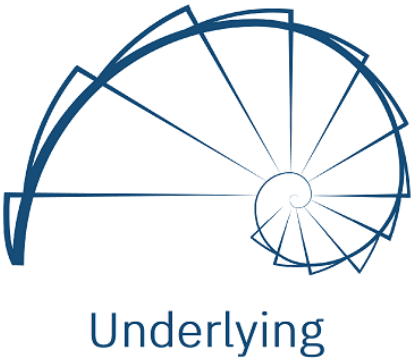


PROJETO 2022.02 - UNDERLYING
Mineração de Dados em Investimentos – Produtos Similares

Versão 0.1:
Casos de teste e modelos

Equipe de Projeto Underlying:

Adriel Douglas Nogueira Carlos - 2018012346

Ivan Leoni Vilas Boas - 2018009073

Lucas Tiense Blazzi - 2018003310

COM923 - Tópicos Especiais em Inteligência Artificial

Vanessa Cristina Oliveira Souza

https://github.com/lucasblazzi/market-datalake

UNIFEI
Universidade Federal de Itajubá

IMC - Instituto de Matemática e Computação

Av. BPS, 1303 - Caixa postal 50 - 37500-903

Itajubá - MG - Brasil Telefone: 35-3629-1135

E-mail:imc@unifei.edu.br

1. Projeto

Acesso dos notebooks de preparação dos casos de teste e aplicação dos modelos em: GitHub

2. Casos de teste

Para a elaboração do trabalho foram definidos sete casos de teste. Os casos de teste variam de acordo com os atributos selecionados e a aplicação ou não do balanceamento dos tipos de produtos coletados na etapa anterior. Para a execução dos casos de teste foram utilizados a primeiro momento três modelos: Clusterização hierárquica (AgglomerativeClustering), K-Means e DBSCAN.

A partir desse critério foram definidos os seguintes casos de teste:

1. **Caso de Teste 1**
    a. Remoção de atributos identificadores (index, id, name)
2. **Caso de teste 2**
    a. Remoção de atributos identificadores (index, id, name)
    b. Balanceamento das amostras por mercado
3. **Caso de teste 3**
    a. Remoção de atributos identificadores (index, id, name)
    b. Avaliação individual por mercado (aplicação do modelo de forma individual em produtos de renda fixa, fundos de investimento e renda variável)
    c. Remoção de caracterizadores de produto (market_type e market)
4. **Caso de teste 4**
    a. Remoção de atributos identificadores (index, id, name)
    b. Balanceamento das amostras por mercado
    c. Remoção de caracterizadores de produto (market_type e market)
5. **Caso de teste 5**
    a. Remoção de atributos identificadores (index, id, name)
    b. Balanceamento das amostras por mercado
    c. Remoção de caracterizadores de produto (market_type e market)
    d. Remoção do benchmark (correlação com strategy)
6. **Caso de teste 6 (sugestão Prof. Adriana)**
    a. Remoção de atributos identificadores (index, id, name)

b. Remoção de caracterizadores de produto (market_type e market)

c. Utilização de apenas um atributo para definir um contexto de negócio (features: risk", "liquidity", "return", "volatility", "sharpe", "minimum_application", “strategy”)

7. **Caso de teste 7 (sugestão Prof. Adriana)**

   a. Remoção de atributos identificadores (index, id, name)

   b. Remoção de caracterizadores de produto (market_type e market)

   c. Utilização de apenas um atributo para definir um contexto de negócio (features: risk", "liquidity", "return", "volatility", "sharpe", "minimum_application", “strategy”)

