U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Professor Joaquim Costa

Métodos Estatísticos em *Data Mining*

Primeiro Projeto: Análise de Componentes Principais, Análise de *Clusters* e *Internal Measure*

Abril de 2023

Trabalho realizado por:

Pedro Leite - 201906697

Pedro Carvalho - 201906291

Índice:

1. Introdução

O objetivo deste projeto é fazer uma análise dos atributos do *dataset* (2.) "Telco Customer Churn", para perceber a sua influência nos clientes que mudaram de operadora. Para fazer esta análise vamos utilizar vários métodos estatísticos de *data mining*, nomeadamente: (4.1.) Análise de Componentes Principais (ACP) e (4.2.) Análise de *Cluster*s. Estes métodos permitem-nos reduzir a dimensão dos dados e identificar padrões. No final, iremos apresentar uma discussão e conclusão dos resultados.

2. Background

O *dataset* que escolhemos para este projeto foi o [1] "Telco Customer Churn". Este *dataset* contém 7043 clientes e 21 atributos, 11 dos clientes não foram utilizados porque têm atributos com valores não atribuídos. Os atributos são:

- "CustomerID" é uma *string* única que identifica o cliente.
- "Gender" é uma *string* que identifica o cliente como "Female" ou "Male".
- "SeniorCitizen" é um número binário se o cliente é ou não idoso.
- "Partner" é uma *string* "Yes" ou "No" que diz se o cliente está casado.
- "Dependents" é uma *string* "Yes" ou "No" que diz se o cliente tem filhos.
- "Tenure" é o número de meses que o cliente está na operadora atual.
- "PhoneService" é uma *string* "Yes" ou "No" que diz se o cliente tem serviço de telefone.
- "MultipleLines" é uma *string* "Yes", "No" ou "No phone service" que diz se o cliente tem múltiplas linhas de telemóvel.
- "InternetService" é uma *string* "DSL", "Fiber optic" ou "No" que diz que tipo de serviço de internet o cliente possui.
- "OnlineSecurity" é uma *string* "Yes", "No" ou "No internet service" que diz se o cliente tem segurança online.
- "OnlineBackup" é uma *string* "Yes", "No" ou "No internet service" que diz se o cliente tem *backup* online.
- "DeviceProtection" é uma *string* "Yes", "No" ou "No internet service" que diz se o cliente tem proteção do dispositivo .

- "TechSupport" é uma *string* "Yes", "No" ou "No internet service" que diz se o cliente tem suporte técnico.
- "StreamingTV" é uma *string* "Yes", "No" ou "No internet service" que diz se o cliente tem *streaming* de canais televisivos.
- "StreamingMovies" é uma *string* "Yes", "No" ou "No internet service" que diz se o cliente tem *streaming* de filmes.
- "Contract" é uma *string* "Month-to-month", "One year" ou "Two year" que diz que tipo de contrato o cliente tem.
- "PaperlessBilling" é uma *string* "Yes" ou "No" que diz se o o cliente tem faturamento sem papel.
- "PaymentMethod" é uma *string* "Electronic check", "Mailed check", "Bank transfer (automatic)" ou "Credit card (automatic)" que diz o tipo de pagamento que o cliente utiliza.
- "MonthlyCharges" é o montante que o cliente paga mensalmente.
- "Total Charges" é o montante que o cliente pagou até agora.
- "Churn" é o atributo objetivo que quer dizer se um cliente mudou de operadora, no último mês.

3. Materiais

O projeto foi feito utilizando a linguagem R, com as bibliotecas: "tidyverse", "readr", "dplyr", "tidyr", "caret", "ggplot2", "reshape2", "cluster", "tidyverse", "tidymodels", "pheatmap", "tidyr", "janitor", "factoextra", "cluster" e "fpc". Os resultados foram corridos num computador com as seguintes especificações: AMD Rysen 5 2600X SixCore Processor 3.6 GHz, 16GB RAM, Asus GeForce GTX 1660 Ti.

