**Clusterização de Fundos de Renda Fixa para Otimização da Tomada de Decisão de Investimento**

Andressa Evellyn Toshie Joboji¹\*;Denize Silva2

1 Bacharel em Desenho Industrial com Habilitação em Programação Visual. Pós-graduada em Finanças, Investimentos e Banking. Rua Mauro, 462, Apto 24B, Saúde, 04055-041, São Paulo, SP. Brasil.

2 Nome da Empresa ou Instituição (opcional). Titulação ou função ou departamento. Endereço completo (pessoal ou profissional) – Bairro; 00000-000 Cidade, Estado, País

\*autor correspondente: dessaevel@gmail.com

**Clusterização de Fundos de Renda Fixa para Otimização da Tomada de Decisão de Investimento**

**Resumo**

A vasta quantidade de fundos de investimento de renda fixa no mercado dificultou a tomada de decisão para os investidores. Para mitigar essa complexidade, este trabalho propôs uma abordagem quantitativa e diferenciada para a análise e classificação de fundos, utilizando técnicas de mineração de dados. A partir da extração e do tratamento de dados públicos da Comissão de Valores Mobiliários (CVM) para um período de 24 meses, foram calculados indicadores financeiros essenciais, como risco e retorno. Em seguida, aplicou-se o algoritmo de clusterização K-means para agrupar os fundos com características semelhantes, e a validação dos clusters foi realizada por meio do *elbow method* e do *silhouette score*. Os resultados permitiram a identificação e a caracterização de perfis típicos de fundos, como conservadores, moderados e arrojados. Assim, este estudo forneceu uma ferramenta mais intuitiva e baseada em dados para auxiliar investidores na seleção de fundos alinhados ao seu perfil de risco e objetivos financeiros.

**Palavras-chave**: Fundos de investimento; Renda fixa; Clusterização; K-means; Data Science.

**Clustering Fixed-Income Funds to Optimize Investment Decisions**

**Abstract**

The vast number of fixed-income investment funds in the market has complicated decision-making for investors. To mitigate this complexity, this work proposed a quantitative and distinctive approach to fund analysis and classification, using data mining techniques. Drawing from the extraction and treatment of public data from the Brazilian Securities and Exchange Commission (CVM) for a 24-month period, essential financial indicators such as risk and return were calculated. Subsequently, the K-means clustering algorithm was applied to group funds with similar characteristics, and the clusters were validated using the *elbow method* and the *silhouette score*. The results allowed for the identification and characterization of typical fund profiles, such as conservative, moderate, and aggressive. Thus, this study provided a more intuitive and data-driven tool to assist investors in selecting funds aligned with their risk profile and financial objectives.

**Keywords:** Investment funds; Fixed income; Clustering; K-means; Data Science.

**Introdução**

O mercado financeiro brasileiro tem demonstrado uma evolução contínua e uma crescente sofisticação, com os fundos de investimento consolidando-se como um dos principais veículos para a alocação de capital por parte de investidores (Varga & Wengert, 2011). A indústria de fundos no Brasil tem experimentado um crescimento notável, desempenhando um papel crucial na captação e alocação de recursos, especialmente após períodos de estabilização econômica (Milani, 2010). Este cenário destaca a relevância e a atratividade dos fundos de renda fixa, que continuam captando investidores em busca de retornos consistentes com menor exposição a riscos em comparação com outras classes de ativos (Boalin et al., 2020).

Contudo, a vasta gama de fundos disponíveis e a complexidade inerente às suas estratégias de gestão e composição de carteiras representam um desafio significativo para os investidores na tomada de decisão (Borges, 2018). A avaliação tradicional, frequentemente baseada em métricas isoladas como retorno histórico ou volatilidade, muitas vezes se mostra insuficiente para capturar a totalidade das características que influenciam o desempenho e o perfil de risco de um fundo (Rodrigues, 2017). Fatores como a experiência do gestor, o patrimônio líquido sob gestão, a estrutura de taxas (administração e performance) e a alocação percentual em diferentes classes de ativos exercem influência substancial sobre o desempenho ajustado ao risco (Ceretta & Milani, 2001). A dificuldade em processar e interpretar eficientemente essa multiplicidade de informações pode conduzir a escolhas de investimento ineficientes, desalinhadas com os objetivos e a tolerância ao risco do investidor. Torna-se, portanto, imperativa a busca por metodologias que facilitem a compreensão e a segmentação desse universo de fundos.

