

SCC-5871

Clusterização de clientes e planejamento de *marketing*



Customer Personality Analysis

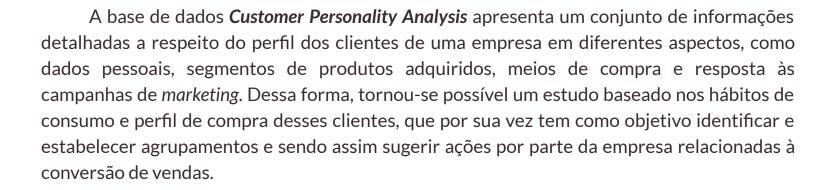


Profa. Dra. Roseli Aparecida Romero

Julyana Flores de Prá

Thiago Rafael Mariotti Claudio







Tópicos abordados



O1 Pré-processamento

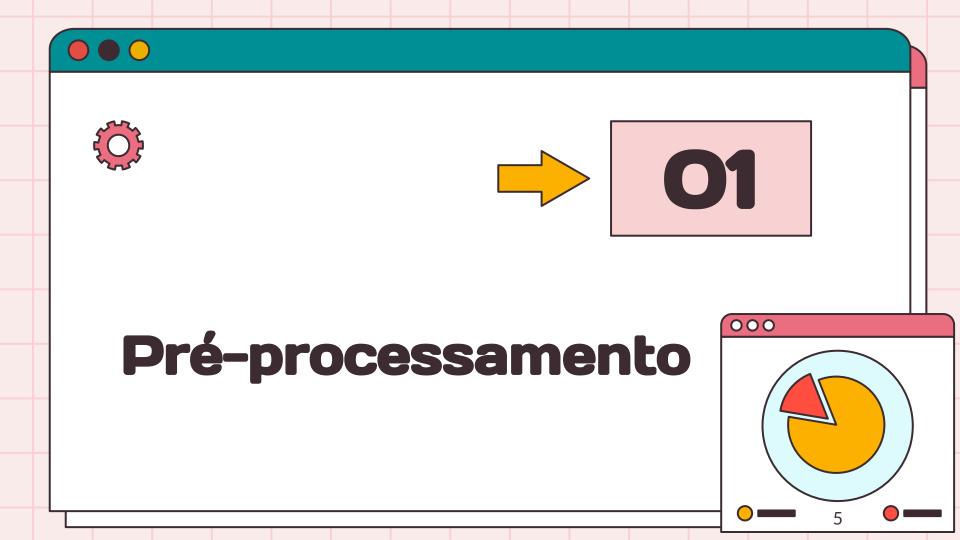
02

Análise Exploratória

O3 Experimentos

04

Conclusão





Características gerais



O dataset apresenta:

- 29 colunas;
- 2240 registros;
- Os dados do dataset foram divididos em 4 grupos pelo autor: People, Products, Promotion e Place;
- Cada registro representa um ID de cliente único, o que significa que não há dados repetidos.



Tranformações realizadas





Inicialmente só havia a *feature* "Year_Birthday" que foi utilizada para descobrir a idade de cada cliente, criando uma nova *feature* "Age".

Tratamento de valores faltantes

A única feature que continha valores *NaN* era a "*Income*" (renda do cliente) e isso foi tratado com o método de interpolação.

Formatação de data

A data de cadastro do cliente, disponível na feature "Dt_Customer", era um object e foi formatada para datetime.



