Relatório
De R

Exploração do Tempo Consumido por um Utilizador nas Redes Sociais: Comportamentos e Tendências

Ana Catarina Pacheco Pinheiro

Data Analyst

09/04/2024

O seguinte trabalho vai explorar o dataset "Average Time Spent By A User On Social Media", o qual permite explorar a utilização por parte dos utilizadores de várias redes sociais, tendo em conta o seu gênero, idade, localização, rendimentos, área profissional, interesses, entre outras variáveis. O objetivo é identificar padrões e tendências no comportamento dos utilizadores.

```
social.o                  1000 obs. of 12 variables
    $ age          : int  56 46 32 60 25 38 56 36 40 28 ...
    $ gender       : chr  "male" "female" "male" "non-binary" ...
    $ time_spent   : int  3 2 8 5 1 3 8 4 7 2 ...
    $ platform     : chr  "Instagram" "Facebook" "Instagram" "Instagram" ...
    $ interests    : chr  "Sports" "Travel" "Sports" "Travel" ...
    $ location     : chr  "United Kingdom" "United Kingdom" "Australia" "United Kingdom" ...
    $ demographics: chr  "Urban" "Urban" "Sub_Urban" "Urban" ...
    $ profession   : chr  "Software Engineer" "Student" "Marketer Manager" "Student" ...
    $ income       : int  19774 10564 13258 12500 14566 19179 16881 13636 16030 10223 ...
    $ indebt       : chr  "True" "True" "False" "False" ...
    $ isHomeOwner  : chr  "False" "True" "False" "True" ...
    $ Owns_Car     : chr  "False" "True" "False" "False" ...
```

Tendo em conta estes dados, seria interessante fazer as seguintes questões de BA para podermos analisar os dados presentes:

- Será que existe uma relação entre a plataforma que determinado utilizador prefere e as suas características, ou seja, podem algumas características como idade, rendimento, etc; ter influência na escolha dos utilizadores entre Facebook, Instagram e Youtube?
- Haverá alguma correlação entre os rendimentos de um utilizador e o tempo que o mesmo despende em redes sociais?

Estas questões são levantadas uma vez que seria importante perceber se determinadas características do utilizador podem ditar qual a probabilidade de um utilizador preferir utilizar determinada plataforma.

Relativamente a questões de BI, seria importante perceber:

- Qual é a plataforma de redes sociais mais utilizada entre os utilizadores, com base nas suas faixas etárias e em média quanto tempo as mesmas gastam nessas redes?
- Quais as redes sociais mais populares com base nas profissões e interesses dos utilizadores?

Estas questões poderão ser relevantes se numa perspetiva de estratégia de marketing, perceber de que forma poderíamos direcionar os conteúdos e o desenvolvimento de produtos, ajudando a entender melhor o comportamento dos utilizadores nas redes sociais e a adaptar as abordagens de acordo com as características etárias e interesses específicos.

```
> summary(social.f)
      age              gender      time_spent        platform       interests             location
 Min.   :18.00   female    :331   Min.   :1.000   Facebook :307   Lifestlye:341   Australia     :352
 1st Qu.:29.00   male      :337   1st Qu.:3.000   Instagram:363   Sports   :331   United Kingdom:329
 Median :42.00   non-binary:332   Median :5.000   YouTube  :330   Travel   :328   United States :319
 Mean   :40.99                    Mean   :5.029
 3rd Qu.:52.00                    3rd Qu.:7.000
 Max.   :64.00                    Max.   :9.000
    demographics             profession      income        indebt    isHomeOwner  Owns_Car
 Rural    :340   Marketer Manager :355   Min.   :10012   False:503   False:492   False:461
 Sub_Urban:335   Software Engineer:336   1st Qu.:12402   True :497   True :508   True :539
 Urban    :325   Student          :309   Median :14904
                                         Mean   :15015
                                         3rd Qu.:17674
                                         Max.   :19980
```

Questões BI

Plataformas - Faixa Etárias e Tempo Gasto

Primeiramente, averiguamos quais as plataformas existentes no nosso dataset, bem como se distribuem os utilizadores pelas mesmas. Olhando para o summary do nosso dataset, percebemos que a rede Instagram é a preferida (363), seguida pelo Youtube (330) e finalmente o Facebook(307),

Passemos a explorar então algumas das características dos utilizadores de acordo com as plataformas.