4. Estudo e Aplicação

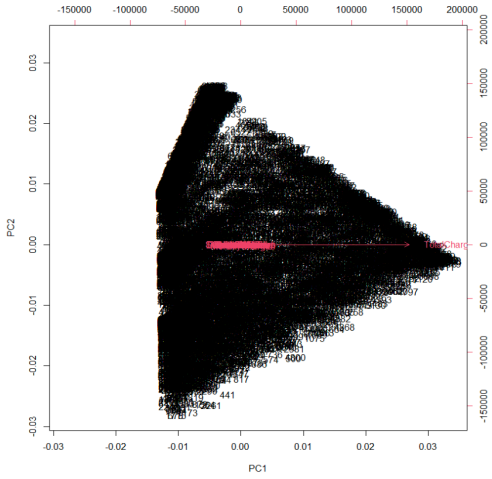
4.1. Análise de Componentes Principais (ACP)

Como temos um *dataset* com um elevado número de atributos e valores, utilizamos o método ACP, que nos permite reduzir a dimensão dos dados, mantendo a maior parte das informações relevantes. Originando vários componentes principais (CP), que representam a combinação linear dos valores

dos atributos com maior variância, por ordem de grandeza, por exemplo, o CP1 é mais importante que o CP2.

Para analisar a aplicação do método ACP, utilizamos o *biplot* que apresenta a relação entre o CP1 e o CP2, os componentes principais mais importantes. Utilizamos também a matriz de correlação, que mostra as correlações entre os pares de atributos.

Primeiramente, aplicamos *one hot encoding* ao nosso *dataset*, que consiste em substituir todos os valores das variáveis categóricas, por números, por exemplo, no "Multiple Lines" substituímos o "Yes", "No" e "No phone service" por "1", "2" e "3", respetivamente. Gerando o seguinte *biplot* e resultados do ACP:

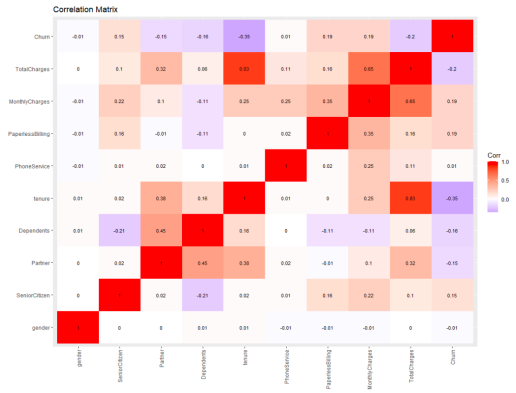



```
> summary(pca)
Importance of components:
                          PC1      PC2     PC3   PC4   PC5    PC6    PC7    PC8 PC9   PC10   PC11   PC12   PC13  PC14
Standard deviation     2266.9469 25.07315 9.25969 1.159 1.064 0.6101 0.5275 0.5077 0.5 0.4475 0.4075 0.3894 0.3741 0.372
Proportion of Variance    0.9999  0.00012 0.00002 0.000 0.000 0.0000 0.0000 0.0000 0.0 0.0000 0.0000 0.0000 0.0000 0.000
Cumulative Proportion     0.9999  0.99998 1.00000 1.000 1.000 1.0000 1.0000 1.0000 1.0 1.0000 1.0000 1.0000 1.0000 1.000
                          PC15   PC16   PC17   PC18   PC19   PC20
Standard deviation     0.3627 0.3522 0.3318 0.3118 0.1929 0.1048
Proportion of Variance 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
Cumulative Proportion  1.0000 1.0000 1.0000 1.0000 1.0000 1.0000
```

Fizemos uma matriz de correlação com todos os atributos, e verificamos uma correlação alta ou baixa entre certos atributos não binários, e aplicamos o método *chi-squared test* nesses atributos, para verificar a associação significativa entre as duas variáveis, por exemplo no InternetService e PhoneService:

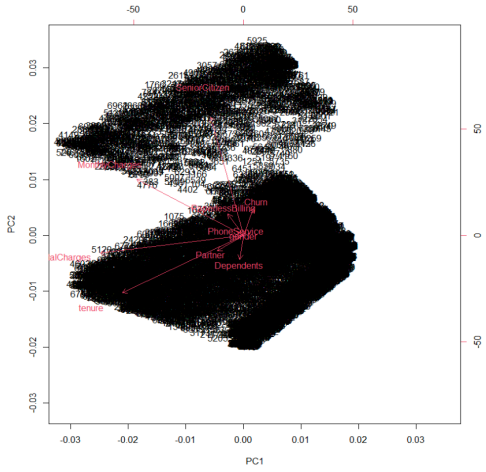
```
        Pearson's Chi-squared test

data:  churn_gender_table
X-squared = 1438.3, df = 2, p-value < 2.2e-16
```

Como podemos ver pelo *biplot* não é possível tirar conclusões, por isso decidimos aplicar *one hot encoding* apenas aos atributos categóricos binários. Gerando a seguinte matriz de correlação:




```
> summary(pca)
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10
Standard deviation     1.5009 1.0630 0.8522 0.52916 0.49996 0.45138 0.37742 0.33272 0.28316 0.23847
Proportion of Variance 0.4305 0.2159 0.1388 0.05351 0.04777 0.03893 0.02722 0.02116 0.01532 0.01087
Cumulative Proportion  0.4305 0.6464 0.7852 0.83874 0.88650 0.92544 0.95266 0.97381 0.98913 1.00000
```

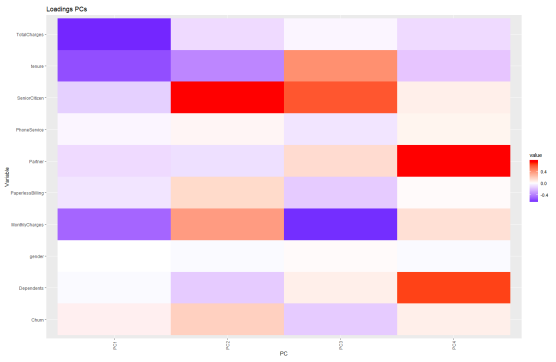
A partir destes novos resultados, já conseguimos tirar conclusões e provar a eficácia dos métodos, já que os atributos que demonstram uma correlação elevada devem demonstrar uma correlação elevada, por exemplo o TotalCharges e o MonthlyCharges, e os que demonstram uma correlação baixa devem demonstrar uma correlação baixa, por exemplo o Dependents e o SeniorCitizen, que demonstram uma correlação baixa, já que os cidadãos idosos, normalmente não têm dependentes. A partir do *biplot* podemos tirar as mesmas conclusões.



A partir do *biplot*, podemos ver a correlação entre os atributos, se o ângulo que 2 atributos formam for menor que 90º, significa que têm correlação positiva, se o ângulo que 2 atributos formam for maior que 90º, significa que têm correlação negativa. Os atributos: Churn e TotalCharges, Churn/SeniorCitizen e Tenure/ Partner/Dependents , PaperlessBilling/MonthlyCharges e Dependents, têm um ângulo maior que 90º, logo têm correlação negativa. O comprimento dos

vetores, representam a importância do atributo na justificação da variância dos dados no *dataset*. Os que apresentam os maiores vetores são: SeniorCitizen, MonthlyCharges, TotalCharges e Tenure.

