

Metodologia

# SELETOR DE CASAS

---

Organização e apoio



15 DE NOVEMBRO

---

WOB20

Mariano et al.



# Introdução

Visando aperfeiçoar as interações entre participantes do WOB20 (1º Workshop Online de Bioinformática), estabeleceu-se um método para divisão dos inscritos em grupos. A estratégia de agrupamento visou:

- 1) Construir quatro grupos com um número de indivíduos aproximado
- 2) Preferencialmente, agrupar indivíduos com interesses em comum
- 3) Apresentar um baixo custo computacional para calcular os membros dos grupos

Para criação de grupos mais atrativos foram nomeados de acordo com as ilustrações estabelecidas pelo departamento de divulgação do Comitê de Organização do Curso de Verão da UFMG:



Os grupos são:

- Cobra
- Jaguatirica
- Jacaré
- Capivara

## Metodologia de agrupamento

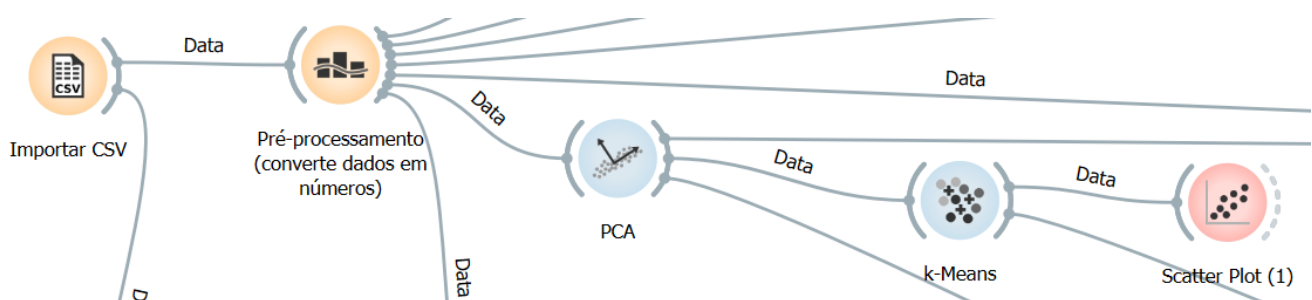
Para realização dos agrupamentos foi necessária uma série de passos desde o uso de algoritmos de agrupamento, redução dimensional e a cálculo de distância. Por fim, decidimos utilizar árvores de decisão para estabelecer métricas simples para agrupamento.

Inicialmente, uma amostra com 2436 inscrições foi coletada. Para essa análise, o único campo utilizado correspondeu aos temas de interesse dos inscritos. Cada inscrito poderia escolher até cinco temas de interesse em uma lista com 25 opções:

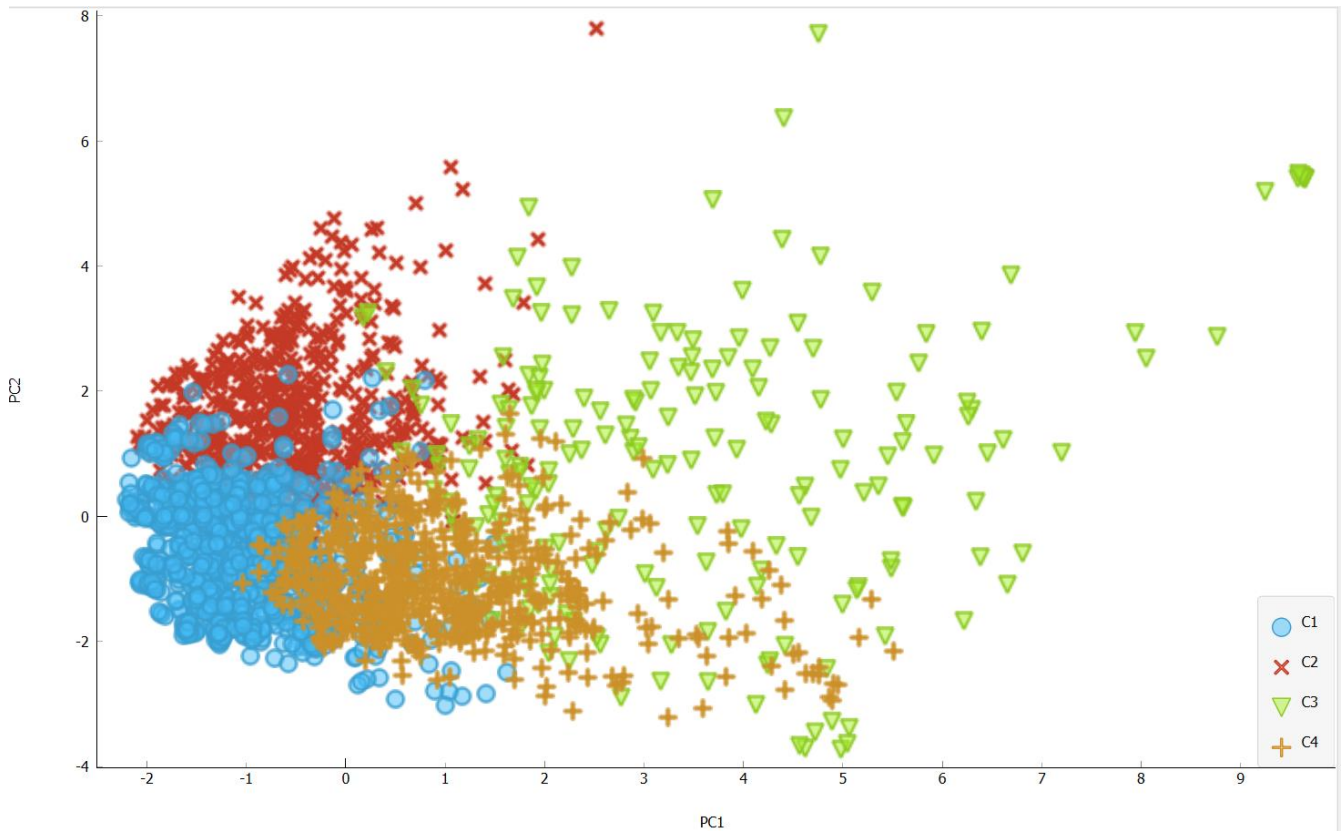
```
temas = ['R','Python','Perl','C++','Java','MatLab','Linux','WSL','Divulgação da ciência','Genômica/Metagenômica','Transcriptômica','Proteômica','Sequenciamento','Georreferenciamento','Empreendedorismo','Diagnóstico','Ecologia','Evolução','Vírus','Bactérias','Fungos','Insetos','Animais','Plantas','Aprendizado de Máquina']
```

A partir das respostas dos inscritos, foi construída usando Python uma matriz binária de 2436 linhas por 25 colunas, onde 1 representa que o usuário tinha interesse no tema e 0 que ele não tinha interesse. A matriz foi exportada no formato CSV.

Os dados foram então analisados usando a ferramenta Orange Data Mining. Primeiro realizou-se um pré-processamento, seguido por uma análise de componentes principais (PCA). O agrupamento foi realizado usando o algoritmo k-means, de acordo com a estrutura abaixo:

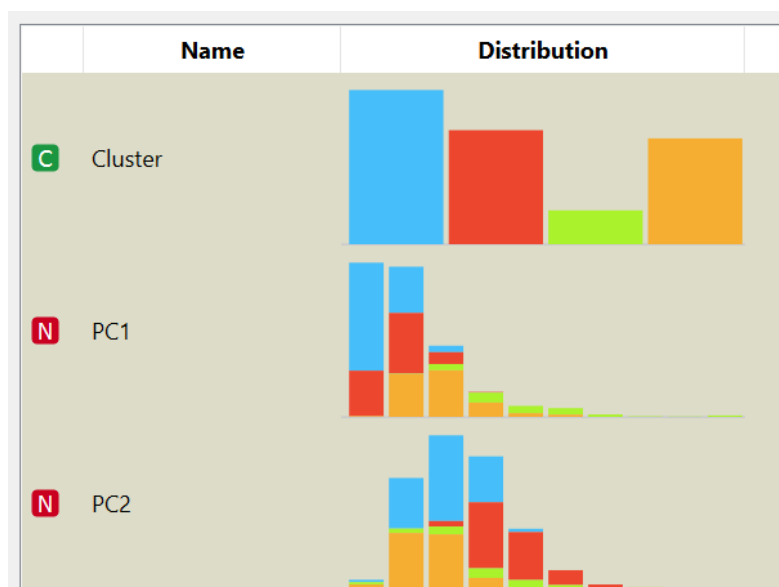


Abaixo está o resultado do agrupamento:



Cor e forma foram definidas de acordo com o grupo estabelecido por k-means. Observe que, como estamos analisando um agrupamento usando 25 dimensões, não é possível visualizar todo o resultado. Por isso, aplicou-se a análise de componentes principais para que fosse possível ver os grupos em um gráfico bidimensional.

