Universidade Federal de São Carlos

Centro de Ciências Exatas e de Tecnologia

Departamento de Estatística

Análise da relação entre características demográficas e hábitos de consumo em jovens

Trabalho 2 de Machine Learning

Grupo: Gabrielly Castilho Guimarães – 805757
Luiza Neves Graça – 791328
Miguel Antonio de Oliveira – 772180
Rubens Neto – 759550'

Prof. Dr. Murilo Naldi

Agosto, 2023

Sumário

Capítulo 1

Introdução

Em um mundo movido por enormes quantidades de dados e em constante evolução tecnológica, aprimorar os modelos de negócios das empresas e efetivamente delinear o perfil de seu público-alvo assume um caráter crucial para garantir a consolidação em um mercado crescentemente competitivo. Nesse sentido, a segmentação de perfis de consumidor com base em interesses comuns tem emergido como um componente central.

Nesse trabalho, buscaremos traçar uma relação hábitos de consumo e o seu perfil demográfico a fim de identificar agrupamentos de possíveis segmentos de mercado; uma tarefa particularmente útil para personalizar estratégias de marketing, aprimorar a compreensão do público-alvo e direcionar de maneira eficaz produtos e serviços que atendam às necessidades específicas de cada grupo. Em suma, o estudo de dados demográficos transcende a análise de mercado e passa a transbordar para áreas como a tecnologia e ciência de dados.

1.1 Objetivo

Dessa forma, o objetivo desse trabalho é ilustrar de forma didática como as tarefas de agrupamento vistas em sala de aula podem ser usadas para agrupar indíviduos em segmentos de mercado com base em seu perfil socieconômico.

1.2 A base de dados

Os dados utilizados foram retirados do Kaggle[1] e baseados em um estudo de 2013 que entrevistou 1010 jovens de 15 a 30 anos e coletou ao todo dados sobre 150 variáveis: (i) 17 atributos referentes ao gosto musical dos respondentes; (ii) 12 atributos referentes às preferências de filmes; (iii) 32 atributos sobre hobbies e interesses; (iv) 10 atributos em relação a fobias; (v) 3 atributos sobre os hábitos de saúde; (vi) 57 atributos referentes à personalidade e opiniões; (vii) 7 atributos referentes aos hábitos de consumo; e (viii) 10 atributos referentes ao perfil demográfico dos participantes.

Dado que o nosso objetivo é apenas traçar a relação entre o perfil demográfico e os hábitos

de consumo dos participantes, iremos focar apenas nessas duas categorias ignorando as outros, o que nos deixa com apenas 17 atributos. Abaixo segue a descrição de cada um deles.

Atributos referentes ao perfil demográfico:

- Idade (assume valores inteiros de 15 a 30);
- Altura (assume valores inteiros em cm);
- Peso (assume valores inteiros);
- Quantos irmãos a pessoa tem;
- Gênero (binário);
- Qual a mão dominante da pessoa (binário);
- Se é um filho único (binário);
- Em que lugar a pessoa cresceu (em uma cidade ou uma vila);
- Em que casa a pessoa morou na infância (casa térrea ou apartamento).

Atributos referentes aos hábitos de consumo:

Para cada item abaixo, as pessoas responderam o quanto concordam com a frase em uma escala de 1 a 5, sendo 1 "discordo totalmente"e 5 "concordo totalmente".

- Eu economizo todo o dinheiro que posso;
- Eu gosto de frequentar Shoppings e grandes centros comerciais;
- Prefiro roupas de marca;
- Gasto muito dinheiro com festas e eventos sociais;
- Gasto muito dinheiro com a minha aparência;
- Gasto muito dinheiro com *gadgets* e aparatos tecnológicos;
- Eu não me incomodaria de pagar mais por uma comida boa, de qualidade ou saudável.

Capítulo 2

Pre-processamento e visualização da base de dados

Para dar início ao pré-processamento, importamos o arquivo .csv no Python através da biblioteca Pandas. e trabalhamos para transformar todos os atributos categóricos em numéricos. Dado que nossos atributos categóricos eram binários, não tivemos problemas nesse sentido.

2.1 Tratamento de outliers e valores faltantes

É importante mencionar que identificamos a presença de alguns valores ausentes e um valor absurdo, em que a pessoa reportou ter uma altura de 68 cm. No entanto, dada a quantidade reduzida desses valores em nossa amostra, decidimos pela sua remoção completa, a fim de preservar a integridade dos dados.

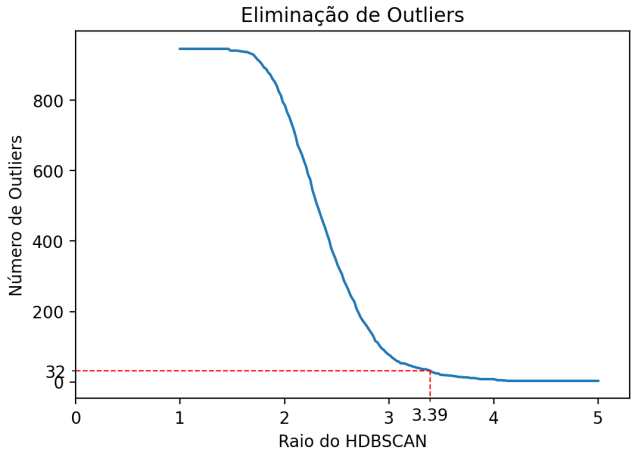

Figura 2.1: Gráfico de HDBSCAN

Quanto à identificação e remoção de possíveis outliers, utilizamos a técnica HDBSCAN, ou *Hierarchical Density-Based Spatial Clustering of Applications with Noise*, que é capaz de distinguir regiões de alta densidade daquelas de baixa densidade em um espaço de dados multi-

dimensional, de forma que os pontos que possuem densidades locais mais baixas são classificados como ruído ou outliers.

