UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ENGENHARIA DE SÃO CARLOS

LUCAS TEÓFILO DE CASTRO

Análise de Indicadores Financeiros e Clusterização das Principais Empresas do Ibovespa via *k-means*

São Carlos
2023

LUCAS TEÓFILO DE CASTRO

Análise de Indicadores Financeiros e Clusterização das Principais Empresas do Ibovespa via *k-means*

Monografia submetida ao Curso de Engenharia de Materiais e Manufatura, na renomada Escola de Engenharia de São Carlos da Universidade de São Paulo, como critério parcial para a obtenção do título de Engenheiro de Materiais e Manufatura.

Orientador: Prof. Dr. Lucas Gabriel Zanon

São Carlos

2023

Ficha catalográfica elaborada pela Biblioteca Prof. Dr. Sérgio Rodrigues Fontes da EESC/USP com os dados inseridos pelo(a) autor(a).

Teofilo de Castro, Lucas
T355a Análise de Indicadores Financeiros e Clusterização das Principais Empresas do Ibovespa via k-means / Lucas Teofilo de Castro; orientador Lucas Gabriel Zanon. São Carlos, 2023.

Monografia (Graduação em Engenharia de Materiais e Manufatura) -- Escola de Engenharia de São Carlos da Universidade de São Paulo, 2023.

1. K-means. 2. Análise de Cluster. 3. Mercado Financeiro. 4. B3. 5. Indicadores Financeiros. I. Título.

Eduardo Graziosi Silva - CRB - 8/8907

FOLHA DE AVALIAÇÃO OU APROVAÇÃO

Candidato / *Student*: Lucas Teofilo de Castro
Título do TCC / *Title*: Análise de Indicadores Financeiros e Clusterização das Principais Empresas do Ibovespa via k-means
Data de defesa / *Date*: 13/12/2023

Comissão Julgadora / *Examining committee*	**Resultado** / *Result*
Professor Lucas Gabriel Zanon (orientador)	Aprovado
Instituição / *Affiliation*: EESC - SEP	
Doutora Ana Carolina Bertassini	Aprovado
Instituição / *Affiliation*: EESC - SEP	
Doutor Luiz Cesar Ribeiro Carpinetti	Aprovado
Instituição / *Affiliation*: EESC - SEP	

Presidente da Banca / *Chair of the Examining Committee*

Professor Lucas Gabriel Zanon

AGRADECIMENTOS

Aos meus pais Marisa e Ricardo, expresso minha profunda gratidão por todo o apoio e ajuda, que foram cruciais para a realização deste trabalho.

À minha irmã Julia, agradeço por me incentivar nos momentos desafiadores e por me auxiliar na motivação e dedicação ao longo deste trabalho.

Aos amigos, em especial Igor Cunha, Gabriel Lobato, Thales Arbache, Carlos Esteves, Vinicius Rocha e Isadora Francisco agradeço pela amizade incondicional e pelo apoio demonstrado ao longo de todo o período de dedicação a este curso e especialmente a este trabalho.

À República Zero Bala, pelos momentos memoráveis e suporte durante minha jornada acadêmica.

Ao Professor Dr. Lucas Gabriel Zanon, expresso minha gratidão por aceitar ser meu orientador e desempenhar essa função com tanta dedicação e amizade.

Aos demais professores do departamento de Engenharia de Materiais e Manufatura, agradeço pelas correções e ensinamentos que me permitiram evoluir pessoal e profissionalmente, contribuindo para meu desempenho no processo de formação ao longo do curso.

A todos que participaram, direta ou indiretamente, no desenvolvimento deste trabalho de pesquisa, enriquecendo meu processo de aprendizado, minha sincera gratidão.

A todos os alunos da minha turma, agradeço pelo ambiente amistoso no qual convivemos e solidificamos nossos conhecimentos, o que foi fundamental na elaboração deste trabalho de conclusão de curso.