   d. Balanceamento das amostras por mercado

3. Metodologia

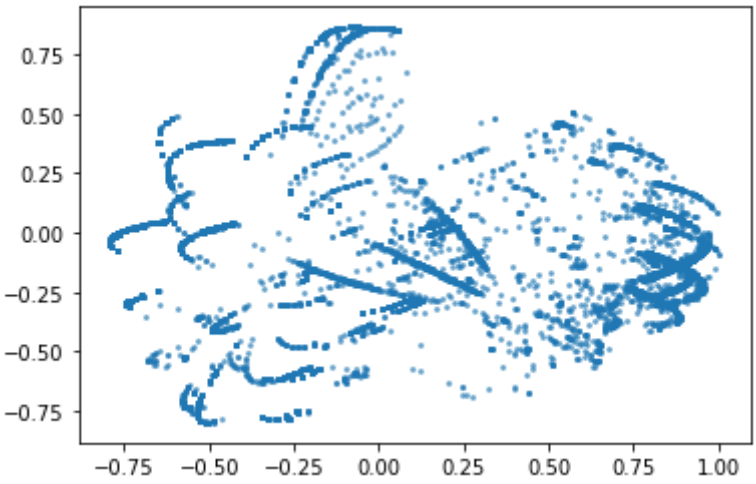
A aplicação dos modelos foi distribuída entre os integrantes e cada um ficou responsável por aplicar um modelo. O desenvolvimento das etapas de geração dos testes de caso e aplicação dos modelos pode ser acessada no GitHub.

3.1 Agglomerative Clustering

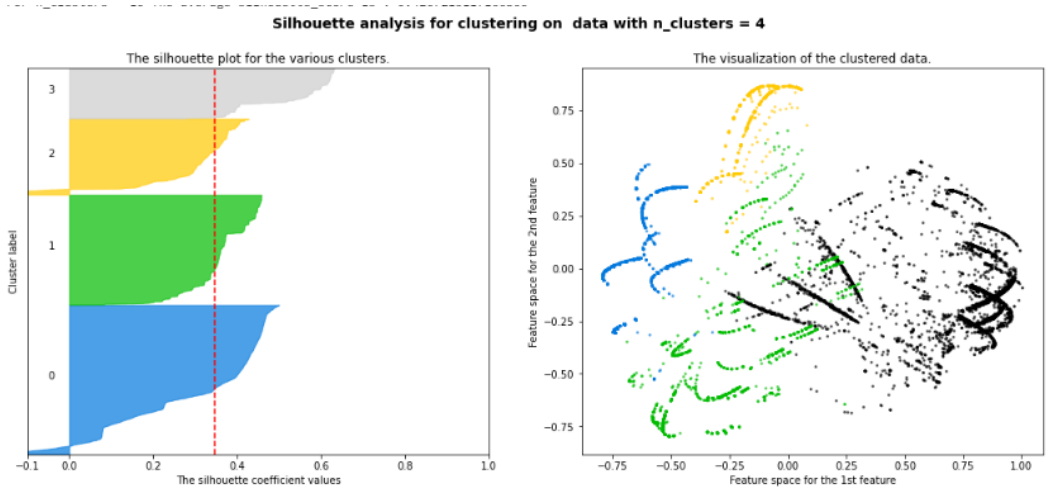
3.1.1 Metodologia

Para a execução do Agglomerative Clustering foram adotados os seguintes passos:

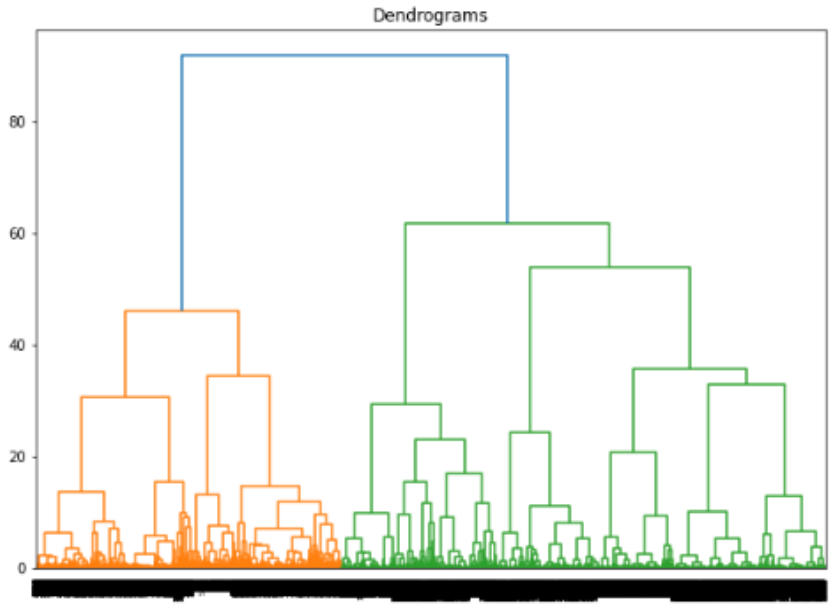
1. Decomposição das features no plano bidimensional para verificação da distribuição (tentativa de encontrar algum possível problema na seleção de features)



