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

Introdução

A base de dados ***Customer Personality Analysis*** apresenta um conjunto de informações detalhadas a respeito do perfil dos clientes de uma empresa em diferentes aspectos, como dados pessoais, segmentos de produtos adquiridos, meios de compra e resposta às campanhas de *marketing*. Dessa forma, tornou-se possível um estudo baseado nos hábitos de consumo e perfil de compra desses clientes, que por sua vez tem como objetivo identificar e estabelecer agrupamentos e sendo assim sugerir ações por parte da empresa relacionadas à conversão de vendas.

Tópicos abordados

01 Pré-processamento

02 Análise Exploratória

03 Experimentos

04 Conclusão

01

Pré-processamento

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

Idade dos clientes

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.

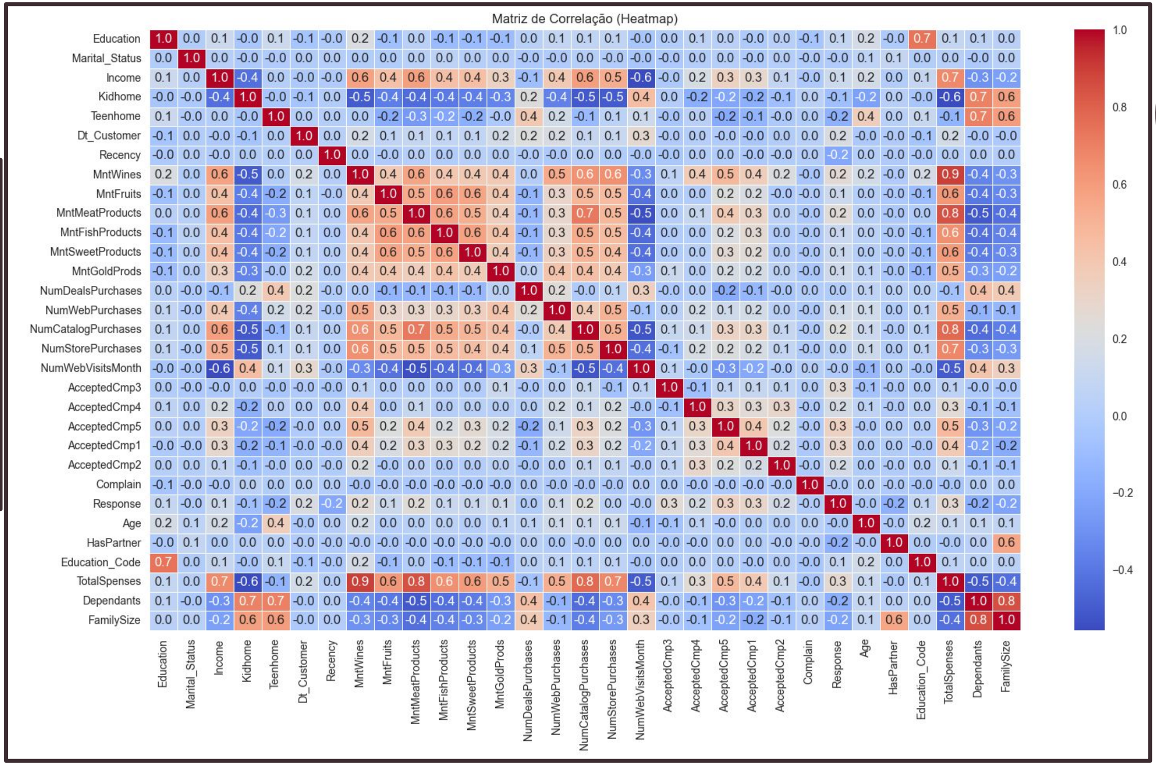
```
#coluna tratando a info de pessoal vivendo sozinha ou com parceirx
df["HasPartner"] = df["Marital_Status"].replace({
                                                "Single": 1,
                                                "Divorced": 1,
                                                "Widow":1,
                                                "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["MntGoldProds"]
#coluna somando os pirralhos
df["Dependants"]=df["Kidhome"]+df["Teenhome"]
#coluna total de pessoas em casa
df["FamilySize"] = df["HasPartner"] + df["Dependants"]
```