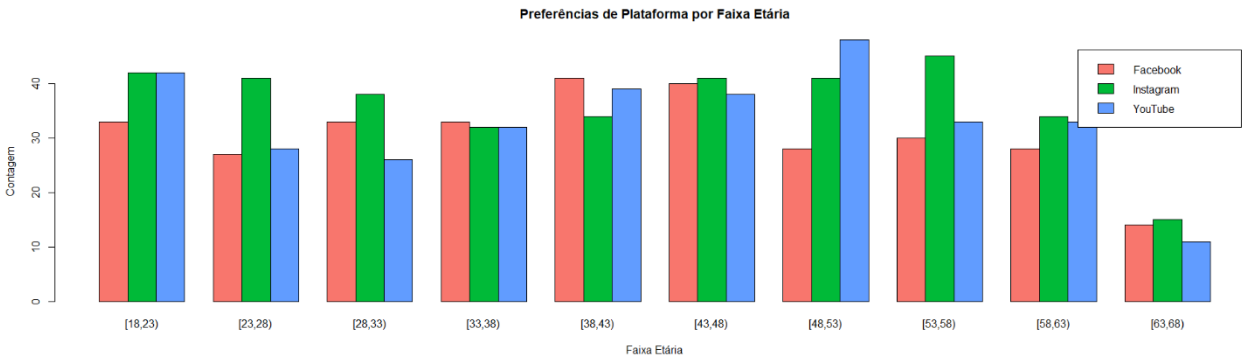
Sabendo que a idade mínima e máxima do dataset em questão, foi possível estabelecer intervalos de idades de 5, desde os 18 aos 64, e perceber como se distribuíam os utilizadores, por idade, pelas plataformas em análise.

intervalos_idade	Freq
[43,48)	119
[18,23)	117
[48,53)	117
[38,43)	114
[53,58)	108
[28,33)	97
[33,38)	97
[23,28)	96
[58,63)	95
[63,68)	40

Olhando para esta distribuição vemos que a faixa etária mais presente no nosso dataset é a faixa dos 43 aos 47, seguido da faixa dos 18 aos 22 e dos 48 aos 52.

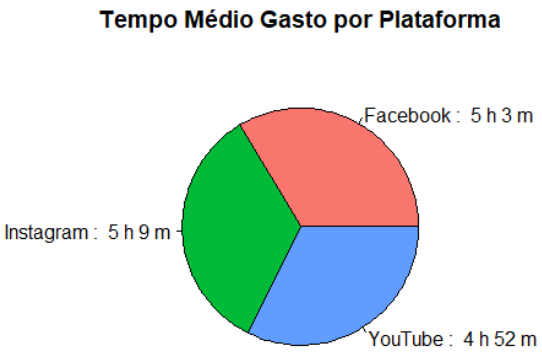
intervalos_idade	Facebook	Instagram	YouTube
[18,23)	33	42	42
[23,28)	27	41	28
[28,33)	33	38	26
[33,38)	33	32	32
[38,43)	41	34	39
[43,48)	40	41	38
[48,53)	28	41	48
[53,58)	30	45	33
[58,63)	28	34	33
[63,68)	14	15	11

Comparando agora a distribuição pelas diferentes plataformas, conseguimos ver que a plataforma do Facebook é a preferida dos utilizadores na faixa etária dos 33 aos 42 anos; o Instagram é a plataforma preferida nos utilizadores dos 18 aos 32, bem como na faixa dos 43 aos 47 e nas faixas etárias com idades superior a 53; relativamente ao Youtube, apenas é a preferência dominante na faixa etária dos 48 aos 52, sendo que na faixa dos 18 aos 22 anos também é a preferida, mas partilha o 1º lugar com o Instagram.



De seguida vamos visualizar a tabela com a média de tempo gasto nas plataformas sociais por intervalo de idade e tipo de plataforma, bem como o tempo médio gasto por plataforma através de um gráfico de pizza.

