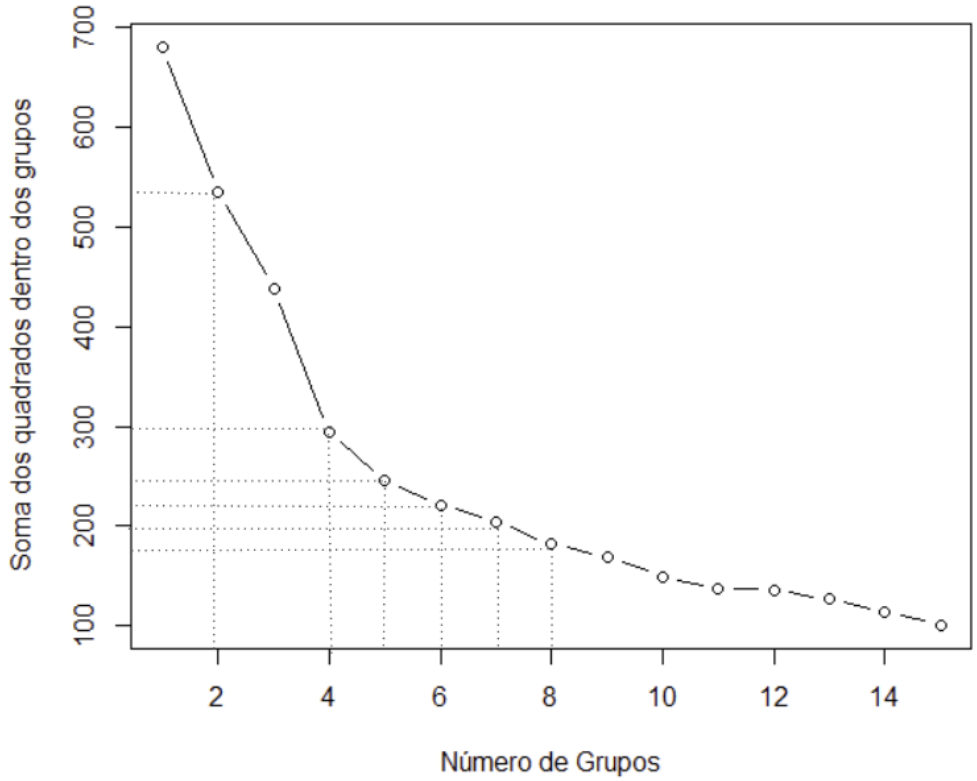
Técnicas de Aprendizado de Máquina Aplicadas a Negócios

