Segmentação de clientes e análise dos atributos mais relevantes dos *clusters*

Nuno Lopes[1], Luís Cavique[2]

[1] Universidade Aberta de Portugal, Lisboa, Portugal, 803204@estudante.uab.pt
[2] Universidade Aberta de Portugal, Lisboa, Portugal, luis.cavique@uab.pt

Resumo

Tendo por base um conjunto de dados dos clientes de uma empresa de produtos alimentares, tentamos implementar duas estratégias de *data mining* com o objetivo de compreender quais os atributos que melhor podem segmentar estes consumidores. Aplicamos primeiro um algoritmo de segmentação (*k-means*) para agrupar estes clientes e, seguidamente, utilizamos um algoritmo de classificação (árvore de decisão) para análise visual dos atributos que definiram os clusters da segmentação. Através da análise visual dos gráficos resultantes da indução de árvores de decisão conseguimos verificar que só o valor do salário dos clientes pode segmentar este conjunto de dados.

Palavras-chave: segmentação, classificação, prospeção de dados, k-means, árvores de decisão

Title: Customers Segmentation: Decision Tree Induction for Visual Analysis of Cluster Attributes

Abstract: From a dataset of customers of a food company, we tried to implement two data mining strategies to understand which attributes can best segment these consumers. First, we applied a segmentation algorithm (k-means) to segment these customers and then we applied a classification algorithm (decision tree) for visual analysis of the attributes that defined the segmentation clusters. Through the visual analysis of the graphs resulting from the decision tree induction, we were able to verify that only the value of the customers' salary can segment this dataset.

Keywords: clustering, classification, data mining, k-means, decision trees

1. Introdução

A segmentação de mercado é uma estratégia de marketing que consiste em dividir esse mercado – alvo em subconjuntos de consumidores que têm necessidades comuns. As empresas não conseguem satisfazer as necessidades de todos os clientes, utilizando as mesmas formas de marketing. Os consumidores têm diferentes preferências e raramente um produto ou serviço satisfaz completamente todos. Como nem todos os consumidores são iguais, a segmentação tem por objetivo agrupar consumidores em grupos homogéneos de acordo com as suas características. Estas características podem ser demográficas, geográficas, baseadas na personalidade, comportamento como consumidor ou relação com o produto.

No caso da segmentação demográfica, o mercado pode ser dividido tendo por base algumas variáveis como: idade, sexo, rendimentos, ocupação, estado civil, tamanho da família ou nacionalidade. Trata-se de uma metodologia popular de segmentação, já que dependem de variáveis fáceis de recolher, medir e pouco mutáveis ao longo do tempo. No caso da segmentação baseada no comportamento como consumidor, esta tem por base variáveis relativas às compras realizadas pelo cliente, nomeadamente valores, volume, frequência, e por isso, mais mutáveis ao longo do tempo.

O sucesso de uma ação de marketing direcionado para determinados grupo-alvo, depende da qualidade dos segmentos construídos ou identificados. Assim, a metodologia utilizada é um fator fundamental para o sucesso da segmentação do mercado (Bendle, Farris, Pfeifer, & Reibstein, 2017; Berry & Linoff, 2011; Celeste & Moniz, 2019; Rodrigues & Oliveira, 2013; Cavique 2003; & Cavique 2007).

As técnicas de *data mining* podem auxiliar na melhoria dos processos de segmentação dos clientes. O *data mining* é um processo não trivial de descoberta de conhecimento novo, implícito, útil e abrangente a partir de uma grande quantidade de dados. Neste trabalho iremos utilizar duas estratégias de *data mining:* a segmentação e a classificação (Bose & Chen, 2009; Tan, Steinbach, Karpatne, & Kumar, 2018 & Witten & Frank, 2005).

2. Metodologia

Neste artigo vamos tentar compreender as vantagens da utilização de um algoritmo de classificação de indução de árvores de decisão para análise visual dos atributos utilizados na segmentação dos dados de consumidores. Para tal realizamos um estudo de caso, empregando um conjunto de dados públicos sobre clientes de uma empresa de venda de produtos alimentares. Neste estudo tentamos compreender que tipo de segmentação consegue definir melhor estes consumidores: os seus atributos demográficos ou os seus comportamentos como consumidores?

Para responder a este problema, seguimos o subsequente procedimento: 1- Engenharia dos dados, 2- Segmentação dos clientes e 3- Análise da classificação dos atributos. Todo o procedimento foi realizado utilizando a linguagem de programação R.

2.1. Engenharia dos dados

Antes de se aplicar modelos de *data mining* a um conjunto de dados, é fundamental analisar esse mesmo conjunto de dados, recorrendo para tal a técnicas estatísticas e de visualização. Compreender a distribuição dos dados é essencial para se escolher as técnicas de modelação que auxiliarão a responder ao problema.

Após o conhecimento inicial do conjunto dos dados, é fundamental preparar estes para serem posteriormente modelados. Esta fase de limpeza e pré-processamento é o estádio que mais ocupa tempo e esforço num projeto de Ciência de Dados, podendo por vezes corresponder a 90% do tempo despendido. O objetivo é conseguir que os dados apresentem a qualidade necessária para análise e aplicação de técnicas de *data mining*. Esta fase, pode envolver tarefas como a eliminação, a deteção, a limpeza ou a imputação de dados incompletos, inconsistentes, redundantes, omissos e/ou *outliers.* Em muitas situações é necessário a integração de dados, a amostragem de dados, a conversão de dados, a transformação de atributos numéricos e a redução da dimensionalidade (Aragon, et al., 2022; Gama, et al., 2012; & Santos & Ramos, 2017).

Apesar da literatura distinguir as fases de análise exploratória dos dados e a limpeza e pré-processamento destes, na generalidade das situações práticas, estes processos decorrem de forma iterativa e não sequencial.

2.2. Segmentação dos clientes

A segmentação é um método de aprendizagem não supervisionada, isto é, o algoritmo aprende, não por rótulos pré-existentes, mas através da análise dos atributos, encontrando relações que permitem a criação de agrupamentos. O método procura as maiores similaridades dentro dos mesmos *clusters* e as maiores dissemelhanças entre clusters diferentes. Um dos algoritmos mais populares de segmentação é o *K-means*. Este algoritmo executa a construção de partições do conjunto dos dados em k classes, sendo este k um valor de entrada. O processo de identificação das classes passa por um conjunto de interações que inicia pelo posicionamento aleatório dos centroides de cada cluster, passando por reajustamentos dos pontos aos centroides mais próximos, finalizando apenas quando já não se verificarem alterações na formação dos clusters (Aragon, Guha, Kogan, Muller, & Neff, 2022; Gama, Carvalho, Faceli, Lorena, & Oliveira, 2012; & Santos & Ramos, 2017; Tan, et.al., 2018 & Witten & Frank, 2005).

Neste estudo, aplicamos o algoritmo de segmentação *k-means* para tentar compreender que atributos definem melhor estes clientes: demográficos ou relativos ao seu comportamento como consumidores. Antes da aplicação do algoritmo de segmentação *k-means* é fundamental compreender o número de clusters que se irá utilizar. Duas das metodologias mais comumente utilizadas para estimar o número de k ótimo são: o método de cotovelo (*elbow*) e o método silhueta (*silhouette*). Ambos os métodos serão utilizados neste estudo.

2.3. Classificação dos *clusters*

A classificação é um método de aprendizagem supervisionada em que o modelo aprende tendo por base um rótulo previamente definido e um conjunto de atributos com esse rótulo, que servem para supervisionar a aprendizagem para realizar futuras previsões. Os algoritmos de indução de árvores de decisão, são muito utilizados como método de classificação em *data mining.* Este algoritmo, tal como indica o nome, desenvolve estruturas em forma de árvore que representa um conjunto de decisões. Este permite gerar regras de classificação dos dados, integrando nós, ramos e folhas. Nos nós encontramos os atributos a classificar, nos ramos são descritos os valores possíveis dos atributos e as folhas indicam as diversas classes em que cada registo pode ser classificado. Ao induzir uma árvore de decisão, podemos obter ramos que espelham os *outliers* ou o ruído dos dados utilizados na fase de treino. Por esse motivo, pode ser necessário proceder-se à "poda" destes ramos para evitar os sobre-ajustamento dos dados e melhorar o desempenho da classificação. A visualização gráfica das arvores de decisão resultantes da aplicação destes algoritmos são muito fáceis de interpretar pelos utilizadores (Aragon, et al., 2022; Gama, et al., 2012; & Santos & Ramos, 2017).

