IX SEMANA UNIVERSITÁRIA DA URCA
XXVII Semana de Iniciação Científica da URCA

04 a 08 de NOVEMBRO de 2024

Tema: "CIÊNCIA, TECNOLOGIA E AMBIENTE: MÚLTIPLOS SABERES E FAZERES"

ISSN 1983-8174

ANÁLISE DE RISCO E RETORNO DE EMPRESAS QUE OPERAM NO MERCADO DE AÇÕES NO BRASIL

Tamires de Aquino Silva[1], Luis Abel da Silva Filho[2]

Resumo: Um dos principais fatores analisados para escolha de um ativo é o retorno que ele pode oferecer. Com isso o objetivo principal deste trabalho, é fazer previsões de risco e retorno para as empresas que operam na B3-Ibovespa, tendo em vista a importância do modelo machine learning como mecanismo de previsões em precificação, foi utilizar métodos de aprendizado de máquina na perspectiva de melhorar as previsões em curto, médio e longo prazo para auxiliar na escolha dos melhores ativos. Para tal finalidade foram analisados dados diários do ano de 2021 e 2022 de 83 empresas listadas na B3-ibovespa por meio do método clusterização hierárquica e k-means. Os resultados indicam um bom desempenho para análise de conglomerados para as empresas inscritas no Ibovespa. Conclui-se que o desempenho obtido pelo algoritmo de clusterização hierárquica é relevante para a base de dados e para os grupos de indicadores de risco e retorno.

Palavras-chave: machine learning. Clusterização. Ibovespa. Risco. Retorno.

1. Introdução

O machine learning é amplamente utilizado em diversas áreas no mundo à fora. De tal modo, que treina algoritmos em conjunto de dados em busca de alcançar resultados e propor soluções para problemas. Ou seja, é o processo de otimização do modelo para que se possa prever a resposta correta com base nas amostras de dados de treinamento. Agarwal e Dhar (2014) existem oportunidades que surgem pela disponibilidade de dados e grandes avanços em machine learning permitem realizar análises eficientes e concede uma compreensão mais precisa dos dados.
Para o mercado financeiro o machine learning atua como um importante mecanismo de previsões em precificação de ações, com capacidade de previsão superior à dos mecanismos utilizados e destacando-se na perspectiva de melhora nas previsões a curto, médio e longo prazo, (WEIGAND, 2019).

Para realização deste estudo, optou-se pelo processo de clusterização hierárquica, que consiste em um conjunto de técnicas hierárquicas que tem por objetivo formar grupos homogêneos internamente e heterogêneos entre si. As observações de cada grupo devem ser semelhantes entre si, em relação às variáveis inseridas na análise, e consideravelmente diferentes das observações de outros grupos, Favero e Belfiore (2017).

A contribuição deste trabalho consiste na aplicação de metodologias de aprendizado de máquina em um banco de dados de empresas com capital aberto no Brasil, pois verificou que a aglomeração por hierarquia mostra um bom desempenho para os clusters formados das empresas que atuam no mercado de ações. Situação essa que

[1] Universidade Regional do Cariri, email: tamires.aquino@urca.br
[2] Universidade Federal do Cariri, email: luis.abel@ufca.br

ISSN 1983-8174

foi comprovada através do uso dos índices de validação internos, tais como métricas de índice de silhueta, índice de Dunn e índice de Dawes Boudin como forma de comparar o desempenho por clusterização hierárquica e por k-means.

2. Objetivo

O objetivo desse estudo foi agrupar as análises de risco e retorno diário, mensal, semestral e anual por meio de técnicas de clusterização hierárquica por machine learning .

3. Metodologia

Foram analisados dados diários de 83 empresas que operam na b3-ibovespa entre o período de 2021 a 2022. Utilizou-se técnicas de aprendizado de máquinas para mostrar o comportamento do risco e retorno diário, mensal, semestral e anual através do modelo de clusterização hierárquica e k-means. Para isso, utilizou-se métricas de análise interna visando identificar em qual dos algoritmos melhores se adequa aos clusters. Assim, considerou-se métricas de índice de silhueta, índice de Dunn e índice de Davies-Bouldin.

Foram utilizadas a clusterização hierárquica e k-means, procurando se verificar o melhor método de agrupamento a partir de métricas de avaliação internas dos algoritmos. Para esse grupo de dados o método que melhor se adequou foi o de clusterização hierárquica, com base em testes indicados pela literatura atual.