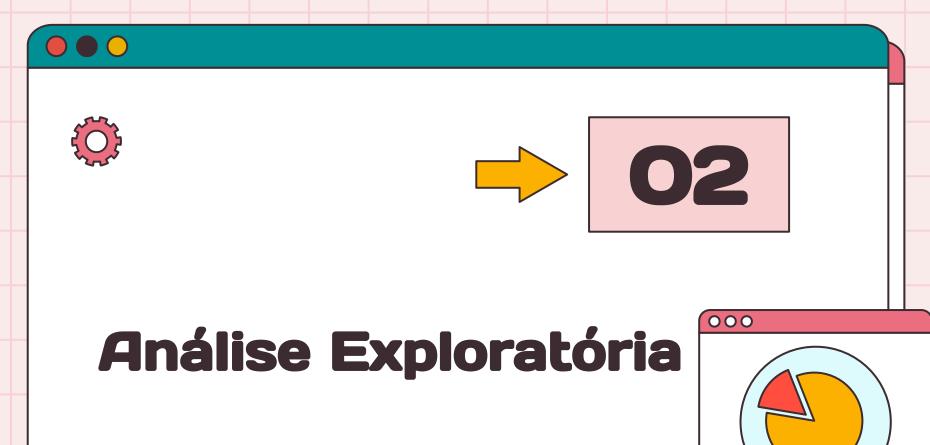
Outras transformações

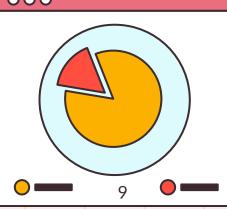


Geração de dados relevantes

Somatório de gastos, quantidade de dependentes (crianças e adolescentes), tamanho da familia e outras transformações.

```
#coluna tratando a info de pessoal vivendo sozinha ou com parceirx
df["HasPartner"] = df["Marital_Status"].replace({
                                                                                                                                                                                "Single": 1,
                                                                                                                                                                                "Divorced": 1.
                                                                                                                                                                                "Widow":1,
                                                                                                                                                                                "Together": 2,
                                                                                                                                                                                "Married": 2
                                                                                                                                                                 }).infer objects(copy=False)
#tratando dados formação escolar
df["Education Code"]=df["Education"].replace({
                                                                                                                                                    "Basic":1,
                                                                                                                                                     "Graduation":2,
                                                                                                                                                     "2n Cycle":3,
                                                                                                                                                     "Master":4,
                                                                                                                                                     "PhD":5
                                                                                                                                                  }).infer_objects(copy=False)
#criando coluna total gasto
df["TotalSpenses"] = df["MntWines"]+ df["MntFruits"]+ df["MntMeatProducts"]+ df["MntFishProducts"]+ df["MntSweetProducts"]+ df
#coluna somando os pirralhos
df["Dependants"]=df["Kidhome"]+df["Teenhome"]
#coluna total de pessoas em casa
df["FamilySize"] = df["HasPartner"] + df["Dependents"]
```







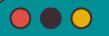
Correlação dos dados

Corr	elação
0.	891839
0.	849596
0.	842965
0.	778577
e 0.	739934
0.	723827
0.	698433
0.	689971
0.	674669
0.	667576