Após termos segmentados os clientes com o algoritmo *K-means*, aplicamos um algoritmo de árvore de decisão para compreender, de forma visual, quais os atributos (nós) que melhor classificam os clusters obtidos na fase de segmentação (folhas).

3. Implementação computacional

Para realizar este trabalho utilizamos como conjunto de dados o ficheiro "Marketing Campaign[1]" que inclui os dados de 2240 clientes de uma empresa que vende produtos alimentares, através de loja, por catálogo e através da web (tabela 1). Neste tipo de análise é vulgar a dicotomia entre atributos demográficos e atributos de consumo. Os atributos de 2 a 7 são dados demográficos (ou socioeconómicos) e os atributos 8 a 29 são atributos de consumo.

Tabela 1- Atributos do Conjunto de Dados

#	Nome	Tipo	Descrição
1	ID	num	Identificador único de registo
2	Year_Birth	num	Data de nascimento
3	Education	chr	Habilitações literárias
4	Marital_Status	chr	Estado Civil
5	Income	num	Salário anual em dólares $
6	Kidhome	num	Nº de crianças a viver no local de residência
7	Teenhome	num	Nº de adolescentes a viver no local de residência
8	Dt_Customer	chr	Data em que começou a ser cliente
9	Recency	num	Nº de dias desde a última compra
10	MntWines	num	valor gasto produtos vitivinícolas últimos 2 anos
11	MntFruits	num	valor gasto em fruta nos últimos 2 anos
12	MntMeatProducts	num	valor gasto em carne nos últimos 2 anos
13	MntFishProducts	num	valor gasto em peixe nos últimos 2 anos

[1] https://www.kaggle.com/datasets/rodsaldanha/arketing-campaign?select=marketing_campaign.xlsx

14	MntSweetProducts	num	valor gasto em doces nos últimos 2 anos
15	MntGoldProds	num	valor gasto produtos de ouro nos últimos 2 anos
16	DealsPurchases	num	Nº de compras feitas com desconto
17	NumWebPurchases	num	Nº de compras feitas através do site da empresa
18	NumCatalogPurchases	num	Nº de compras feitas usando o catálogo
19	NumStorePurchases	num	Nº de compras feitas diretamente nas lojas
20	NumWebVisitsMonth	num	Nº de visitas ao site da empresa no último mês
21	AcceptedCmp3	num	se o cliente aceitou a oferta na ª3 campanha
22	AcceptedCmp4	num	se o cliente aceitou a oferta na ª4 campanha
23	AcceptedCmp5	num	se o cliente aceitou a oferta na ª5 campanha
24	AcceptedCmp1	num	se o cliente aceitou a oferta na ª1 campanha
25	AcceptedCmp2	num	se o cliente aceitou a oferta na ª2 campanha
26	Z_CostContact	num	Sem explicação metadados (todas com valor 3)
27	Z_Revenue	num	Sem explicação metadados (todas com valor 11)
28	Response	num	se o cliente aceitou a oferta da última campanha
29	Complain	num	se o cliente reclamou nos últimos 2 anos

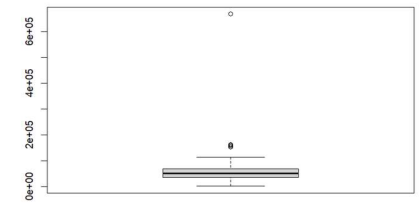
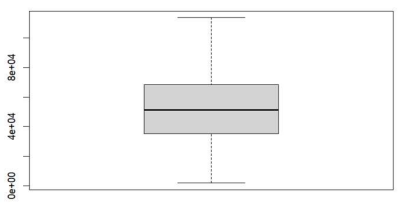
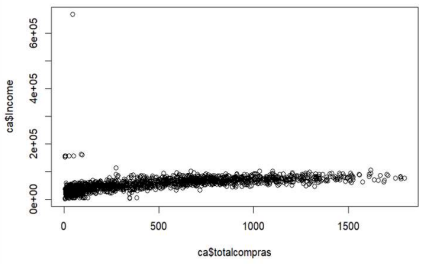
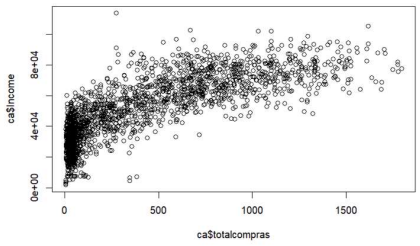
3.1. Engenharia dos dados

Na fase de engenharia dos dados são realizados procedimentos realizados de limpeza e transformação (ver bloco de código 1):
- Eliminação dos atributos *Z_CostContact* , *Z_Revenue* que para além de não serem explicados nos metadados, são iguais em todas as observações;
- Transformação do atributo *Dt_Customer*, passando de uma data em que a pessoa passou a ser cliente da empresa, para o número de dias que passou desde essa data;
- Transformação do atributo *Marital_Status* diminuindo o número de fatores. Conversão dos fatores: "*Absurd*", "*Alone*", "*Single*" e "*YOLO*" em apenas um fator denominado "*Single*". Conversão dos fatores "*Married*" e "*Together*" em apenas um fator denominado de "*Together*";
- Conversão do atributo *Education* em fator;
- Transformação do atributo *Year_Birth* num novo denominado *Year,* que se refere à idade do cliente, subtraído o ano atual ao ano de nascimento. Verificou-se que existiam 3 *outliers* de clientes que teriam mais de 120 anos e, por isso, apagou-se essas 3 observações.
- Eliminação dos 24 valores omissos do conjunto de dados, que estavam todos no atributo *Income*;
- Ainda relativamente ao atributo *Income*, verificou-se, através de um gráfico "*boxplot*" a existência de alguns *outliers* superiores. Um deles consideramos logo que seria um erro de registo, já que o valor do salário anual deste cliente era 6 vezes superior ao segundo cliente e sendo o mesmo dígito repetido várias vezes (666666). Para verificar se os outros *outliers* superiores eram também erros de registo, criou-se um novo atributo "totalcompras" resultante da soma do valor de compras de todos os produtos ("MntWines", "MntFruits", "MntFishProducts", "MntSweetProducts" e "MntGoldProds"). Ao criar um gráfico de dispersão que compara este novo atributo com o salário, verificou-se que todos os salários superiores a 150000 deveriam ser erros de registo, já que o valor das compras de todos era muito baixo, não correspondendo com a proporcionalidade verificada nos outros clientes. Assim, foram eliminados todos os registos que os clientes indicavam ter salário superior a 150000 que deveria ser um erro de registo (tabela 2).

```
#Apagar atributos:
ca$Z_CostContact<-NULL
ca$Z_Revenue<-NULL

#Converter em dias Dt_Customer
date_1 = as.Date("2022-11-01")
ca$Dt_Customer<-as.Date(ca$Dt_Customer)
ca$Dt_Customer<- as.numeric(difftime(date_1,ca$Dt_Customer, units =
"days"))


#reformular fatores estado cívil
ca$Marital_Status<-as.factor(ca$Marital_Status)
levels(ca$Marital_Status)
## "Absurd","Alone","Divorced","Married","Single","Together",
## "Widow","YOLO"

levels(ca$Marital_Status)<-
c('Single','Single','Divorced','Together','Single','Together','Widow
','Single')

#Converter em factor os chr:
ca$Education<-as.factor(ca$Education)
# apagamos valores omissos
ca<-na.omit(ca)

# Convertemos o ano de nascimento na idade e apagamos valores
anormais:
ca$Year_Birth<-(2022-ca$Year_Birth)
colnames(ca)[2] <- "Year"
ca <- subset(ca,ca$Year < 100 )

# deteção e correção outliers no salário:
boxplot(ca$Income)

#criamos um novo atributo que somamos as compras totais:
ca$totalcompras<-
ca$MntWines+ca$MntFruits+ca$MntFishProducts+ca$MntSweetProducts+ca$M
ntGoldProds
plot(ca$Income~ca$totalcompras)
summary(ca$Income)

#verificamos que existe um conjunto de outliers(>150000) que devem
resultar de erros de inserção:
ca=ca[(ca$Income < 150000),]
boxplot(ca$Income)
plot(ca$Income~ca$totalcompras)
summary(ca$Income)
```

Bloco de Código 1- Procedimentos de limpeza e pré-processamento

Tabela 2- Análise do atributo salário (*Income*) pré e pós eliminação de um *outlier*

	Pré-eliminação do *outlier*	Pós-eliminação do *outlier*
boxplot(ca$Income)		
plot(ca$Income~ca $totalcompras)		
summary(ca$Income)	Min. 1st Qu. Median Mean 3rd Qu. Max. 1730 35246 51373 52237 68487 666666	Min. 1st Qu. Median Mean 3rd Qu. Max. 1730 35196 51287 51622 68281 113734

3.2. Análise Exploratória

Para melhor compreensão do nosso conjunto de dados, no âmbito do nosso problema, podemos observar os gráficos 1 e 2 relativos aos atributos demográficos destes clientes. Este conjunto de clientes tem idade compreendida entre os 26 e os 82 anos, com uma média de 53 anos, sendo que em média o salário anual é de 51.959$, a grande maioria está em situação de "junto" (casado, união de facto ou situações semelhantes) e tem como habilitações "*graduation*", em que 58% não tem crianças, 51% não têm adolescentes a viver na mesma casa, 40% tem uma criança a viver na mesma casa e 46% um adolescente a viver na mesma casa.