2. Aplicação do Agglomerative Clustering em diferentes números de clusters (sendo utilizado: 4, 5, 7, 8, 10) calculando os respectivos silhoutte score's fazendo o plot da visão bidimensional para verificação inicial da distribuição dos clusters (verificação de um potencial número de clusters a ser utilizado)



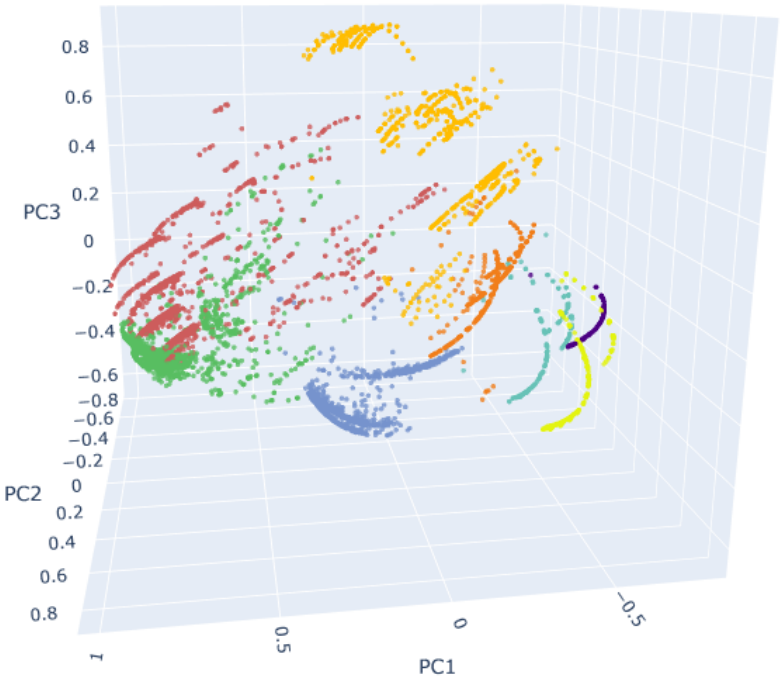
3. Plot do dendrograma (auxílio na escolha do número de clusters a ser utilizado)



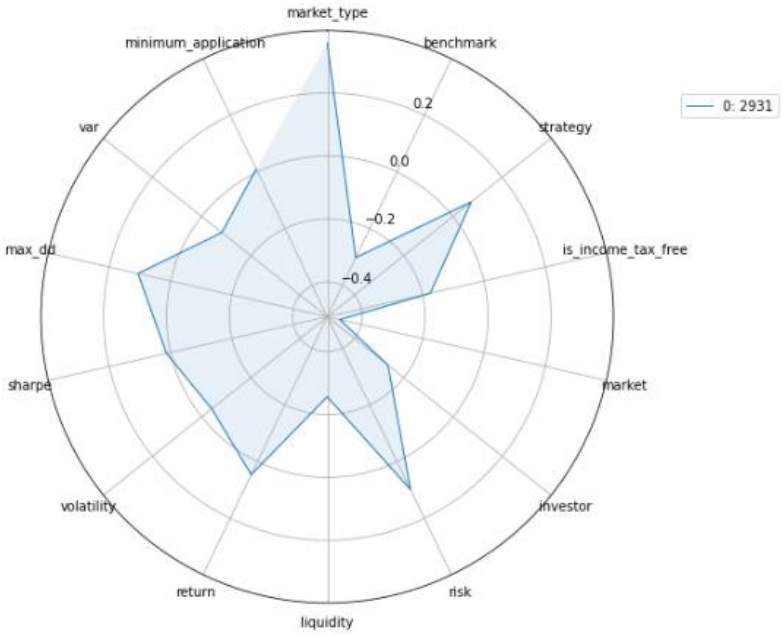
4. Definição do número de clusters

5. Aplicação do Agglomerative Clustering e definição do agrupamento gerado no dataset

6. Plot da visão tridimensional (verificação da distribuição dos clusters)



7. Plot de radar para cada cluster (verificação da distribuição dos atributos, validação inicial da eficiência dos clusters quanto a lógica de negócio)