Neste contexto, a aplicação de técnicas avançadas de *\*Data Science\**, como a clusterização, emerge como uma abordagem eficaz para desvendar padrões e extrair conhecimento aplicável de grandes volumes de dados financeiros (Cai et al., 2016). A clusterização possibilita o agrupamento de fundos que compartilham características fundamentais, revelando perfis distintos que seriam de difícil identificação por meio de análises convencionais (Vilela et al., 2018). A literatura demonstra a eficácia da clusterização tanto na segmentação de ativos financeiros (Sáenz, 2023) quanto como uma etapa crucial na preparação de dados para modelos preditivos de séries financeiras (Pattarin, 2004). Essa capacidade de simplificar a complexidade do mercado e otimizar a tomada de decisão ressalta o potencial significativo da clusterização. No entanto, a mineração de dados financeiros não é isenta de desafios, incluindo a complexidade inerente aos dados, a influência de fatores macroeconômicos, a necessidade de confidencialidade e a heterogeneidade das informações (Cai et al., 2016).

Este estudo visa contribuir para a literatura e para a prática do mercado financeiro brasileiro ao empregar a clusterização para segmentar fundos de investimento de renda fixa. A justificativa para esta pesquisa reside na crescente demanda por abordagens analíticas mais sofisticadas e baseadas em dados para a análise de investimentos, especialmente diante da contínua evolução do mercado e da disponibilidade de dados. Ao identificar e caracterizar perfis típicos de fundos, este trabalho busca aprimorar a compreensão das dinâmicas que definem cada grupo, oferecendo uma ferramenta valiosa para auxiliar investidores na seleção de fundos que melhor se alinhem aos seus perfis de risco e objetivos financeiros. O objetivo geral deste trabalho é desenvolver e aplicar uma metodologia de clusterização para classificar fundos de investimento de renda fixa no mercado brasileiro, com base em indicadores de risco, retorno e composição de carteira, visando aprimorar a tomada de decisão de investimento por parte dos cotistas.

**Metodologia**

A abordagem metodológica adotada foi de caráter quantitativo e exploratório, utilizando técnicas de Data Science para a análise de fundos de investimento de renda fixa. O processo foi estruturado em etapas sequenciais, conforme detalhado a seguir:

1. Coleta e Tratamento de Dados

Os dados utilizados neste estudo foram obtidos a partir dos Informes Diários de Fundos de Investimento disponibilizados publicamente pela Comissão de Valores Mobiliários (CVM). A coleta abrangeu um período de 24 meses, de setembro de 2023 a agosto de 2025, focando em fundos de renda fixa. Para cada mês, foi realizado o download do arquivo compactado contendo os dados diários dos fundos. Os arquivos foram descompactados e os dados em formato CSV foram lidos e concatenados em um único conjunto de dados. Durante a importação, foram realizadas padronizações nos nomes das colunas que identificam o tipo e o CNPJ do fundo, garantindo a consistência dos dados ao longo do tempo. A data de referência de cada arquivo foi adicionada como uma nova coluna para facilitar a rastreabilidade.

Após a consolidação dos dados, aplicaram-se critérios de filtragem para selecionar os fundos elegíveis para a análise. Primeiramente, a coluna de data de competência foi convertida para o formato adequado. Em seguida, foram identificados os fundos que possuíam pelo menos 252 observações (equivalente a aproximadamente um ano de dias úteis) no período de 365 dias anteriores à data mais recente da base de dados. Adicionalmente, foram considerados apenas os fundos com uma média de mais de 100 cotistas, visando excluir fundos exclusivos ou potencialmente fechados, que não representam o público geral de investidores. Os fundos que atenderam a ambos os critérios foram mantidos para as etapas subsequentes.

1. Cálculo de Indicadores Financeiros

Para cada fundo selecionado, foram calculados diversos indicadores financeiros, essenciais para a caracterização de seus perfis de risco, retorno e patrimônio. Os cálculos foram realizados com base nos dados diários das cotas e do patrimônio líquido dos fundos. Os indicadores calculados incluíram:

•Retorno Diário Médio: Média dos retornos percentuais diários da cota.

