VIII SEMANA UNIVERSITÁRIA DA URCA
XXVI Semana de Iniciação Científica da URCA
04 a 09 de dezembro de 2023
Tema: "INTERIORIZAÇÃO DA CIÊNCIA E REDUÇÃO DE ASSIMETRIAS: O PAPEL DOS PIBIC´S COMO EXPERIÊNCIA DE ARTICULAÇÃO DA PESQUISA NA GRADUAÇÃO E NA PÓS GRADUAÇÃO"

ISSN 1983-8174

RISCO E RETORNO EM AÇÕES DE EMPRESAS QUE OPERAM NA BOLSA DE VALORES BRASILEIRA: UMA ABORDAGEM EM MACHINE LEARNING.

Tamires de Aquino Silva[1], Gustavo Henrique Leite de Castro[2], Luis Abel da Silva Filho[3].

Resumo: este estudo analisa algoritmo de Machine Learning para aglomerar empresas que operaram na B3_ibovespa entre o período de janeiro de 2021 a dezembro de 2022. Os dados são dos dias úteis de cada ano. Recorre-se a técnicas de clusterização hierárquica, depois de testados outros algoritmos, e os resultados indicaram aglomerar o conjunto de empresas em quatros clusters, conforme o desempenho, semana, mensal semestral e anual, de acordo com os riscos e os retornos oferecidos pelas ações comercializadas. O estudo está dividido em introdução, a explanação da metodologia e métodos utilizados, os resultados e discussões, considerações finais e referências.
Palavras chave: Machine learning. Clusterização hierárquica. Risco e retorno

1. Introdução

No mercado financeiro o Machine Learning mostra-se importante mecanismo de previsões em precificação, tendo capacidade preditiva muito além dos mecanismos já utilizados e evoluindo substancialmente na perspectiva de melhora nas previsões em curto, médio e longo prazo ao longo dos anos (GU, KELLY e CHIU, 2019). Sua capacidade em resolver problemas práticos em análises de riscos e retornos, com ampla perspectiva de análise e baixo custo computacional, em alguns casos, faz das técnicas aplicadas de Machine Learning um importante aliado dos estudiosos de mercado financeiro ao redor do mundo (WEIGAND, 2019; RYLL e SEIDENS, 2019; NDIKUN, 2020; GU, KELLY e XIU, 2020).

Nessa perspectiva, o agrupamento de ativos diferentes em grupos ordenados por suas características de risco e retorno pode auxiliar os tomadores de decisão de investimento no mercado financeiro e ser de relevante importância como uma nova ferramenta de uso disponível no mercado. Ademais, algumas técnicas de Machine Learning, por ser uma ferramenta de potencial preditivo aceito pela literatura, dado sua capacidade de minimização de variâncias e de seletividade por características dos ativos, tem uma grande utilidade para previsões. Isso pode contribuir sobremaneira para uma análise exploratória de boa acurácia acerca de um conjunto de ativos disponibilizados no mercado..

2. Objetivo

Agrupar as análises de risco e retorno diário, semanal, mensal, semestral e anual por meio de algoritmos de clusterização hierárquica. Visando mostrar a importância do modelo machine learning nas previsões em precificação e sua

[1] Universidade Regional do Cariri – URCA – e-mail: tamires.aquino@urca.br
[2] Universidade de São Paulo – USP – e-mail: gustavocastro@usp.br
[3] Universidade Regional do Cariri – URCA – e-mail: luis.abel@urca.br

ISSN 1983-8174

capacidade de resolver problemas práticos em análise de risco e retorno no mercado financeiro.

