Definição do problema

Em um supermercado na maioria das vezes os clientes que realizam as compras não possuem cadastro com o estabelecimento, com isso não é possível associar os produtos adquiridos com o cliente que está realizando a compra, com isso o perfil de compra do cliente fica inexplorado, deixando a possibilidade de melhorar o atendimento ao cliente.

A utilização de machine learning para segmentação de clientes já foi adotada em outros trabalhos como segmentação de marketing[1], e criação de score de crédito[2]. Ambos os trabalhos utilizam de machine learning para criação de um modelo para segmentar os clientes. De forma similar aos trabalhos citados, o objetivo é criar grupos que possuam comportamento similar de compra, no conjunto de dados adotado não existe nenhum atributo que pode ser utilizado para identificar esses segmentos. Para criar os segmentos de clientes será adotado o algoritmo de aprendizagem não supervisionado de clustering, utilizando o dataset composto pelos seguintes dados: frios_lacteos_congelados, alimento_basico, alimento_industrializado, material_de_limpeza, perfumaria_higiene_pessoal e bebidas. Sendo o atributo alvo será o segmento do cliente que será representado pelo cluster que o dado foi agrupado.

Para realizar a avaliação do avaliação para o modelo por se tratar de uma aprendizagem não supervisionada será utilizado o coeficiente de silhueta que é o algoritmo que consegue avaliar através da distância entre os pontos qual é o melhor número de clusters.

Análise do problema

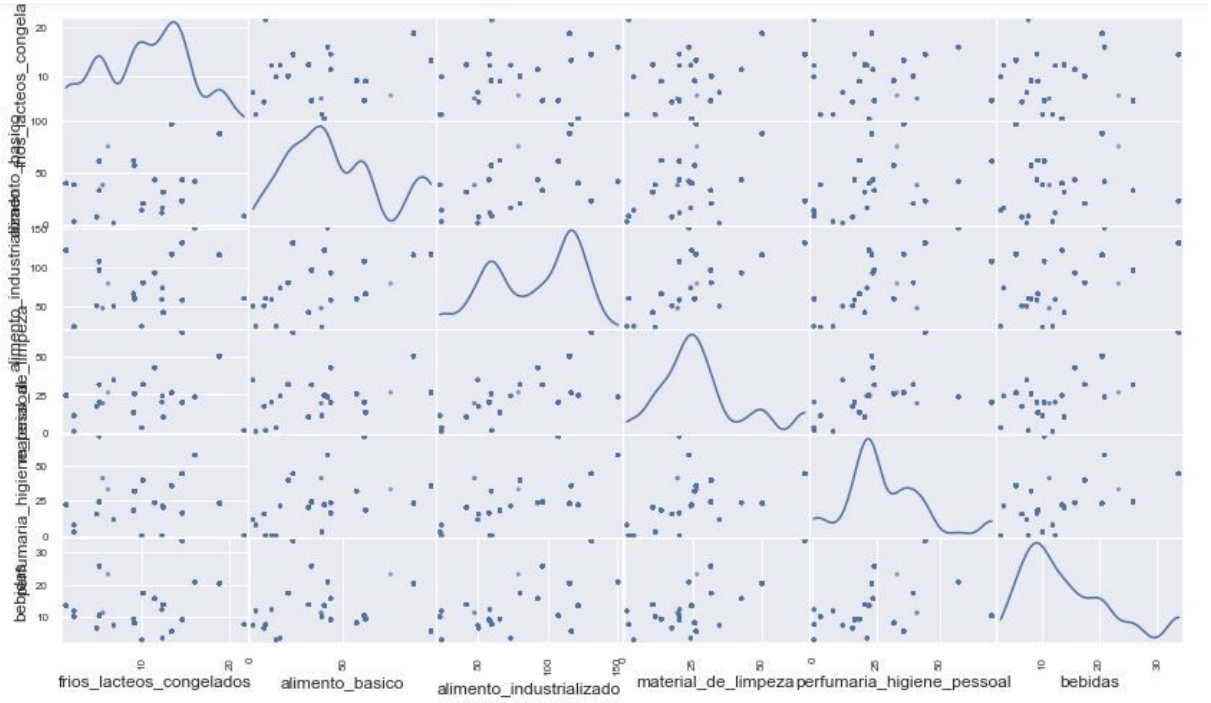
O conjunto de dados disponível em https://www.kaggle.com/owirth/supermarket-sales, é formado por um total de 1000 vendas, os produtos de cada venda foram agrupados em 6 categorias, cada categoria possui a soma do valor dos produtos que compõem a venda, assim o conjunto de dados está organizado com os atributos:

- frios_lacteos_congelados;
- alimento_basico;
- alimento_industrializado;
- material_de_limpeza;
- perfumaria_higiene_pessoal;
- bebidas;

Estatísticas referente a distribuição dos dados podem ser observadas no quadro abaixo.

	frios_lacteos_congelados	alimento_basico	alimento_industrializado	material_de_limpeza	perfumaria_higiene_pessoal	bebidas
count	1000.000000	1000.000000	1000.00000	1000.00000	1000.000000	1000.00000
mean	10.564910	46.941700	89.49046	28.90430	29.231080	14.88266
std	5.295723	27.374487	33.64815	16.66262	16.937287	8.40738
min	1.390000	1.690000	22.52000	1.09000	0.350000	2.97000
25%	5.160000	23.130000	59.39000	20.04000	20.440000	8.97000
50%	10.120000	41.670000	96.33000	25.62000	23.730000	13.55000
75%	14.560000	61.930000	116.98000	31.57000	39.630000	20.46000
max	21.600000	96.930000	150.46000	65.45000	70.610000	33.69000

A distribuição dos dados parece tender a origem.



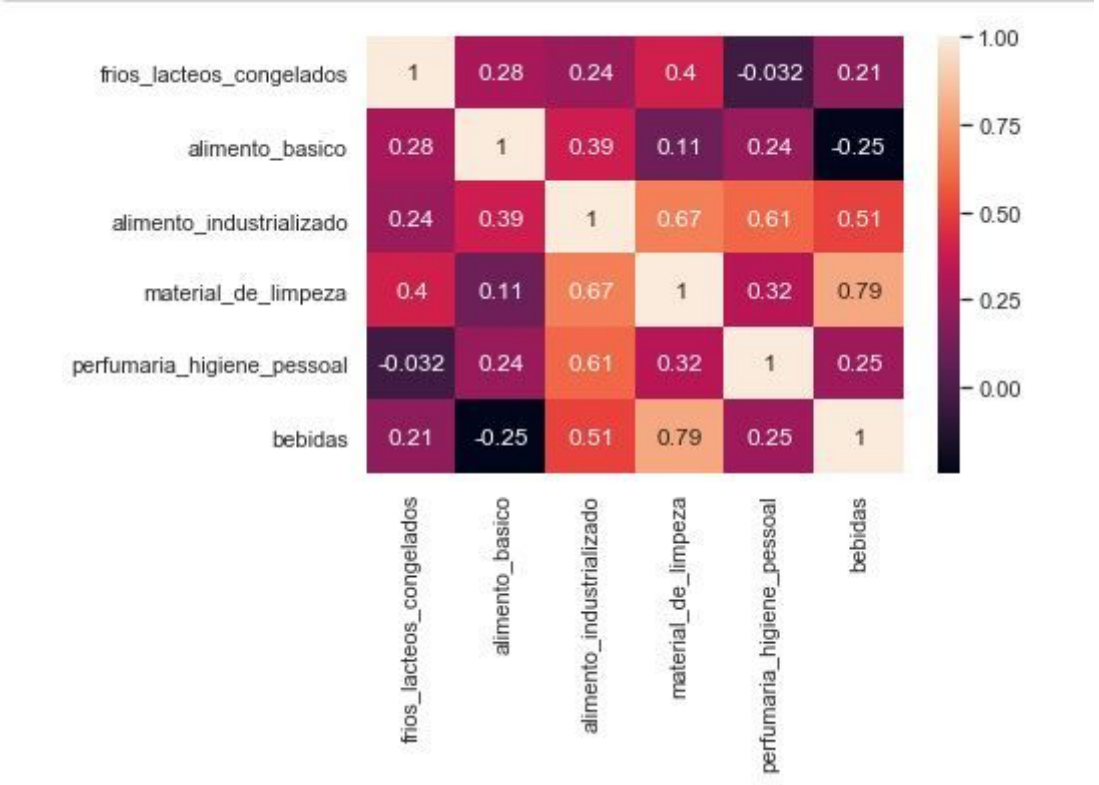
Todos os atributos são numéricos, o atributo alvo não faz parte do conjunto de dados base pois o objetivo é a criação de clusters, o conjunto de dados não possui dados faltantes ou nulos, existe a possibilidade de outliers no conjunto de dados a validação da existência dos outliers será analisada na metodologia.
Para criar os segmentos de clientes irei utilizar algum algoritmo de clustering, pretendo realizar testes utilizando os algoritmos K-Means e Modelo de Mistura Gausiano para ver qual se comporta melhor com os dados. Como modelo de referência será utilizado o algoritmo K-Means com a sua configuração padrão.

