# Trabalho Final

Felipe Rocha

2022-1-6

## Tarefa 1

**Utilize o conjunto BreastCancer para realizar análises de reconhecimento de padrões. Para tanto suponha que a variável Class (última variável do conjunto) não é conhecida. Faça um relatório. Tal relatório deve ser estregue de maneira organizada discutindo os resultados encontrados.**

O bando de dados *BreastCancer* possui 699 observações e 9 variáveis a serem analisadas:

- *Cl.thickness*: Espessura do aglomerado
- *Cell.size*: Uniformidade do tamanho da célula
- *Cell.shape*: Uniformidade da forma da célula
- *Marg.adesão*: Adesão Marginal
- *Epith.c.size*: Tamanho de célula epitelial única
- *Bare.nuclei*: Núcleos nus
- *Bl.cromatina*: Cromatina sem graça
- *Normal.nucleoli*: Nucléolos Normais
- *Mitoses*: Mitoses

Com posse dos dados e sem nenhum conhecimento prévio sobre grupos ao qual gostaríamos de classificar os dados, ou seja, não possuímos informações de categorias já existentes para então classificarmos os dados nestas, optou-se pela realização de uma análise não supervisionada.

```
library(mlbench) #banco de dados
library(dplyr)
library(tidyr)
library(factoextra)
library(cluster)
data(BreastCancer)
dados=BreastCancer%>%dplyr::select(-c(Id,Class))%>%dplyr::mutate(
  dplyr::across(.cols=everything(),~as.numeric(.)));rm(BreastCancer)
summary(dados)
```

```
##   Cl.thickness      Cell.size        Cell.shape     Marg.adhesion
##  Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000
##  1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.: 1.000
##  Median : 4.000   Median : 1.000   Median : 1.000   Median : 1.000
##  Mean   : 4.418   Mean   : 3.134   Mean   : 3.207   Mean   : 2.807
##  3rd Qu.: 6.000   3rd Qu.: 5.000   3rd Qu.: 5.000   3rd Qu.: 4.000
##  Max.   :10.000   Max.   :10.000   Max.   :10.000   Max.   :10.000
```

```
##
##   Epith.c.size     Bare.nuclei      Bl.cromatin     Normal.nucleoli
##  Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000
##  1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 2.000   1st Qu.: 1.000
##  Median : 2.000   Median : 1.000   Median : 3.000   Median : 1.000
##  Mean   : 3.216   Mean   : 3.545   Mean   : 3.438   Mean   : 2.867
##  3rd Qu.: 4.000   3rd Qu.: 6.000   3rd Qu.: 5.000   3rd Qu.: 4.000
##  Max.   :10.000   Max.   :10.000   Max.   :10.000   Max.   :10.000
##                   NA's   :16
##     Mitoses
##  Min.   :1.000
##  1st Qu.:1.000
##  Median :1.000
##  Mean   :1.569
##  3rd Qu.:1.000
##  Max.   :9.000
##
```

Após uma breve análise exploratória e observando que todos os dados são numéricos e não foram observados *outliers*, escolheu-se o método **k-means** de clusterização não-supervisionado.

## K-means

Como passo inicial ao processo de clusterização, observou-se a presença de 16 valores faltantes na variável *Bare.nuclei* e, após a remoção das linhas com esses dados, sobraram 683 observações a serem utilizadas no modelo.

```
dados=dados%>%tidyr::drop_na();nrow(dados)
```