No que respeita aos atributos relativos aos comportamentos destas pessoas como clientes desta empresa, tal como podemos observar nos gráficos 3 e 4, verifica-se uma cauda longa (assimetria positiva) relativamente ao valor gasto em compras de diferentes produtos: fruta (média 26$, mediana 8$), carne (média 167$, mediana 68$), peixe (média 38$, mediana 12$), doces (média 27$, mediana 8$), vinho (média 376$, mediana 174$) e ouro (média 44$, mediana 25$). Verifica-se também que é nas lojas físicas que se apuram maior número médio de vendas por cliente (6), seguindo-se as vendas online (4) e as vendas por catálogo (2), sendo que o número médio por cliente de visitas do site no último mês foi de 5 (mediana 6), sendo que em média a última compra foi realizada à 48 dias. O número de clientes que responderam a ofertas nas várias campanhas é muito residual, a rondar os 160 clientes, com exceção da 2ªcampanha que foram apenas 30 clientes e na última que conseguiram atingir 333 clientes. Foram 20 os clientes que apresentaram alguma reclamação nos últimos 2 anos (0.9%).

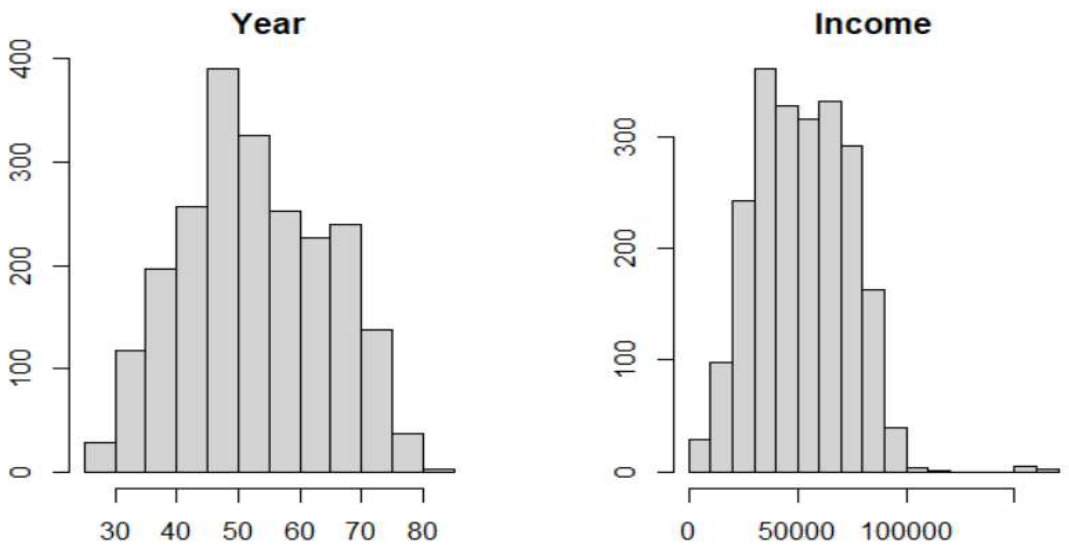


Gráfico 1- Histogramas dos atributos demográficos contínuos

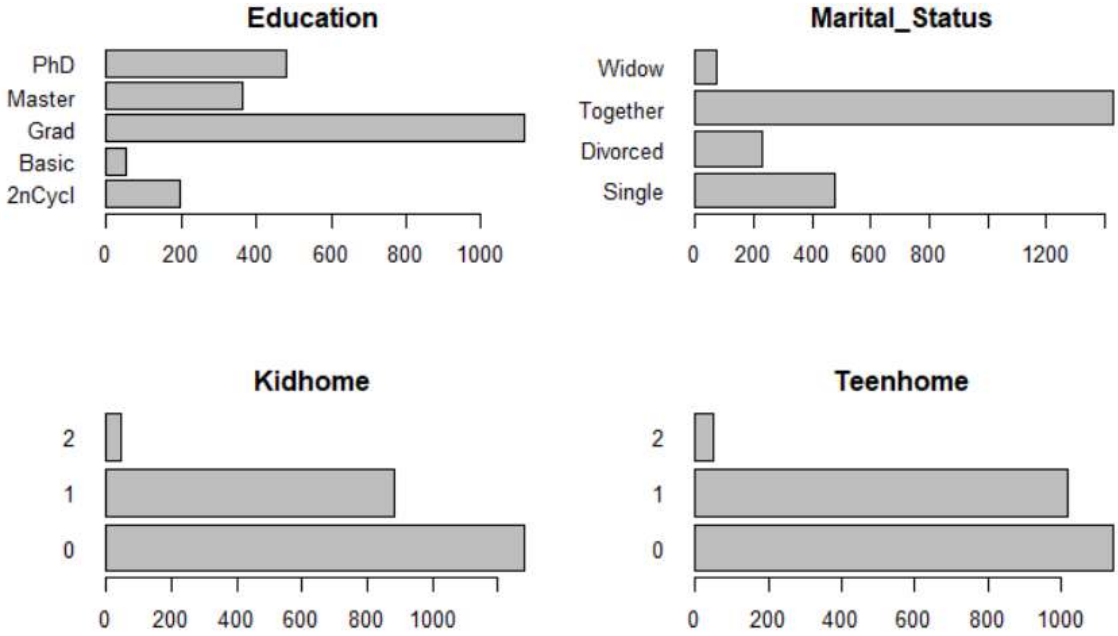


Gráfico 2 - Gráficos de Barras atributos demográficos categóricos

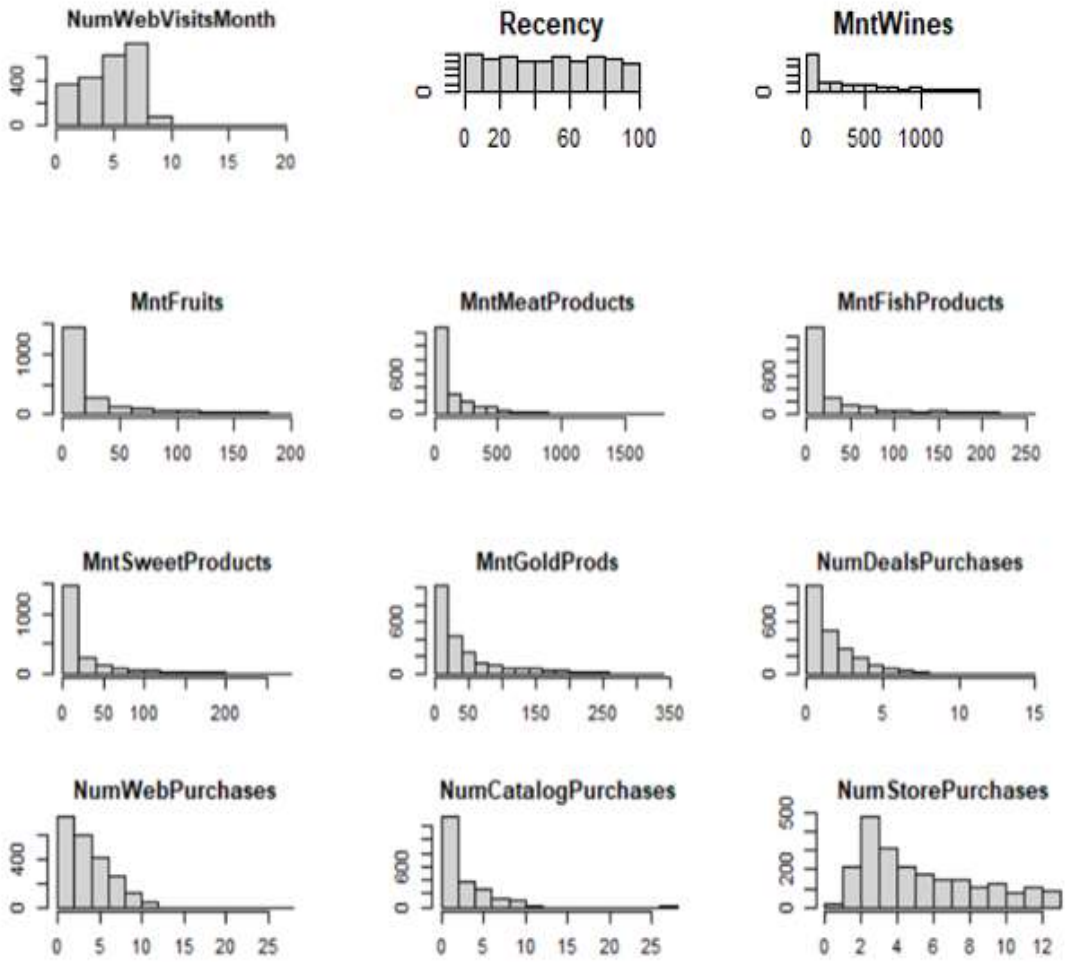


Gráfico 3-Histogramas dos atributos comportamentais contínuos

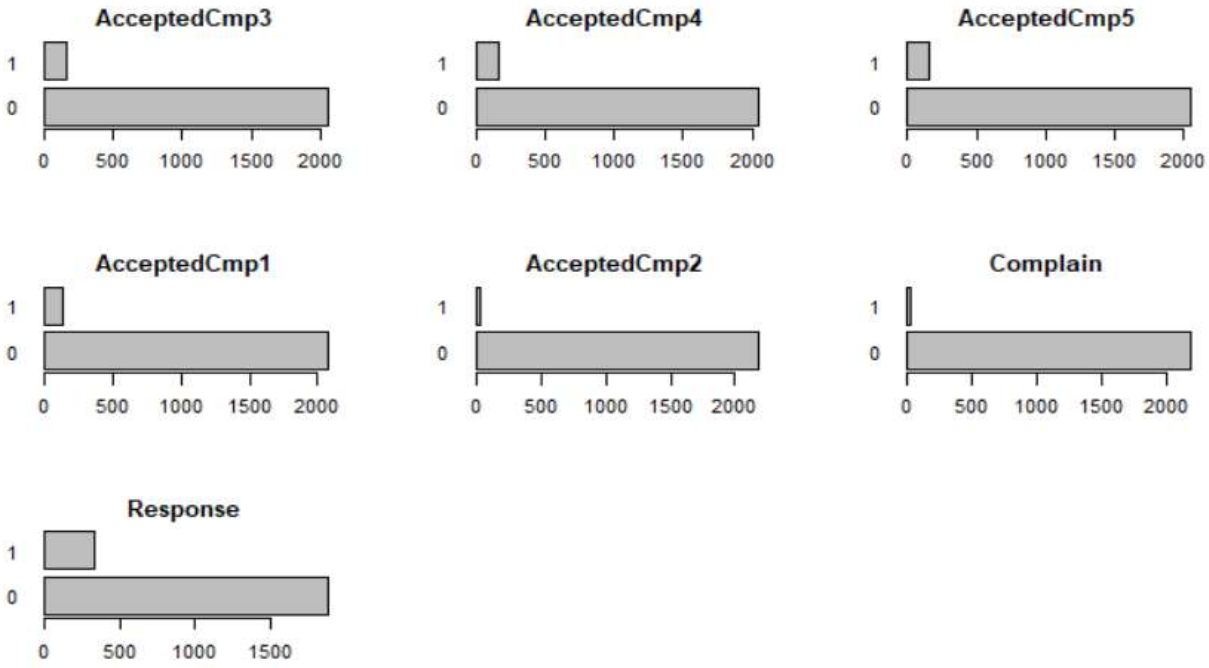


Gráfico 4 -Gráficos de Barras atributos comportamentais categóricos

3.3. Escolha do número de Clusters

Nos gráficos 5 e 6 podemos observar, respetivamente, a aplicação do método cotovelo e silhueta (ver bloco de código 2). No caso do método cotovelo, apesar de assinalado o k 4 verificamos que essa forma de cotovelo pode ser observada entre os k 2, 3 ou 5, já no caso do método de silhueta este aponta claramente para o 2. Tendo em conta estas possibilidades, iremos aplicar o algoritmo *k-means* com 2, 3, 4 e 5 clusters e verificar qual o que melhor responde aos nossos objetivos.

```
library(factoextra)
# metodo do cotovelo
fviz_nbclust(ca[2:27], kmeans, method = "wss") +
  geom_vline(xintercept = 4, linetype = 2) +
  labs(subtitle = "Elbow method")
# metodo da silhueta
fviz_nbclust(ca[2:27], kmeans, method = "silhouette") +
  labs(subtitle = "Silhouette method")
"Elbow method")
```