RESUMO

No contexto do mercado financeiro brasileiro, caracterizado pela sua complexidade e dinamismo, este trabalho explora a aplicação de técnicas de análise de dados e aprendizado de máquina. Com uma abordagem quantitativa, o estudo emprega o algoritmo *K-means* para segmentar empresas listadas na bolsa de valores (B3), utilizando indicadores financeiros como Margem *EBITDA*, *CAGR* de receitas de 5 anos, *ROE*, *EV/EBITDA* e Dívida Líquida/*EBITDA*. Esta análise identifica padrões e características comuns entre as empresas, resultando na formação de *clusters* distintos, que refletem diferentes perfis de eficiência, crescimento e valorização no mercado. Estes resultados oferecem perspectivas sobre a eficiência e crescimento no mercado de ações brasileiro, contribuindo para o entendimento das estratégias financeiras e operacionais das empresas listadas, e fornecendo uma ferramenta útil para investidores e analistas na tomada de decisões de investimento baseadas em dados.

Palavras-chave: *K-means*, Análise de Cluster, Mercado Financeiro, B3, Indicadores Financeiros.

ABSTRACT

In the context of the Brazilian financial market, characterized by its complexity and dynamism, this work explores the application of advanced data analysis techniques and machine learning. With a quantitative approach, the study employs the K-means algorithm to segment companies listed on the stock market, using financial indicators such as EBITDA Margin, 5-year Revenue CAGR, ROE, EV/EBITDA, and Net Debt/EBITDA. This analysis identifies common patterns and characteristics among companies, resulting in the formation of distinct clusters that reflect different profiles of efficiency, growth, and market valuation. These results offer valuable perspectives on efficiency and growth in the Brazilian stock market, contributing to the understanding of the financial and operational strategies of the listed companies, and providing a useful tool for investors and analysts in data-based investment decision-making.

Keywords: K-means, Cluster Analysis, Financial Market, B3, Financial Indicators.

LISTA DE ILUSTRAÇÕES

Figura 1 - Gráfico CAPM....26
Figura 2 - Aprendizado de Máquina: Subáreas e Aplicações....27
Figura 3 - Dados agrupados utilizando k-means....30
Figura 4 - Exemplo Gráfico do Método do Cotovelo....31
Figura 5 – Exemplo de Demonstrativo do Resultado do Exercício (DRE)....36
Figura 6 - Fluxograma do processo da coleta de dados....37
Figura 7 - Diagrama representando bibliotecas utilizadas no software Python e suas funcionalidades....38
Figura 8 - Gráfico de Dispersão do CAGR Receita 5 anos vs Margem EBITDA....39
Figura 9 - Exemplo de outliers em um gráfico....40
Figura 10 - Desvio padrão numa amostra....40
Figura 11 - Método do Cotovelo gerado pelo código....42
Figura 12 - Dados antes e depois da clusterização por k-means....43
Figura 13 - Representação esquemática do método aplicado na pesquisa....43
Figura 14 - Clusters CAGR Receita 5 anos X Margem EBITDA....46
Figura 15 - Clusters EV/EBITDA e ROE (%)....49
Figura 16 - Clusters Dívida Líquida/EBITDA e Margem EBITDA....53
Figura 17 - Clusters EV/EBITDA e CAGR Receita 5 Anos....58