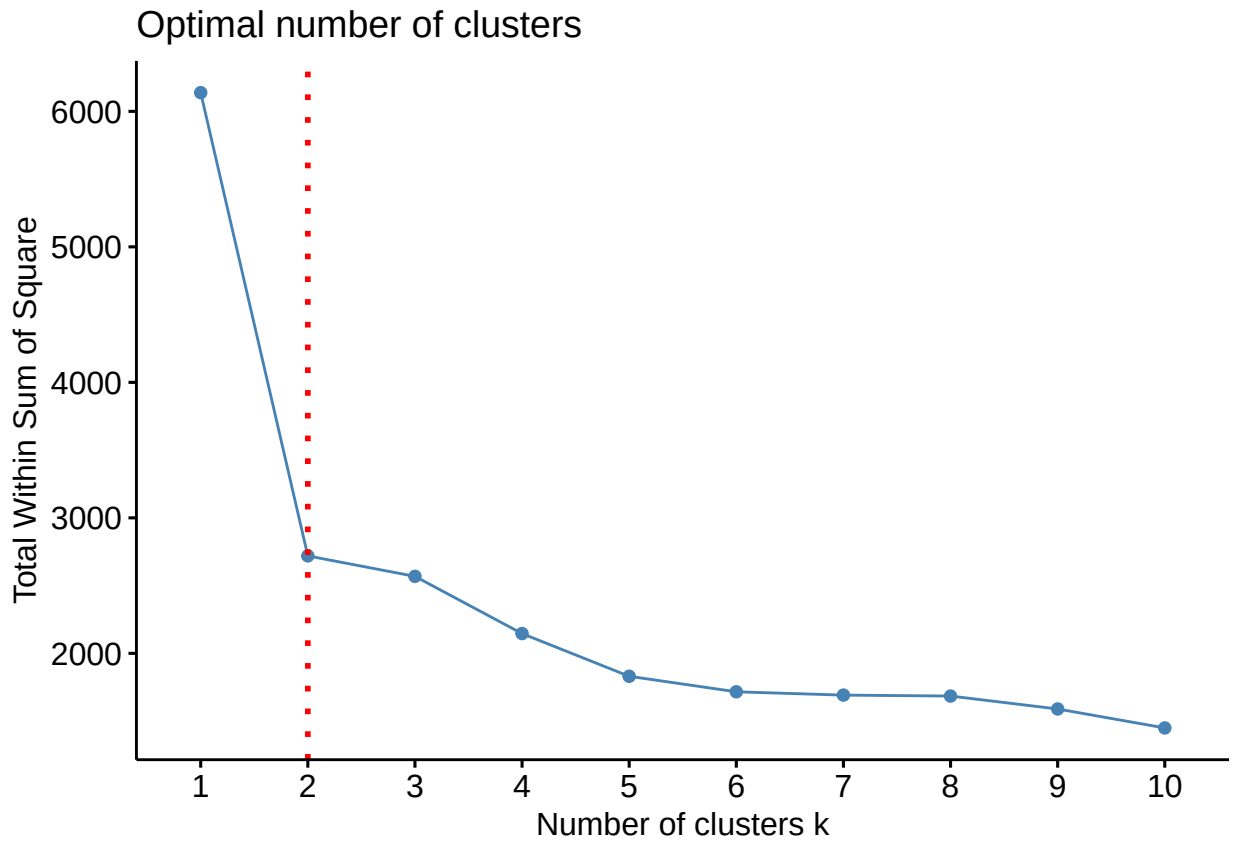
```
## [1] 683
```

Após a remoção, normalizamos os dados afim de garantir que todas as variáveis possuíssem média 0 e desvio-padrão 1.

```
dados <- scale(dados)%>%data.frame()
```

Utilizando o princípio do cotovelo, escolhemos dois como número ideal de clusters: por meio do gráfico a seguir, é possível observar que de 1 para 2, a diferença da soma dos quadrados é significativa, mas de 2 para 3 não, fazendo assim a curva e gerando o "cotovelo". Por tanto, pela parcimónia, optou-se por 2 clusters.

```
factoextra::fviz_nbclust(dados, kmeans, method = "wss")+
  geom_vline(xintercept = 2,linetype="dotted",color="red",size=1)
```




Por fim, criamos nosso modelo, que resultou em 2 clusters com 452 observações classificadas no *cluster 1* e 231 no *cluster 2*.

```
set.seed(573)#número da disciplina no PAVNET
km <- kmeans(dados, centers =2)
km$size
```

```
## [1] 452 231
```

```
paste0("(between_SS / total_SS =  ",round(km$betweenss/km$totss*100,1),"%)")
```