Bloco de Código 2 - Aplicação dos métodos cotovelo e silhueta

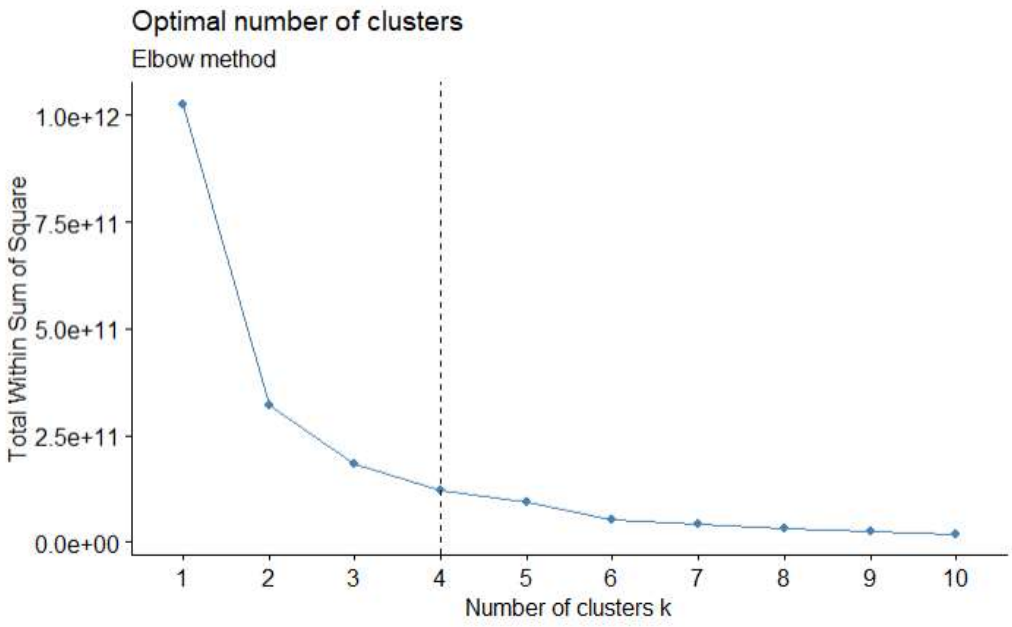


Gráfico 5- Aplicação dos métodos Cotovelo

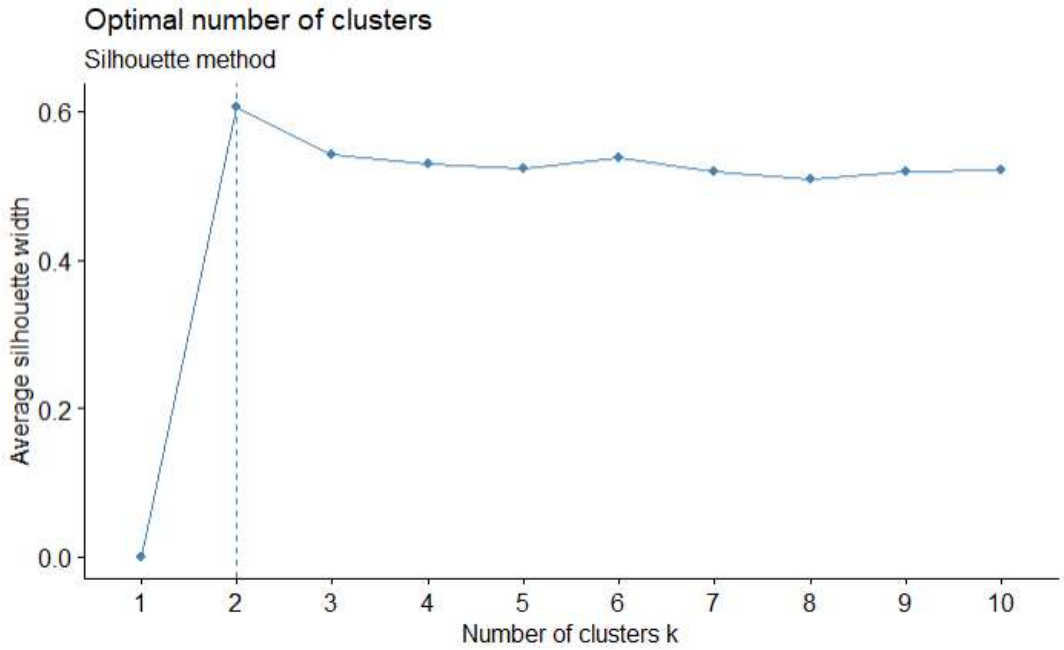


Gráfico 6 -Aplicação dos métodos Silhueta

3.4. Aplicação do algoritmo K-means

No bloco de código 3 podemos observar o script em R e os resultados relativos à aplicação do algoritmo *K-means* a 2, 3, 4 e 5 clusters. Verificamos que quanto maior o número de clusters utilizado, maior é a percentagem de variância total dos dados explicados por esta segmentação (*betwenn SS/ total SS*). No caso de se aplicar o algoritmo a 5 clusters, este resultado chega aos 94%.

```
#cluster 2:
km_2 <- kmeans(ca[2:27], 2, nstart = 20)
km_2
## K-means clustering with 2 clusters of sizes 1083, 1122
##
## Cluster means:
##       Year Education Marital_Status   Income   Kidhome Teenhome Dt_Customer
## 1 55.29086  3.499538       2.533703 69511.09 0.1542013 0.544783    3402.050
## 2 50.97683  3.289661       2.462567 34354.91 0.7201426 0.469697    3399.433
##    Recency  MntWines MntFruits MntMeatProducts MntFishProducts MntSweetProducts
## 1 49.06187 544.39058 46.478301       299.48569        65.39335        47.741459
## 2 48.95811  76.21925  7.025847        35.80214        11.08021         7.231729
##   MntGoldProds NumDealsPurchases NumWebPurchases NumCatalogPurchases
## 1     66.28532          2.223453        5.528163           4.5235457
## 2     22.60160          2.409982        2.722816           0.8324421
##   NumStorePurchases NumWebVisitsMonth AcceptedCmp3 AcceptedCmp4 AcceptedCmp5
## 1          8.108957          3.965836   0.06555863   0.12742382 0.1477377655
## 2          3.617647          6.660428   0.08199643   0.02317291 0.0008912656
##   AcceptedCmp1 AcceptedCmp2    Complain Response
## 1  0.127423823  0.024007387 0.006463527 0.189289
## 2  0.003565062  0.003565062 0.011586453 0.114082
## Within cluster sum of squares by cluster:
## [1] 129162659860 135617674193
##  (between_SS / total_SS =  72.0 %)
#################################################################################
```

```r
#cluster 3:
km_3 <- kmeans(ca[2:27], 3, nstart = 20)
km_3
## K-means clustering with 3 clusters of sizes 724, 715, 766
##
## Cluster means:
##       Year Education Marital_Status   Income    Kidhome  Teenhome Dt_Customer
## 1 48.79282  3.165746       2.446133 28077.80 0.81215470 0.3080110    3399.544
## 2 54.62937  3.467133       2.511888 75516.22 0.09230769 0.3776224    3391.547
## 3 55.73107  3.537859       2.532637 51572.19 0.41906005 0.8146214    3410.389
##    Recency  MntWines MntFruits MntMeatProducts MntFishProducts MntSweetProducts
## 1 48.41436  29.86188  5.872928        25.30525        9.063536         6.022099
## 2 49.00839 623.22797 56.630769       386.06853       81.426573        59.416783
## 3 49.57180 271.36423 17.592689        91.58355       24.113577        16.938642
##   MntGoldProds NumDealsPurchases NumWebPurchases NumCatalogPurchases
## 1     17.50276          2.146409        2.142265           0.5290055
## 2     71.04755          1.661538        5.535664           5.3272727
## 3     43.96214          3.093995        4.612272           2.1422977
##   NumStorePurchases NumWebVisitsMonth AcceptedCmp3 AcceptedCmp4 AcceptedCmp5
## 1          3.067680          6.911602   0.08425414  0.004143646  0.000000000
## 2          8.495105          3.306294   0.07132867  0.135664336  0.220979021
## 3          5.934726          5.744125   0.06657963  0.083550914  0.003916449
##   AcceptedCmp1 AcceptedCmp2    Complain  Response
## 1  0.001381215   0.00000000 0.015193370 0.1160221
## 2  0.179020979   0.02657343 0.006993007 0.2363636
## 3  0.016971279   0.01436031 0.005221932 0.1044386
## Within cluster sum of squares by cluster:
## [1] 50289421127 49077906162 36944318444
##  (between_SS / total_SS =  85.6 %)
#################################################################################################

#cluster 4:
km_4 <- kmeans(ca[2:27], 4, nstart = 20)
km_4