3. Material e Métodos

Conforme Digrius e Keen (2002), Assaf Neto (2010), Dimos, Groves e Powell (2011), o retorno de um investimento pode ser definido como a quantidade de dinheiro que um investidor pode ganhar ou perder, mediante determinado investimento por ele realizado. O cálculo de indicador de retorno bruto de investimento em mercado de ações construído neste estudo, desenvolveu-se por meio da seguinte formula:

$$R_i = \left[\left(\frac{r_{t,0}}{r_{t-1}}\right) - 1\right] * 100 \quad (1)$$

Em que $r_{t,0}$ corresponde ao retorno em t_0 de um ativo; e, r_{t-1} o seu retorno do período anterior, multiplicando por 100 para a apresentação do valor em termos percentuais, conforme utilizado neste estudo.

No que diz respeito ao risco das ações apresentadas neste estudo, o seu cálculo ocorreu por meio da variância do valor dos ativos no período analisado, tendo como base a teoria de Markowitz (1952), de forma que:

$$s_i = \sqrt{\frac{(X_1 - \bar{X})^2}{n-1}} \quad (2)$$

O processo de agrupamento por meio da clusterização hierárquica, que também pode ser designado como classificação não supervisionada, dar-se por meio de agrupamento dos indivíduos similares, de acordo com sua similaridade, de conjunto de dados, de forma que sejam apresentadas características similares dentro dos *clusters*, e a máxima dissimilaridade entre os *clusters* compostos pelo algoritmo. (JAIN et al., 1999; MINGOTI, 2007; THEODORIDIS & KOUTROUMBAS, 2009). A aglomeração do algoritmo por meio da clusterização hierárquica aqui apresentada permite a maximização da variância entre os *clusters* e minimização dela dentro dos *clusters*. De acordo com Mingoti (2007), a medida da variabilidade total pode ser representada pela equação seguinte:

$$SSi = \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)'(X_{ij} - \bar{X}_i) \quad (3)$$

Na fórmula, n_i representa um número de elementos que pertencem a um mesmo conglomerado *Ci*, onde se encontra no passo *K* do processo de agrupamento da análise; X_{ij} denota o vetor de variável observada (risco e retorno), onde o *j-ésimo* elemento amostral representa o *i-ésimo* conglomerado. Com isso, $\bar{X}_i$ representa o centroide do conglomerado *Ci,* onde SSi corresponde a soma dos quadrados que se identificam no conglomerado *Ci*. De acordo Mingoti (2007) acentua que a distância entre os *clusters* a serem formados, C_1 e C_i, assume a expressão matemática, a seguir:

$$d(C_1 C_i) = \left[\frac{n_1 n_i}{n_1 + n_i}\right] (\bar{X}_{1.} - \bar{X}_{i.})'(\bar{X}_{1.} - \bar{X}_{i.}) \quad (4)$$

Nesse sentido, a soma dos quadrados entre cada um dos *clusters* C_1 e C_i é combinada de tal forma que seja realizada a minimização entre cada um dos passos do algoritmo de agrupamento.

4. Resultados e discussões

ISSN 1983-8174

Os dados da Tabela 1 apresentam-se relevantes para observar o comportamento de retorno e risco diário, mensal, semestral e anual. O retorno máximo foi de 2% e o risco máximo foi de 3%, portanto, ainda superior. No que diz respeito às estatísticas mensais, percebe-se que a média e a mediana do retorno foram negativas e em 0.5% a 1%, aproximadamente, para mediana e média, respectivamente, no período analisado. O risco, que sempre assume valores positivos, foi de 1,1% e 1,5%, aproximadamente, no mesmo período. O retorno máximo mensal foi de 5% e o risco máximo foi de 7,4%.

Tabela 1: Medidas de dispersão para as empresas que operaram na B3 Ibovespa entre janeiro de 2021 a dezembro de 2022