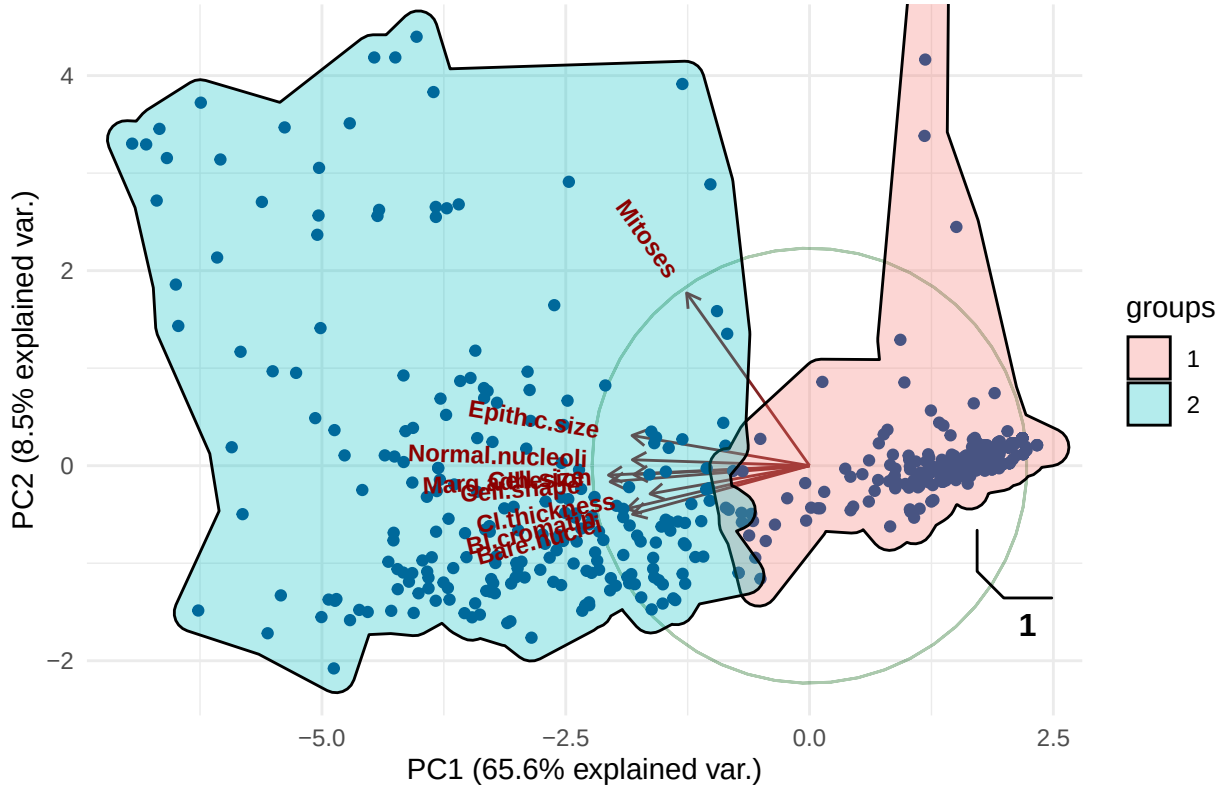
```
## [1] "(between_SS / total_SS =  55.7%)"
```