02

Análise Exploratória

Correlação dos dados

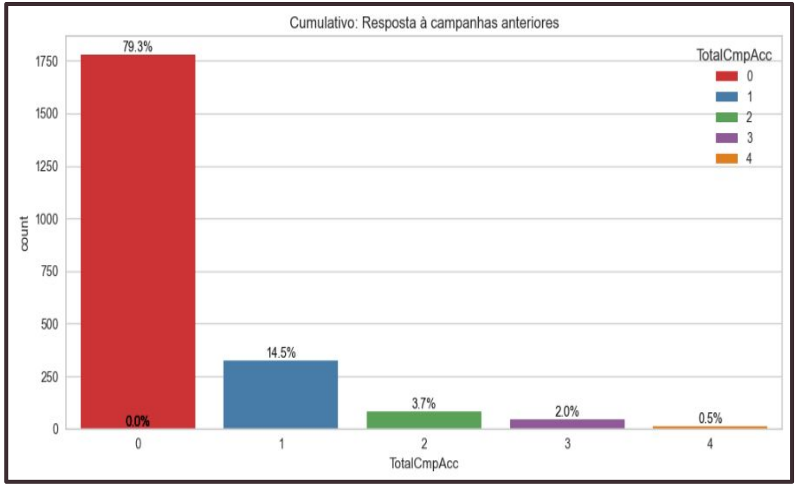
		Correlação
TotalSpenses	MntWines	0.891839
Dependants	FamilySize	0.849596
TotalSpenses	MntMeatProducts	0.842965
	NumCatalogPurchases	0.778577
Education	Education_Code	0.739934
MntMeatProducts	NumCatalogPurchases	0.723827
Teenhome	Dependants	0.698433
Dependants	Kidhome	0.689971
NumStorePurchases	TotalSpenses	0.674669
TotalSpenses	Income	0.667576



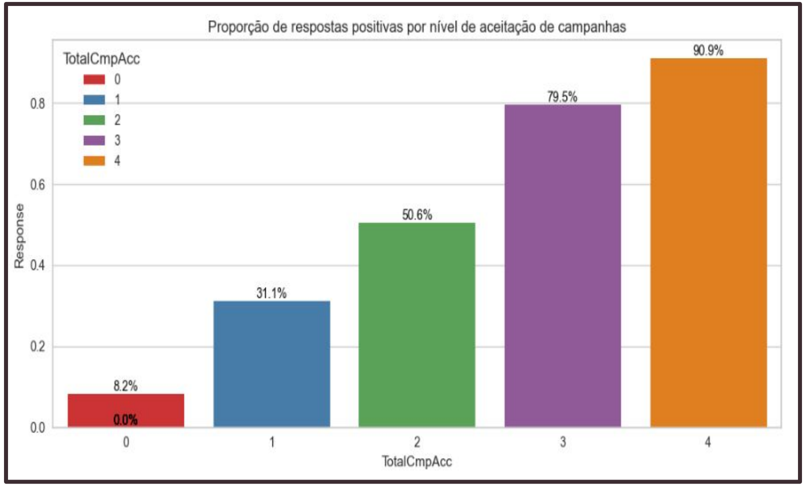
Figuras 1: Matriz de correlação.

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.