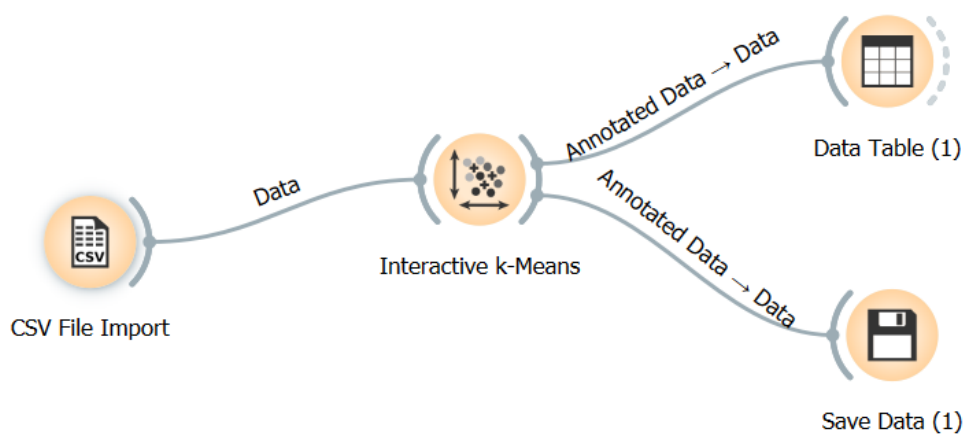
Entretanto, o resultado não foi satisfatório, uma vez que a quantidade de indivíduos em cada grupo foi desproporcional.



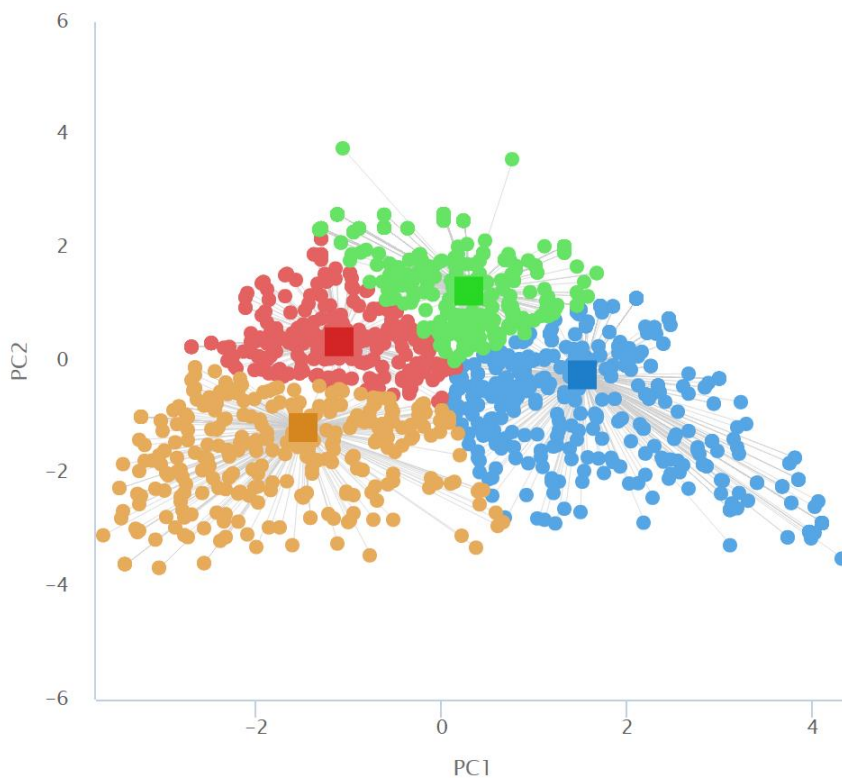
Veja acima que a barra azul indica que esse grupo possui muito mais membros do que o grupo representado pela barra verde.

Por isso, apesar de não ser a melhor estratégia, decidimos aplicar o agrupamento apenas nos dados de k-means e manipular manualmente os centroides usando a ferramenta de k-means interativo:

Para essa análise usamos apenas 1496 amostras.