## K-means clustering with 4 clusters of sizes 656, 483, 454, 612
## Cluster means:
##       Year Education Marital_Status   Income    Kidhome  Teenhome Dt_Customer
## 1 53.29421  3.527439       2.471037 41717.60 0.68597561 0.6463415    3392.056
## 2 54.22981  3.399586       2.486542 79659.11 0.07453416 0.2650104    3383.994
## 3 47.45374  2.929515       2.440529 23259.45 0.78193833 0.2070485    3409.300
## 4 56.17320  3.586601       2.576797 61151.68 0.21895425 0.7696078    3416.837
##    Recency  MntWines MntFruits MntMeatProducts MntFishProducts MntSweetProducts
## 1 49.14787 112.29878  7.993902        45.59299       13.278963         8.253049
## 2 49.75983 650.13458 63.372671       445.59213       91.200828        68.204969
## 3 48.51101  17.10793  5.526432        20.79956        8.011013         5.742291
## 4 48.63725 456.93301 32.446078       179.64052       43.880719        30.807190
##   MntGoldProds NumDealsPurchases NumWebPurchases NumCatalogPurchases
## 1     26.76677          2.670732        3.230183           1.0792683
## 2     74.03313          1.335404        5.331263           5.7826087
## 3     16.26432          2.030837        1.947137           0.4537445
## 4     59.55065          2.929739        5.660131           3.4738562
##   NumStorePurchases NumWebVisitsMonth AcceptedCmp3 AcceptedCmp4 AcceptedCmp5
## 1          4.071646          6.371951   0.08536585  0.033536585   0.00000000
## 2          8.372671          2.832298   0.08902692  0.149068323   0.30641822
## 3          2.887665          7.088106   0.07929515  0.002202643   0.00000000
## 4          7.867647          4.905229   0.04575163  0.112745098   0.02124183
##   AcceptedCmp1 AcceptedCmp2    Complain  Response
## 1  0.006097561  0.006097561 0.012195122 0.1219512
## 2  0.244306418  0.031055901 0.004140787 0.2898551
## 3  0.000000000  0.000000000 0.011013216 0.1035242
## 4  0.032679739  0.017973856 0.008169935 0.1078431
## Within cluster sum of squares by cluster:
## [1] 19681524873 22557732077 20892936323 17903385335
##  (between_SS / total_SS =  91.4 %)
#################################################################################################
```

```
#cluster 5:
km_5 <- kmeans(ca[2:27], 4, nstart = 20)
km_5
## K-means clustering with 5 clusters of sizes 499, 481, 362, 342, 521
##
## Cluster means:
##       Year Education Marital_Status   Income    Kidhome  Teenhome Dt_Customer
## 1 55.49499  3.505010       2.531062 66553.23 0.12825651 0.6232465    3411.044
## 2 56.53015  3.571726       2.553015 51410.43 0.41995842 0.8565489    3399.638
## 3 54.51934  3.433702       2.477901 82167.55 0.07458564 0.2292818    3382.279
## 4 47.22515  2.801170       2.456140 20832.66 0.76023392 0.1725146    3413.830
## 5 50.49136  3.479846       2.454894 36504.44 0.80998081 0.4836852    3396.031
##    Recency  MntWines MntFruits MntMeatProducts MntFishProducts MntSweetProducts
## 1 48.91984 534.44088 45.809619       260.37876       62.767535        44.951904
## 2 49.36590 269.39917 13.203742        80.92100       20.503119        14.534304
## 3 49.61050 670.22652 64.875691       472.23757       93.698895        68.914365
## 4 49.99415  12.58187  6.008772        21.35965        8.263158         6.456140
## 5 47.70058  61.23033  6.658349        33.40883       10.220729         6.220729
##   MntGoldProds NumDealsPurchases NumWebPurchases NumCatalogPurchases
## 1     68.41483          2.376754        5.715431           4.2304609
## 2     41.82952          3.245322        4.671518           2.1268191
## 3     71.81215          1.207182        5.339779           6.0027624
## 4     17.64327          2.096491        1.956140           0.4912281
## 5     20.83877          2.324376        2.573896           0.6871401
##   NumStorePurchases NumWebVisitsMonth AcceptedCmp3 AcceptedCmp4 AcceptedCmp5
## 1          8.372745          4.184369   0.04408818   0.10821643  0.048096192
## 2          5.889813          5.835759   0.07276507   0.09355509  0.004158004
## 3          8.430939          2.698895   0.09116022   0.16298343  0.372928177
## 4          2.821637          7.122807   0.06725146   0.00000000  0.000000000
## 5          3.479846          6.641075   0.09596929   0.01151631  0.000000000
##   AcceptedCmp1 AcceptedCmp2    Complain  Response
## 1  0.068136273   0.01202405 0.006012024 0.1242485
## 2  0.012474012   0.02079002 0.006237006 0.1018711
## 3  0.279005525   0.03867403 0.005524862 0.3314917
## 4  0.000000000   0.00000000 0.011695906 0.1111111
## 5  0.001919386   0.00000000 0.015355086 0.1228407
## Within cluster sum of squares by cluster:
## [1]  9154415536  9267124787 13288866316 12613016905  9222764974
##  (between_SS / total_SS =  94.3 %)
```