## Componentes Principais (PCA)

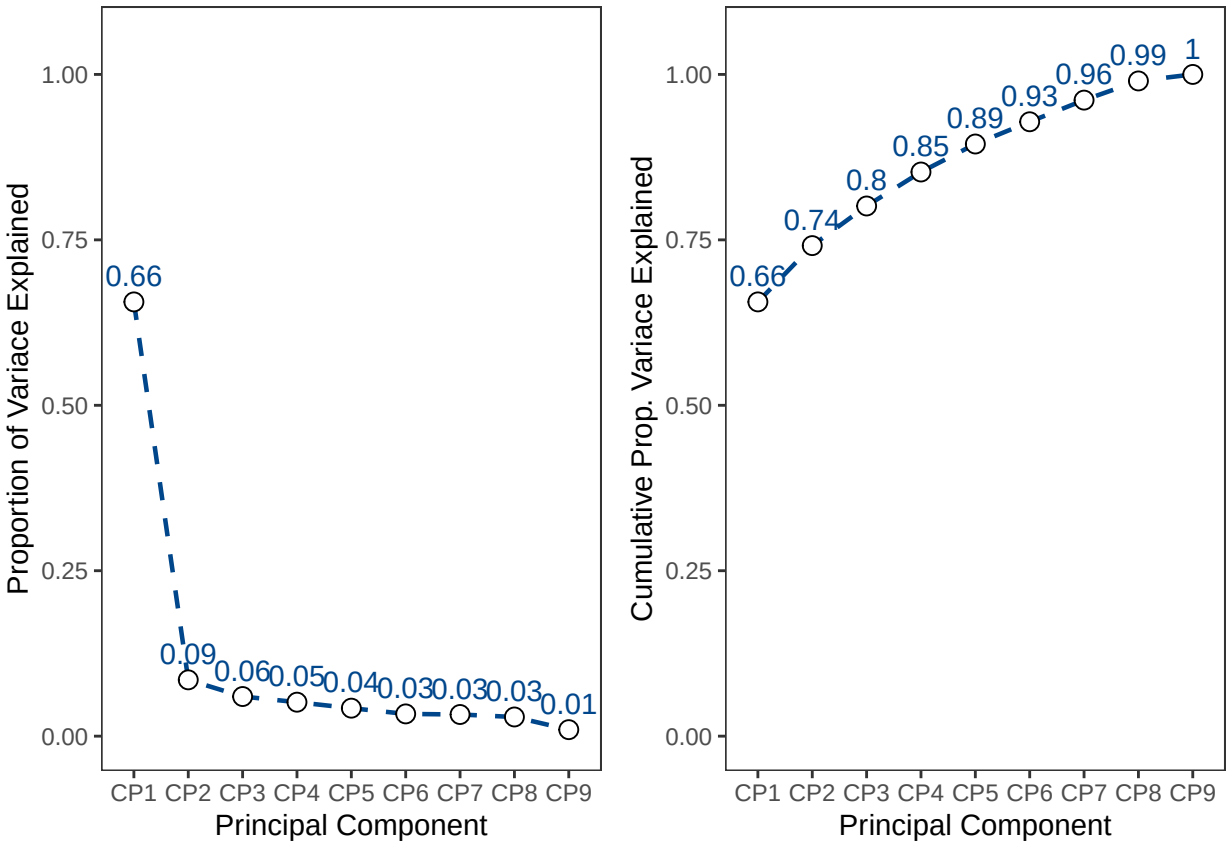
Utilizando o método dos **Componentes Principais** (PCA) para visualizar os dados, podemos observar que a variável *Mitoses* compõem uma componente (*PC2*, que explica 8,5%) e as demais compõem o *PC1* (a componente mais explicativa, 65,6%). E observa-se que o *PC1* é o que mais influência na classificação dos indivíduos: baixos valores das variáveis dessa componente resultam na classificação no *grupo 1* (vermelho), enquanto altos valores resultam na classificação em *grupo 2* (azul).

```
source("F:/Estudo/Estudo---UFV/UFV---GitHub/00.Functions/Function_PCA.R") #[1]
dados2=dados%>%dplyr::mutate(classificado=km[["cluster"]])
pca=prcomp(dados2%>%dplyr::select(-classificado), scale=TRUE)
```

```
ggbiplot(pcobj=pca,
         obs.scale = 1, var.scale = 1,
         groups = dados2$classificado%>%as.character(),ellipse = TRUE,
         circle = TRUE,labels = dados$Class ) +
  scale_color_discrete(name = '') +
  theme(legend.direction = 'horizontal', legend.position = 'top')+theme_minimal()
```




```
PCA_Prop_Var(pca)
```




Sobre a interpretação desses grupos e componentes, caberia a um especialista mais familiarizado com esses dados. Um exemplo hipotético de interpretação seria o *grupo 2* ser pessoas com alto risco de possuírem câncer de mama e o *grupo 1* pessoas com baixo risco (nesse exemplo hipotético, indivíduos com baixo valores no *componente 1*, cuja variáveis estão correlacionadas, resultariam em altas chances de possuir o câncer).

*Function_PCA.R está disponível em https://github.com/felipe179971/felipe179971-UFV-Inteligencia_Artificial_e_Computacional/blob/main/00.Functions/Function_PCA.R*