	Matriz de Correlação (Heatmap)																																
Education	1.0	0.0	0.1	-0.0	0.1	-0.1	-0.0	0.2	-0.1	0.0	-0.1	-0.1	-0.1	0.0	0.1	0.1	0.1	-0.0	0.0	0.1	0.0	-0.0	0.0	-0.1	0.1	0.2	-0.0	0.7	0.1	0.1	0.0		1.0
Marital_Status	0.0	1.0	0.0	-0.0	-0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.0	-0.0	0.0	-0.0	-0.0	-0.0	0.0	0.0	-0.0	0.0	-0.0	-0.0	0.1	0.1	0.0	0.0	-0.0	0.0		
Income	0.1	0.0	1.0	-0.4	0.0	-0.0	-0.0	0.6	0.4	0.6	0.4	0.4	0.3	-0.1	0.4	0.6	0.5	-0.6	-0.0	0.2	0.3	0.3	0.1	-0.0	0.1	0.2	0.0	0.1	0.7	-0.3	-0.2		
Kidhome	-0.0	-0.0	-0.4	1.0	-0.0	-0.1	0.0	-0.5	-0.4	-0.4	-0.4	-0.4	-0.3	0.2	-0.4	-0.5	-0.5	0.4	0.0	-0.2	-0.2	-0.2	-0.1	0.0	-0.1	-0.2	0.0	-0.0	-0.6		0.6		0.0
Teenhome	0.1	-0.0	0.0	-0.0	1.0	0.0	0.0	0.0	-0.2			-0.2	-0.0	0.4	0.2	-0.1	0.1	0.1	-0.0	0.0	-0.2	-0.1	-0.0	0.0	-0.2	0.4	0.0	0.1	-0.1	0.7	0.6		0.8
Dt_Customer	-0.1	0.0	-0.0	-0.1	0.0	1.0	0.0	0.2	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.1	0.1	0.3	-0.0	0.0	-0.0	-0.0	0.0	0.0	0.2	-0.0	-0.0	-0.1	0.2	-0.0	-0.0		
Recency	-0.0	0.0	-0.0	0.0	0.0	0.0	1.0	0.0	-0.0	0.0	0.0	0.0	0.0	-0.0	-0.0	0.0	0.0	-0.0	-0.0	0.0	0.0	-0.0	-0.0	0.0		0.0	-0.0	-0.0	0.0	0.0	0.0		
MntWines	0.2	0.0	0.6	-0.5	0.0	0.2	0.0	1.0	0.4	0.6	0.4	0.4	0.4	0.0	0.5		0.6	-0.3	0.1	0.4	0.5	0.4	0.2	-0.0	0.2	0.2	-0.0	0.2	0.9	-0.4	-0.3		0.6
MntFruits	-0.1	0.0	0.4	-0.4	-0.2	0.1	-0.0	0.4	1.0	0.5	0.6	0.6	0.4	-0.1	0.3	0.5	0.5	-0.4	0.0	0.0	0.2	0.2	-0.0	-0.0	0.1	0.0	-0.0	-0.1	0.6	-0.4	-0.3		0.6
MntMeatProducts	0.0	0.0	0.6	-0.4	-0.3	0.1	0.0	0.6	0.5	1.0	0.6	0.5	0.4	-0.1	0.3	0.7	0.5	-0.5	0.0	0.1	0.4	0.3	0.0	-0.0	0.2	0.0	-0.0	0.0	8.0	-0.5	-0.4		
MntFishProducts	-0.1	0.0	0.4	-0.4		0.1	0.0	0.4	0.6	0.6	1.0	0.6	0.4	-0.1	0.3	0.5	0.5	-0.4	0.0	0.0	0.2	0.3	0.0	-0.0	0.1	0.0	-0.0	-0.1	0.6	-0.4	-0.4		
MntSweetProducts	-0.1	0.0	0.4	-0.4	-0.2	0.1	0.0	0.4	0.6	0.5	0.6	1.0	0.4	-0.1	0.3	0.5	0.4	-0.4	0.0	0.0	0.3	0.2	0.0	-0.0	0.1	0.0	-0.0	-0.1	0.6	-0.4	-0.3		0.4
MntGoldProds	-0.1	0.0	0.3	-0.3	-0.0	0.2	0.0	0.4	0.4	0.4	0.4	0.4	1.0	0.0	0.4	0.4	0.4	-0.3	0.1	0.0	0.2	0.2	0.0	-0.0	0.1	0.1	-0.0	-0.1	0.5	-0.3	-0.2		0.4
NumDealsPurchases	0.0	-0.0	-0.1	0.2	0.4	0.2	-0.0	0.0	-0.1	-0.1	-0.1	-0.1	0.0	1.0	0.2	-0.0	0.1	0.3	-0.0	0.0	-0.2	-0.1	-0.0	0.0	0.0	0.1	0.0	0.0	-0.1	0.4	0.4		
NumWebPurchases	0.1	-0.0	0.4	-0.4	0.2	0.2	-0.0	0.5	0.3	0.3	0.3	0.3	0.4	0.2	1.0	0.4	0.5	-0.1	0.0	0.2	0.1	0.2	0.0	-0.0	0.1	0.1	0.0	0.1	0.5	-0.1	-0.1		
NumCatalogPurchases	0.1	0.0	0.6	-0.5	-0.1	0.1	0.0	0.6	0.5	0.7	0.5	0.5	0.4	-0.0	0.4	1.0	0.5	-0.5	0.1	0.1	0.3	0.3	0.1	-0.0	0.2	0.1	-0.0	0.1	8.0	-0.4	-0.4		0.2
NumStorePurchases	0.1	-0.0	0.5	-0.5	0.1	0.1	0.0	0.6	0.5	0.5	0.5	0.4	0.4	0.1	0.5	0.5	1.0	-0.4	-0.1	0.2	0.2	0.2	0.1	-0.0	0.0	0.1	0.0	0.1	0.7	-0.3	-0.3		0.2
NumWebVisitsMonth	-0.0	-0.0	-0.6	0.4	0.1	0.3	-0.0	-0.3	-0.4	-0.5	-0.4	-0.4		0.3	-0.1	-0.5	-0.4	1.0	0.1	-0.0	-0.3		-0.0	0.0	-0.0	-0.1	0.0	-0.0	-0.5	0.4	0.3		