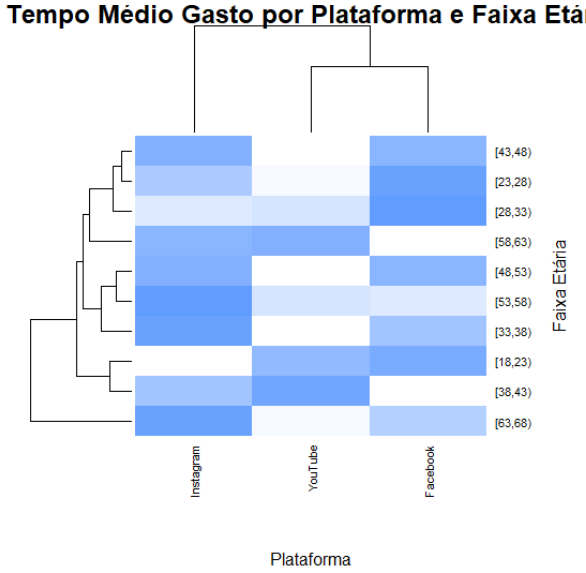
platform	intervalos_idade	time_spent	time_spent_convertido
Instagram	[63,68)	6.200000	6 h 12 m
Instagram	[33,38)	5.781250	5 h 47 m
Facebook	[18,23)	5.727273	5 h 44 m
YouTube	[38,43)	5.717949	5 h 43 m
YouTube	[18,23)	5.642857	5 h 39 m
Instagram	[38,43)	5.529412	5 h 32 m
Instagram	[48,53)	5.439024	5 h 26 m
Facebook	[48,53)	5.392857	5 h 24 m
Facebook	[28,33)	5.333333	5 h 20 m
Instagram	[53,58)	5.288889	5 h 17 m
Facebook	[38,43)	5.195122	5 h 12 m
Facebook	[33,38)	5.151515	5 h 9 m
Instagram	[18,23)	5.119048	5 h 7 m
YouTube	[53,58)	4.969697	4 h 58 m
YouTube	[58,63)	4.969697	4 h 58 m
YouTube	[48,53)	4.958333	4 h 57 m
Facebook	[53,58)	4.933333	4 h 56 m
Instagram	[58,63)	4.911765	4 h 55 m
Instagram	[43,48)	4.878049	4 h 53 m
Facebook	[43,48)	4.875000	4 h 52 m
Facebook	[23,28)	4.814815	4 h 49 m
Instagram	[23,28)	4.609756	4 h 37 m
YouTube	[28,33)	4.576923	4 h 35 m
Instagram	[28,33)	4.526316	4 h 32 m
YouTube	[43,48)	4.526316	4 h 32 m
Facebook	[63,68)	4.500000	4 h 30 m
YouTube	[23,28)	4.357143	4 h 21 m
YouTube	[33,38)	4.218750	4 h 13 m
Facebook	[58,63)	4.178571	4 h 11 m
YouTube	[63,68)	3.000000	3 h 0 m



Percebemos, pelo gráfico que a plataforma onde os utilizadores passam mais tempo é o Instagram, com uma média de 5h e 9 minutos, não ficando muito atrás as médias do Facebook (5h03) e por fim Youtube (4h52).

Olhando agora para as faixas etárias, concluímos que a faixa etária que passa mais tempo no Instagram é a faixa etária dos 63 aos 67, gastando uma média de 6 h 12 m nesta plataforma; no Facebook, a faixa etária que mais tempo gasta a nesta plataforma é a dos 18 aos 22, com uma média de 5 h 44 m; por fim, no Youtube é a faixa dos 38 aos 42 que mais tempo passa nesta plataforma com uma média de 5h 43 m,