Em seguida, considerando que nossos dados possuem magnetudes e escalas diferentes, aplicamos a técnica de padronização em que cada variável é normalizada, isto é, tem sua média subtraída, e o resultado da normalização é dividido pelo desvio padrão da variável. Assim, garantimos maior homogeneidade, uma vez que todos os atributos passam a possuir média e desvio padrão iguais a zero e um, respectivamente.

2.2 Seleção e redução de atributos

Nessa etapa do projeto, nos concentramos em identificar redundâncias nos atributos e como podemos deixar a base de dados mais limpa para a aplicação das técnicas de agrupamento.

Começamos plotando o *heatmap* do índice de correlação de pearson entre todas as variáveis escolhidas, que pode ser visto na figura 2.2 abaixo. De início, notamos que as variáveis possuem, no geral, baixas correlações entre si, sendo 0.59 o maior valor encontrado e -0.61 o menor.

A fim de explorar todas as potenciais correlações nos dados, também calculamos os coeficientes de Spearman e Kendall, que são capazes de capturar algumas relações não lineares mais simples entre variáveis. Contudo, após essa análise adicional, constatamos que os resultados não diferiram substancialmente das correlações lineares obtidas previamente. Por essa razão, optamos por não incorporar esses coeficientes no relatório em questão.

No entanto, embora as correlações entre variáveis possam ser fracas, as técnicas de agrupamento podem revelar conjuntos com características semelhantes, permitindo a exploração de relações mais complexas e não lineares que podem ser significantes para a compreensão do fenômeno em estudo. Dessa forma, não interpretamos as baixas correlações como um impeditivo para a análise.

2.3 Agregação

Com isso dito, antes de prosseguirmos para a próxima etapa, acreditamos que a base de dados possa ser aprimorada, através da remoção de atributos redundantes ou que não fazem sentido para análise. Desse modo:

- Optamos por remover a variável que indica se a pessoa é filha única ou não, dado que ela já está sendo explicada pela variável número de irmãos;
- Optamos por remover a variável que indica se a pessoa é destra ou canhota, por não acreditarmos que faz sentido para o problema, e qualquer relação encontrada não irá indicar uma relação causal com os hábitos de consumo e, sim, uma coincidência;

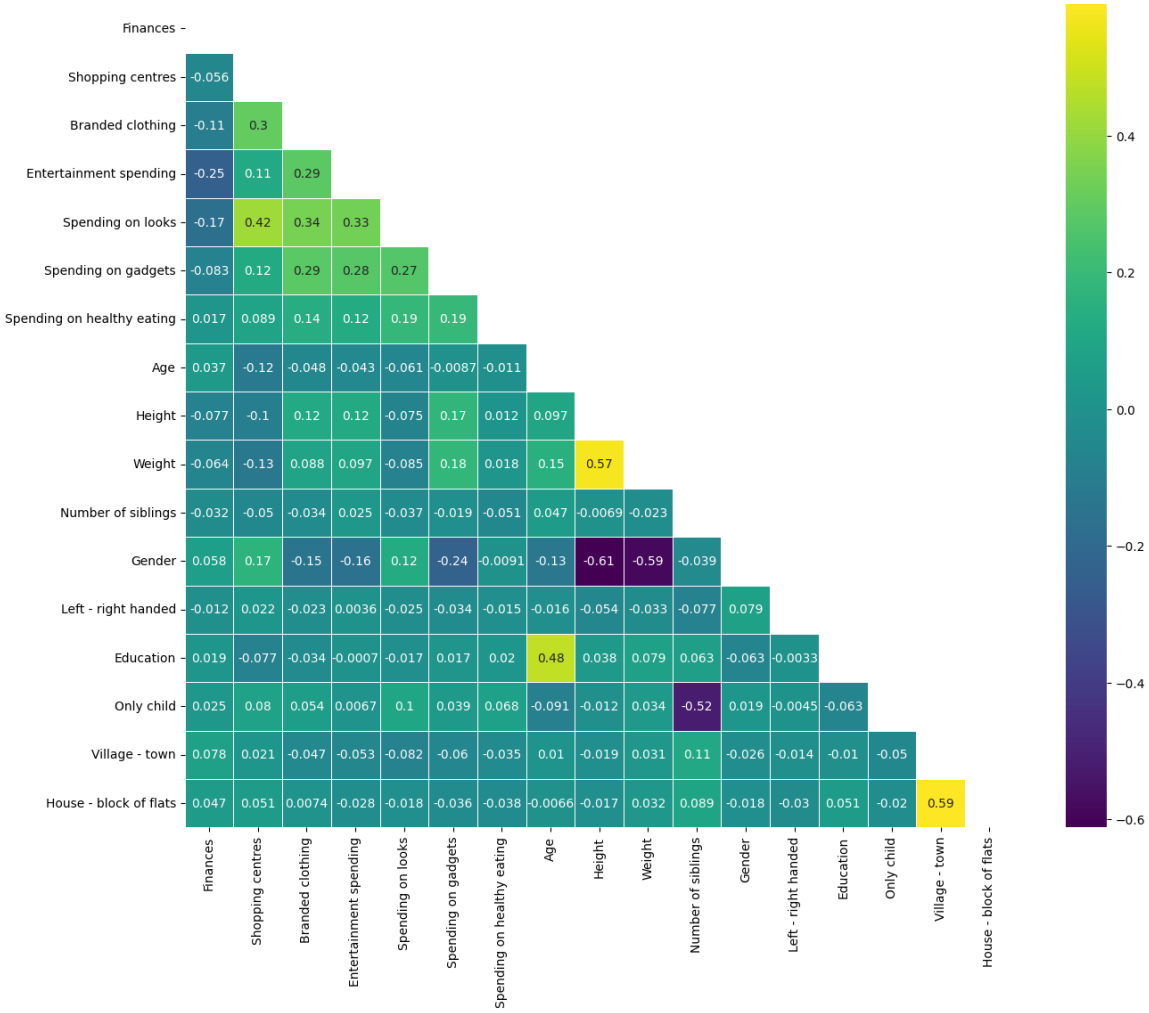

Figura 2.2: Gráfico de correlação entre as variáveis

- Removemos as variáveis de altura e peso, pelo mesmo motivo que a variável indicativa da mão dominante dos participantes. Além disso, essas variáveis já estão sendo parcialmente explicadas pela variável gênero;
- Por fim, removemos também a variável que indica onde a pessoa morou na infância (casa ou apartamento), por acreditarmos que estaria reduntante junto com o atributo que indica se ela morava na cidade ou no campo.