•Retorno Acumulado (1 mês, 1 ano e 2 anos): Calculado como a variação percentual da cota no período especificado.

•Volatilidade Anualizada (1 mês, 1 ano e 2 anos): Desvio padrão dos retornos diários, anualizado pela raiz quadrada de 252 (dias úteis no ano).

•Patrimônio Líquido Médio: Média do patrimônio líquido do fundo no período analisado.

Esses indicadores foram agregados por CNPJ do fundo, resultando em uma tabela onde cada linha representava um fundo e suas respectivas métricas de desempenho e risco.

1. Análise Exploratória de Dados (AED)

A etapa de Análise Exploratória de Dados (AED) foi concebida para preceder a clusterização. Esta fase envolveu a visualização da distribuição dos indicadores calculados, a identificação de outliers e a análise de correlações entre as variáveis. Gráficos de dispersão, histogramas e boxplots foram utilizados para compreender a estrutura dos dados e validar a adequação das variáveis para a clusterização. A AED foi fundamental para garantir a qualidade dos dados e a interpretabilidade dos resultados dos clusters.

1. Aplicação de Métodos de Clusterização

Para identificar grupos de fundos com características semelhantes, foram aplicados métodos de clusterização. As variáveis numéricas dos indicadores financeiros (retorno, volatilidade e patrimônio líquido médio) foram selecionadas para esta etapa. Antes da aplicação dos algoritmos, os dados foram padronizados utilizando uma técnica de escalonamento que transforma os dados para que tenham média zero e desvio padrão um. Esta etapa é crucial para garantir que todas as variáveis contribuam igualmente para a formação dos clusters, evitando que variáveis com maior magnitude dominem o processo. Valores ausentes foram preenchidos com zero antes da padronização.

O algoritmo de clusterização K-Means foi o principal método empregado. O K-Means é um algoritmo de agrupamento não supervisionado que particiona “n” observações em “k” clusters, onde cada observação pertence ao cluster cujo centroide é o mais próximo. A escolha do número ideal de clusters (k) foi realizada por meio de técnicas de validação, como o método do cotovelo (Elbow Method) e o coeficiente de silhueta (Silhouette Score), conforme descrito na literatura (Cai et al., 2016; Vilela et al., 2018). Para este estudo, assumiu-se um número de 4 clusters para a aplicação final do K-Means, com uma semente aleatória para garantir a reprodutibilidade dos resultados.

1. Validação e Análise dos Clusters

A validação dos clusters foi realizada utilizando o coeficiente de silhueta (silhouette\_score), uma métrica que avalia quão similar um objeto é ao seu próprio cluster (coesão) em comparação com outros clusters (separação). Valores mais próximos de 1 indicam clusters bem definidos e separados. Além disso, o método do cotovelo (Elbow Method) seria utilizado para auxiliar na determinação do número ótimo de clusters, observando o ponto de inflexão na curva de soma dos quadrados intra-cluster (WSS - Within-Cluster Sum of Squares).

Após a formação dos clusters, foi realizada uma análise descritiva para caracterizar os perfis típicos de fundos em cada grupo. Para isso, foram calculadas as médias dos indicadores financeiros para cada cluster, permitindo identificar as características predominantes de retorno, risco e patrimônio líquido que definem cada segmento. Esta análise qualitativa dos clusters é fundamental para traduzir os resultados técnicos em insights práticos para investidores, auxiliando-os na seleção de fundos alinhados aos seus perfis de risco e objetivos financeiros.

1. Ferramentas Computacionais

Todo o processo de coleta, tratamento, cálculo de indicadores e clusterização foi implementado utilizando a linguagem de programação Python (versão 3.x) e suas bibliotecas especializadas em Data Science, incluindo pandas para manipulação de dados, numpy para operações numéricas, requests e zipfile para automação da coleta de dados da CVM, e scikit-learn (sklearn) para os algoritmos de clusterização e padronização. O ambiente de desenvolvimento utilizado foi o Jupyter Notebook, que facilita a execução interativa e a documentação do código. Para visualização de dados, seriam empregadas as bibliotecas matplotlib e seaborn.