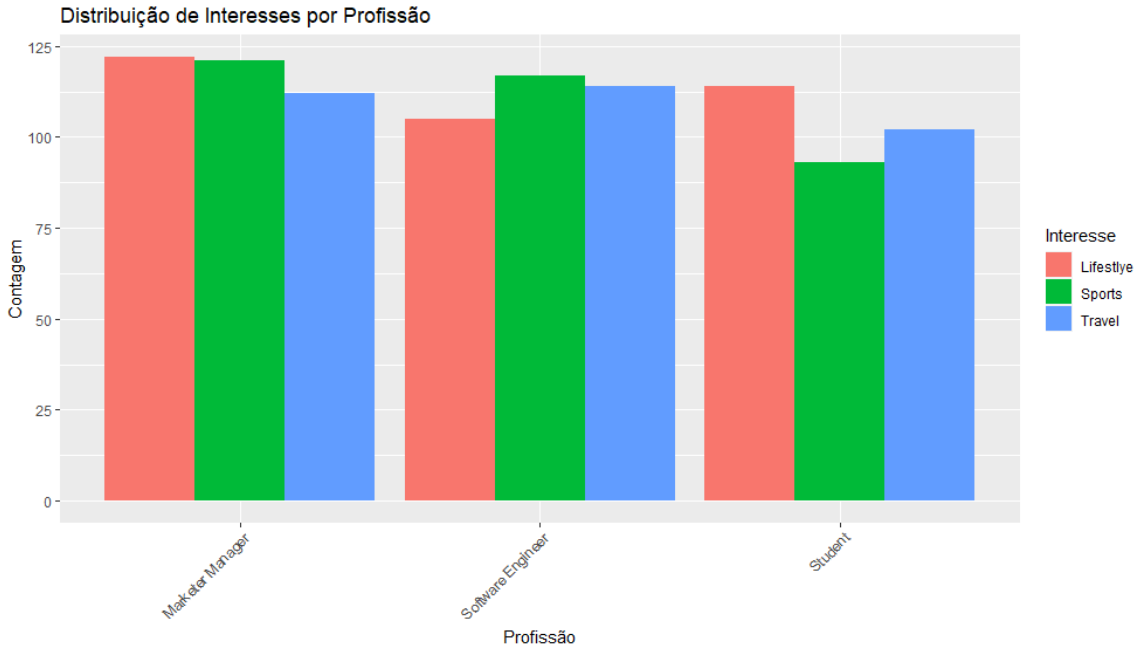
Podemos ver estes resultados de forma gráfica, olhando para o seguinte mapa de heatmap, em que as zonas em azul mais escuro representam um maior número de médio de horas gastas nas redes sociais, enquanto que quanto mais branco for, menos horas passa. Numa primeira análise, percebemos que o Youtbe é a plataforma onde menos horas os seus utilizadores gastam, contrapondo com o Instagram.



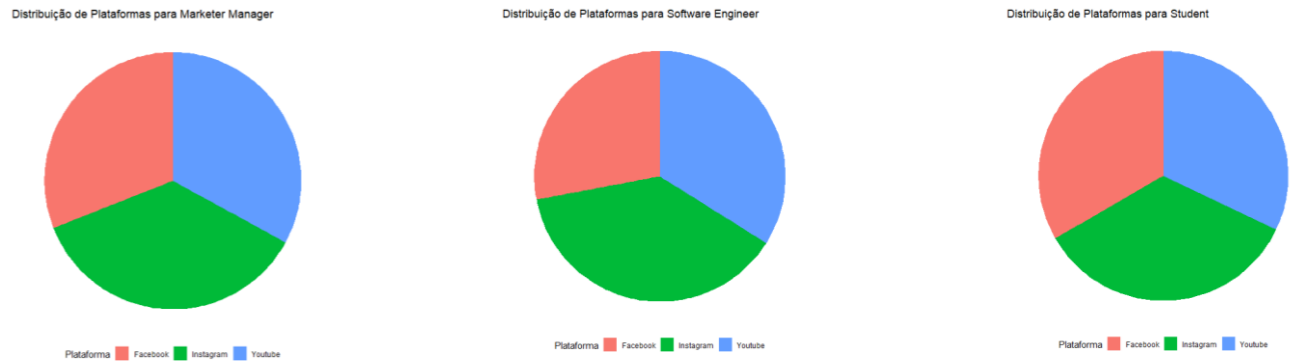
Plataformas - Interesses e Profissão

Inicialmente tínhamos feito o summary, onde podemos identificar três tipos de interesse, sendo o Lifestyle o interesse que era mais popular entre os utilizadores, com 341 a preferiram este interesse, seguido de Sports (331) e Travel por último (328). Relativamente à profissão, existem também três categorias: os utilizadores que são Marketer Manager são os mais presentes na dataset, com 355 utilizadores; seguido de Software Engineer com 336; e com menor representação os Student, com 309.