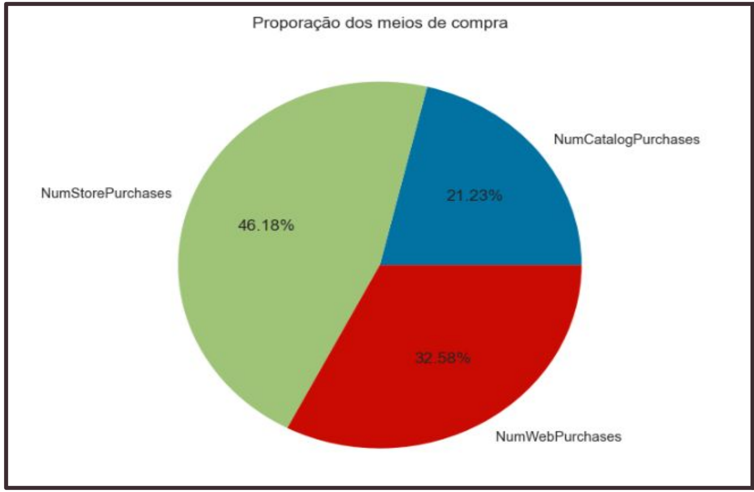
Figuras 2: Distribuição de clientes que aceitaram campanhas cumulativamente.

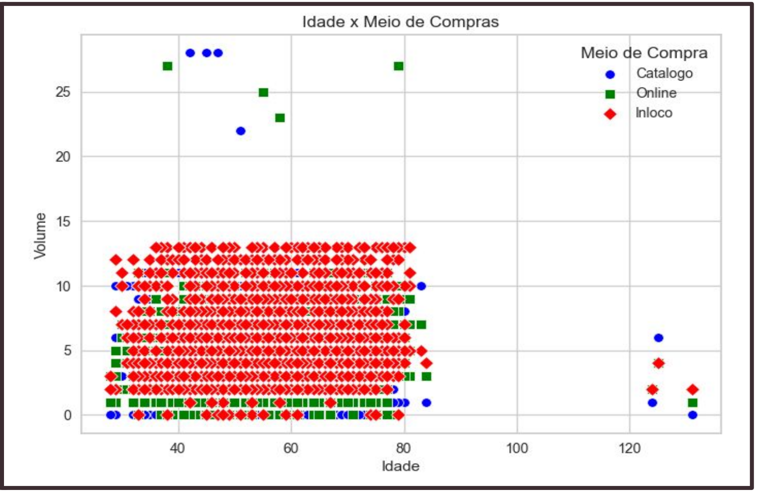


Figuras 3: Proporção de respostas positivas à campanha vigente do cumulativo.

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.

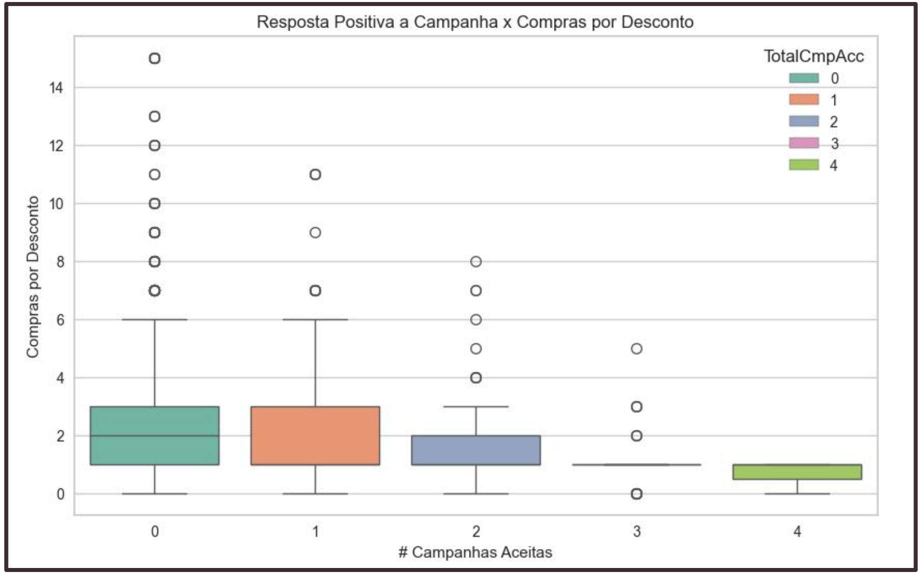


Figura 4: Boxplot: resposta à campanha x volume de compras em campanhas de desconto.