Bloco de Código 3 –Aplicação do algoritmo *k-means* (com 2, 3, 4 e 5 K) e resultados

3.5. Aplicação do algoritmo C5.0

Para compreendermos quais os atributos que melhor classificam a segmentação realizada com o algoritmo *K-means,* aplicamos um algoritmo de indução de árvores de decisão (C5.0) a todos os atributos utilizados na segmentação. Primeiro dividimos o nosso conjunto de dados em dois: um conjunto para treino e outro para teste. Seguidamente aplicamos esse algoritmo 5 vezes, respetivamente para classificar o conjunto de dados em função da segmentação com 2, 3 ,4 e 5 clusters (ver procedimento no bloco de código 4). As árvores de decisão obtidas em cada um destes procedimentos podem ser observadas nos gráficos 7, 8, 9 e 10.

```
library(C50)
#treino e teste:
set.seed(160)
s<-sample(1:2240,1680)
train_data<-ca[s,]
test_data<-ca[-s,]
#K2
tree_mod_tk2 <- C5.0(x = train_data[, c(2:27)], y = train_data$k2)
plot(tree_mod_tk3)
```

```
#K3
tree_mod_tk2 <- C5.0(x = train_data[, c(2:27)], y = train_data$k3)
plot(tree_mod_tk2)
#K4
tree_mod_tk2 <- C5.0(x = train_data[, c(2:27)], y = train_data$k4)
plot (tree_mod_tk4)
#K5
tree_mod_tk5 <- C5.0(x = train_data[, c(2:27)], y = train_data$k5)
plot (tree_mod_tk5)
```