Aqui está o resultado do agrupamento:



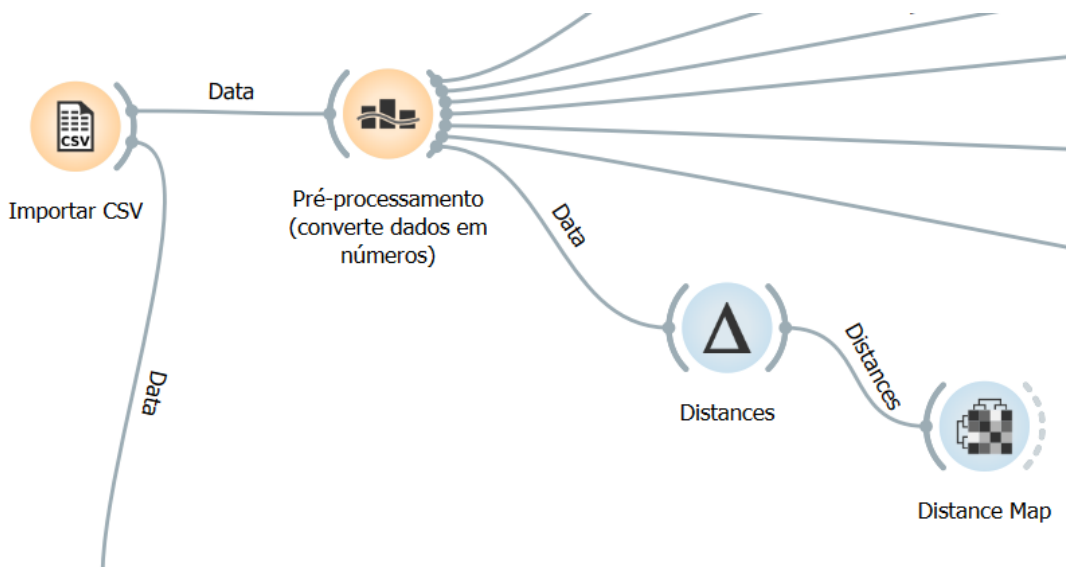
Perceba que agora os grupos apresentam uma quantidade de elementos mais parecida.

Agora possuíamos uma forma de agrupar nossos indivíduos em grupos com base em suas características mantendo uma quantidade parecida de indivíduos no mesmo grupo. Entretanto, devido a limitações técnicas não foi possível estabelecer um modelo para implementação em ambiente de produção e assim classificar novas entradas.