Capítulo 3

Modelos de agrupamento

Com a base de dados pronta, iremos dar início ao ajuste de modelos de agrupamento. Nessa etapa, faremos uso de dois modelos distintos: (i) K-Means e (ii) via Agrupamento Hierárquico.

3.1 K-Means

O K-Means é um método de aprendizado de máquina não supervisionado que organiza dados em grupos, chamados de clusters, onde cada cluster é representado por um ponto central chamado centróide. Inicialmente, seleciona-se um número de clusters (k) e pontos iniciais como centroides. O algoritmo itera alternando entre duas etapas: atribuir cada ponto ao centróide mais próximo e recalcular os centroides com base nos pontos atribuídos. Essas etapas são repetidas até que os centroides se estabilizem. O resultado final é uma divisão dos dados em clusters, onde idealmente os pontos dentro de cada cluster são semelhantes entre si e diferentes dos pontos em outros clusters.

Para definir a quantidade ótima de clusters do nosso modelo, utilizaremos o método do cotovelo

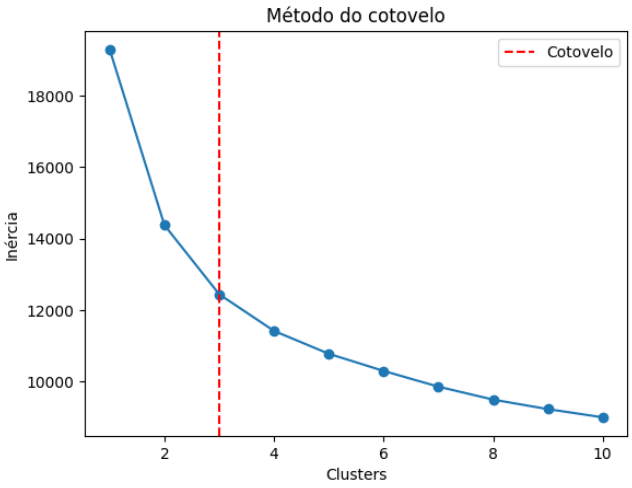

Figura 3.1: Gráfico de correlação entre as variáveis

Com base no gráfico 3.1 acima, ajustamos o modelo com o valor ótimo encontrado de K = 3. Abaixo seguem os resultados obtidos.