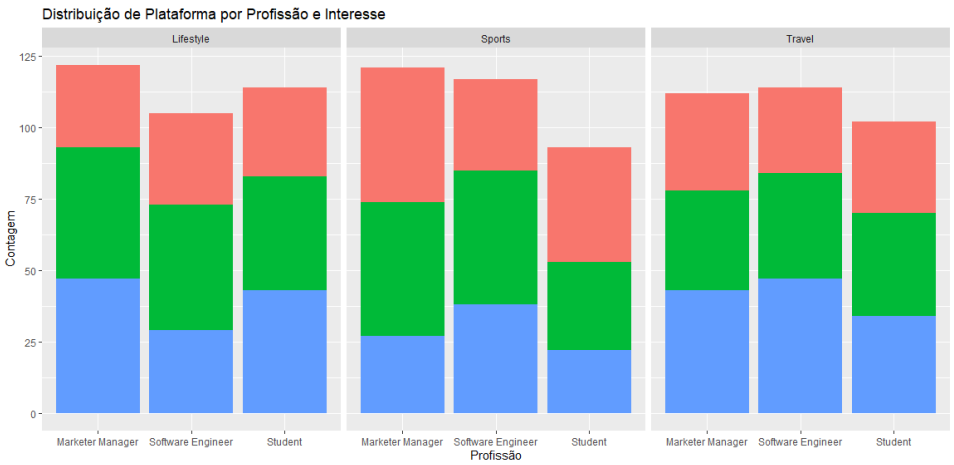
	Lifestlye	Sports	Travel
Marketer Manager	122	121	112
Software Engineer	105	117	114
Student	114	93	102



Partimos agora para uma análise dos resultados, para tal fiz uma crosstable e um historgrama. Olhando para ambas estas fontes podemos ver que nos Marketing Managers, observa-se que existe um maior interesse em lifestyle e sports, enquanto o interesse em travel é ligeiramente menor; quanto aos software engineers, as contagens para os três interesses são semelhantes, havendo uma leve preferência por sports; já para os estudantes, verifica-se um maior interesse em lifestyles, seguido por travel e, por último, sports.



Tentando agora ver como os utilizadores se distribuem pelas redes sociais, tendo em conta a sua profissão, vimos pelos gráficos pizza que parece não haver uma preferência clara por uma plataforma específica com base na profissão, parecendo a distribuição semelhante entre as profissões, com variações ténues.




```
                    1  2  3
Marketer Manager   29 47 34
Software Engineer  32 32 30
Student            31 40 32

,  , = 2

                    1  2  3
Marketer Manager   46 47 35
Software Engineer  44 47 37
Student            40 31 36

,  , = 3

                    1  2  3
Marketer Manager   47 27 43
Software Engineer  29 38 47
Student            43 22 34
```

Juntando as duas variáveis, Interesses e Profissão, podemos ver mudanças ténues, mas mais uma vez não parece haver uma preferência clara por uma plataforma específica com base na profissão ou interesse. Podemos dizer, por exemplo, que o Youtube tem uma contagem mais baixa em comparação com o Facebook e Instagram, mas ainda é relevante em várias situações; o Instagram também é bastante popular, especialmente entre Marketer Managers e Software Engineers em todos os interesses; o Facebook parece ser bastante popular quando consideramos utilizadores com interesse em Sports.

Apesar do seguinte método se enquadrar numa abordagem BA, considerei fazer a seguinte análise para finalizar esta secção. Fiz o teste de Cochran-Mantel-Haenszel (CMH) para considerar se existe uma correlação entre as variáveis Profissão e Interesse e a escolha da plataforma. Com o resultado obtido, vemos que o p-value associado ao teste CMH é de 0.4885, isto indica que com base neste teste, não temos evidências estatisticamente significativas para afirmar que existe uma relação entre a escolha da plataforma social e as variáveis de profissão e interesse. O valor-p alto sugere que as variáveis são independentes entre si após o controle da terceira variável.