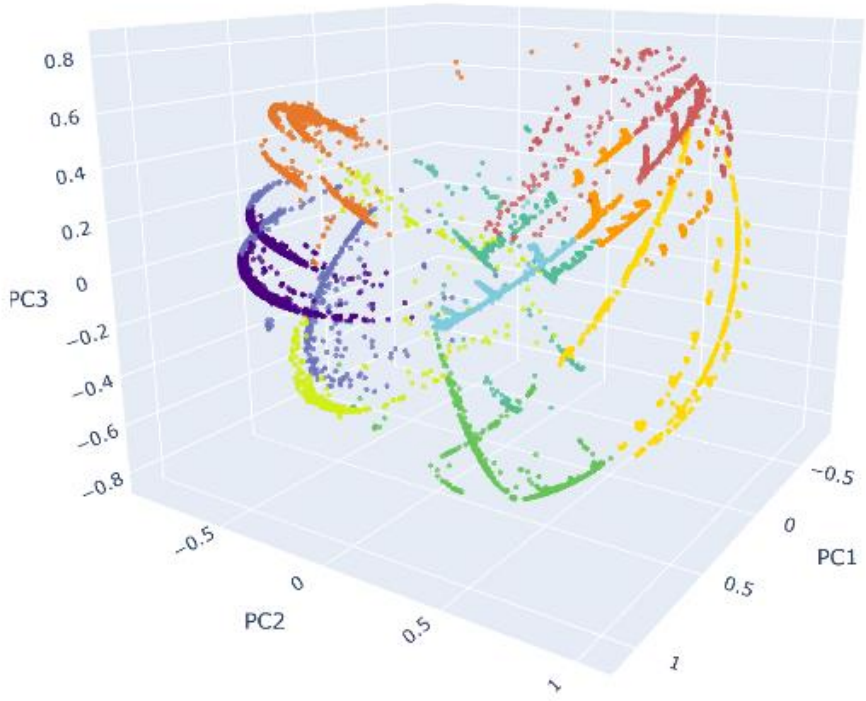
8. Describe e cálculo da mediana das features de cada cluster gerado (validação lógica)

```
_________________ CLUSTER 0 _________________

            level_0         risk    liquidity       return   volatility  \
count   2931.000000  2931.000000  2931.000000  2931.000000  2931.000000
mean    3266.224497     3.364722    56.531559     0.024753     0.092594
std     2009.024214     1.334385   232.594035     0.155207     0.126423
min      786.000000     1.000000     0.000000    -0.999056     0.000378
25%     1844.000000     2.000000     1.000000    -0.049749     0.003381
50%     2984.000000     4.000000     3.000000     0.080202     0.043146
75%     4234.500000     4.000000    31.000000     0.105860     0.184528
max    14551.000000     5.000000  3253.000000     1.695191     3.515663

             sharpe       max_dd          var  minimum_application  cluster
count  2931.000000  2931.000000  2931.000000         2.931000e+03   2931.0
mean      1.094868    -0.100081     0.012484         1.913431e+05      0.0
std      10.472041     0.131015     0.016762         2.002267e+06      0.0
min    -106.247714    -0.999578    -0.000553         0.000000e+00      0.0
25%      -1.355389    -0.192942     0.000021         1.000000e+00      0.0
50%      -0.714675    -0.032766     0.005726         1.000000e+03      0.0
75%       0.911044    -0.000473     0.025665         1.000000e+04      0.0
max     103.074220     0.000000     0.451675         5.000000e+07      0.0
_____________________________________________
```

9. Aplicação do pandas profiling em cada cluster gerado (validação lógica)

3.1.2 Caso principal

Após a análise realizada conforme descrito na etapa anterior foi aprofundado a visão de negócio referente ao teste de caso considerado mais relevante para o projeto, sendo, no caso, analisado o **teste de caso 6**. No teste de caso 6 foi realizada a aplicação da clusterização hierárquica com 10 clusters, sendo a visualização 3D e os resultados exibidos abaixo:



Test Case 06 - Média (Mediana)

cluster	count	%	risk	liquidity	return	volatility	sharpe	minimum_application	strategy
0	1859	12,61%	2,30 (2,5)	1514,39 (1461)	0,1090 (0,119)	0,000743 (0,00048)	-883,81 (50,18)	6759,45 (10000)	Pré-fixado / Pós-fixado
1	1943	13,18%	4,07 (4,0)	19,66 (3)	-0,1417 (-0,138)	0,2242 (0,2097)	0,4214 (-1,0872)	29748,22 (500)	Renda Variável / Pós-fixado
2	1817	12,33%	1,78 (2,0)	929,59 (752)	0,1116 (0,1101)	0,0012 (0,0016)	24,847 (8,0157)	19343,22 (5000)	Pós-fixado / Pré-fixado
3	1925	13,06%	2,81 (2,5)	2310,97 (2068)	0,1225 (0,1213)	0,0033 (0,00336)	7,876 (7,533)	6668,83 (1137,77)	Inflação
4	477	3,24%	3,08 (3,0)	14,58 (3,0)	0,2677 (0,099)	0,2431 (0,042)	0,3058 (0,198)	1655743,71 (1000)	Multimercado / Renda Variável / Pós-Fixado
5	1266	8,59%	2,55 (2,5)	396,69 (365)	0,0958 (0,1382)	0,04180 (0,0033)	10,3306 (12,28)	8909,17 (1081,58)	Inflação / Alternativo
6	1921	13,03%	1,55 (2,0)	134,94 (32)	0,0893 (0,0962)	0,00819 (0,0017)	-0,4296 (0,071)	47605,86 (2000)	Pós-fixado / Multimercado
7	1575	10,68%	2,5 (2,5)	293,86 (338)	0,1354 (0,137)	0,00051 (0,0005)	75,6529 (80,25)	7201,89 (10000)	Pré-fixado
8	507	3,44%	4,03 (4,0)	7,40 (2)	-0,2027 (-0,169)	0,4093 (0,306)	-0,8672 (-0,895)	7462,22 (101)	Internacional / Multimercado / Moeda
9	1451	9,84%	4,42 (4,0)	21,21 (22)	0,0900 (0,085)	0,0675 (0,053)	0,1112 (-0,19)	36776,34 (1000)	Multimercado / Moeda
	14741	100,00%							

Test Case 06 - Final

cluster	risk	liquidity	return	volatility	sharpe	minimum_application	strategy
0	Moderado	Ilíquido	Médio	Baixa	Alto	Alto aporte	Pré-fixado / Pós-fixado
1	Sofsticado	Liquidez média	Baixo	Alta	Baixo	Baixo aporte	Renda Variável / Pós-fixado
2	Conservador	Ilíquido	Médio	Baixa	Alto	Alto aporte	Pós-fixado / Pré-fixado
3	Moderado	Ilíquido	Médio	Baixa	Médio	Médio aporte	Inflação
4	Moderado	Liquidez baixa	Alto	Alta	Baixo	Médio aporte	Multimercado / Renda Variável / Pós-Fixado
5	Moderado	Ilíquido	Médio	Baixa	Alto	Médio aporte	Inflação / Alternativo
6	Conservador	Ilíquido	Médio	Baixa	Baixo	Médio aporte	Pós-fixado / Multimercado
7	Moderado	Ilíquido	Alto	Baixa	Alto	Alto aporte	Pré-fixado
8	Sofisticado	Liquidez baixa	Muito Baixo	Muito alta	Baixo	Baixo aporte	Internacional / Multimercado / Moeda
9	Sofisticado	Liquidez média	Médio	Média	Baixo	Baixo aporte	Multimercado / Moeda

A partir da análise dos resultados foi possível nomear os clusters confome os critérios abaixo:

- **Cluster 0: aposentadoria**, composto por produtos confiáveis (retorno razoável com baixa volatilidade), bom para cenário de pouca preocupação com liquidez e alta disponibilidade de renda.
- **Cluster 1: corrompidos**, composto por produtos que mesmo com baixo a porte não valem a pena pela aversão ao risco, já que possuem um baixo retorno ao mesmo tempo que uma alta volatilidade.
- **Cluster 2: aposentadoria***, semelhante ao cluster 0, mas com menor nível de risco
- **Cluster 3: inflação,** concentrou produtos de inflação.
- **Cluster 4: recompensadores**, composto por produtos que pagam a risco assumido, já que apresentam alto retorno e alta volatilidade, não necessitando de muita renda para investimento.
- **Cluster 5: seguros**, composto por produtos de boa relação risco retorno com baixo risco, garantindo um retorno adequado em cenário de baixa volatilidade, mas que travam a recuperação do dinheiro em caso de necessidade por serem produtos de alta liquidez.
- **Cluster 6: seguros flexíveis**, semelhante ao cluster 5, mas possuem uma menor liquidez, permitindo resgate em prazo menor.
- **Cluster 7: assertivos limitados**, composto por produtos de alto retorno e baixa volatilidade (melhor relação risco x retorno), bons para curto prazo, sendo limitados por seu valor de entrada muito alto.
- **Cluster 8: day traders**, composto por produtos de baixa liquidez que trazem a pior relação risco x retorno, apresentando baixo retorno e alta volatilidade e possuindo baixo valor de entrada.
- **Cluster 9: diversificadores / protetores,** composto por produtos que apresentam níveis intermediários de retorno e volatilidade, sendo que apresentam concentração em investimentos que contribuem na diversificação e proteção das carteiras (multimercado / moeda).

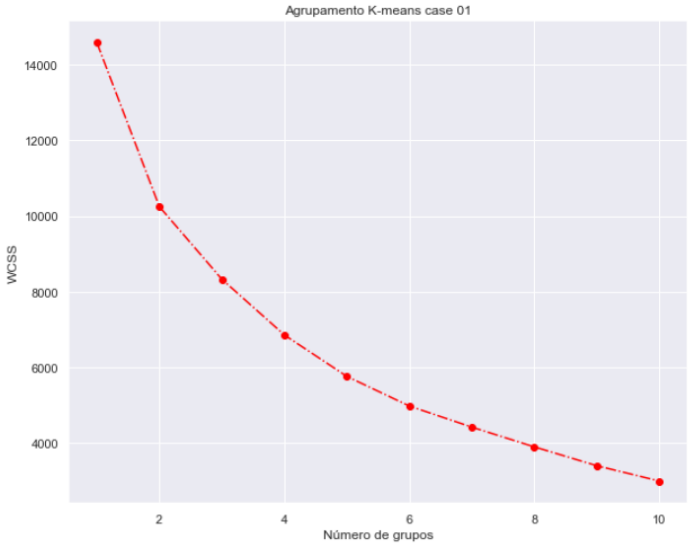
Os resultados desse caso de teste sugerem que possa ocorrer uma melhor análise lógica com a redução do número de clusters, já que alguns se apresentaram similares, sendo o próximo passo a reexecução do modelo a partir de 8 clusters.

3.2 K-means

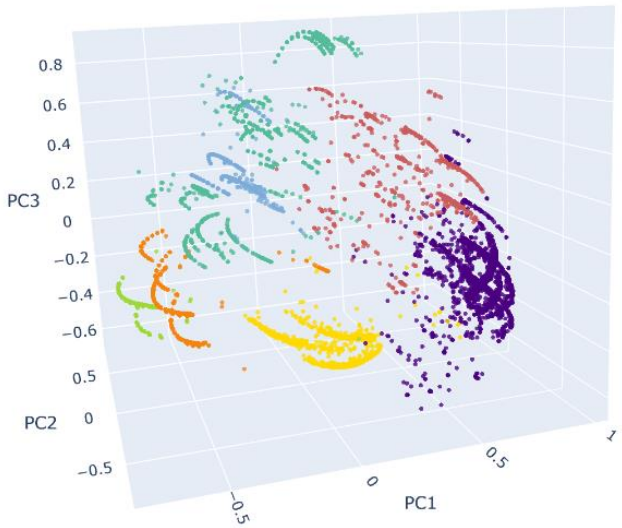
3.2.1 Metodologia

Para a execução do K-means foram adotados os seguintes passos:

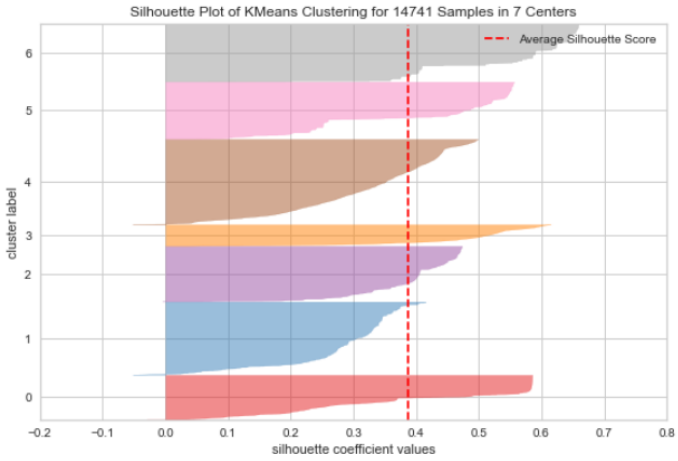
- Plot do cotovelo de 1 a 11 clusters para definição do número de clusters



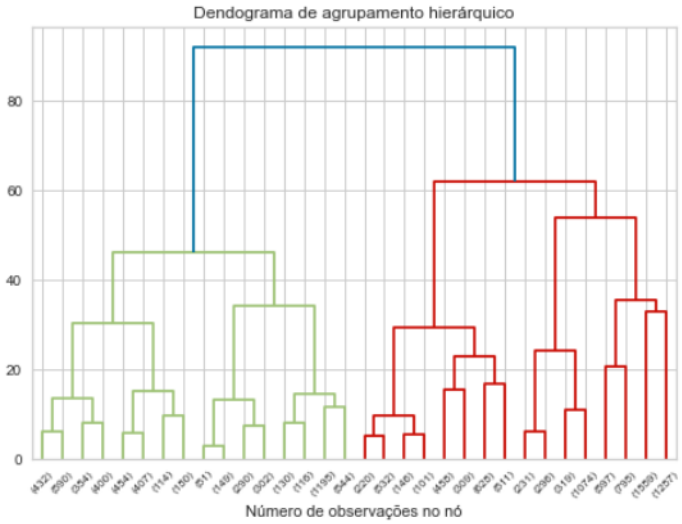
- Aplicação do K-means com base no número de clusters encontrado
- Aplicação dos clusters gerados no dataset base
- Plot da visão tridimensional para visualização da distribuição dos clusters gerados



- Plot do silhouette para validação do número de clusters utilizado



- Plot do dendrograma para validação do número de clusters utilizado



- Aplicação do pandas profilling para cada um dos clusters gerados pelo k-means para visualização das características do cluster

3.3 OPTICS

3.3.1 Metodologia

Foi aplicada a técnica OPTICS presente na biblioteca scikit learn e testou-se variações nos parâmetros xi, min_samples e min_cluster_size, em busca de maior diferenciação entre os clusters, o melhor resultado foi encontrado com os valores: xi de -0.01, min_samples de 100 e min_cluster_size de 0.005, de modo a ter uma avaliação mais igualitária entre os casos foi utilizado os mesmos valores encontrados para todos os casos mesmo quando variações neles geravam melhor resultados em algum caso.

3.3.2 Análise

Usando a técnica OPTICS o número de clusters variou a cada caso porem na maioria deles (5 de 7) o número de clusters ficou abaixo de 7, o melhor caso em nossa avaliação foi o caso 7, ele apresento 5 clusters sendo que 4 destes ficam bem definidos no DBSCAN em 0.5 eps, não somente a definição é clara como pela análise das medias dos valores de cada coluna fica evidente que a volatilidade é um dos fatores que mais ajudam a caracterizar o cluster, de forma que organizar os clusters entendendo que maior volatilidade em geral envolve maior risco, a caracterização dos clusters fica mais clara.