AcceptedCmp3	0.0	-0.0	-0.0	0.0	-0.0	-0.0	-0.0	0.1	0.0	0.0	0.0	0.0	0.1	-0.0	0.0	0.1	-0.1	0.1	1.0	-0.1	0.1	0.1	0.1	0.0	0.3	-0.1	-0.0	0.0	0.1	-0.0	-0.0		
AcceptedCmp4	0.1	0.0	0.2	-0.2	0.0	0.0	0.0	0.4	0.0	0.1	0.0	0.0	0.0	0.0	0.2	0.1	0.2	-0.0	-0.1	1.0	0.3	0.3	0.3	-0.0	0.2	0.1	-0.0	0.0	0.3	-0.1	-0.1		0.0
AcceptedCmp5	0.0	0.0	0.3	-0.2	-0.2	-0.0	0.0	0.5	0.2	0.4	0.2	0.3	0.2	-0.2	0.1	0.3	0.2	-0.3	0.1	0.3	1.0	0.4	0.2	-0.0	0.3	-0.0	0.0	0.0	0.5	-0.3	-0.2		0.0
AcceptedCmp1	-0.0	-0.0	0.3	-0.2	-0.1	-0.0	-0.0	0.4	0.2	0.3	0.3	0.2	0.2	-0.1	0.2	0.3	0.2		0.1	0.3	0.4	1.0	0.2	-0.0	0.3	0.0	0.0	-0.0	0.4	-0.2	-0.2		
AcceptedCmp2	0.0	0.0	0.1	-0.1	-0.0	0.0	-0.0	0.2	-0.0	0.0	0.0	0.0	0.0	-0.0	0.0	0.1	0.1	-0.0	0.1	0.3	0.2	0.2	1.0	-0.0	0.2	0.0	-0.0	0.0	0.1	-0.1	-0.1		
Complain	-0.1	-0.0	-0.0	0.0	0.0	0.0	0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	0.0	-0.0	-0.0	-0.0	0.0	0.0	-0.0	-0.0	-0.0	-0.0	1.0	-0.0	0.0	-0.0	-0.0	-0.0	0.0	0.0		-0.2
Response	0.1	-0.0	0.1	-0.1	-0.2	0.2	-0.2	0.2	0.1	0.2	0.1	0.1	0.1	0.0	0.1	0.2	0.0	-0.0	0.3	0.2	0.3	0.3	0.2	-0.0	1.0	-0.0	-0.2	0.1	0.3	-0.2	-0.2		
Age	0.2	0.1	0.2	-0.2	0.4	-0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.1	-0.1	-0.1	0.1	-0.0	0.0	0.0	0.0	-0.0	1.0	-0.0	0.2	0.1	0.1	0.1		
HasPartner	-0.0	0.1	0.0	0.0	0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	0.0	0.0	-0.0	0.0	0.0	-0.0	-0.0	0.0	0.0	-0.0	-0.0	-0.2	-0.0	1.0	0.0	-0.0	0.0	0.6		
Education_Code	0.7	0.0	0.1	-0.0	0.1	-0.1	-0.0	0.2	-0.1	0.0	-0.1	-0.1	-0.1	0.0	0.1	0.1					0.0	-0.0	0.0	-0.0	0.1	0.2	0.0	1.0	0.1	0.0	0.0		-0.4
TotalSpenses	0.1	0.0	0.7	-0.6	-0.1	0.2	0.0	0.9	0.6	8.0		0.6	0.5	-0.1	0.5	8.0	0.7	-0.5	0.1	0.3	0.5	0.4	0.1	-0.0	0.3	0.1	-0.0	0.1	1.0	-0.5	-0.4		
Dependants	0.1	-0.0	-0.3	0.7	0.7	-0.0	0.0	-0.4	-0.4	-0.5	-0.4	-0.4	-0.3	0.4	-0.1	-0.4	-0.3	0.4	-0.0	-0.1	-0.3	-0.2	-0.1	0.0	-0.2	0.1	0.0	0.0	-0.5	1.0	8.0		
FamilySize	0.0	0.0	-0.2	0.6	0.6	-0.0	0.0	-0.3	-0.3	-0.4	-0.4	-0.3	-0.2	0.4	-0.1	-0.4	-0.3	0.3	-0.0	-0.1	-0.2	-0.2	-0.1	0.0	-0.2	0.1	0.6	0.0	-0.4	8.0	1.0		
	ution	atns	Income	Kidhome	ome	mer	Recency	ines	ruits	MntMeatProducts	MntFishProducts	ncts	MntGoldProds	ses	NumWebPurchases	NumCatalogPurchases	NumStorePurchases	onth	AcceptedCmp3	mp4	AcceptedCmp5	mp1	mp2	Complain	ouse	Age	tner	opo	Ses	Dependants	Size		
	Education	Marital Status	ll c	Kidh	Feenhome	Customer	Rec	MntWines	MntFruits	Proc	Proc	MntSweetProducts	Aplo	NumDealsPurchases	urch	urch	urch	NumWebVisitsMonth	tedC	AcceptedCmp4	tedC	AcceptedCmp1	AcceptedCmp2	Com	Response		HasPartner	Education_Code	TotalSpenses	pend	FamilySize		
	ш	farit			1	d		~		Meat	Fish	weet	IntG	alsP	ebP	OgP	oreP	sbVis	deoc	deoc	deoc	coep	deoc		LE.		포	icatii	Total	Del	Lo		
		2								Mntl	Mnt	ntS	2)De	Mm	atal	nStc	nWe	A	A	A	A	A					Edi					
										_		Σ		Nun	ž	Dmn	N	Nu															
																Ż																	