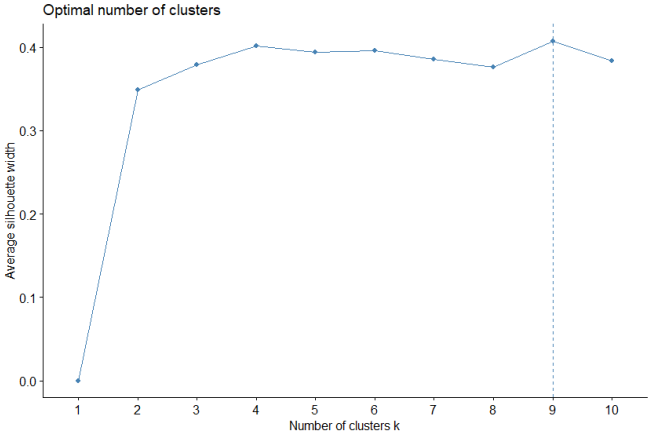
```
        Cochran-Mantel-Haenszel test

data:  array_counts
Cochran-Mantel-Haenszel M^2 = 3.4307, df = 4, p-value = 0.4885
```

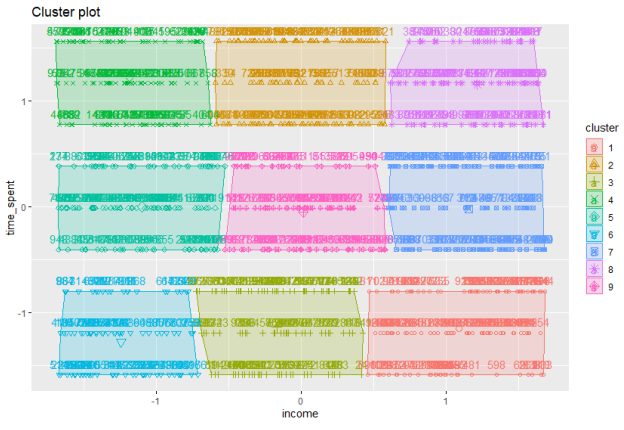
Questões BA

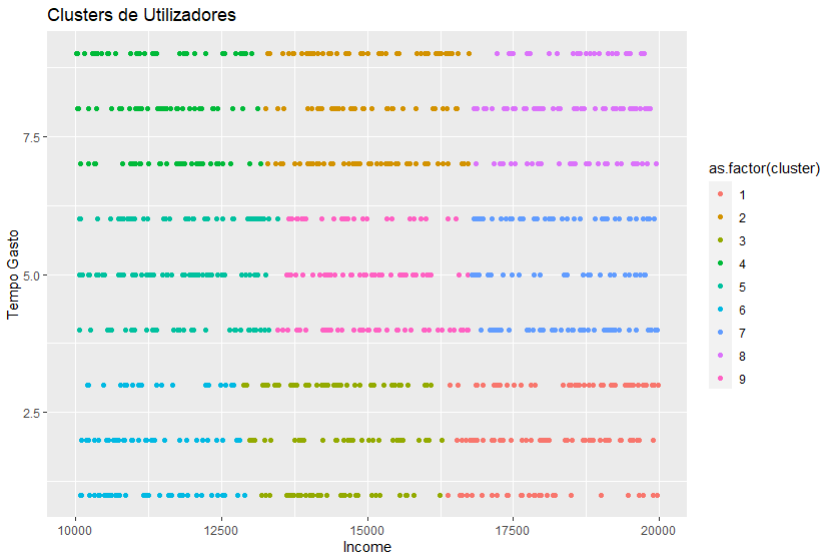
Rendimento & Tempo Gasto

Passamos agora às questões de BA, querendo explorar se haverá alguma correlação entre os rendimentos de um utilizador e o tempo que o mesmo despende em redes sociais.

Para tal comecei por fazer uma análise por cluster, gerando um gráfico de cotovelo para encontrar o número de clusters ideal, o qual foi 9.







Analisando o gráfico de clusters, podemos ver que os utilizadores com rendimentos abaixo dos, aproximadamente 13000, pertencem aos clusters 4, 5 e 6; utilizadores com rendimentos acima dos 16000 e que passem menos tempo nas redes sociais, são cluster 1.

Apesar serem todos muito semelhantes, podemos ver a comparação dos utilizadores de income mais baixo, nomeadamente cluster 4 e 6 , não são utilizadores que prefiram o Facebook; já nos utilizadores dos cluster que coincidem com income mais alto (7,8, e 9) não parece haver uma grande diferença na preferência de plataformas, apenas no cluster 7 em que há um ligeiro desfavorecer em relação ao instagram.