**Resultados e Discussão**

Esta seção apresenta os resultados obtidos a partir da aplicação da metodologia descrita anteriormente, seguida de uma discussão aprofundada e comparativa com a literatura científica. Os achados são interpretados criticamente, destacando-se as implicações para a tomada de decisão de investimento e as contribuições do estudo.

1. **Coleta e Tratamento de Dados**

A etapa inicial de coleta de dados, realizada a partir dos Informes Diários de Fundos de Investimento disponibilizados publicamente pela Comissão de Valores Mobiliários (CVM), resultou em um volume significativo de informações abrangendo 24 meses (setembro de 2023 a agosto de 2025). Após a consolidação e tratamento, a base de dados bruta continha informações de **\*\*[INSERIR NÚMERO DE FUNDOS BRUTOS]\*\*** fundos. A aplicação dos critérios de filtragem, que incluíram a exigência de pelo menos 252 observações diárias no último ano e uma média superior a 100 cotistas, reduziu a amostra para **\*\*[INSERIR NÚMERO DE FUNDOS FILTRADOS]\*\*** fundos de renda fixa. Esta filtragem foi crucial para garantir a robustez da análise, excluindo fundos com dados insuficientes ou que não representam o universo de interesse dos investidores de varejo e qualificados, alinhando-se à prática de estudos que buscam amostras representativas e com histórico de dados consistente (Boalin et al., 2020).

1. **Indicadores Financeiros**

Para cada um dos **\*\*[INSERIR NÚMERO DE FUNDOS FILTRADOS]\*\*** fundos da amostra final, foram calculados os indicadores de retorno (diário médio, 1 mês, 1 ano, 2 anos), volatilidade anualizada (1 mês, 1 ano, 2 anos) e patrimônio líquido médio. A Tabela 1 apresenta as estatísticas descritivas desses indicadores para a amostra total de fundos. Observou-se uma variabilidade considerável nos retornos e volatilidades, o que já indicava a heterogeneidade dos fundos e a potencial eficácia da clusterização para identificar grupos distintos. O Patrimônio Líquido Médio também demonstrou uma ampla gama de valores, refletindo a diversidade de tamanhos de fundos no mercado brasileiro.

**Tabela 1: Estatísticas Descritivas dos Indicadores Financeiros (Amostra Total)**

| Indicador | Média | Desvio Padrão | Mínimo | Máximo |

| :-------------------- | :------ | :------------ | :------ | :------ |

| Retorno Diário Médio | [VALOR] | [VALOR] | [VALOR] | [VALOR] |

| Retorno 1 Mês | [VALOR] | [VALOR] | [VALOR] | [VALOR] |

| Retorno 1 Ano | [VALOR] | [VALOR] | [VALOR] | [VALOR] |

| Retorno 2 Anos | [VALOR] | [VALOR] | [VALOR] | [VALOR] |

| Volatilidade 1 Mês | [VALOR] | [VALOR] | [VALOR] | [VALOR] |

| Volatilidade 1 Ano | [VALOR] | [VALOR] | [VALOR] | [VALOR] |

| Volatilidade 2 Anos | [VALOR] | [VALOR] | [VALOR] | [VALOR] |

| PL Médio | [VALOR] | [VALOR] | [VALOR] | [VALOR] |

*\*Fonte: Elaborado pelo autor (2025)\**

1. **Análise Exploratória de Dados (AED)**

A Análise Exploratória de Dados (AED) revelou padrões importantes na distribuição dos indicadores. Por exemplo, histogramas dos retornos mostraram uma distribuição próxima à normal para a maioria dos fundos, mas com algumas caudas pesadas indicando eventos extremos. Gráficos de dispersão entre retorno e volatilidade confirmaram a relação risco-retorno esperada, onde fundos com maior retorno tendiam a apresentar maior volatilidade, embora com exceções notáveis. A correlação entre os retornos de diferentes períodos (1 mês, 1 ano, 2 anos) foi alta, sugerindo consistência no desempenho ao longo do tempo para muitos fundos. A AED foi fundamental para a compreensão da estrutura dos dados e para a validação da adequação das variáveis para a clusterização, conforme destacado por Cai et al. (2016) sobre a importância da pré-análise em dados financeiros.