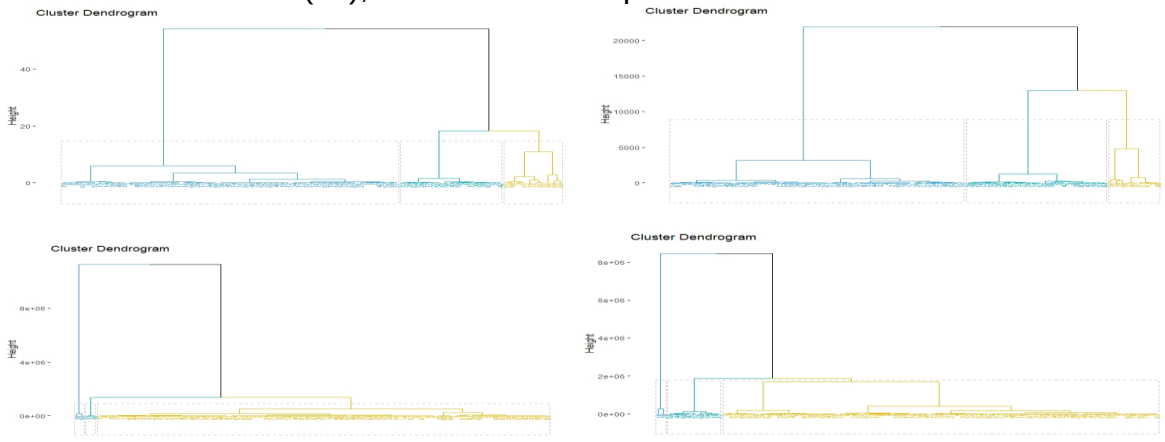
Estatísticas	REMD	RIMD	REMM	RIMM	REMS	RIMS	REMA	RIMA
Min:	-0.300	0.000	-7.500	0.300	-38.400	0.500	-63.800	0.000
1st Qu:	-0.100	0.000	-2.250	0.800	-14.300	2.000	-27.650	1.150
Median:	0.000	0.100	-0.500	1.100	-4.000	5.600	-8.300	3.600
Mean:	-0.042	0.081	-0.866	1.481	-4.592	12.68	-10.710	13.640
3rd Qu:	0.000	0.100	0.800	1.800	3.000	9.500	3.950	12.150
Max:	0.200	0.300	5.000	7.400	53.200	226.500	58.600	216.60

Fonte: elaboração a partir dos resultados da pesquisa

No que diz respeito aos resultados semestrais, tanto a mediana quanto a média dos retornos assumiram valores negativos mostrando haver perdas médias semestrais dos investimentos em ações das companhias analisadas. Os riscos assumiram valores medianos de 5,6% e valores médios de 12,7% no mesmo período. Ademais, o retorno máximo foi de 53,2% e o risco máximo foi de 226,5%, em média, no período semestral. Os valores registrados para o período anual também apresentaram mediana e média negativas, no que diz respeito ao retorno. O risco mediano foi de 3,6% e médio de 13,6%. Ademais, o retorno médio máximo foi de 58,6% e o risco médio máximo foi de 216,6%, aproximadamente igual àquele registrado para o período semestral.

4.1 Clusterização hierárquica

Pela análise do dendrograma na Figura 2, tanto o dendrograma para a clusterização pelo retorno e risco diário (2a) quanto mensal (2b) e semestral (2c) foram indicados à formação de três clusters. Já para a clusterização para o retorno e risco anual (2d), foram indicados quatro clusters.



ISSN 1983-8174

Figura 2: dendrograma da clusterização hierárquica no seu passo a passo para a formação dos clusters diários.
Fonte: elaborado a partir dos dados da pesquisa

Pelos testes de silhueta realizados, os dados da Tabela 3 apresentam o número de clusters formados pelo algoritmo por meio do método de clusterização hierárquica. Os resultados mostram que o melhor desempenho por esta técnica foi o agrupamento de três clusters para o retorno e risco médio das empresas que operam na B3-Ibovespa entre os dias úteis de 01/01/2021 a 31/012/2022, referentes aos períodos, a saber: diário, mensal e semestral. No que diz respeito ao período anual o algoritmo definiu um número de quatros clusters a partir das informações de retorno e risco de 83 empresas que operaram na B3-Ibovespa nos dias úteis do período acima definido.
Para o período diário foi indicado três clusters com maximização da silhueta média igual a 0,47, que significa que em média os grupos encontrados possuem uma estrutura fraca e pode ser artificial. Contudo, cabe ressaltar que o cluster 1 está acima de 0,71 (Figura 6), portanto o algoritmo produziu um grupo com estrutura muito robusta; e, os clusters 2 e 3 produzidos pelo algoritmo possuem estrutura razoável.