Algoritmos de Agrupamento

K-Means

A partir dos scripts foi obtido este gráfico

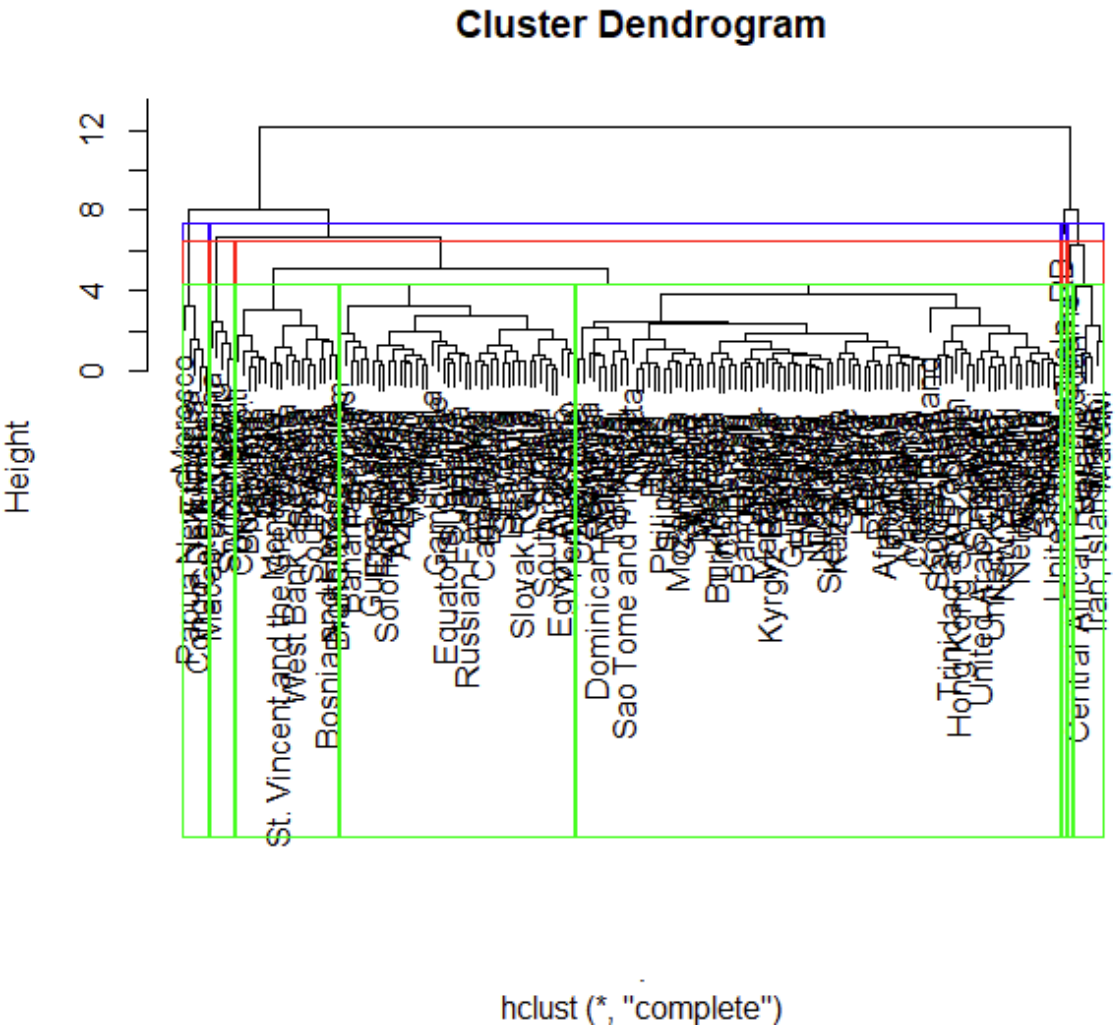


Neste gráfico colocamos linhas tracejadas para ajudar na interpretação, conforme vamos reduzindo o número de grupos menor e a soma dos quadrados dentro dos grupos, mas estamos em uma busca do número ótimo de grupos. Se olhamos para o começo do gráfico pequena diferenças fazer uma redução muito grande na soma dos quadrados:

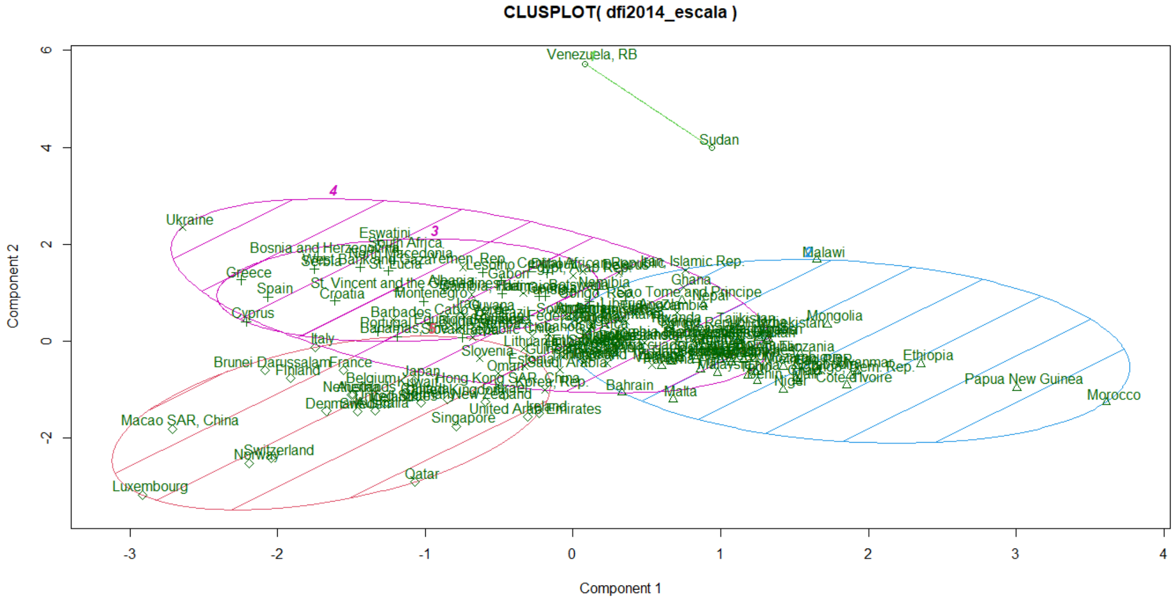
- Indo de 2 para 4 grupos: a diferença na soma é de aproximadamente 233;
- Indo de 4 para 5 grupos: a diferença já caiu bastante e fica em aproximadamente 50;
- Indo de 5 para 6 grupos: neste a diferença é de aproximadamente 30;