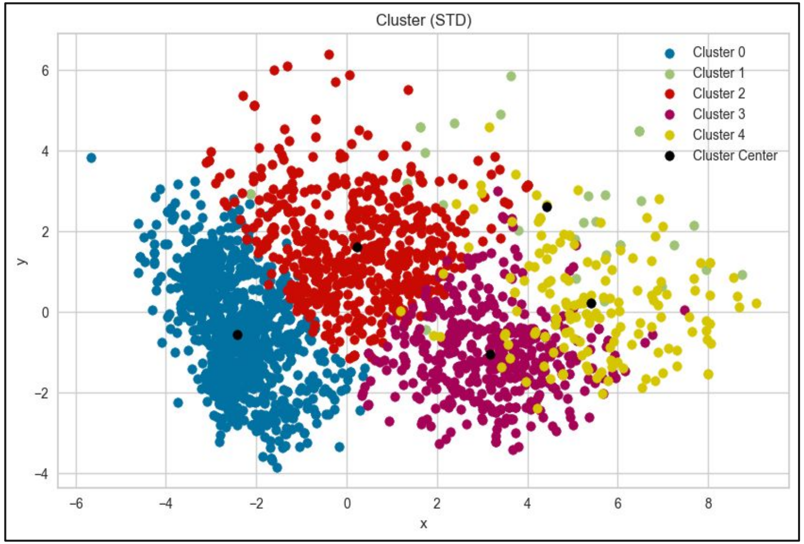
03

Experimentos

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



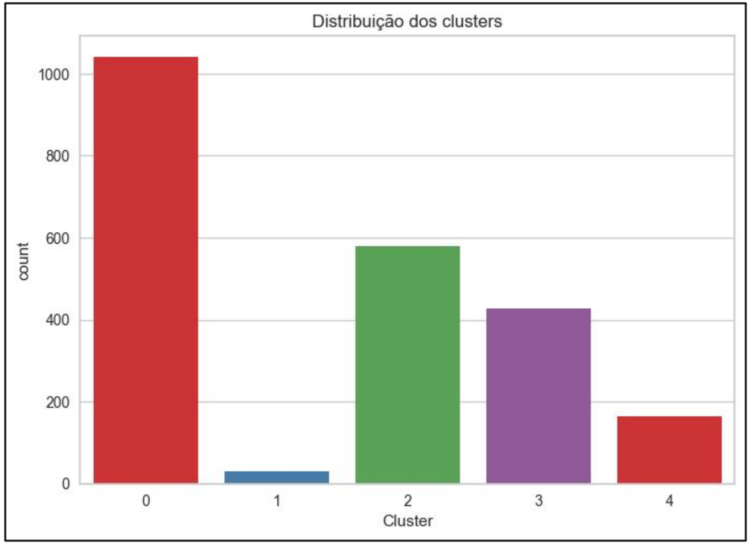


Figura 5: Visualização do cluster - grupo de controle..