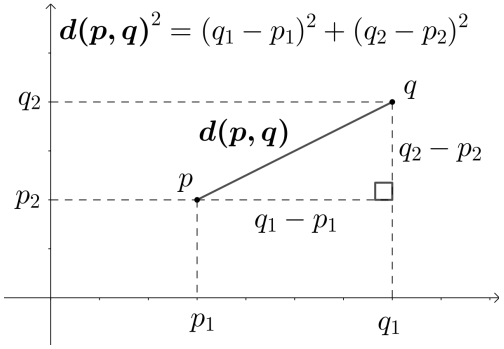
A variância apresentada pelo CP1 é, aproximadamente, 43%, pelo CP2 é, aproximadamente, 21%, pelo CP3 é, aproximadamente, 12%, pelo CP4 é, aproximadamente, 5%, etc. Utilizamos apenas os primeiros 4 CPs porque representam, aproximadamente, 84% da variância. E calculamos a importância dos atributos em cada um dos CPs, porque vai ser relevante para a análise de *clusters*:



4.2. Análise de *Clusters*

A análise de *clusters*, permite-nos agrupar observações semelhantes de um *dataset* em *clusters*, com base nas suas características.

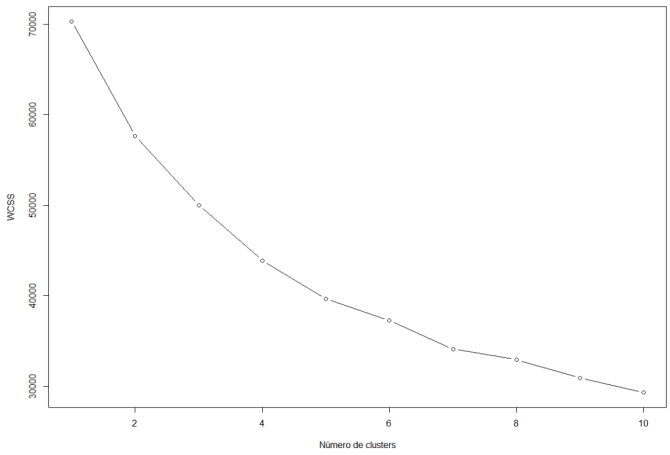
Começamos por fazer a análise dos *clusters* através do dendrograma. Cada nó do dendrograma representa um *cluster*, os *clusters* são definidos com base no grau de similaridade, no nosso caso, utilizamos como medida da distância entre os *clusters* a distância euclidiana:



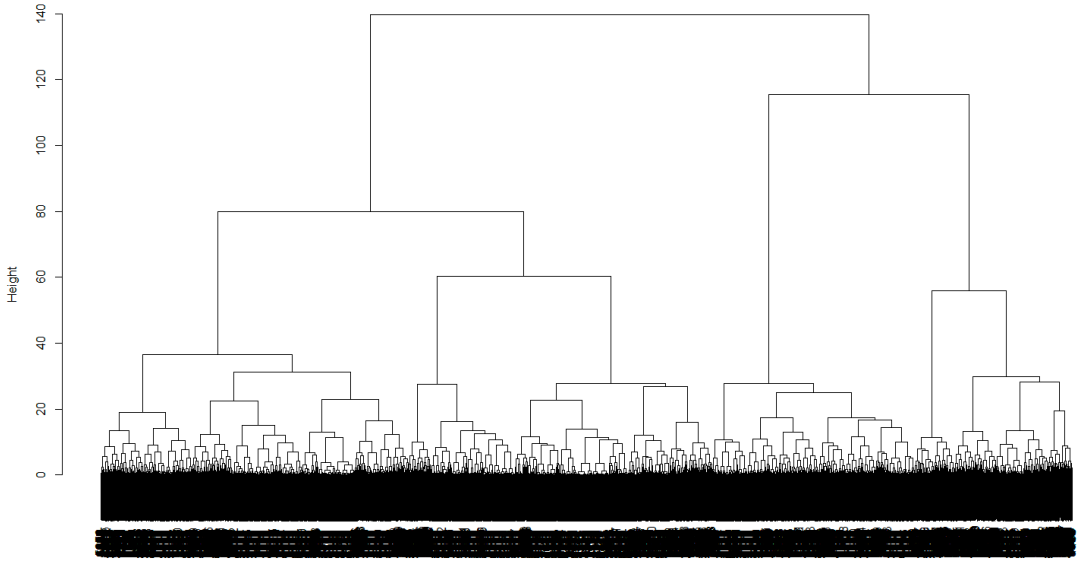
O método para a escolha da melhor distância para agrupar 2 grupos, foi o método de Ward:

$$\begin{aligned}\Delta(A,B) &= \sum_{i\in A\cup B}\|\vec{x}_i-\vec{m}_{A\cup B}\|^2-\sum_{i\in A}\|\vec{x}_i-\vec{m}_A\|^2-\sum_{i\in B}\|\vec{x}_i-\vec{m}_B\|^2\\ &= \frac{n_A n_B}{n_A+n_B}\|\vec{m}_A-\vec{m}_B\|^2\end{aligned}$$

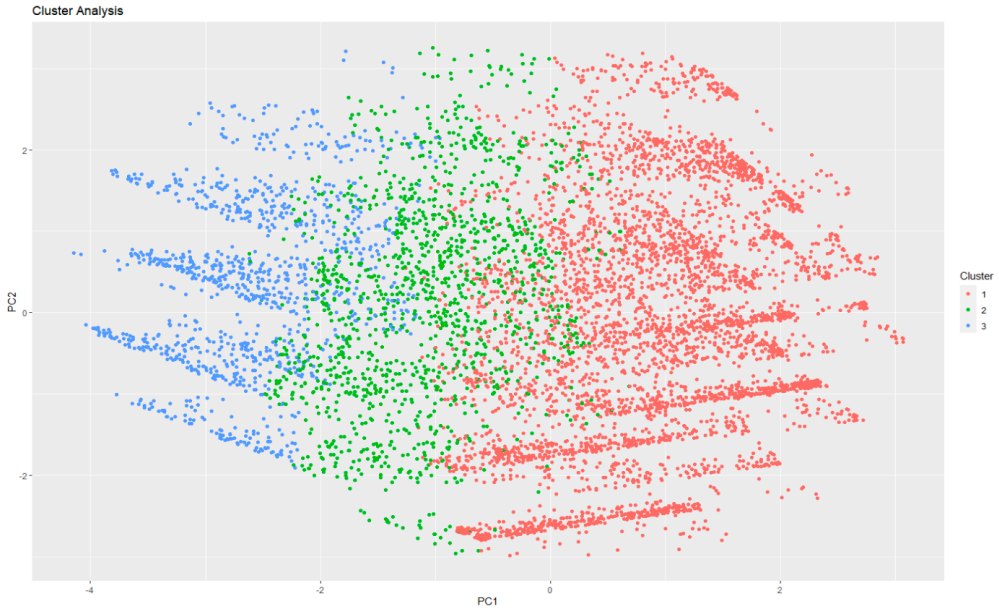
Fizemos um gráfico que apresenta a relação entre o WCSS (*Within Cluster Sum of Squares*) e o número de *clusters*, para determinar o número de *clusters* ideal:



Utilizamos o método do cotovelo, que diz que o número ideal de *clusters* é o ponto do gráfico em que faz uma "dobra", esse ponto é nos 3 *clusters.* O dendrograma que obtemos é o seguinte, a altura de cada *cluster* é a distância entre os mesmos:



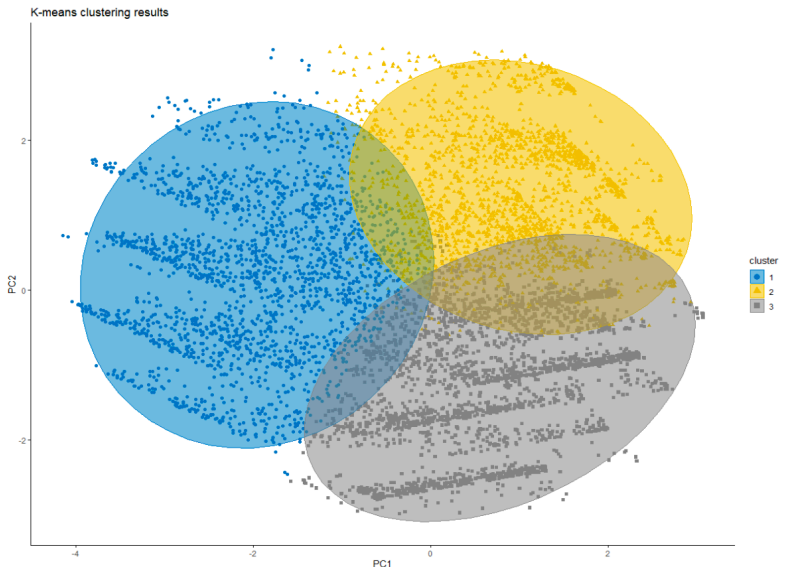
O resultado está ilegível, e por isso fizemos um *scatterplot*, que utiliza as duas componentes principais e atribui cores às observações baseado nos agrupamentos hierárquicos de *clustering*:



Representamos o *k-means*, que é um método, para todas as variáveis quantitativas, que procura minimizar a soma dos quadrados das distâncias ao centro do grupo respetivo:

$$J = \sum_{k=1}^{K} \sum_{C(i)=k} ||X_i - \overline{X}_k||^2$$

Obtendo o seguinte resultado:



Vimos anteriormente, (4.1.) na análise importância dos atributos em cada um dos CPs, que o CP1 é inversamente relacionado com TotalCharges, Tenure e MonthlyCharges e que CP2 é relacionado com SeniorCitizen. Portanto podemos concluir que o *cluster* azul (que tem um valor baixo de CP1) é constituído por clientes que têm um valor elevado de TotalCharges, Tenure e MonthlyCharges. O *cluster* cinzenta tem um valor alto de CP1 e baixo de CP2 por isso tratam-se de clientes jovens que não tem despesas elevadas e o *cluster* amarelo é constituída por idosos com poucas despesas.

4.3. *Internal Measure*

Internal measure é uma medida de qualidade usada para avaliar a validade dos *clusters* obtidos.

Para calcular o *internal measure,* utilizamos o *silhouette coefficient*, o valor do *silhouette* de cada uma das observações varia entre -1 e 1, -1 quando está no *cluster* errado, 0 quando está em 2 *clusters* ao mesmo tempo e 1 quando está no *cluster* correto. Somamos o valor de todas as observações e dividimos pelo número total de observações, para obter uma média. A média do *k-means* é 0.1762239, o que quer dizer que o *clustering* não foi muito bem feito, do *scatterplot* do *cluster* hierárquico é 0.6489913, que é um resultado muito melhor.

5. Discussão e Conclusão

Ao longo do projeto, aplicamos vários métodos estatísticos de *data mining* no *dataset* "Telco Customer Churn", nomeadamente: (4.1.) Análise de Componentes Principais (ACP) e (4.2.) Análise de *Clusters*, com o objetivo de reduzir a dimensão dos dados e identificar padrões.

Ao aplicar o ACP, reduzimos a dimensão do *dataset.* Foi possível identificar *clusters* de clientes que apresentam comportamentos semelhantes, bem como *outliers* que apresentam características únicas.

Ao aplicar métodos de análise de *clusters*, conseguimos construir *clusters* que demonstram a relação entre certos atributos e como o seu valor influência na decisão do cliente de mudar de operador.

Concluindo, foi possível identificar *clusters* de clientes com comportamentos semelhantes e atributos que influenciam a decisão de mudar de operadora.

Através dos métodos aplicados, percebemos que é importante compreender o comportamento do cliente para o sucesso de uma empresa.

6. Bibliografia

[1] https://www.kaggle.com/datasets/blastchar/telco-customer-churn