Metodologia

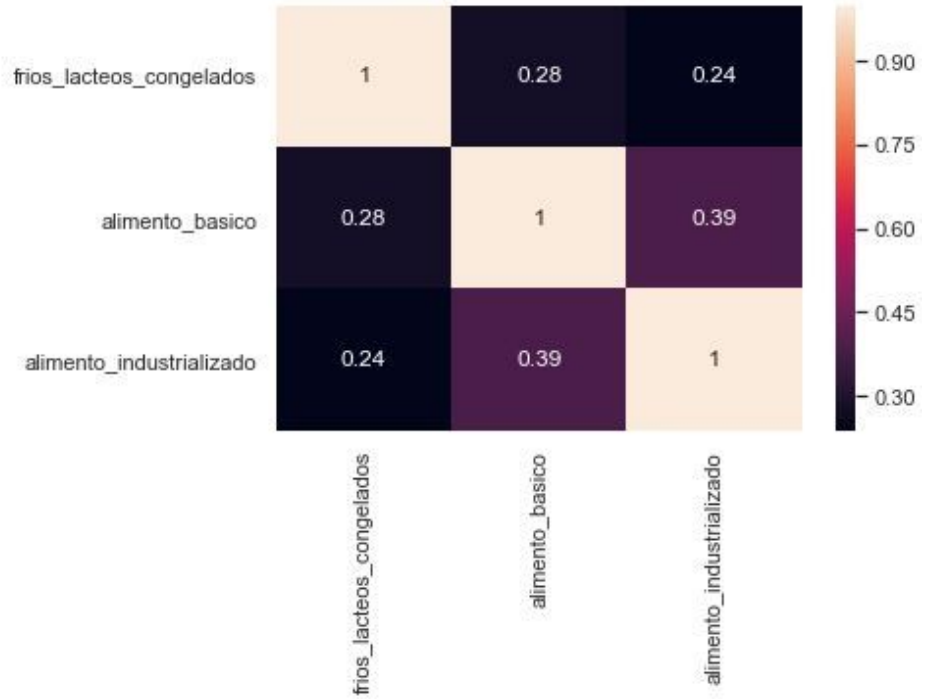
Pré-processamento dos dados

Através da visualização de correlação é possível observar que alguns dos atributos possuem um valor de correlação mais elevado, através dessa visualização e após realizar testes removendo os atributos, todos os atributos que possuem correção superior a .6 foram removidos, os atributos removidos foram:

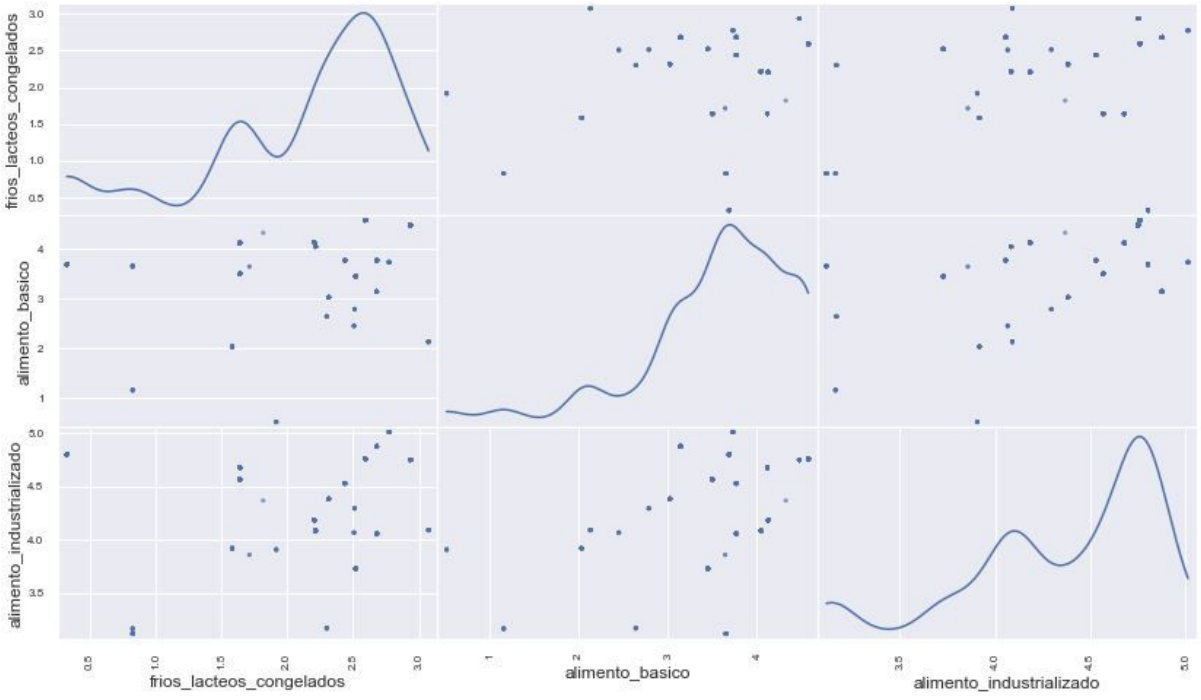
- material_de_limpeza;
- perfumaria_higiene_pessoal;
- bebidas;



Após a remoção dos atributos a matriz de correlação ficou da seguinte forma:



Analisando o conjunto dos dados foi identificado a existência de outliers, os mesmo foram removidos do conjunto de dados, após os dados foram escalonados utilizando o algoritmo natural, após esse processo a distribuição dos dados ficou da seguinte maneira:

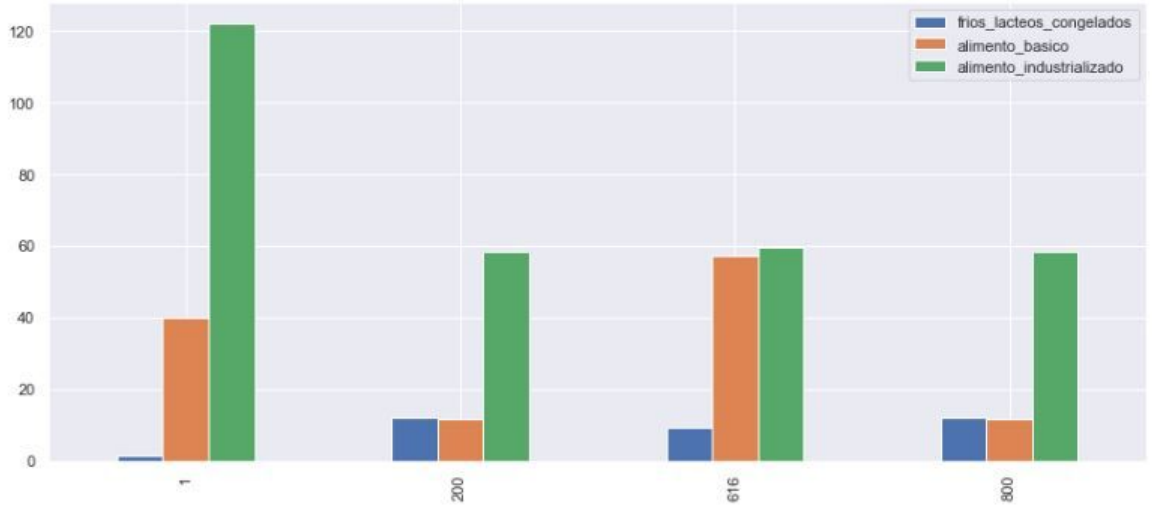


Implementação e Refinamento

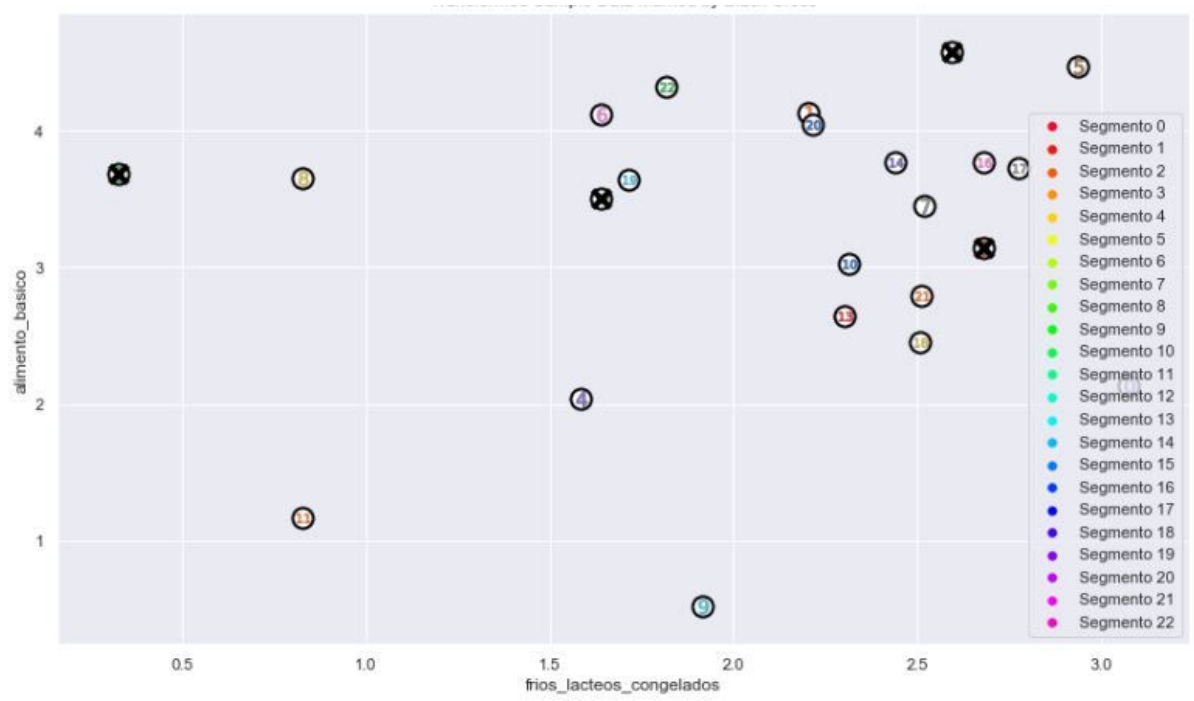
Como implementação base foi utilizado o algoritmo do Kmeans na forma padrão, utilizando 2 clusters, para avaliar o modelo criado foi utilizado o *siluete score*, o modelo obteve *scorre* aproximado a 0.5082. Com o foco em melhorar o scorre realizei a implementação para realizar testes com o Gaussian Mixture e Kmeans alterando o número de clusters utilizados por cada algoritmo, utilizando o intervalo de 2 a 24. Após realizar os testes em ambos os algoritmos ao assumir o valor de 23 para o número de clusters o siluete scorre obteve o valor igual a 1.

Resultados e Conclusão

Os dados utilizados para a criação do modelo possuem 3 dimensões, a fim de facilitar a visualização final o gráfico gerado considera apenas duas dimensões, antes de realizar o treinamento do modelo 4 amostras aleatórias foram selecionadas, tais amostras foram removidas do conjunto utilizado para o treinamento, na imagem abaixo a visualização dessas amostras.



As amostras selecionadas tem como objetivo visualizar o quão próximo ao cluster a amostra está, para isso foi criada uma visualização onde estão todos os segmentos de clientes e os pontos do conjunto exibido anteriormente, representados por um X no gráfico.



Através do gráfico é possível observar que os dados estão exatamente sobre um determinado cluster, justificando assim o siluete scorre obtido, com o valor de siluete scorre igual a 1, o modelo para esses dados pode ser considerado perfeito, mas acredito que se existissem mais dados ou um maior número de atributos poderiam ser melhor trabalhados o número de clusters.

Referências

[1] Raquel Florez-Lopez, Juan Manuel Ramon-Jeronimo, Marketing Segmentation Through Machine Learning Models: An Approach Based on Customer Relationship Management and Customer Profitability Accounting, Disponível em: https://doi.org/10.1177/0894439308321592

[2] SILVERIO, Murilo. Aplicação de algoritmos de aprendizado de máquina no desenvolvimento de modelos de escore de crédito. Disponível em: http://hdl.handle.net/11224/1503