1.52 1.52 1.52 1.52 1.52 ...
##  $ Na  : num  13.6 13.9 13.5 13.2 13.3 ...
##  $ Mg  : num  4.49 3.6 3.55 3.69 3.62 3.61 3.6 3.61 3.58 3.6 ...
##  $ Al  : num  1.1 1.36 1.54 1.29 1.24 1.62 1.14 1.05 1.37 1.36 ...
##  $ Si  : num  71.8 72.7 73 72.6 73.1 ...
##  $ K   : num  0.06 0.48 0.39 0.57 0.55 0.64 0.58 0.57 0.56 0.57 ...
##  $ Ca  : num  8.75 7.83 7.78 8.22 8.07 8.07 8.17 8.24 8.3 8.4 ...
##  $ Ba  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Fe  : num  0 0 0 0 0 0.26 0 0 0 0.11 ...
##  $ Type: Factor w/ 6 levels "1","2","3","5",..: 1 1 1 1 1 1 1 1 1 1 ...
```

A partir dessa análise, podemos ver que todas as variáveis são do tipo numérico, exceto a variável TYPE, que é um fator com 6 níveis, representando as diferentes categorias de vidro.

Distribuição das Classes

Agora, utilizando a função `table()`, nos permite ver o número de amostras presentes em cada classe de vidro:

```
table(Glass$Type)
```

```
##
##  1  2  3  5  6  7
## 70 76 17 13  9 29
```

Os resultados mostram a quantidade de amostras em cada 1 das 6 classes de vidro, com isso podemos ver que as classes 1 e 2, correspondentes ao vidro de construção de janelas, são as mais frequentes, enquanto as outras classes têm uma quantidade menor de amostras.

Integridade da base de dados

Para verificar a integridade da base de dados e identificar se há valores ausentes (NA), utilizamos de duas formas para essa verificação, a função summary(glass), que fornece um resumo das variáveis, permitindo identificar se há ou não valores NA.

```
summary(Glass)
```

```
##        RI              Na              Mg              Al        
##  Min.   :1.511   Min.   :10.73   Min.   :0.000   Min.   :0.290  
##  1st Qu.:1.517   1st Qu.:12.91   1st Qu.:2.115   1st Qu.:1.190  
##  Median :1.518   Median :13.30   Median :3.480   Median :1.360  
##  Mean   :1.518   Mean   :13.41   Mean   :2.685   Mean   :1.445  
##  3rd Qu.:1.519   3rd Qu.:13.82   3rd Qu.:3.600   3rd Qu.:1.630  
##  Max.   :1.534   Max.   :17.38   Max.   :4.490   Max.   :3.500  
##        Si              K               Ca              Ba        
##  Min.   :69.81   Min.   :0.0000   Min.   : 5.430   Min.   :0.000  
##  1st Qu.:72.28   1st Qu.:0.1225   1st Qu.: 8.240   1st Qu.:0.000  
##  Median :72.79   Median :0.5550   Median : 8.600   Median :0.000  
##  Mean   :72.65   Mean   :0.4971   Mean   : 8.957   Mean   :0.175  
##  3rd Qu.:73.09   3rd Qu.:0.6100   3rd Qu.: 9.172   3rd Qu.:0.000  
##  Max.   :75.41   Max.   :6.2100   Max.   :16.190   Max.   :3.150  
##        Fe          Type      
##  Min.   :0.00000   1:70  
##  1st Qu.:0.00000   2:76  
##  Median :0.00000   3:17  
##  Mean   :0.05701   5:13  
##  3rd Qu.:0.10000   6: 9  
##  Max.   :0.51000   7:29
```

E como segundo método, de uma forma mais direta, usamos a seguinte função para mostrar se há algum valor NA nesta base:

```
sum(is.na(Glass))
```

```
## [1] 0
```

Após essa verificação, pode-se dizer que não há dados ausentes na base de dados, ou seja todas as amostras desta base estão completas. Portanto, não é necessário retirar nenhuma amostra.

Porcentagem dos dados

Além disso, também calculei a porcentagem de cada tipo em relação ao total de amostras. O código utilizado para essa análise foi:

```
round(prop.table(table(Glass$Type)) * 100, 2)
```

```
##
##     1     2     3     5     6     7 
## 32.71 35.51  7.94  6.07  4.21 13.55
```

Com base nos resultados, observamos que a base de dados apresenta um desequilíbrio significativo, uma vez que a classe com a maior porcentagem representa 32,71% do total de amostras, enquanto a classe com a menor porcentagem corresponde apenas a 4,21%.

Balanceamento dos dados

Problemas de Regressão

Análise de Outliers

Análise Descritiva das Variáveis

Análise de Correlação

Pré-processamento e Padrões Esperados

Estudos sobre a base de dados