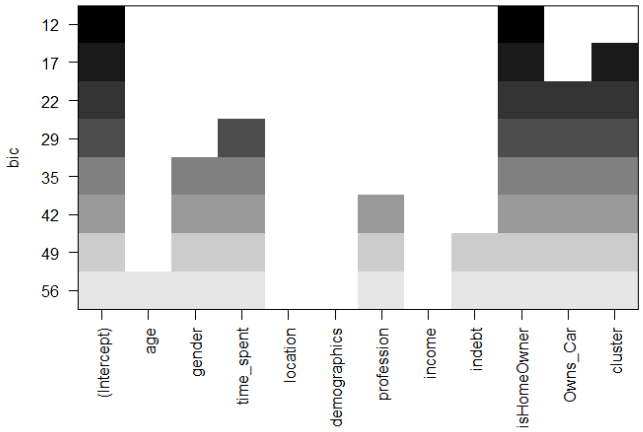
De seguida calculei a correlação entre o rendimento e o tempo gasto nas redes sociais, segundo o método de Pearson.

```
        Pearson's product-moment correlation

data:  social.n$income and social.n$time_spent
t = 0.15029, df = 998, p-value = 0.8806
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.05725272  0.06673065
sample estimates:
        cor
0.004757252
```

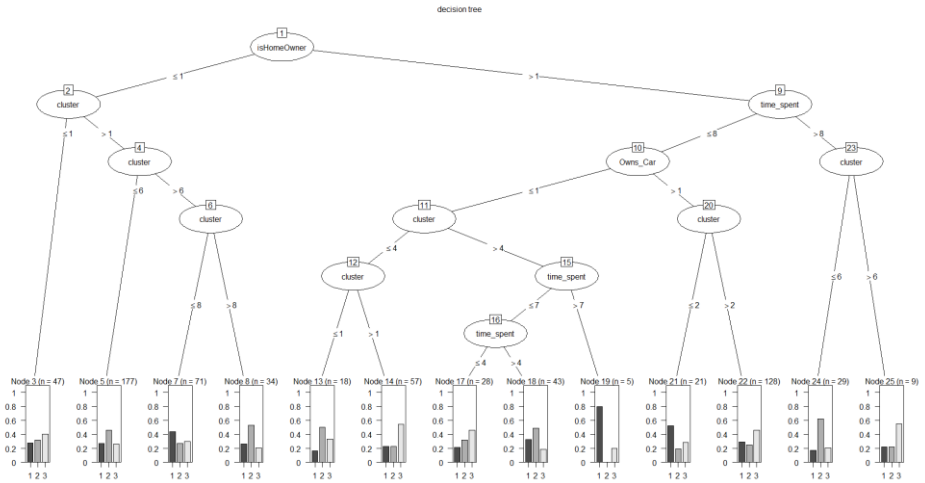
O valor-p de 0.8806 indica que a probabilidade de observar uma correlação entre o rendimento e o tempo gasto em redes sociais é de 88.06%. Como este valor-p é maior do que o nível de significância comum de 0.05, não temos evidências estatísticas suficientes para rejeitar a hipótese nula. Portanto, não podemos concluir que há uma correlação significativa entre o rendimento e o tempo gasto em redes sociais com base nos dados analisados. Em termos de correlação, a correlação estimada é de 0.0048, o que também indica que a correlação entre as duas variáveis é muito próxima de zero.

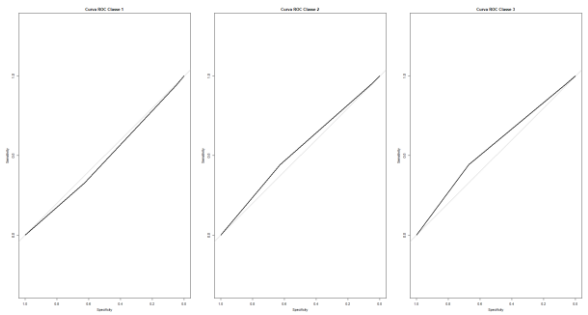
Previsões com base nas características

Para tentar explorar a possível relação entre as características do utilizador e a escolha da plataforma, bem como determinar se é viável prever qual plataforma será escolhida, realizei uma análise dos melhores subsets de variáveis. O objetivo era confirmar se a idade é uma variável que tem uma influência significativa nesse processo.