Bloco de Código 4- Aplicação do algoritmo C5.0 para classificar K2, K3, K4 e K5

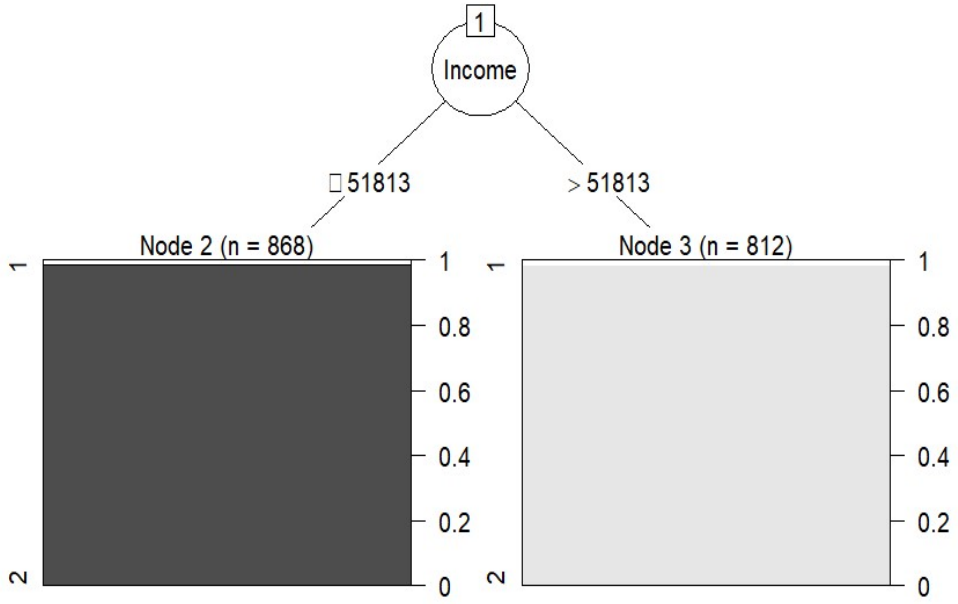


Gráfico 7- Árvore de Decisão da segmentação com 2 clusters

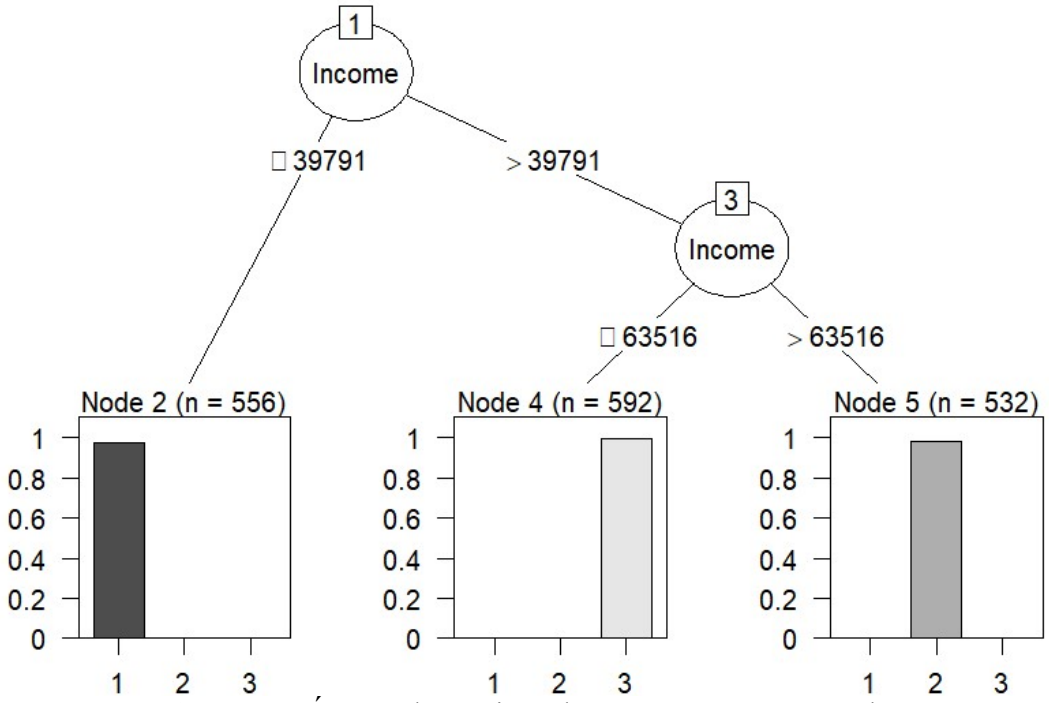


Gráfico 8 -Árvore de Decisão da segmentação com 3 clusters

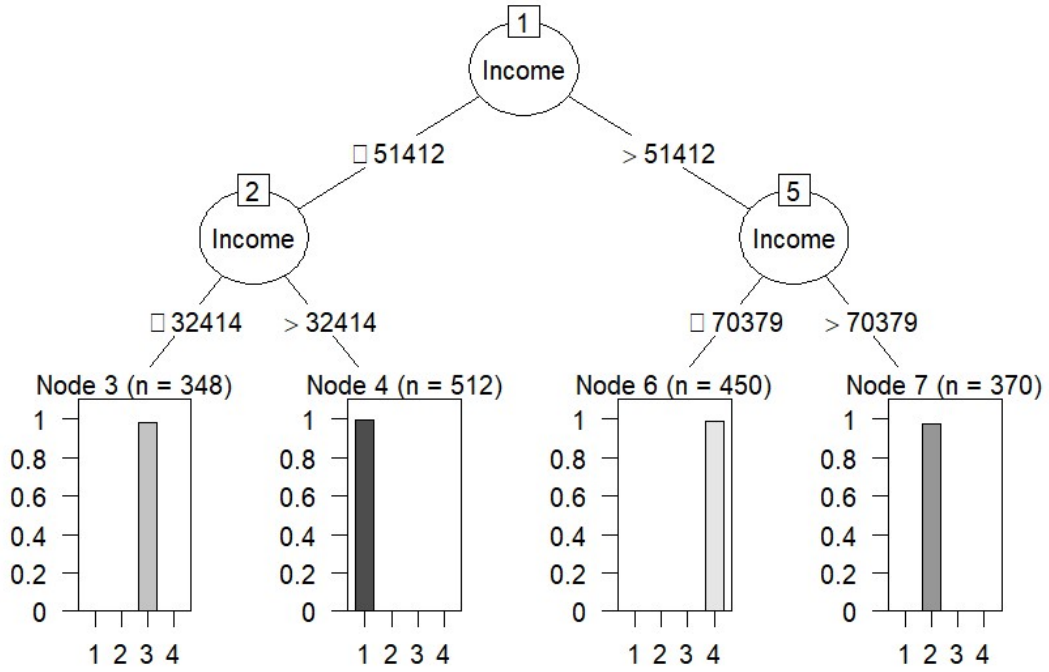


Gráfico 9 -Árvore de Decisão da segmentação com 4 clusters

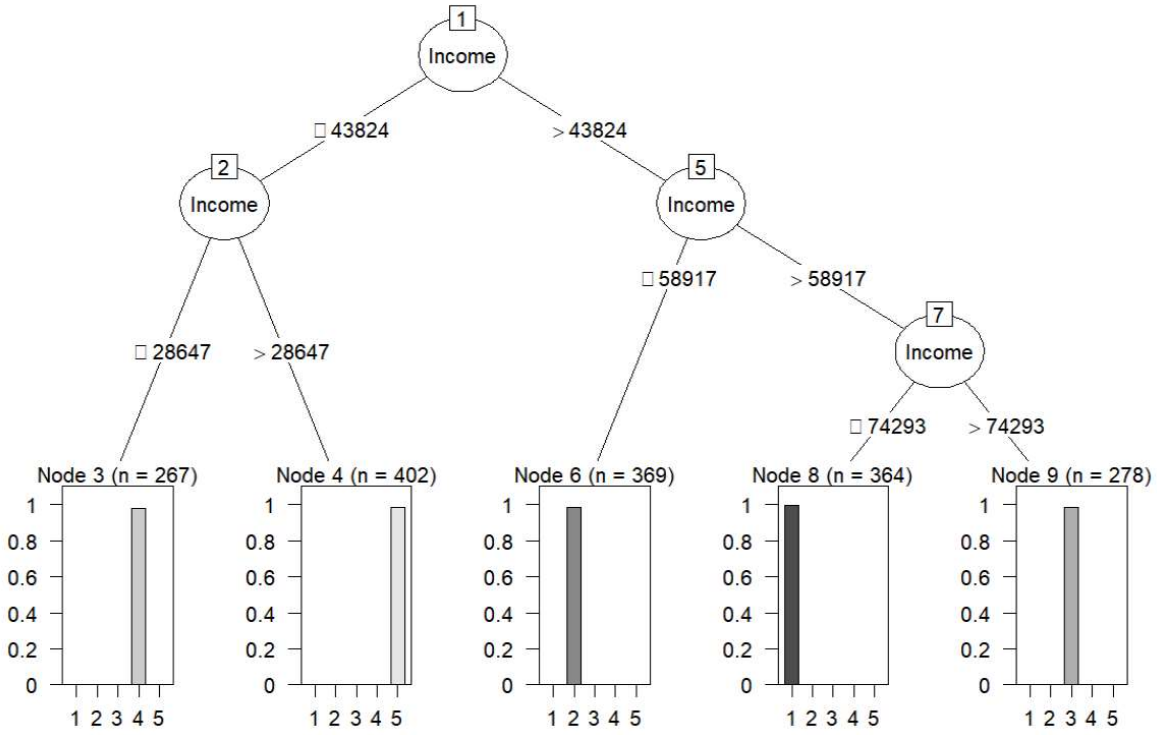


Gráfico 10 -Árvore de Decisão da segmentação com 5 clusters

4. Discussão dos Resultados

Verificamos assim, que independentemente do número de clusters, neste conjunto de dados, apenas um atributo demográfico consegue definir a segmentação dos clientes, mais especificamente o valor do seu salário. Tendo em consideração que na aplicação inicial do algoritmo *k-means* foi com 5 clusters que se obteve uma maior percentagem de

variância total dos dados explicados pela segmentação (*betwenn SS/ total SS*), consideramos que seria a versão de 5 clusters aquela que melhor poderia criar segmentos de clientes relativamente a este conjunto de dados. Assim, para melhor compreendermos a relação desta segmentação com o valor gasto por cada cliente em compras, voltamos a criar um gráfico de dispersão que compara estes dois valores, mas agora utilizamos as cores para verificar os clusters (ver criação do gráfico no bloco de código 5).

Como podemos observar no gráfico 8, verifica-se uma certa relação entre estes dois atributos. O cluster 4, representa os 267 clientes com salários mais baixos, que em média recebem cerca de 20.830$ por ano e gastaram 51$ na empresa nos últimos dois anos. Segue-se o cluster 5 em que estes 402 clientes têm salários médios de aproximadamente 36.500$ ano e gastaram nos últimos dois anos na empresa 101$. O cluster 2 representa 369 clientes que têm em média um salário de 51.410$ e gastaram 360$ nos últimos dois anos. O cluster 1 inclui 364, com salário médio anual de 66.550$ e com um valor médio de compras nos últimos dois anos de 756$. Por último, os clientes mais valiosos correspondem ao cluster 3, em que estes 278 clientes recebem em média 82.170$ salário anual e gastaram nos últimos dois anos em média 970$ na empresa.

```
#criar gráfico
SalCom5<- ggplot()+ geom_point(data = ca,
          aes(x=totalcompras, y=Income,color = k5))+
          labs(title = "Relação salário e valor das compras?",
          x = "Valor das Compras Totais", y = "Salário")
par(mfrow = c(2,2))
SalCom5
```