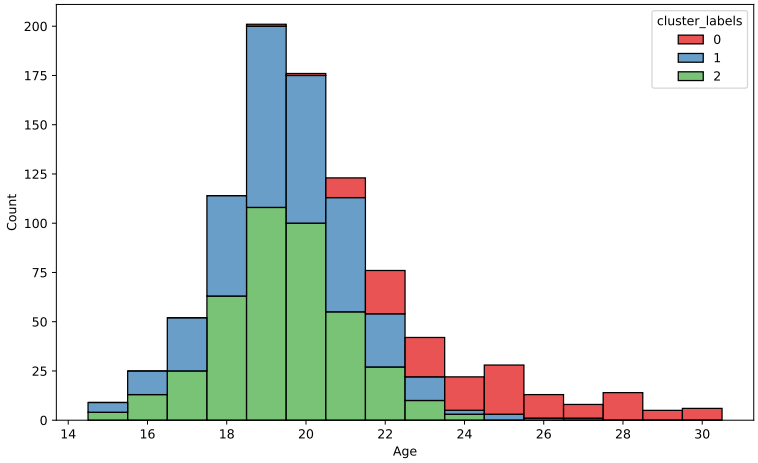

Figura 3.2: Histograma da distribuição da idade por grupo

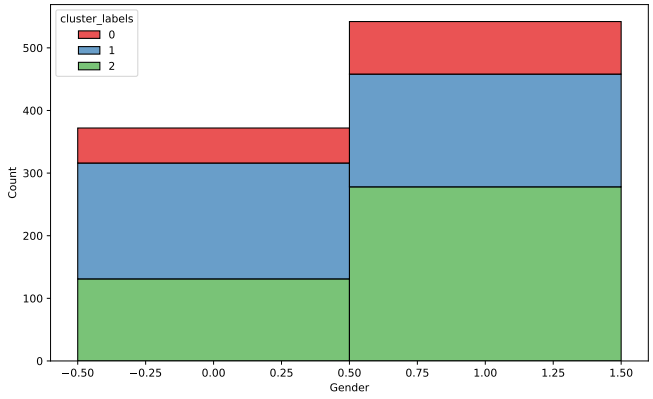

Figura 3.3: Histograma da distribuição do gênero por grupo

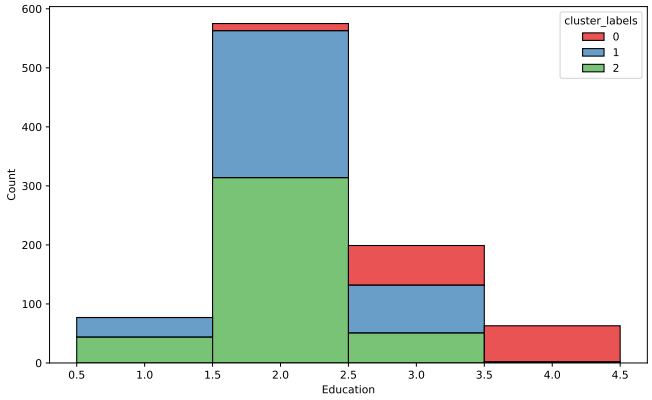

Figura 3.4: Histograma da distribuição da escolaridade por grupo

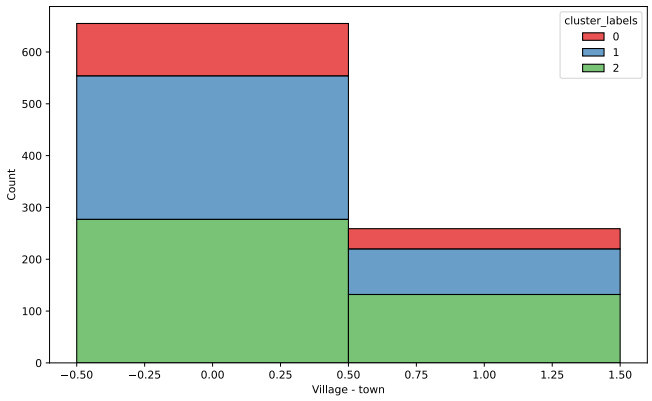

Figura 3.5: Histograma da distribuição de onde a pessoa morou na infânica, cidade ou área rural, por grupo

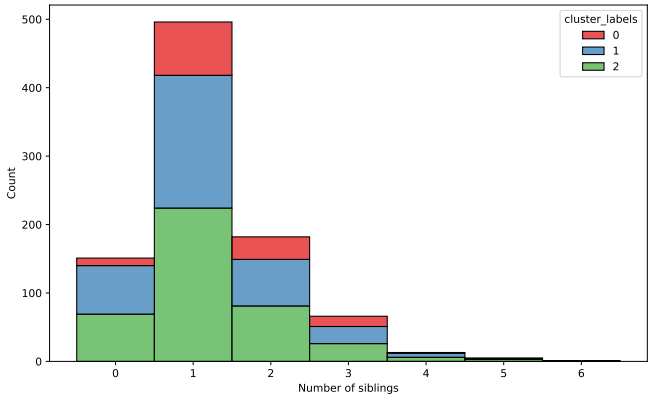

Figura 3.6: Histograma da distribuição do número de irmãos por grupo

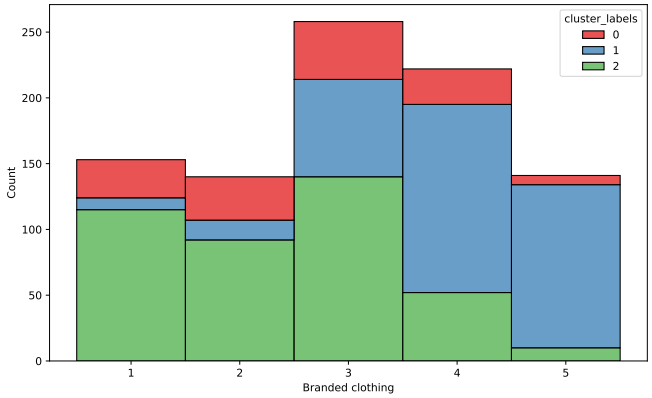

Figura 3.7: Histograma da distribuição da preferência por roupas de marca por grupo

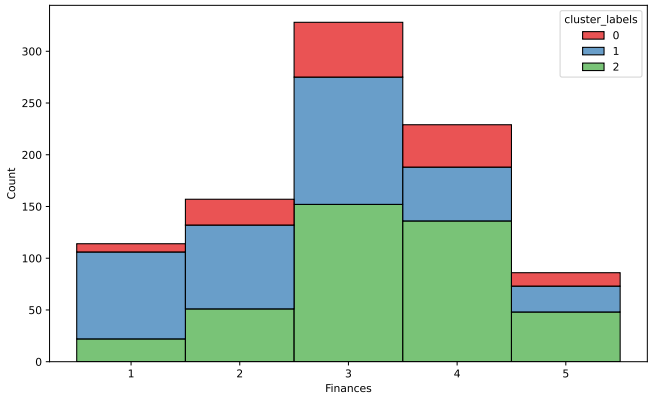

Figura 3.8: Histograma da distribuição da tendência a economizar dinheiro por grupo

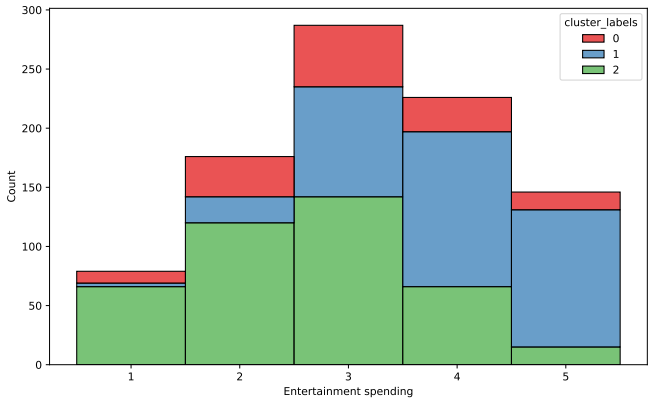

Figura 3.10: Histograma da distribuição da tendência a gastar com festas e socializações por grupo

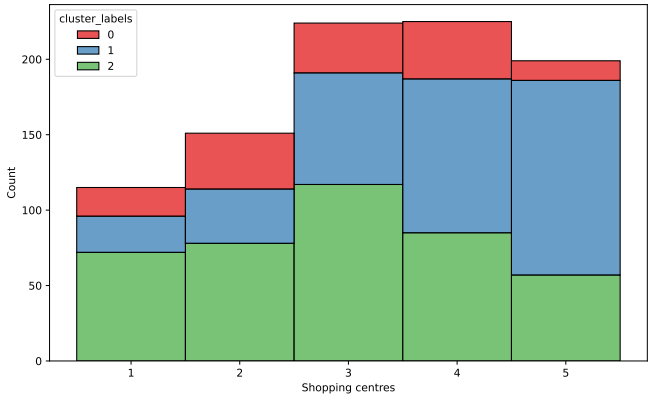

Figura 3.9: Histograma da distribuição da tendência a gostar de Shoppings por grupo