Clusterização

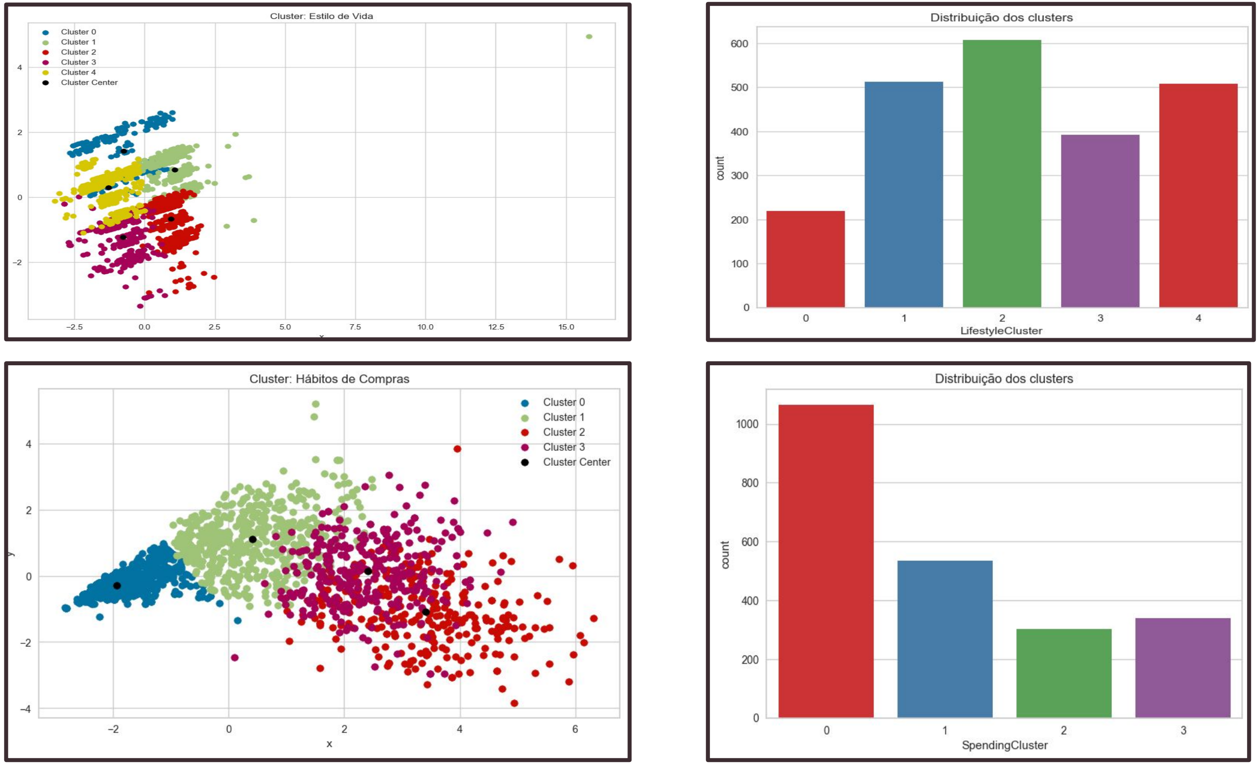


Figura 6: Clusters de mais relevantes - Estilo de Vida e Hábito de Consumo.