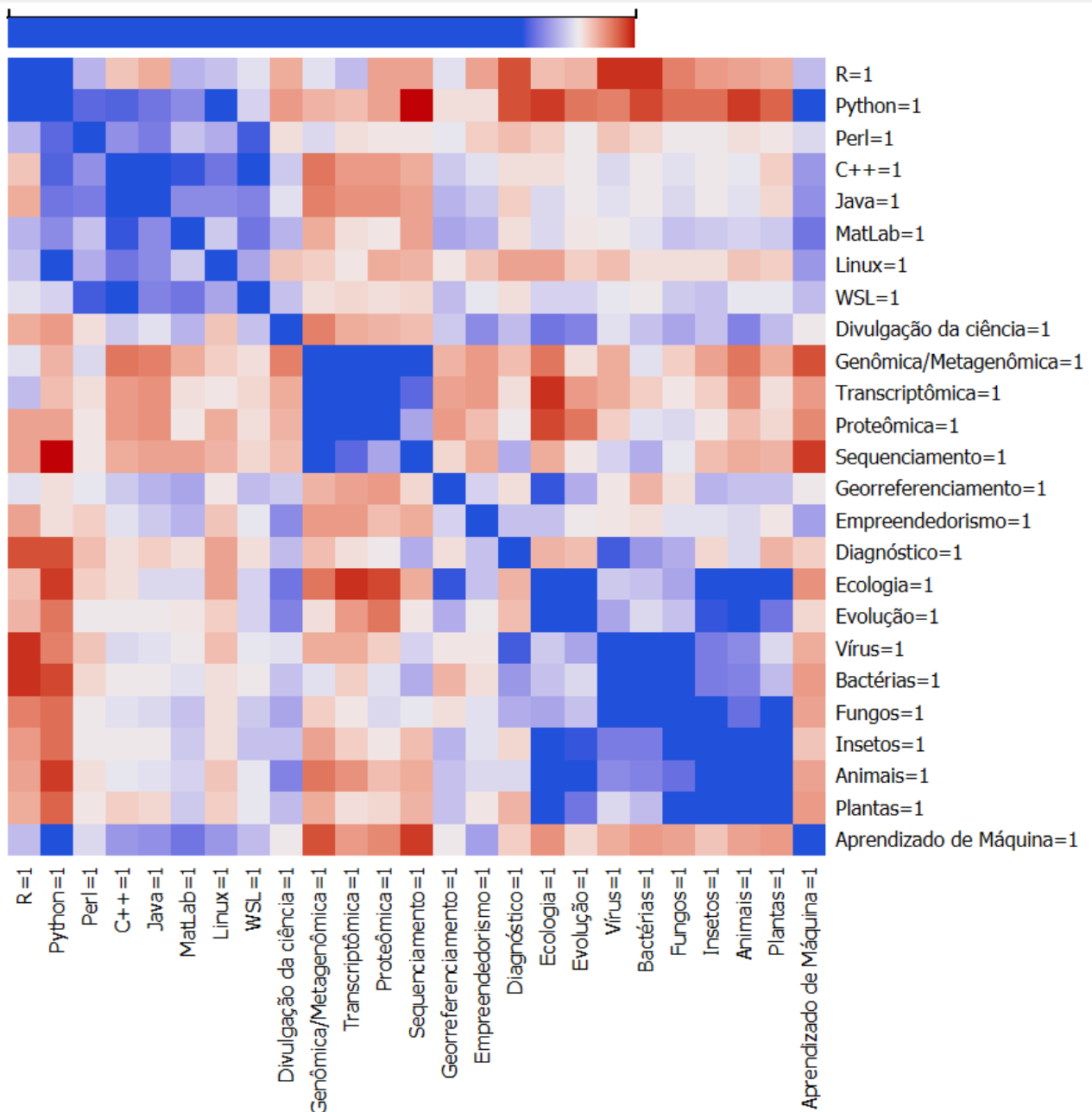
Concluimos que precisávamos de uma estratégia mais simplificada.

## Matrix de distância

A seguir, decidimos verificar correlações entre interesses dos alunos. Para isso, usamos a distância entre respostas para um mesmo campo.



A figura a seguir apresenta um mapa de distâncias entre preferências dos inscritos. Campos mais azuis representam uma maior correlação entre os campos, enquanto vermelho indica uma menor correlação. Interprete da seguinte forma: o aluno que preferiu o campo A (disposto na coluna) também preferiu o campo B (disposto na linha) caso a célula esteja marcada de azul. Quanto mais vermelha se encontra a célula, indica que a maior parte dos inscritos que optaram pelo campo A NÃO optou pelo campo B.



Por exemplo, Python e R apresentam uma alta correlação, ou seja, quem escolheu Python também escolheu R (cada inscrito poderia escolher até 5 opções). Entretanto, Python tem uma baixa correlação com sequenciamento (veja que o ponto de interseção de Python e sequenciamento está vermelho).

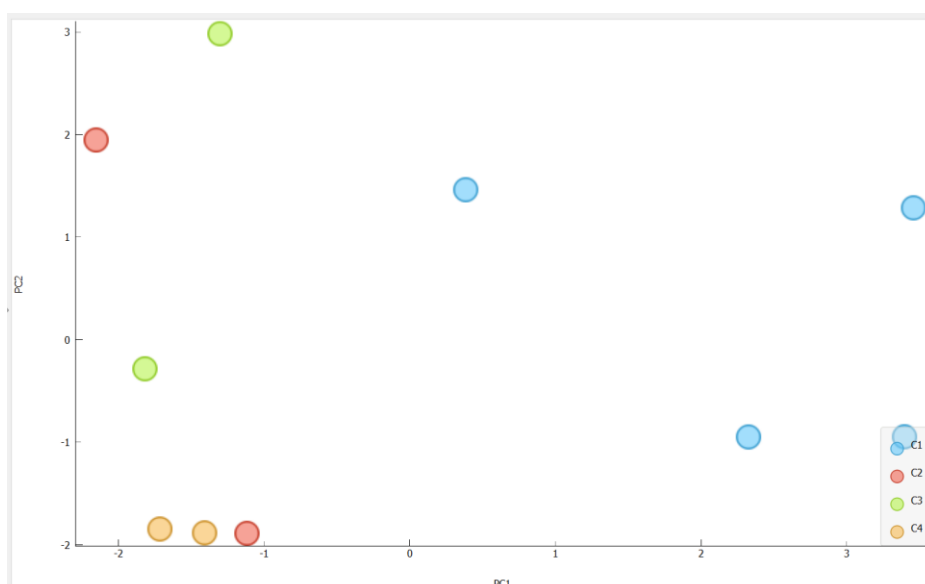
Com base nisso, agrupamentos manualmente as opções, e estabelecemos seis perguntas que poderiam apresentar respostas que englobariam boa parte das opções disponíveis e poderiam resolver possíveis conflitos. São elas:

- (A) Seu nome (necessário para evitar duplicações).
- (B) Escolha uma área: genômica, transcriptômica ou proteômica.
- (C) Você prefere: construir scripts ou usar webtools?

- (D) O que prefere? empreendedorismo ou divulgação científica?
- (E) Escolha um sistema operacional: Windows, Linux, macOS, tanto faz.
- (F) Qual o alvo mais interessante para estudos? Vírus/bactérias, Fungos, Insetos/humanos, ou Plantas.
- (G) Selecione uma linguagem de programação: Python, R, [JavaScript ou PHP], [Perl, Java ou C++].

Como tivemos que reaplicar o questionário, não foi possível estabelecer um alto número de respostas. Para essa análise usamos uma amostra de apenas 10 indivíduos (o que poderá impactar os resultados do estudo).

A seguir, coletamos os resultados, os importamos para o Orange, realizamos a análise de componentes principais e o agrupamento com K-means. Esse foi o resultado:



Podemos ver que o grupo C1 apresenta 4 indivíduos, entretanto, os outros grupos apresentam 2 indivíduos cada. Assim concluímos que as sete perguntas apresentavam uma resposta satisfatória para o nosso questionário.

Entretanto, acreditamos que seria possível reduzir a quantidade de perguntas. Por isso decidimos utilizar uma árvore de decisão para verificar quais eram as perguntas mais significativas para definição dos grupos. Para isso utilizamos o algoritmo J48 da ferramenta WEKA implementado no Google Colab usando R.

Para cada resposta, aplicamos um valor numérico iniciado em 0. Por exemplo, para a pergunta: (B) Escolha uma área. As respostas seriam (0) genômica, (1) transcriptômica ou (2) proteômica.

Após realizar a construção do modelo, obtivemos um total de 100% das instâncias classificadas corretamente:

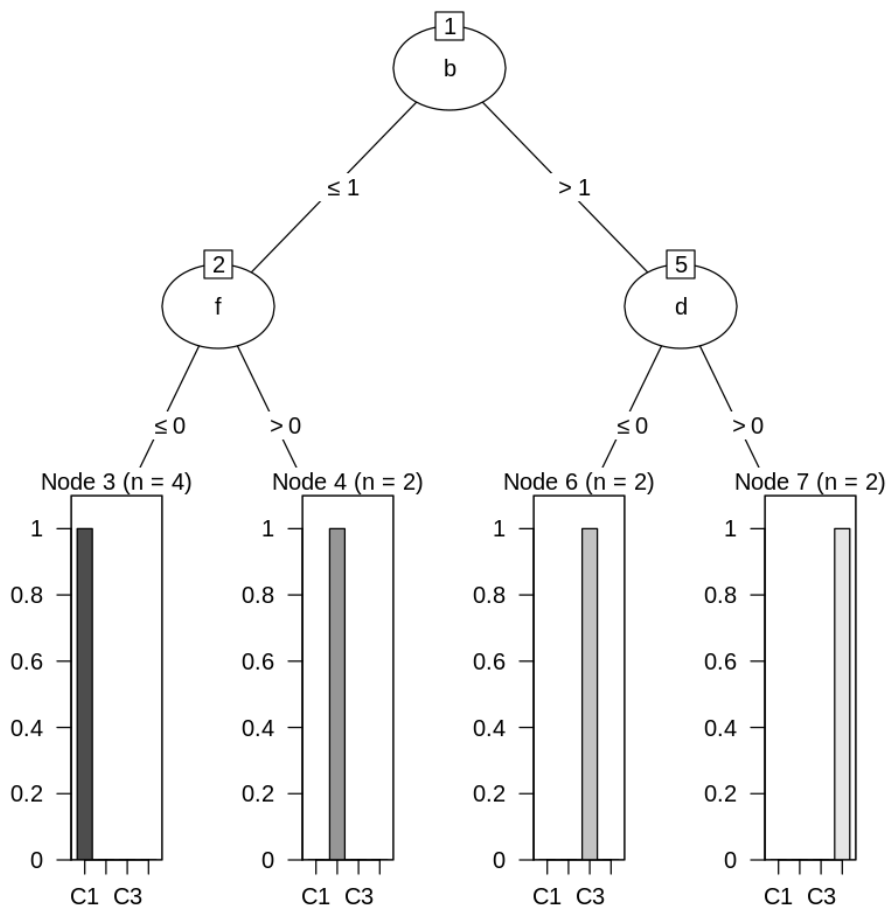
=== Summary ===

Correctly Classified Instances	10	100	%
Incorrectly Classified Instances	0	0	%
Kappa statistic	1		
Mean absolute error	0		
Root mean squared error	0		
Relative absolute error	0	%	
Root relative squared error	0	%	
Total Number of Instances	10		

=== Confusion Matrix ===

```
a b c d <-- classified as
4 0 0 0 | a = C1
0 2 0 0 | b = C2
0 0 2 0 | c = C3
0 0 0 2 | d = C4
```

Ao plotar a árvore de decisão, percebemos que apenas três perguntas são necessárias para estabelecer os grupos: b, d e f.



Entendendo a árvore: o nó raiz avalia se a resposta pra b é menor ou igual a 1 ou maior que 1. Sendo, (0) genômica, (1) transcriptômica ou (2) proteômica. Logo, proteômica poderia levar apenas aos grupos C3 e C4. Enquanto, genômica e transcriptômica poderiam levar aos clusters C1 e C2. A pergunta (d) avalia: O que prefere? (1) empreendedorismo ou (0) divulgação científica? Essa resposta indica se o indivíduo será agrupado nos grupos C3 ou C4.

Por fim, na pergunta (f) definirá se o indivíduo será classificado nos grupos C1 ou C2. Nesse caso, vemos respostas apenas para Vírus/bactérias, ou Insetos/humanos. Portanto, decidimos resumir essa questão para apenas duas opções (0) procariotos ou (1) eucariotos.

Desta forma estabelecemos as três perguntas que classificam nos quatro grupos:

1. Escolha uma área: Genômica/ Transcriptômica/ (Proteômica e/ou Bioinformática estrutural)
2. Qual tópico mais te atrai: Empreendedorismo/ Divulgação científica
3. Qual dos seguintes itens corresponde a uma área mais interessante de estudo?  
Procariotos /Eucariotos

Com base nesse resultado, estabeleceu-se uma matriz de pontuação. Assim, um indivíduo poderá ser classificado em um grupo com base na coluna que obtiver maior pontuação.

#	Respostas	C1	C2	C3	C4
1	Genômica	1	1	0	0
	Transcriptômica	1	1	0	0
	Proteômica e/ou Bioinformática estrutural	0	0	1	1
2	Empreendedorismo	0	0	1	0
	Divulgação científica	0	0	0	1
3	Procariotos	1	0	0	0
	Eucariotos	0	1	0	0

Por exemplo, o indivíduo respondeu:

1. Genômica
2. Divulgação científica

---

### 3. Eucariotos

Logo:

- C1: 1 ponto
- C2: 2 pontos
- C3: 0 pontos
- C4: 1 ponto

Assim, ele será classificado no cluster C2.

Para melhor divulgação, renomeamos os clusters C1 a C4 para:

- C1: Cobra
- C2: Jaguarica
- C3: Jacaré
- C4: Capivara

## Conclusão

Aqui apresentamos o método para classificação de indivíduos em grupo com base nas suas preferências pessoais, visando ainda apresentar uma quantidade aproximada de indivíduos no mesmo grupo. Acreditamos que o resultado apresentado é satisfatório para os objetivos estabelecidos. Entretanto, temos ciência que outras estratégias poderiam apresentar um resultado melhor, entretanto, dada a limitação computacional não podemos estabelecer melhores estratégias.