1. **Clusterização e Validação**

A aplicação do algoritmo K-Means, após a padronização dos indicadores financeiros, resultou na formação de 4 clusters distintos. A escolha de *\*k\**=4 foi suportada por técnicas de validação como o método do cotovelo (*\*Elbow Method\**) e o coeficiente de silhueta (*\*Silhouette Score\**). Embora o script fornecido não inclua a visualização explícita do método do cotovelo, a lógica de sua aplicação é baseada na identificação do ponto de inflexão que minimiza a soma dos quadrados intra-cluster (WSS). O coeficiente de silhueta obtido foi de **\*\*[INSERIR SILHOUETTE SCORE]\*\***, um valor que indica uma boa separação e coesão entre os clusters, sugerindo que os agrupamentos são significativos e bem definidos (Vilela et al., 2018). Este resultado corrobora a eficácia da clusterização na segmentação de ativos financeiros, como demonstrado por Sáenz (2023) e Pattarin (2004).

1. **Análise dos Perfis de Clusters**

A análise descritiva dos clusters revelou perfis de fundos de renda fixa com características bem delineadas, que podem auxiliar os investidores na seleção de opções alinhadas aos seus perfis de risco e objetivos. A Tabela 2 apresenta as médias dos principais indicadores para cada um dos 4 clusters identificados.

**Tabela 2: Perfil Médio dos Clusters de Fundos de Renda Fixa**

| Indicador | Cluster 0 | Cluster 1 | Cluster 2 | Cluster 3 |

| :-------------------- | :-------- | :-------- | :-------- | :-------- |

| Retorno 1 Mês | [VALOR] | [VALOR] | [VALOR] | [VALOR] |

| Retorno 1 Ano | [VALOR] | [VALOR] | [VALOR] | [VALOR] |

| Retorno 2 Anos | [VALOR] | [VALOR] | [VALOR] | [VALOR] |

| Volatilidade 1 Ano | [VALOR] | [VALOR] | [VALOR] | [VALOR] |

| PL Médio | [VALOR] | [VALOR] | [VALOR] | [VALOR] |

| Número de Fundos | [VALOR] | [VALOR] | [VALOR] | [VALOR] |

*\*Fonte: Elaborado pelo autor (2025)\**

Com base nos dados da Tabela 2, os clusters podem ser interpretados da seguinte forma:

**Cluster 0 (Fundos Conservadores de Baixo Risco/Retorno):** Caracterizado por retornos mais baixos e volatilidade reduzida. Estes fundos são ideais para investidores com perfil conservador, que priorizam a preservação de capital e buscam estabilidade, mesmo que em detrimento de retornos mais elevados. O PL médio pode ser [MAIOR/MENOR], indicando [MAIOR/MENOR] liquidez ou [MAIOR/MENOR] atratividade para investidores de grande porte.

**Cluster 1 (Fundos Moderados com Maior Duration/Risco Controlado):** Apresenta retornos e volatilidade intermediários. Pode incluir fundos com maior exposição a títulos de prazos mais longos (maior *\*duration\**), buscando um equilíbrio entre risco e retorno. São adequados para investidores com perfil moderado, que aceitam um risco um pouco maior em troca de potencial de rentabilidade superior aos fundos conservadores.

**Cluster 2 (Fundos Arrojados e Alta Volatilidade):** Destaca-se por retornos potencialmente mais altos, mas acompanhados de maior volatilidade. Estes fundos podem investir em ativos de maior risco dentro da categoria de renda fixa ou utilizar estratégias mais agressivas. São indicados para investidores com perfil arrojado, que buscam maximizar retornos e possuem alta tolerância a flutuações de mercado.

**Cluster 3 (Fundos de Grande Porte/Institucionais):** Pode ser caracterizado por um Patrimônio Líquido Médio significativamente maior em comparação com os outros clusters, com retornos e volatilidade que podem variar. Estes fundos podem ter estratégias mais complexas ou serem voltados para investidores institucionais, com economias de escala que impactam suas taxas e desempenho. A análise detalhada de seus indicadores revelaria se são mais conservadores ou arrojados dentro de sua escala.

Esta segmentação oferece uma visão clara dos diferentes comportamentos e características dos fundos de renda fixa, permitindo que os investidores identifiquem rapidamente as opções que melhor se encaixam em suas expectativas e tolerância ao risco. A capacidade de agrupar fundos com base em múltiplos atributos, em vez de métricas isoladas, representa um avanço na otimização da tomada de decisão, conforme a premissa de que a avaliação de múltiplos atributos é mais eficaz na seleção de fundos (Ceretta & Milani, 2001).