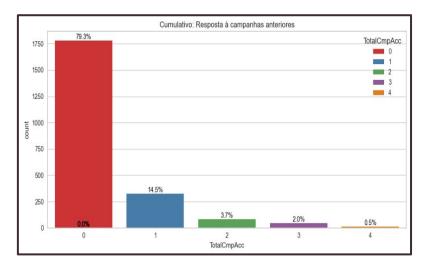


Figuras 1: Matriz de correlação.

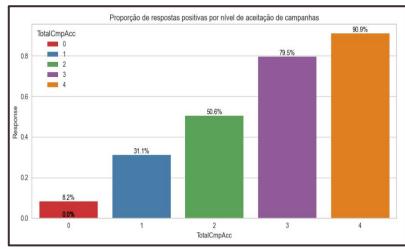


Campanhas de marketing

Podemos observar que a taxa de aceitação de campanhas é cumulativa, isto é, clientes que aceitaram mais campanhas no passado tem maior propensão à resposta positiva no futuro.



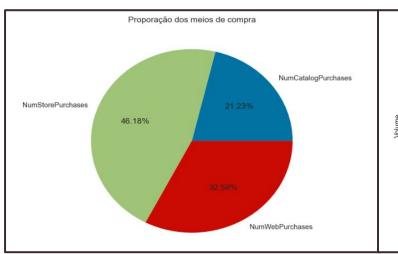
Figuras 2: Distribuição de clientes que aceitaram campanhas cumulativamente.