Como a proporção caiu muito de 5 para 6 e como ainda 4 para 2 é muito grande a diferença, o número ideal parece ser 5 grupos.

A partir dos scripts foi obtido este gráfico em dendograma:

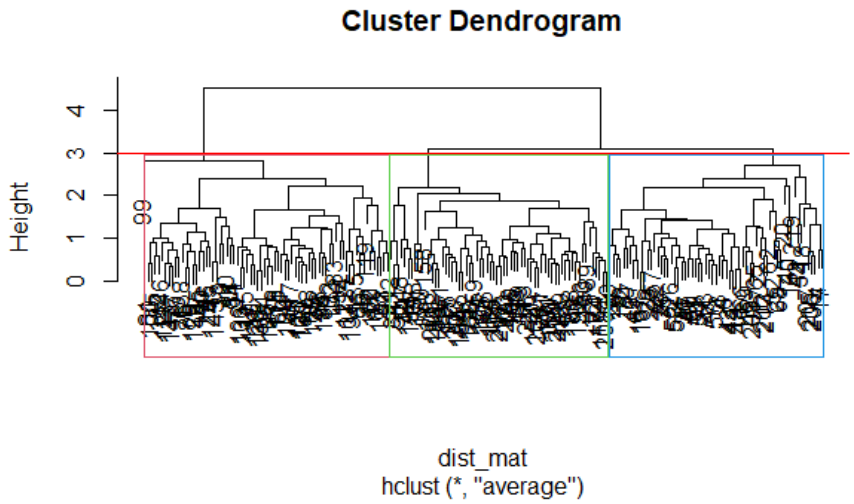


Assim podemos perceber que a divisão em verde que possui k = 8 não seria a ideal, pois há pais que são os únicos no seu cluster, gerando assim um excesso de clusters que atrapalha a análise dos dados. A proposta da divisão em vermelho e azul são claramente melhores, sendo a primeira k = 5 e a segunda k = 4, pensado que buscamos o equilíbrio entre ter o número ótimo de clusters e ao mesmo tempo ter poucos clusters acreditamos que a melhor opção e a vermelha valorizando assim a divisão com mais divisões

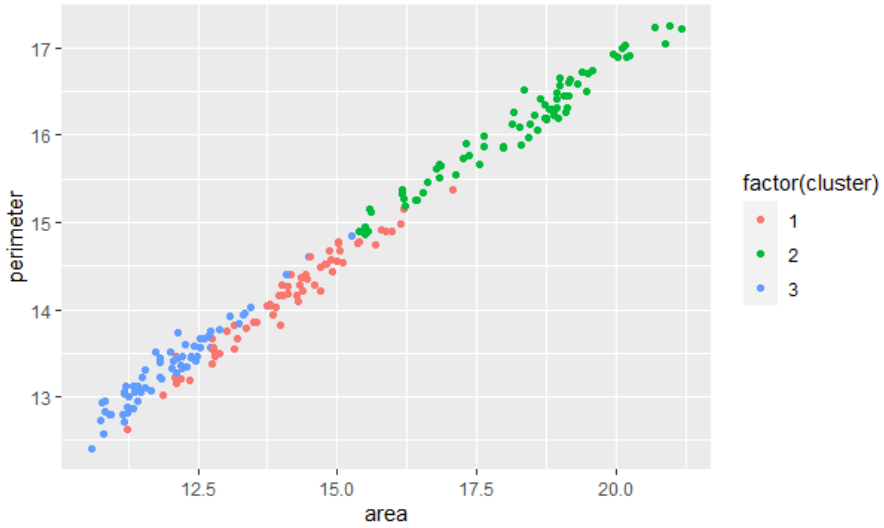


Considerando o agrupamento com 5 clusters, como foi considerado também no laboratório, é possível notar que há uma grande sobreposição dos grupos, principalmente quando observamos, por exemplo, os clusters 3 e 4, o que pode comprometer o alcance de resultados mais claros, precisos e interpretáveis. Porém, é possível pontuar que é bastante coerente a existência de outliers como Venezuela e Sudão, uma vez que, no período analisado, ambos os países estavam enfrentando graves problemas sociais, econômicos e políticos, prejudicando os indicadores.