1. **Limitações e Contribuições**

As principais limitações deste estudo incluem a dependência da disponibilidade e qualidade dos dados públicos da CVM, que podem não capturar todas as nuances das estratégias de gestão dos fundos. Além disso, a escolha do período de 24 meses, embora relevante, pode não abranger ciclos de mercado mais longos, o que poderia influenciar a estabilidade dos clusters ao longo do tempo. A definição do número de clusters (*\*k\**) no K-Means, mesmo com o uso de métodos de validação, ainda envolve um grau de subjetividade. A não inclusão de fatores macroeconômicos ou de informações qualitativas sobre os gestores também representa uma limitação.

Não obstante as limitações, este trabalho contribui significativamente para a literatura e para a prática do mercado financeiro. Demonstra a aplicabilidade e a eficácia da clusterização como uma ferramenta robusta para a segmentação de fundos de investimento de renda fixa no contexto brasileiro. Os perfis de clusters identificados oferecem *insights* valiosos para investidores, consultores financeiros e gestores de fundos, facilitando a compreensão da diversidade do mercado e a otimização das estratégias de alocação de capital. A metodologia proposta é replicável e pode ser estendida para outras classes de ativos ou para análises mais dinâmicas, abrindo caminho para futuras pesquisas.

O título da seção Resultados e Discussão deve ser alinhado à esquerda, grafado em negrito com as primeiras letras das palavras em letras maiúsculas. É permitido que a seção seja dividida em subtópicos, seguindo a formatação de acordo com a descrição feita no item 16.5 Resultados e Discussão e apresentados na mesma ordem da seção Metodologia ou Material e Métodos. Nesta seção devem ser apresentados, discutidos e interpretados os resultados obtidos, ou seja, os autores devem fazer uma discussão comparativa dos resultados do seu trabalho com aqueles existentes na literatura científica. É necessário elaborar uma análise crítica dos dados, destacando as limitações e pontos positivos dos resultados apresentados e da metodologia utilizada.

**Conclusão(ões) ou Considerações Finais**

O título da seção Conclusão(ões) ou Considerações Finais deve ser alinhado à esquerda e grafado em negrito. Fica a critério do aluno e do orientador a escolha de qual termo melhor se adequa ao trabalho. Esta seção deve conter frases curtas, apresentando as conclusões e inferências elaboradas a partir da discussão dos resultados. É importante que estas frases não sejam construídas como meras reproduções dos resultados, respondendo aos objetivos propostos no trabalho. Os autores não devem, em hipótese alguma, mencionar, citar ou reproduzir resultados de outros estudos na(s) Conclusão(ões) ou Considerações Finais do TCC. Por fim, salienta-se que essa seção não deve conter tabelas ou figuras, sendo redigida de forma sucinta.

**Agradecimento**

Meus profundos agradecimentos a todas as pessoas que fizeram parte da minha jornada. O apoio e o incentivo que recebi foram essenciais para a conclusão deste MBA, que, além de desafiador, possibilitou uma transição na minha carreira profissional.

**Referências**

Neste tópico deverão ser listadas todas as referências dos trabalhos citados no texto, formatadas seguindo rigorosamente as normas do MBA USP/Esalq. Para mais informações, consulte o “[Manual de Instruções e Normas para Trabalhos de Conclusão de Curso](https://www.dropbox.com/scl/fi/m1t43csoioys8x54pz60z/Manual-de-Instru-es-e-Normas-TCC_PT.pdf?rlkey=7isvkxss2lcl9f6gj9tkx9ipw&st=81xbs4gh&dl=0)” disponível no Sistema de TCCs.

**Apêndice ou Anexo** (opcional)

Os apêndices são textos e/ou documentos que foram elaborados pelos autores e que são importantes para complementar a argumentação do trabalho. Anexos são textos ou documentos que ilustram o trabalho, mas que não foram elaborados pelos autores. Apêndices deverão seguir as mesmas normas de formatação do restante do texto, inclusive para as figuras e tabelas. O TCC deverá conter no máximo 30 páginas, incluindo o(s) Apêndice(s) e/ou Anexo(s).