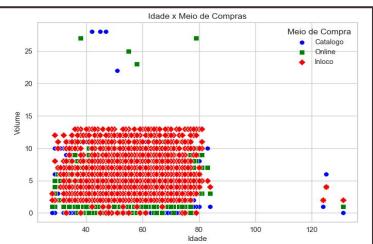


Figuras 3: Proporção de respostas positivas à campanha vigente do cumulativo.

Hábitos de Consumo

Pode-se observar através dos gráficos acima que a faixa etária dos clientes é majoritariamente 40+ e que em geral os clientes costumam fazer suas compras em lojas físicas.





Figuras 4: Proporção de uso dos meios de compras.

Figuras 5: Distribuição dos meios de compras x idade.



🔴 🌑 🔵 💮 Aceitação de Campanha x Compras em Promoções

Através do boxplot abaixo, pode-se observar que quanto mais campanhas aceitas pelo cliente, menos compras com desconto foram realizadas. Esse dado sugere que não necessariamente as ações de marketing estão relacionadas com ações de promoção por parte da empresa.

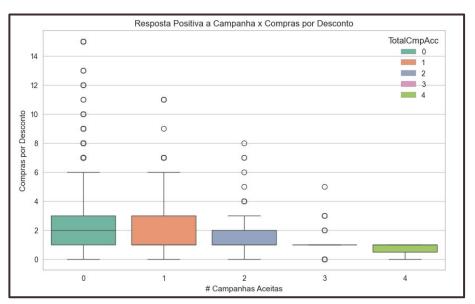
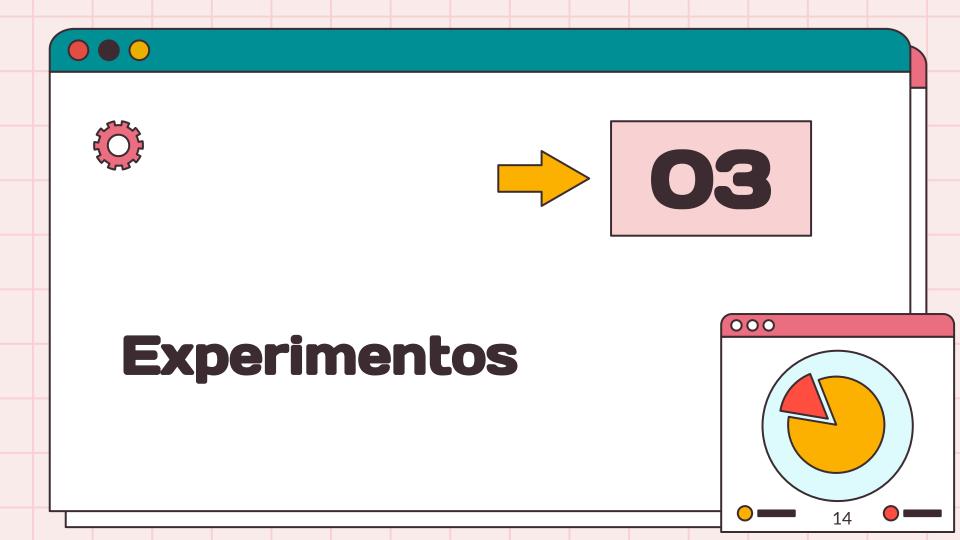


Figura 4: Boxplot: resposta à campanha x volume de compras em campanhas de desconto.



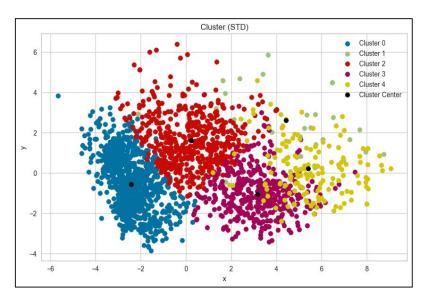




Clusterização

- 1. Controle: sem preferência.
- 2. Por Estilo de Vida: Agrupamento por quantidade de dependentes, estado civil, grau de formação e renda.
- 3. Por Hábitos de Compra: Agrupamento por gastos em produtos, gastos totais, meios de compra, volume de compra, interações on-line e compras em promoções.
- **4. Por Aceitação da Campanha:** Agrupamento por cumulativo de resposta à campanhas, respostas individuais.