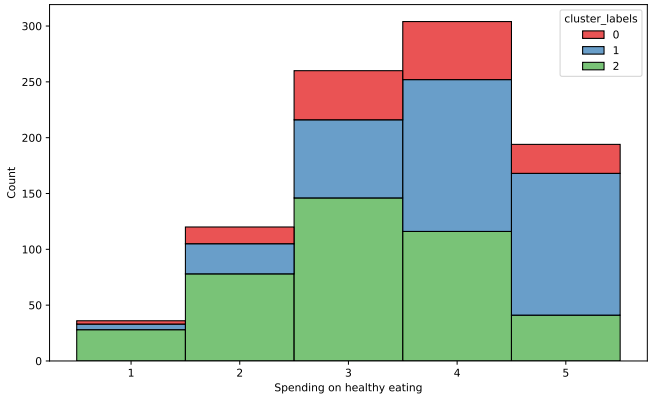

Figura 3.11: Histograma da distribuição da tendência a gastar com comidas mais sudáveis ou gourmet por grupo

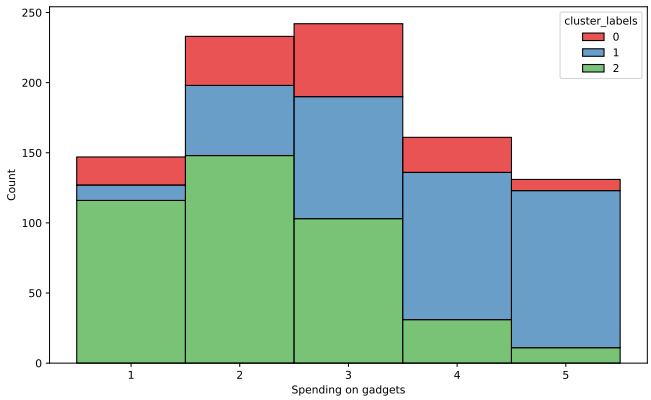

Figura 3.12: Histograma da distribuição da tendência a gastar com gadgets e aparelhos eletrônicos por grupo

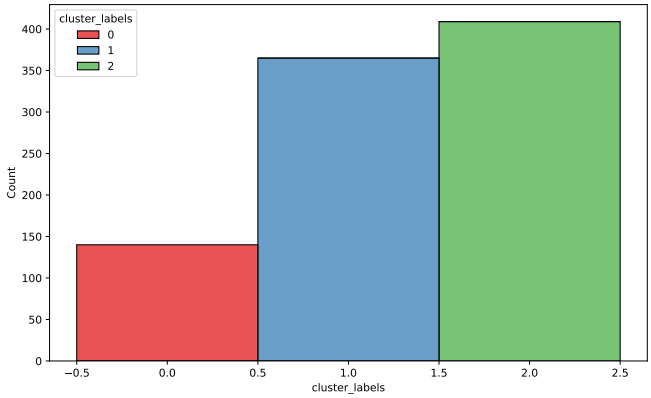

Figura 3.13: Histograma da distribuição de observações por cluster

Com base nos resultados obtidos, conseguimos traçar os seguintes perfis para cada cluster:

- O primeiro grupo, representado em vermelho, é, de modo geral, constituído por pessoas mais velhas e com uma escolaridade maior, que tendem a gostar mais de freqeuntar shoppings e não se incomodam muito em gastar mais para consumir comidas gourmets ou para usar roupas de grife, e valorizam moderadamente a economia de dinheiro e *gadgets* eletrônicos.
- O segundo grupo, em azul, é formado por pessoas mais jovens com um grau de escolaidade médio um pouco menor quando comparado ao grupo acima. Quanto aos seus hábitos de consumo, essas pessoas tendem a valorizar bem menos a economia de dinheiro, e tendem a gastar mais, sobretudo com roupas de marca, festas, idas em Shoppings, comidas, e *gadgets* tecnológicos.
- Por fim, o terceiro grupo, respresentado em verde, é o nosso grupo mais volumoso. Demograficamente, ele é caracterizado por pessoas mais jovens que o primeiro grupo e que também possuem uma escolaridade menor. Quanto aos seus hábitos financeiros, estas são pessoas que valorizam muito mais a poupança de dinheiro e estão menos inclinadas a gastar com supérfluos como roupas de marca, festas e eventos, Shoppings, comidas gourmet, e aparelhos eletrônicos.
- Vale mencionar que as variáveis gênero, local onde a pessoa morou na infância e número de irmãos não aparentaram impactar significativamente nenhum dos grupos.

3.2 Agrupamento Hieráquico

O agrupamento hierárquico é um outro modelo não supervisionado que organiza objetos a partir de uma estrutura de árvore, onde cada nó representa um grupo de objetos e os nós são agrupados em níveis sucessivos. Começando com cada objeto como seu próprio cluster, o algoritmo gradualmente combina os objetos em grupos maiores, criando uma hierarquia de clusters. Existem duas abordagens principais: aglomerativa (começando com clusters individuais

e mesclando-os) e divisiva (começando com um único cluster e dividindo-o).

Nesse trabalho, optamos por realizar o agrupamento hierárquico aglomerativo, dado que temos um grande número de objetos e esse método é menos complexo que o divisivo.

Na figura 3.14 abaixo, conseguimos observar o dendograma inferido pelo algoritmo. Optamos por seccionar o dendograma de modo que forme 3 clusters distintos, dado que acreditamos que esse seja o corte forma grupos mais coesos.