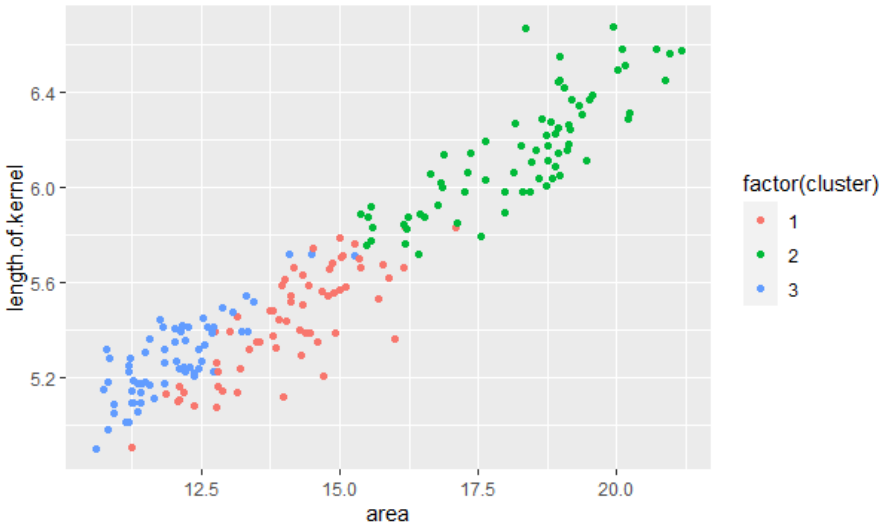
Agrupamento Hierárquico



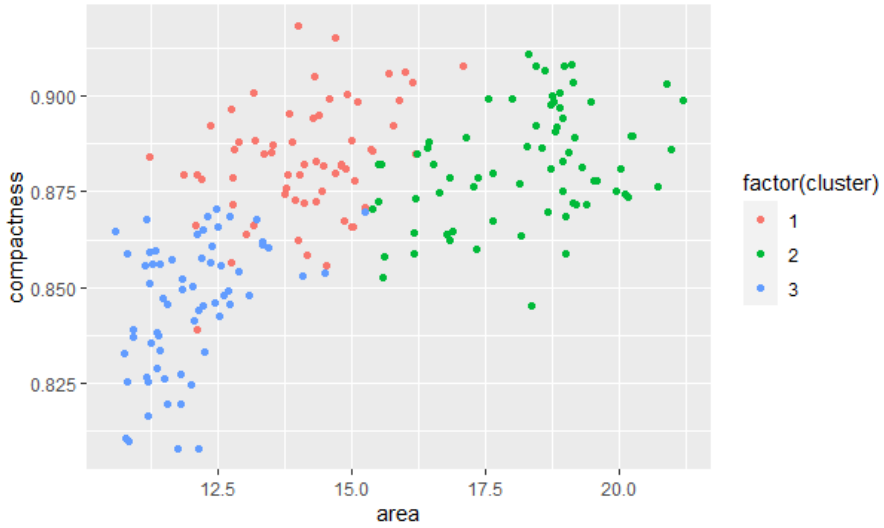
Tendo-se em mente a segunda metade deste laboratório, o desafio seria agrupar os grãos a partir de suas características, como área e perímetro, utilizando a técnica de agrupamento hierárquico, que considera a distância euclidiana entre os diferentes elementos do conjunto de dados (neste caso, as sementes). A figura acima, então, mostra o resultado obtido a partir do agrupamento em três grupos.