Clusterização - Controle



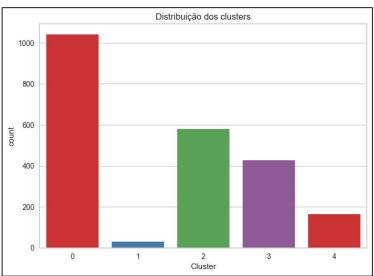


Figura 5: Visualização do cluster - grupo de controle..

Clusterização

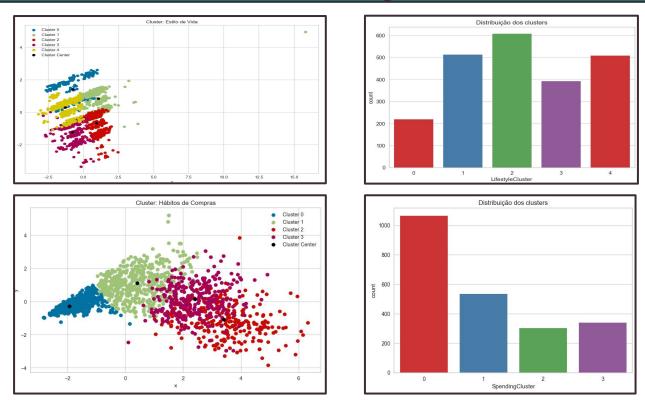
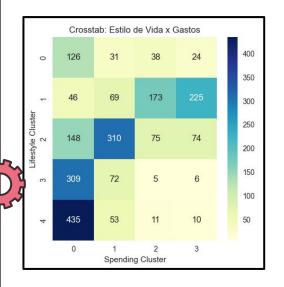
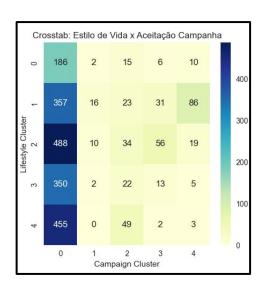


Figura 6: Clusters de mais relevantes - Estilo de Vida e Hábito de Consumo.

Cluster Crosstabing





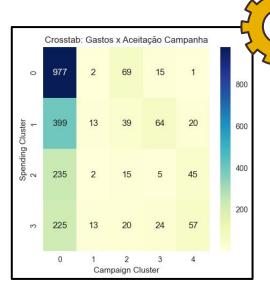


Figura 7: Crosstabing dos clusters implementados.

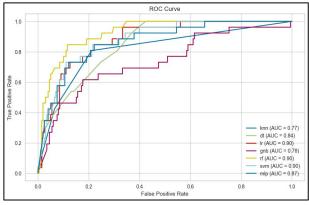
Classificação

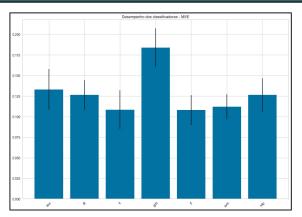
- Atributo Alvo: Response resposta (binária) à campanha vigente.
- Dataframes usados: Clusterês, Controle.
- Adendos: redução dimensional, k-Fold = 2 .. 20.
- Treino/Teste: 80/20, 70/30.

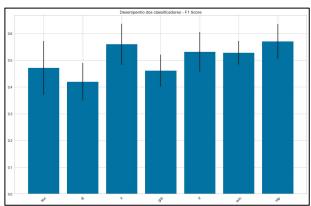
```
knn = KNeighborsClassifier(n_neighbors=3)
dt = DecisionTreeClassifier(criterion='gini', splitter='best', min_samples_split=int(len(data)*0.1))
lr = LogisticRegression(solver='lbfgs', max_iter=1000, n_jobs=-1)
gnb = GaussianNB()
rf = RandomForestClassifier(n_estimators=100, random_state=14)
svm = SVC(kernel='linear', probability=True)
mlp = MLPClassifier(solver='lbfgs', alpha=1e-5, hidden_layer_sizes=(100,), random_state=1)
```

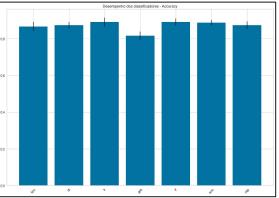
Figura 8: Classificadores implementados e seus respectivos parâmetros.