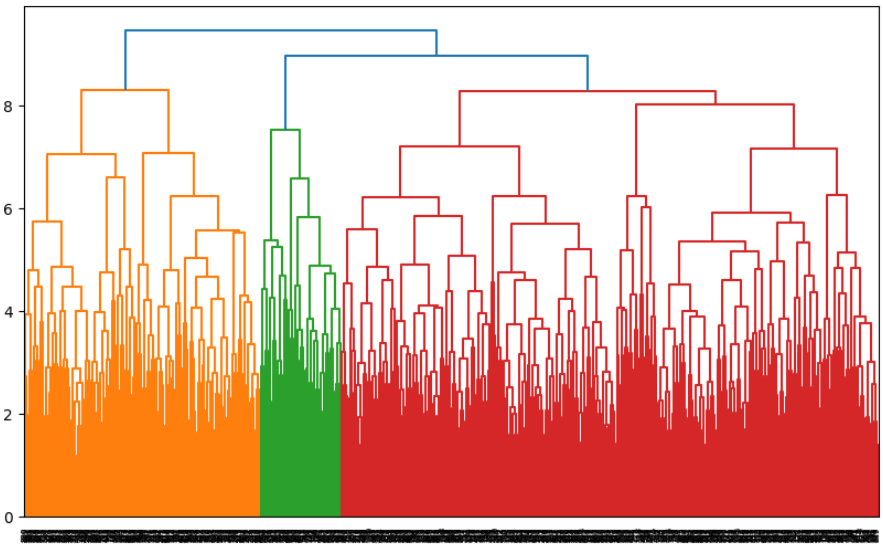

Figura 3.14: Dendograma

A partir da definição dos grupos, obtivemos os resultados abaixo:

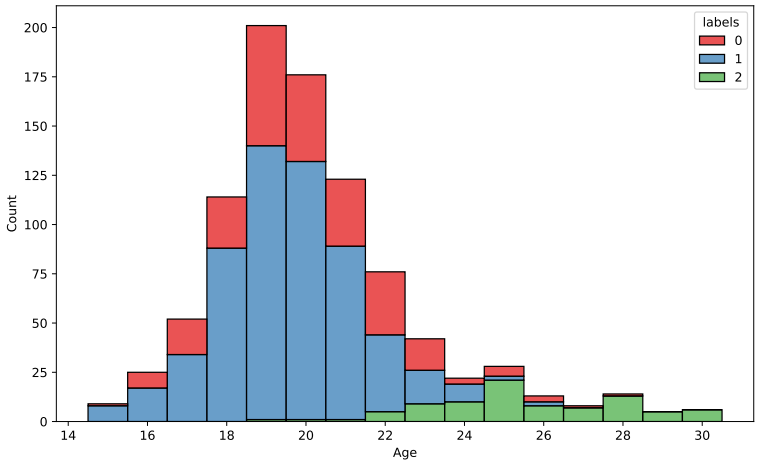

Figura 3.15: Histograma da distribuição da idade por grupo

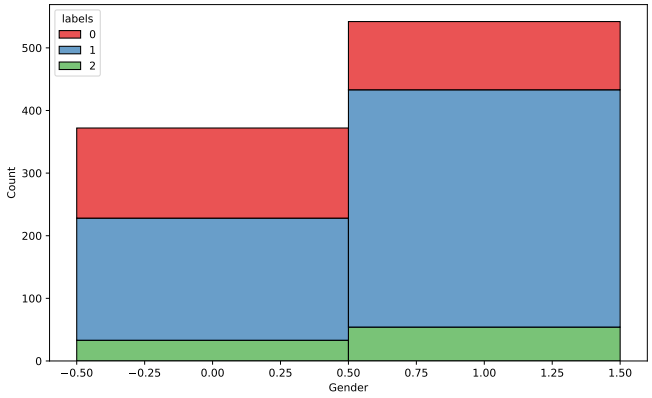

Figura 3.16: Histograma da distribuição do gênero por grupo

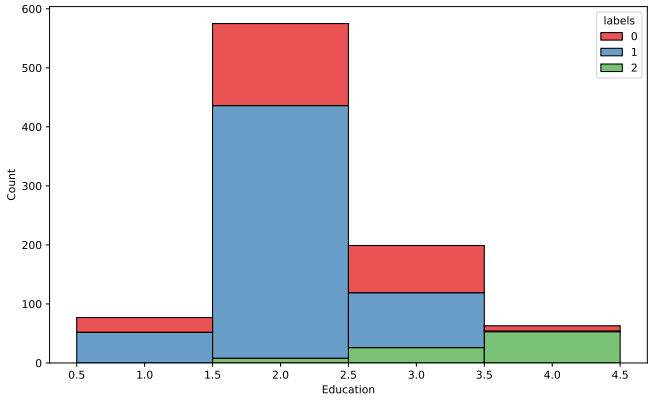

Figura 3.17: Histograma da distribuição da escolaridade por grupo

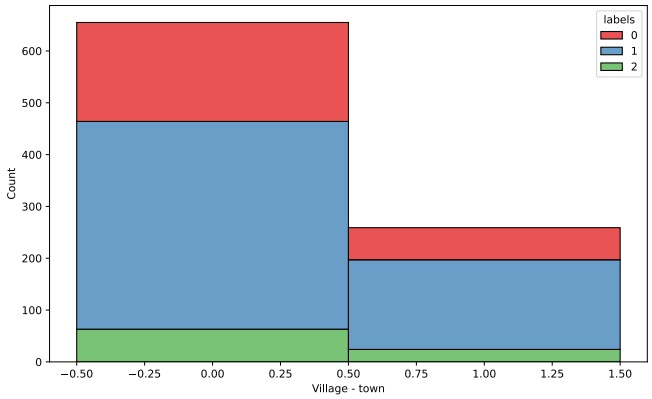

Figura 3.18: Histograma da distribuição de onde a pessoa morou na infânica, cidade ou área rural, por grupo