Cluster Crosstabing

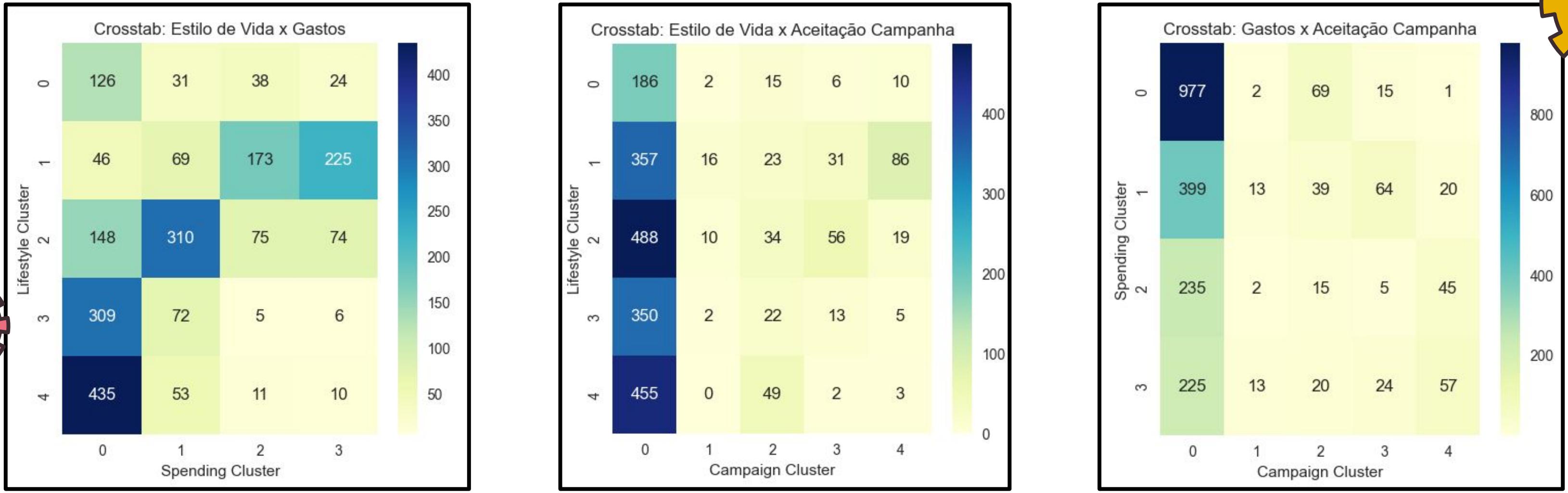


Figura 7: Crosstabing dos clusters implementados.

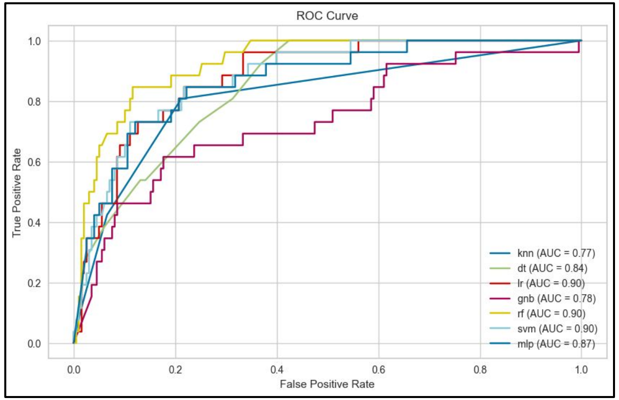
Classificação

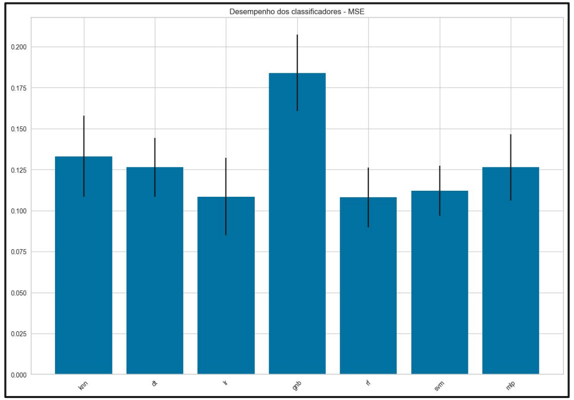
- **Atributo Alvo:** Response - resposta (binária) à campanha vigente.
- **Dataframes usados:** Clusterês, Controle.
- **Adendos:** redução dimensional, k-Fold = 2 .. 20.
- **Treino/Teste:** 80/20, 70/30.