Bloco de Código 5- Edição de gráfico de dispersão com ggplot2

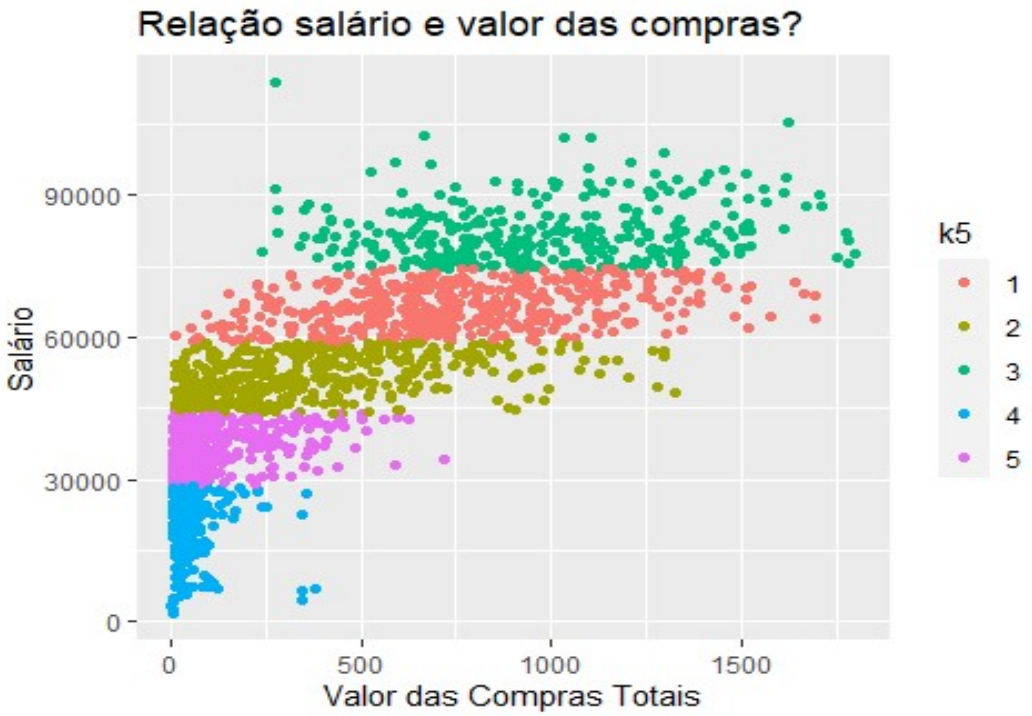


Gráfico 11 –Gráfico de dispersão relação entre o salário e o valor das compras com 5 clusters (k5)

5. Conclusões

A segmentação de clientes é uma técnica muito utilizada no marketing. Contudo, depois de encontrados os *clusters,* deve haver a preocupação de os caracterizar convenientemente com substantivos, adjetivos ou verbos. Na caracterização dos segmentos recorremos a algoritmos de classificação, com vista a encontrar os atributos mais relevantes.

A aplicação de um algoritmo de indução de árvores de decisão permitiu, neste estudo de caso, uma eficaz visualização das opções tomadas pelo algoritmo *k-means* ao segmentar os consumidores. A indução de árvores de decisão é uma das técnicas de classificação mais utilizadas, que permite visualizar os atributos (nós) que melhor espelham a classificação obtida (folhas). No caso de se aplicar este tipo de algoritmos a grupos previamente segmentados, podemos observar nas folhas os vários clusters, nos nós os atributos e nos ramos os valores dos atributos que classificam esses clusters.

Quando trabalhamos com comportamentos humanos, é fundamental ter esta noção e não tomar decisões decorrentes de escolhas provenientes da aplicação de algoritmos "black box" (Aragon, et al., 2022). Outra das vantagens da utilização de algoritmos de indução de árvores de decisão após a segmentação de pessoas, é que a simplicidade da sua visualização gráfica possibilita a análise por uma variedade de possíveis *stakeholders* de diferentes áreas, permitindo assim, uma melhor comunicação de procedimentos e resultados.

Neste estudo de caso tentamos verificar que tipo de segmentação melhor consegue definir os consumidores de um conjunto de dados. Para tal, apresentamos os passos seguidos através da utilização da linguagem R. Após a exploração, limpeza e pré-processamento dos dados, aplicamos duas técnicas de *data mining*: segmentação e classificação. A segmentação tentou reunir os clientes em grupos, para tal utilizou-se o algoritmo *k-means*. Este procedimento de segmentação foi realizado 4 vezes, respetivamente, para 2, 3, 4 e 5 clusters, opção decorrente da pré-aplicação dos métodos cotovelo e silhueta. Para compreender quais os atributos que melhor definem esta segmentação inicial, foi aplicado a estes resultados uma técnica de classificação através do algoritmo de indução de árvores de decisão *C5.0.* As árvores de decisão resultantes demostram que independentemente do número de clusters, o valor do salário é o único atributo que classifica este grupo de consumidores neste conjunto de dados. Importa referir, que neste caso específico não foi necessário realizar um procedimento de "poda" dos ramos, mas em muitas situações a efetivação desse procedimento é fundamental para não comprometer a simplicidade de visualização da informação gráfica.

Neste tipo de análise é vulgar a dicotomia entre atributos demográficos e atributos de consumo. Neste conjunto de dados ficou claro que o atributo salário ('income') é o mais relevante em todos os agrupamentos. Neste caso o atributo demográfico, salário, supera todos os atributos de consumo.

REFERÊNCIAS

Aragon, C., Guha, S., Kogan, M., Muller, M., & Neff, G. (2022). *Human-Centered Data Science: An Introduction*. MIT Press

Cavique, L. (2003). Micro-Segmentação de Clientes com Base em Dados de Consumo: Modelo RM-Similis. Revista Portuguesa e Brasileira de Gestão, volume 2, nº3, 72-77.

Cavique, L. (2007). Network Algorithm to Discover Sequential Patterns, in Progress in Artificial Intelligence, J.Neves, M.Santos and J.Machado (Eds.), EPIA 2007, LNAI 4874, Springer-Verlag Berlin Heidelberg, 406-414.

Gama, J., Carvalho, A., Faceli, K., Lorena, A., & Oliveira, M. (2012). Extração de conhecimento de dados: data mining.

Bendle, N., Farris, P., Pfeifer, P., & Reibstein, D. (2017). Grandes métricas do Marketing: os principais indicadores que todo o gestor deve conhecer. *Coimbra: Conjuntura Atual Editora*

Berry, M., & Linoff, G. (2011). Data Mining Techniques: for Marketing, Sales and Customer Relationship Management 3rd Edition, *John Wiley and Sons*.

Bose, I., & Chen, X. (2009). Quantitative models for direct marketing: A review from systems perspective. *European Journal of Operational Research*, *195*(1), 1-16.

Celeste, P., & Moniz, L. B. (2019). Marketing Performance: 80 métricas de marketing e vendas. Lisboa: Clube do Autor S. A.

Rodrigues, F. & Oliveira, M. (2013). Segmentação, Posicionamento e Targeting. Em Rodrigues, F., Moreira, M. & Vitorino, L. (Ed.). *Comportamento do Consumidor: quando a neurociência, a psicologia, a economia e o marketing se encontram.* Viseu: Psicosoma.

Santos, M. Y., & Ramos, I. (2017). Business Intelligence - Da Informação ao Conhecimento. *FCA–-Livros de Informática, Lisboa*.

Tan, P.; Steinbach, M.; Karpatne, A. & Kumar, V. (2018) Introduction to Data Mining (Second Edition). Pearson

Witten, I. & Frank, F. (2005) Data Mining: Practical Machine Learning Tools and Techniques. Elsevier Inc

Nuno Lopes é Doutorado em Educação (2019) pela Universidade Aberta (UAb) e aluno de doutoramento em Ciência e Tecnologias Web pela UAb e Universidade de Trás-os-Montes e Alto Douro. Mestre em Comunicação Educacional Multimédia pela UAb (2012) e Licenciado em Psicología pela Universidade do Minho (2002). É Gestor da Comunicação e do Impacto Social numa ONG. Tem como principal área de interesse a Ciência de Dados aplicada ao Comportamento Social e Humano.

Luís Cavique, Professor Auxiliar da Secção de Informática, no Departamento de Ciências e Tecnologia (DCeT) da Universidade Aberta e Investigador no LaSIGE, FCUL. Licenciou-se em Engenharia de Informática pela Universidade Nova de Lisboa (FCT-UNL), é Mestre em Investigação Operacional e Engenharia de Sistemas pela Universidade Técnica de Lisboa (IST-UTL) e obteve o grau de Doutor em Engenharia de Sistemas da Universidade Técnica de Lisboa (IST-UTL) em 2002. Tem como área de investigação a Ciências dos Dados, recorrendo à interseção das Ciências da Computação com a Engenharia de Sistemas.

(esta página par está propositadamente em branco)