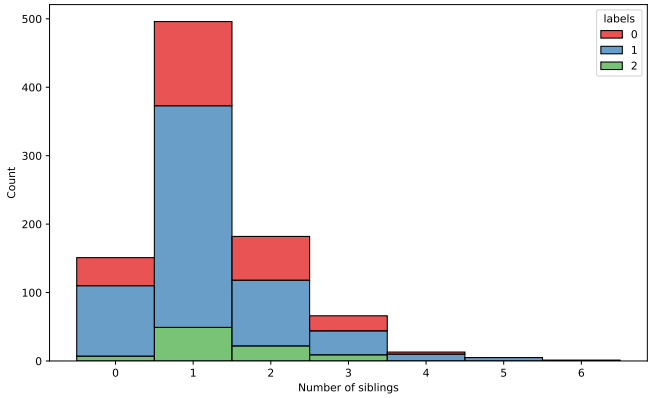

Figura 3.19: Histograma da distribuição do número de irmãos por grupo

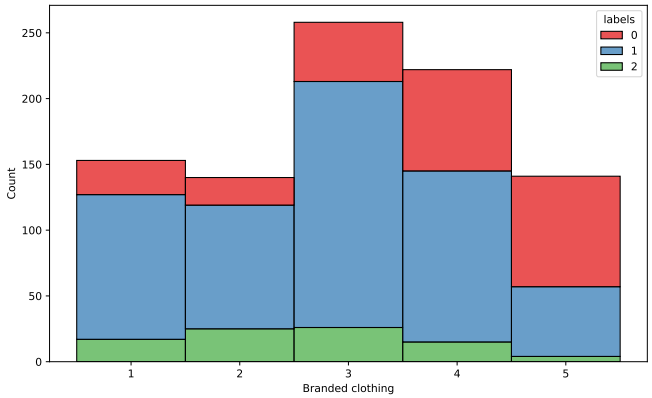

Figura 3.20: Histograma da distribuição da preferência por roupas de marca por grupo

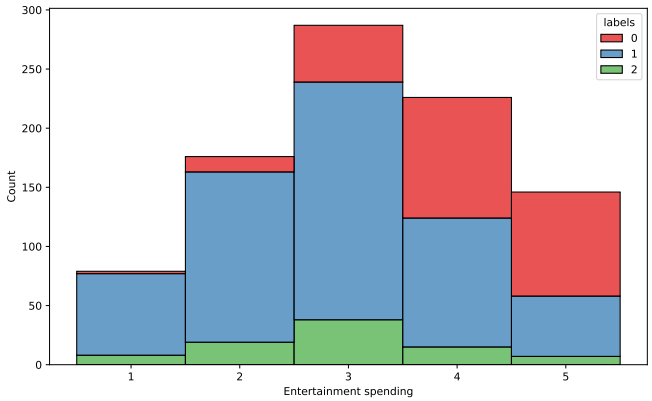

Figura 3.23: Histograma da distribuição da tendência a gastar com festas e socializações por grupo

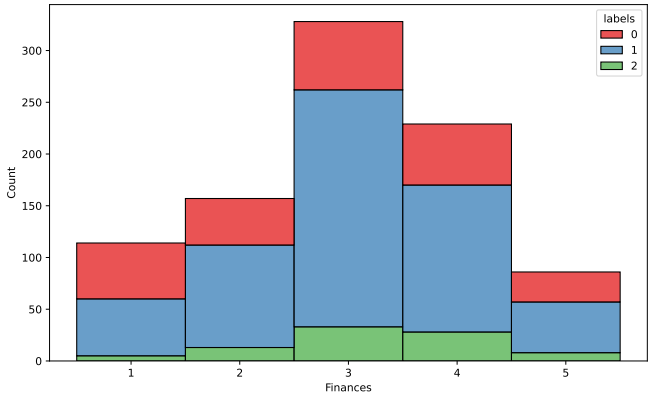

Figura 3.21: Histograma da distribuição da tendência a economizar dinheiro por grupo

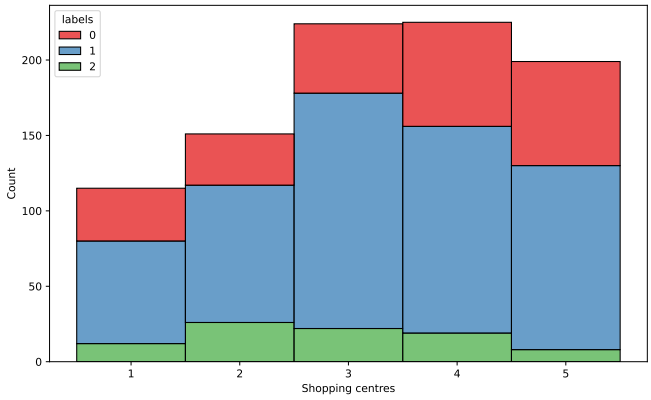

Figura 3.22: Histograma da distribuição da tendência a gostar de Shoppings por grupo

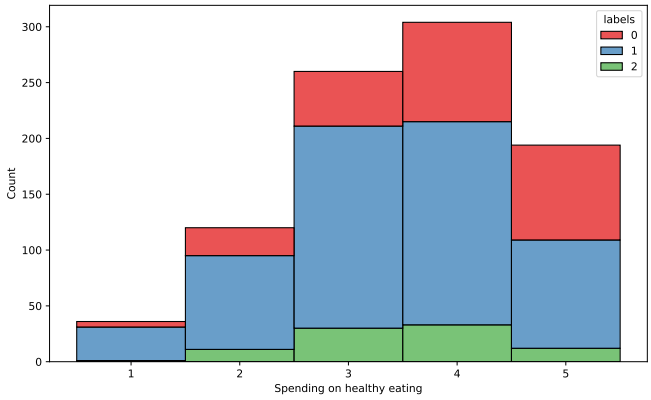

Figura 3.24: Histograma da distribuição da tendência a gastar com comidas mais sudáveis ou gourmet por grupo