Desse modo, o cálculo de indicador de retorno bruto em mercado de ações utilizado neste estudo se deu por meio da seguinte fórmula;

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

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 tocante ao risco das ações apresentadas, o cálculo se deu por meio da variância do valor dos ativos para cada período estudado.

$$s_i = \sqrt{\frac{\left(X_1 - \underline{X}\right)^2}{n-1}}$$

Tal escolha teve como base a teoria desenvolvida por Markowitz (1952). Na qual, argumentou que ao combinar ativos que não estão perfeitamente correlacionados, um investidor pode obter um portfólio com menor risco que a média ponderada dos riscos dos ativos individuais.

O agrupamento do algoritmo por meio da clusterização hierárquica aqui apresentada mostra a maximização da variância entre os clusters e a minimização dela dentro dos clusters.

$$SSi = \sum_{j=1}^{n_i} (X_{ij} - \underline{X_i})'(X_{ij} - \underline{X_i})$$

ISSN 1983-8174

De acordo com o modelo expresso acima, n_1 representam um número de elementos que pertencem a um mesmo grupo C_i onde se encontra no passo K do processo de agrupamento em análise; Xij denota o vetor de variável observada (risco e retorno), onde o *j-ésimo* elemento amostral representa o *i-ésimo* conglomerado. Desse modo, $\underline{X_j}$ representa o centroide do conglomerado *Ci,* onde SSi corresponde à soma dos quadrados que se identificam no conglomerado *Ci.*

4. Resultados

Para o método de clusterização hierárquica do retorno e risco médio diário, mensal, semestral e anual foram realizados quatro testes para determinar a quantidade de clusters, amparados na literatura atual. Foram realizadas as escolhas do número de clusters com base no dendrograma a seguir.

Em torno da análise do dendrograma para a clusterização pelo retorno e risco diário (2a) quanto mensal (2b) e semestral (2c) apontou-se a formação de três clusters. No entanto, para a clusterização para retorno e risco anual (2d) foram apontados quatro clusters.

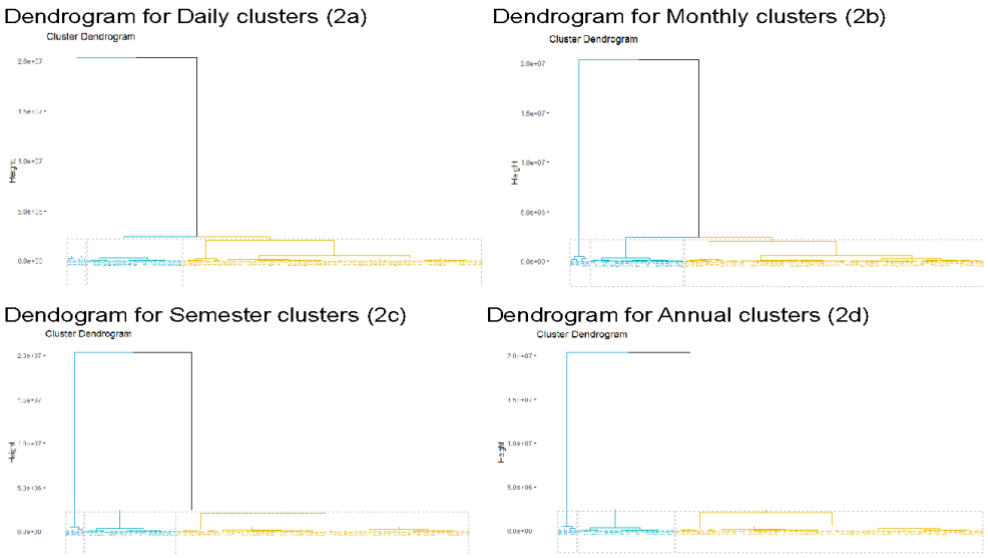

Figura 2: dendrograma da clusterização hierárquica no seu passo a passo para a formação dos clusters diários. Fonte: elaborado a partir dos dados da pesquisa

De acordo com os testes de silhueta efetuados, os dados descritos na tabela abaixo mostram a quantidade de clusters gerada através do método de clusterização hierárquica. Os indicadores mostram que o melhor desempenho por esse procedimento foi o agrupamento de retorno e risco médio das empresas que operam com capital aberto no país, para os períodos diário, mensal e semestral. No tocante ao período anual, o algoritmo definiu uma quantidade de quatro clusters.