Tabela 2: Clusterização hierárquica para as empresas que operaram na B3 Ibovespa entre janeiro de 2021 a dezembro de 2022 pela estatística silhueta

Cluster 1	Número de Empresas	Média dos Clusters
1	60	0.51
2	19	0.50
3	4	-0.16
Média		0,47
Cluster 2	Número de Empresas	Média dos Clusters
1	60	0.49
2	19	0.43
3	4	-0.03
Média		0,45
Cluster 3	Número de Empresas	Média dos Clusters
1	60	0.43
2	19	0.54
3	4	-0.02
Média		0,44
Cluster 4	Número de Empresas	Média dos Clusters
1	52	0.47
2	19	0.70
3	8	0.16
4	4	-0.31
Média		0,46

Fonte: elaboração a partir dos dados da pesquisa

ISSN 1983-8174

Outrossim, é importante destacar que os clusters 1, 2 e 3 apresentaram o mesmo número de empresas agrupadas pelo algoritmo, sendo 60, 19 e 4, respectivamente. Todavia, o que mudou foram os valores médios dos clusters, conforme coluna 3 (Tabela 2). O cluster 4 aglomerou 52, 19, 8 e 4 empresas nos clusters 1, 2, 3 e 4, respectivamente.

5. Considerações Finais

Foram realizadas métricas de análises internas da clusterização pelos dois algoritmos, considerando-se as métricas de índice de silhueta, índice de Dunn e índice de Davies-Bouldin, como mecanismos de comparação do desempenho por clusterização hierárquica e por k-means nas periodicidades diárias, mensais, semestrais e anuais. Os resultados mostram que a aglomeração por hierárquica apresenta bom desempenho na aglomeração das empresas que operaram na B3-Ibovespa em todas as periodicidades analisadas, conforme os índices de validação internos utilizados. Há que se ressaltar que o desempenho apresentado pelos algoritmos é útil para esta base de dados, para esta periodicidade e para a aglomeração pelos indicadores de retornos e riscos médios para estas empresas em cada um dos períodos em análises. Outrossim, não é possível afirmar pela manutenção dos resultados com inclusão das variáveis, já que ambos os algoritmos são não supervisionados e que os seus resultados mudam, conforme acrescentam-se mais dados à amostra.

Referências Bibliográficas

Davies, D. L.; Bouldin, D. W. A Cluster Separation Measure. In: IEEE Transaction on Pattern Analysis and Machine Intelligence, v.1, no 2, p. 224-227, 1979.

Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. Journal of cybernetics, 4(1), 95-104.

Digrius, B., & Keen, J. (2002). Making technology investments profitable: ROI road map to better business cases.

Dimos, J., Groves, S., & Powell, G. (2011). Retorno sobre o investimento em mídias sociais: como definir, medir e avaliar a eficácia das redes sociais. Elsevier Brasil. 320 páginas. ISBN : 9788535247794.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. The Review of Financial Studies, 33(5), 2223-2273.

Hocking, A.; Geach, J. E.; Sun, Y.; Davey, N. 2018. An automatic taxonomy of galaxy morphology using unsupervised machine learning. Monthly Notices of the Royal Astronomical Society 473(1): 1108-1129.

Hruschka, E. R., Campello, R. J. G. B., and Castro, L. N. (2004). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. In IEEE International Conference on Data Mining, pages 403–406.Jaggi, M. (2013, February). Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In International Conference on Machine Learning (pp. 427-435). PMLR.

Jain, Anil K. . Data clustering; a review. ACM computing surveys (CSUR), v. 31, n. 3, p. 264-323, 1999.