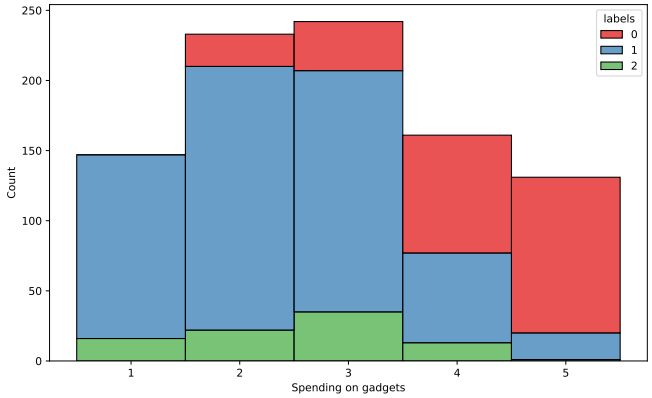

Figura 3.25: Histograma da distribuição da tendência a gastar com gadgets e aparelhos eletrônicos por grupo

Com base nos resultados obtidos, podemos traçar os seguintes perfis para cada cluster:

- O primeiro cluster, representado em vermelho, é caracterizado por pessoas mais jovens, com 20 anos em média e é formado majoritariamente por homens que não chegaram a cursar o ensino superior. Quanto aos seus hábitos financeiros, identificamos que o grupo tende a se importar menos com suas despesas financeiras no geral, e, de todos os grupos, é o que menos se importa com pagar mais caro para roupas de marca, festas, comidas gourmet ou saudáveis e, sobretudo, tendem a gastar mais com aparelhos eletrônicos;
- Em relação ao segundo cluster, indicado em azul, temos um perfil demográfico similar ao anterior com excessão de ser composto majoritariamente por mulheres. Quanto aos suse hábitos de consumo, esse grupo valoriza moderadamente a poupança de dinheiro e se opõe a pagar mais caro por roupas de marca, festas e comidas gourmet ou saudáveis e, acima de tudo, preferem não pagar muito por aparelhos eletrónicos.
- Por fim, o terceiro grupo, indicado em verde, é o grupo em menor volume de observações. Quanto às suas características demográficas, eles tendem a representar pessoas mais velhas, em média com 25 anos, todos com grau de escolaridade superior aos outros grupos. Já em relação aos seus hábitos de consumo, esse grupo tende a se importar mais com poupar dinheiro e tendem a se opor a pagar mais caro por roupas de marca, festas, idas em Shoppings e aparelhos eletrônicos. No entando, não se importam tanto e pagar mais por comidas saudáveis ou gourmet.
- Vale mencionar que os atributos referentes ao número de irmãos, local onde a pessoa morou na infância e gostar ou não gostar de Shoppings não aparentaram ser tão significativos entre os grupos.

Capítulo 4

Validação e comparação de modelos

Com os modelos elaborados, prosseguiremos para a parte final do projeto, em que avaliaremos os resultados a partir de métricas de desempenho.

De início, selecionamos as métricas relativas Silhouette e Davies-Bouldin, cujos resultados podem ser observados abaixo:

Critérios Relativos		
	K- Means	Agrupamento Hierárquico
Silhouette	0.1481	0.094
Davies-Bouldin	2.09	2.50

Em um primeiro momento notamos os baixos valores para o índice Silhouette, o que é esperado, uma vez que ele apenas capta separações lineares nos dados, resultando em baixos valores quando temos outros tipos de separação ou muitas dimensões. Já para o índice Davies-Bouldin obtivemos um valor mais próximo de zero, o que indica uma melhor separação e coesão dos clusters. De modo geral, o K-Means obteve um resultado levemente superior.

Em seguida, para auxiliar na comparação dos dois modelos, também calulamos o índice Rand utilizando os grupos de um modelo para avaliar o outro.

Validação Externa	
Métrica	Valor
Rand	0.094

O resultado obtido indica que os grupos formados pelos dois métodos não são parecidos. No entanto esse método também está suscetível a limitações, dado que possui um viés de favorecer a comparação de partições com níveis mais elevados de granularidade

Capítulo 5

Conclusão

Concluindo, neste trabalho desenvolvemos e avaliamos dois diferentes modelos de agrupamento para agrupar uma população de jovens com base em seus dados demográficos e seus hábitos financeiros. A partir de uma base de dados com mais de 1010 observações e 150 atributos, realizamos o pre-processamento dos dados, incluindo a remoção de valores faltantes e outliers, a agregação de atributos e a seleção e redução de atributos relevantes para a análise.

Em seguida, aplicamos dois modelos de classificação: o K-Means e o Agrupamento Hierárquico Aglomerativo. Para o K-Means, testamos diferentes valores de K e escolhemos o valor que apresentou melhor inércia pelo método do cotovelo. Já no modelo via Agrupamento Hierárquico Aglomerativo, utilizamos o dendograma para identificar possíveis partições.

Após o treinamento dos modelos, realizamos a validação e comparação dos mesmos. Os resultados obtidos indicam que os modelos apresentaram resultados similares entre si nas métricas de validação interna, ao mesmo tempo que não aparentam ter formado clusters parecidos pela métrica de validação extrena Rand.

Portanto, concluímos que os dois modelos são igualmente adequados para o agrupamento de indivíduos com base em sey perfil socioeconômico. No entanto, é importante ressaltar que todos os modelos apresentaram métricas modestas, indicando que outros fatores podem influenciar esses resultados e que a diferença entre os clusters identificados em ambos os modelos pode não ser significativa.

Referências Bibliográficas

[1] MIROSLAV SABO. Young People Survey. kaggle, 2016. DOI:https://www.kaggle.com/datasets/miroslavsabo/young-people-survey.