Para o período diário obteve-se a indicação de três clusters com maximização da silhueta média igual a 0,47, isso mostra que em média os grupos possuem estrutura fraca e pode ser artificial. No entanto, pode-se ressaltar que o cluster 1 está acima de 0,71, significando a formação de um grupo com estrutura sólida; e os clusters 2 e 3 apresentam estrutura razoável.

ISSN 1983-8174

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	Cluster 2	Número de Empresas	Média dos Clusters
1	60	0.51	1	60	0.49
2	19	0.50	2	19	0.43
3	4	-0.16	3	4	-0.03
Média		0,47	Média		0,45
Cluster 3	Número de Empresas	Média dos Clusters	Cluster 4	Número de Empresas	Média dos Clusters
1	60	0.43	1	52	0.47
2	19	0.54	2	19	0.70
3	4	-0.02	3	8	0.16
			4	4	-0,31
Média		0,44	média		0,46

Fonte: elaboração a partir dos dados da pesquisa

Ademais, observa-se que os clusters 1,2 e 3 apresentam o mesmo número de empresas agrupadas, sendo 60,19 e 4, respectivamente. Contudo, o que mudou foram os valores médios do cluster. O cluster 4 aglomerou 52,19,8 e 4 empresas no cluster 1,2,3 e 4, respectivamente.

No tocante ao risco e retorno médio mensal, os grupos produzidos pelo algoritmo, em média, possuem estrutura fraca ou razoável, sabendo que o valor da silhueta média foi de 0,45. No entanto, o cluster 1 possui estrutura robusta com silhueta média superior a 0,71. Somado a isso o cluster 2 possui estrutura razoável, pois o valor da silhueta é superior a 0,50. Contudo, o cluster 3 possui uma estrutura fraca e pode ser artificial.

Ao analisar o aspecto dos clusters para retorno e riscos médios semestrais também apresentaram silhueta média de 0,44. Isso mostra que o algoritmo produziu grupos com estrutura fraca e pode ser artificial. Contudo, ao analisar separadamente o valor médio assumido por cada grupo identifica que os clusters 1 e 2 possuem estrutura razoável e o grupo 1 possui estrutura fraca que pode ser artificial. O cluster 2 apresenta estrutura robusta. O cluster 1 possui estrutura razoável. Já o cluster 3 possui estrutura fraca podendo ser artificial, e o cluster 4 não é possível identificar nenhuma estrutura. Porém, a composição dos clusters 4 deu-se devido ao fato da quantidade de cluster que maximiza o valor da silhueta média.

5. Conclusão

O presente trabalho teve como objetivo utilizar o método de clusterização para auxiliar no mercado dos investimentos dando informações precisas sobre os riscos e retornos em investimentos. Para isso, foi agrupado para análises de risco e retorno

ISSN 1983-8174

dados diário, mensal, semestral e anual de empresas que atuam na B3-Ibovespa entre 2021 e 2022.

Trabalhou-se sobre métricas de análise internas da clusterização pelos dois algoritmos, levando em conta as métricas de índice de silhueta, índice de Dunn e índice de Davis-Bouldin, como mecanismo de contrapor o desempenho por clusterização hierárquica e por k-means nas periodicidades diárias, mensais, semestrais e anuais.

Os resultados indicam que a aglomeração por hierárquica expõe um bom desempenho para os agrupamentos das empresas que operam na bolsa de valores em todas as periodicidades estudadas, de acordo com os índices de validação internos utilizados.

Pode-se enfatizar que o desempenho apresentado pelos algoritmos é relevante para essa base de dados, para o período e para a aglomeração dos indicadores de retorno e riscos médios para essas empresas.Ademais, 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.

6. Referências

Agarwal, R., & Dhar, V. (2014). Big data, data science, and analytics: The opportunity

Everitt, B. S. (1993). Cluster Analysis. Edward Arnold, 3 edition. Citadonas páginas 19, 29, 38, e 47.

Fávero, L. P., & Belfiore, P. (2017). Manual de análise de dados: estatística e modelagem multivariada com Excel®, SPSS® e Stata®. Elsevier Brasil.

Kassambara, A. (2017). Practical guide to cluster analysis in R: Unsupervised machine learning (Vol. 1). Sthda.

Markowitz, H. (1952). A utilidade da riqueza. Revista de economia política, 60 (2), 151-158

Weigand, A. (2019). Aprendizado de máquina em precificação de ativos empíricos. Mercados Financeiros e Gestão de Portfólios, 33 , 93-104.