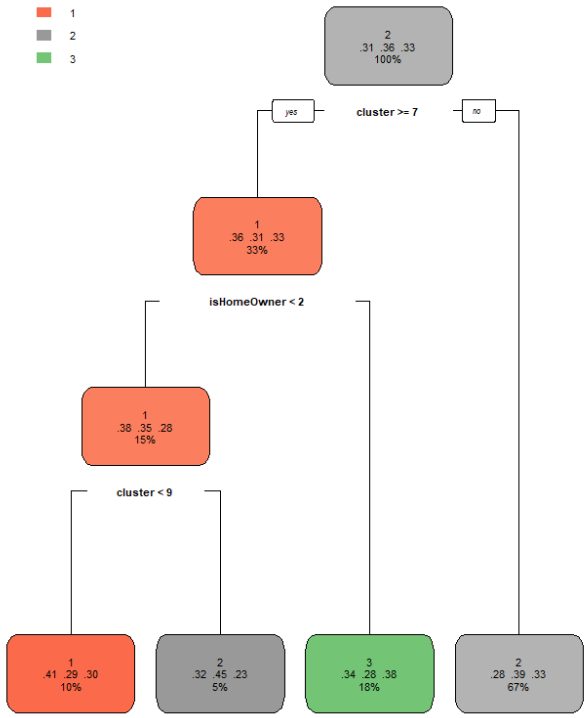
Ao contrário das expectativas, a variável idade não se mostrou a mais adequada para construir modelos preditivos, pelo que optei por seguir um modelo em que as variáveis time_spent, isHomeOwner, Owns_Car e cluster são selecionadas.





O modelo de classificação multiclasse foi treinado e avaliado usando uma abordagem de Árvore de Decisão (C5.0), tendo depois avaliado o desempenho do modelo usando a curva ROC. A AUC média foi calculada para todas as classes combinadas como uma medida geral do desempenho do modelo - o valor da AUC média para todas as classes combinadas foi de aproximadamente 0.5228, o que indica um desempenho moderado do modelo na tarefa de classificação.

Quis experimentar uma segunda interpretação, para ter uma compreensão mais facilitada do modelo, pelo que optei por experimentar uma abordagem diferente de árvore de decisão utilizando o rpart.plot. Esta escolha foi motivada pela busca de uma representação visual mais simples de interpretação, o que poderia facilitar na leitura das relações entre as variáveis e as plataformas escolhidas pelos utilizadores.



Com esta abordagem, podemos interpretar o seguinte: se o utilizador não pertencer a um cluster superior ou igual a 7, utilizará a plataforma 2 (Instagram); caso contrário, se for maior ou igual a 7, e se não for HomeOwner (1 = False; 2 =True), utiliza a plataforma 3 (Youtube); caso seja HomeOwner, e seja do cluster 9, então utiliza a plataforma 2 (Instagram); e se for de um cluster abaixo de 9 utiliza a plataforma 1(Facebook).

Para finalizar, quis saber qual o percentagem de erro deste modelo, tento chegado ao resultado de um percentagem de erro de 60,36%, Isso significa que, ao fazer previsões sobre a plataforma com base nas variáveis time_spent, isHomeOwner, Owns_Car e cluster, o modelo está a errar cerca de 60.36% das vezes.

Conclusão

Fazendo um balanço geral deste projeto, constatamos que, embora tenhamos identificado algumas tendências, não encontramos uma relação direta entre as características dos utilizadores e as suas preferências pelas plataformas de redes sociais. Contudo, por exemplo, observamos que o Instagram é popular entre os utilizadores mais jovens, enquanto o Facebook é mais utilizado por pessoas na faixa etária dos 33 aos 42 anos. Além disso, não encontramos evidências significativas de correlação entre rendimento e tempo gasto nas redes sociais. Ao explorar a possibilidade de prever a plataforma escolhida com base nas características do utilizador, desenvolvemos um modelo com desempenho moderado, indicando que outros fatores não considerados poderiam influenciar essa decisão. É importante ressaltar que o dataset utilizado foi gerado aleatoriamente por meio da biblioteca 'NumPy' do Python, conforme indicado pelo seu proprietário, com o propósito de treinar um modelo AI. Essa informação pode justificar o facto de não chegarmos a conclusões sólidas, quando estavam a ser explorados os modelos preditivos.