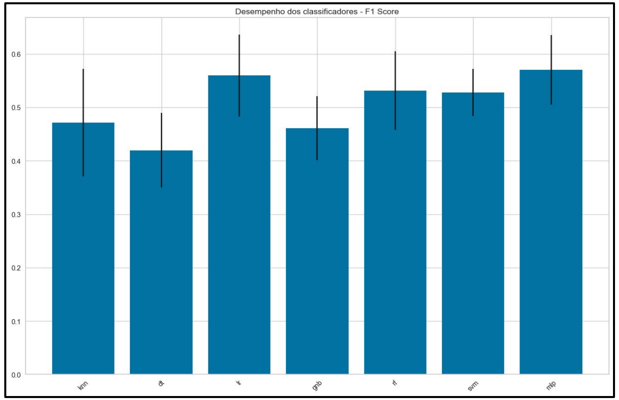
```
knn = KNeighborsClassifier(n_neighbors=3)
dt = DecisionTreeClassifier(criterion='gini', splitter='best', min_samples_split=int(len(data)*0.1))
lr = LogisticRegression(solver='lbfgs', max_iter=1000, n_jobs=-1)
gnb = GaussianNB()
rf = RandomForestClassifier(n_estimators=100, random_state=14)
svm =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

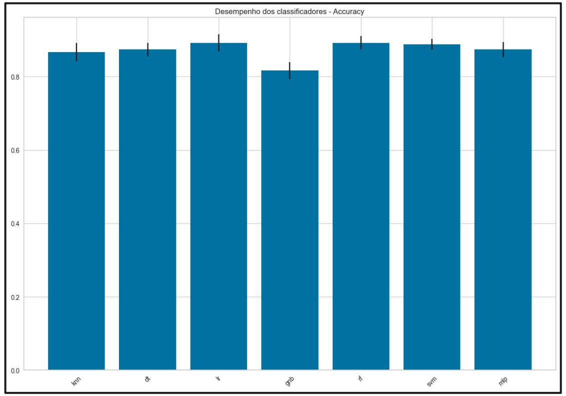
Figura 8: Classificadores implementados e seus respectivos parâmetros.

Classificação: Controle