Classificação: Controle

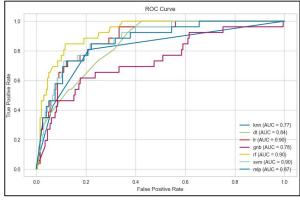


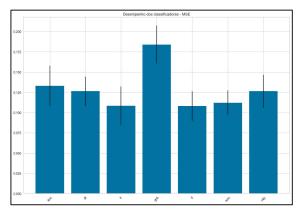


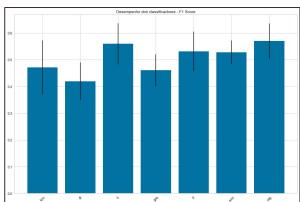


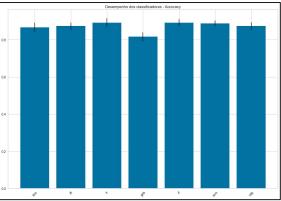


Classificação: Estilo de Vida

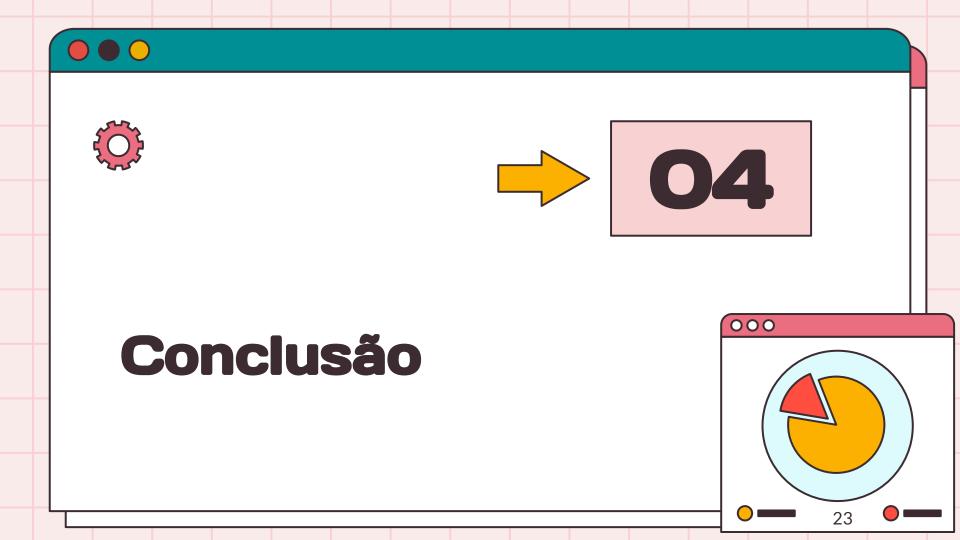








Classificação: Hábitos de Compra knn (AUC = 0.77) svm (AUC = 0.90) mlp (AUC = 0.87) False Positive Rate





- Observa-se bons resultados na classificação da Resposta à Campanha
 Vigente ao utilizar técnicas de clusterização.
 - Métricas que consideram segmentação no Estilo de Vida e Hábitos de Consumo.
- Algoritmo de Random Forest apresentou melhores resultados em detrimento aos outros.
- Justifica-se a segmentação dos clientes para classificação com a premissa de identificação e direcionamento de campanhas de venda.
- Melhoria no clustering: balanceamento dos agrupamentos e clusterização por Random Forest Embedding (literatura).
- Melhoria na classificação: Experimentos com outras técnicas.