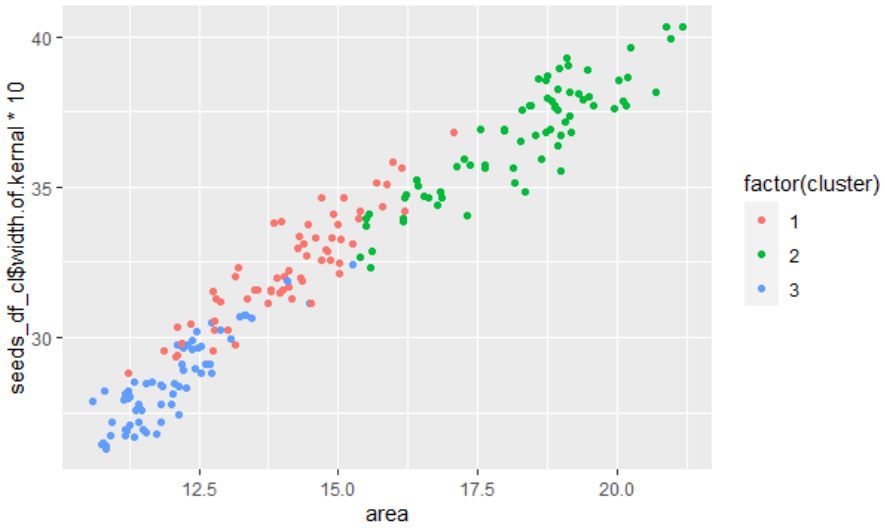
O gráfico acima, por sua vez, mostra a relação entre área e perímetro vis-à-vis a classificação atribuída às sementes. Assim, pode-se ver claramente que o cluster 2 consiste de sementes com área e perímetros maiores, enquanto o cluster 1 e 3 contam com sementes com área menor. O que diferencia os clusters 1 e 3, no que se refere ao gráfico acima, é que as sementes do grupo 1 possuem perímetros maiores, em geral, e menor área.



A figura acima apresenta a mesma lógica da anterior, mas substituímos a variável perímetro pelo comprimento da semente. Pode-se observar que a divisão se manteve a mesma, de maneira geral: cluster 2 com os maiores valores de área e comprimento e os clusters 1 e 2 com áreas menores, sendo que o primeiro conta com maiores valores de comprimento, em geral, e menor área.



O mesmo ocorre em relação à variável de compactabilidade.



E também em relação à largura da semente.

Conclusão

A partir deste laboratório, foi possível notar as diferentes aplicações das duas técnicas de agrupamento: K-means e Agrupamento Hierárquico. No caso do agrupamento hierárquico, não é necessário determinar um número muito específico de clusters com antecedência, de modo que o agrupamento é feito de acordo com o recorte do dendograma, selecionando um determinado nível de altura no dendograma para delimitar quantos grupos serão formados. Já o K-means permite a definição prévia do número de clusters desejados, sendo mais adaptável, nesse sentido, porém há uma menor estabilidade de resultados quando comparado ao outro método

Além disso, observa-se que há um *tradeoff* entre a complexidade do modelo e a interpretação dos resultados, no sentido que modelos com um número muito grande de grupos perdem o aspecto de generalizar os padrões presentes nos dados a partir da classificação, porém um número muito pequeno de clusters torna a análise muito ampla, não conseguindo diferenciar claramente os padrões que diferenciam cada grupo, dificultando a interpretação dos resultados. Nesse sentido, a avaliação dos modelos deve levar esses aspectos em consideração, o que acaba tornando-a subjetiva.

No caso dos dados sobre as sementes, analisando o agrupamento considerando 3 clusters, foi possível notar claramente os padrões de cada grupo, de modo que o modelo colocou no grupo 3 os menores valores, no 2, os intermediários e, no 3, os mais altos quanto a área e ao perímetro dos grãos. Porém, no modelo formado a partir do K-means, foi mais difícil de avaliar se o melhor caso seria considerar k = 5 ou k = 4 ou k = 6 como a melhor alternativa, e mesmo considerando k = 5 para fins de análise, em alguns casos os padrões de cada grupo não ficaram muito claros, como foi pontuado nos problemas do resultado relatados na Conclusão do próprio relatório (Página 17).

Por fim, a execução e análise do laboratório foi interessante no sentido de entender como ocorre a aplicação de cada método de agrupamento, como também como eles podem ser utilizados em diferentes contextos, sejam eles socioeconômicos ou mais relacionados a agricultura, como no caso dos dados do Banco Mundial e dos dados sobre propriedades geométricas de certos grãos, respectivamente. Uma sugestão, nesse caso, seria aplicar ambos os algoritmos para bancos de dados de tamanhos diferentes e calcular o tempo de execução nas etapas de modelagem, a fim de avaliar na prática o desempenho de cada método em contextos que exijam diferentes níveis de complexidade computacional.