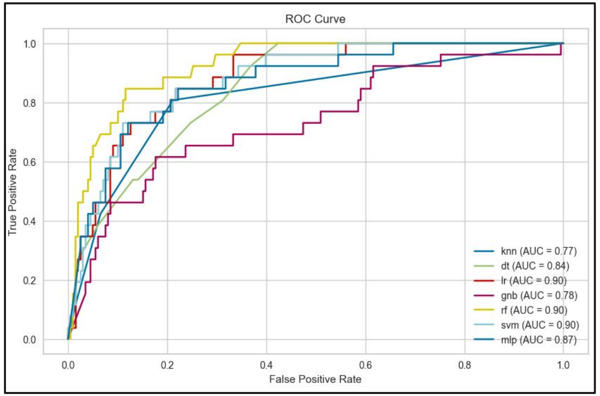


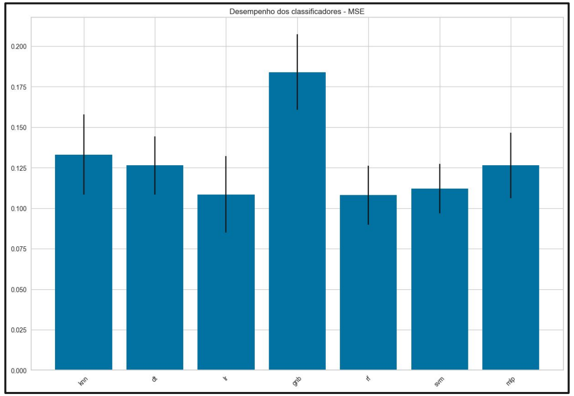


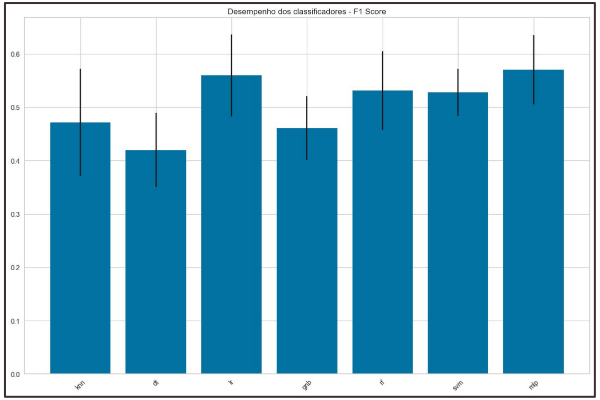


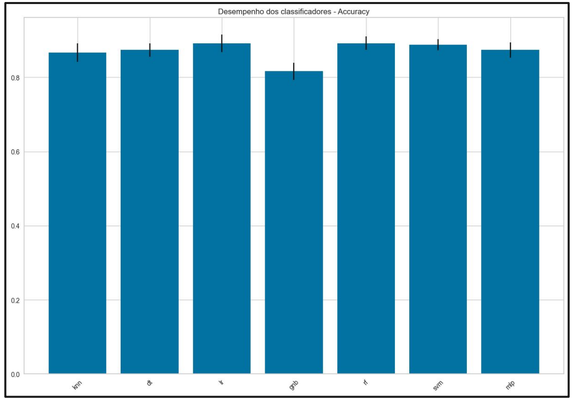


Classificação: Estilo de Vida

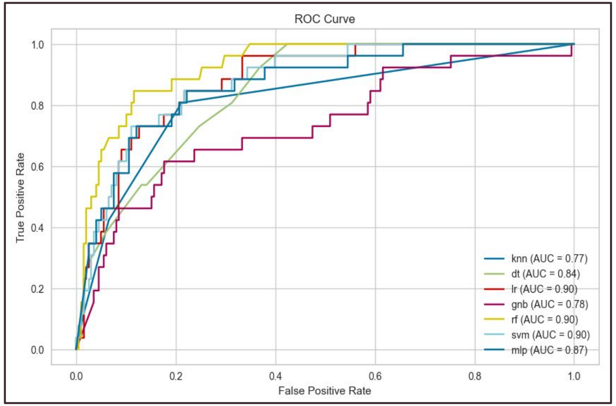


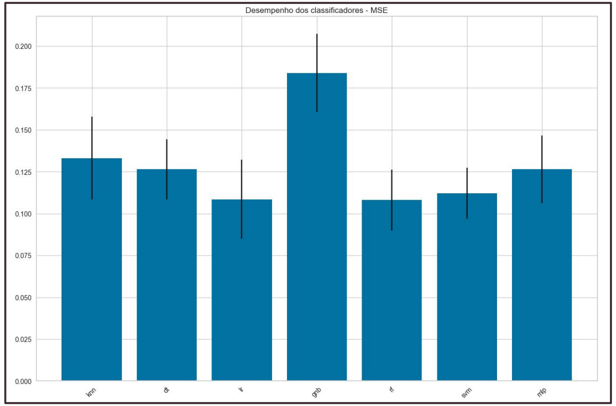


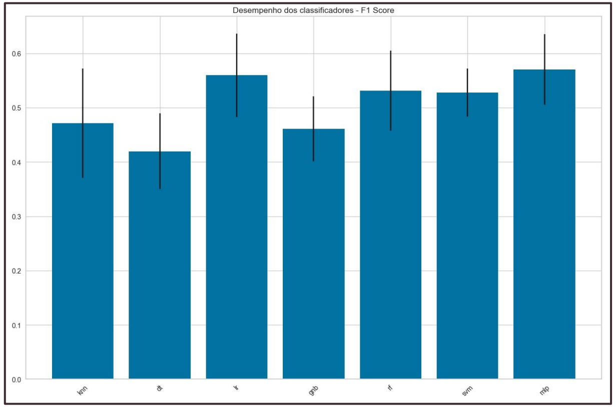


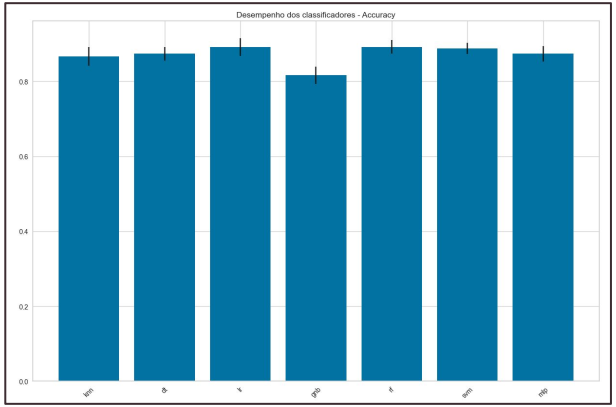


Classificação: Hábitos de Compra









04

Conclusão

- 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.

ICMC - USP

Obrigado!

Julyana Flores de Prá
Thiago Rafael Mariotti Claudio
Junho, 2024