LISTA DE TABELAS

Tabela 1: Tabela com as ações estudadas junto de seus indicadores.....44

LISTA DE ABREVIATURAS E SIGLAS

EBITDA - Earnings Before Interest, Taxes, Depreciation, and Amortization

CAGR - Compound Annual Growth Rate

ROE - Return on Equity

EV/EBITDA - Enterprise Value/EBITDA

ML - Machine Learning

CAPM - Capital Asset Pricing Model

RBS - Revisão Bibliográfica Sistemática

DRE - Demonstrativo do Resultado do Exercício

CVM - Comissão de Valores Mobiliários

ITUB4, PETR4, VALE3, etc. - Códigos de ações específicas na B3

SUMÁRIO

1 INTRODUÇÃO..... 22
1.1 Contextualização..... 22
1.2 Objetivos..... 23
2 REFERENCIAL TEÓRICO..... 25
2.1 Aspectos dos Investimentos em Renda Variável..... 25
2.2 Aprendizado de Máquina (Machine Learning) e Análise de Clusters..... 26
2.3 O Algoritmo k-means..... 29
3 MÉTODO DE PESQUISA..... 33
3.1 Escolha de indicadores..... 34
3.2 Coleta de dados..... 35
3.3 Implementação e Descrição do Algoritmo de Clusterização..... 37
4 RESULTADOS E DISCUSSÃO..... 44
Tabela 1: Tabela com as ações estudadas junto de seus indicadores..... 44
4.1 Eficiência x Crescimento..... 46
4.2 Valorização x Rentabilidade..... 49
4.3 Eficiência x Endividamento..... 52
4.4 Crescimento x Valorização..... 57
CONCLUSÕES..... 62
REFERÊNCIAS..... 64
APÊNDICE..... 66

1 INTRODUÇÃO

1.1 Contextualização

O mercado financeiro brasileiro, particularmente representado pelo Ibovespa, é uma área rica para a análise de indicadores financeiros e seu impacto nas decisões de investimento. Este trabalho explora a relevância desses indicadores na avaliação da saúde econômica das empresas listadas no Ibovespa. Adotando uma abordagem que combina análises fundamentais e técnicas, o estudo se concentra em métricas-chave como lucratividade, receita, fluxo de caixa e endividamento. Esta abordagem é apoiada por estudos como o de Damodaran (2012) em *'Investment Valuation*', que enfatiza a importância da análise fundamental, e o de Murphy (1999) em '*Technical Analysis of the Financial Markets*', destacando técnicas de análise de mercado.

O uso de ferramentas avançadas de análise financeira é crucial para interpretar dados de mercado de forma precisa. O estudo aprofunda-se na análise de relatórios financeiros, uma etapa fundamental para entender a posição financeira das empresas. Isso é estudado por Penman (2013) em *'Financial Statement Analysis and Security Valuation*', que ilustra a importância da análise de relatórios financeiros.

Além disso, este trabalho incorpora a técnica de clusterização via algoritmo *k-means* para agrupar empresas com perfis financeiros semelhantes, uma metodologia que oferece perspectivas sobre as dinâmicas do mercado financeiro brasileiro. Esta abordagem está alinhada com as ideias apresentadas por James et al. (2013) em '*An Introduction to Statistical Learning*', que discute o uso de métodos de aprendizado de máquina, como *k-means*, em aplicações financeiras.

A análise fundamental examina a saúde financeira e o desempenho das empresas, considerando fatores como receita, lucro, dívida e mais, fornecendo uma visão da estabilidade e do potencial de crescimento de uma empresa (ARAÚJO, 2021). Por outro lado, a análise técnica utiliza padrões históricos de dados de mercado, como preços de ações e volumes de negociação, para prever movimentos futuros do mercado (Domingues et al., 2022).

A integração dessas duas abordagens permite uma compreensão mais aprofundada das tendências do mercado e das empresas individualmente. Este estudo visa explorar como essa combinação pode ser usada para identificar oportunidades de investimento no mercado acionário brasileiro, particularmente no Ibovespa, que é um dos principais índices e reflete o desempenho das maiores empresas do país (ARAÚJO, 2021; DOMINGUES et al., 2022).

Dentro da contextualização da pesquisa, o algoritmo k-means desempenha um papel importante ao agrupar empresas com características financeiras similares, revelando padrões e segmentações no mercado acionário que não são imediatamente óbvios. No estudo de Nunes (2016), intitulado "Um breve estudo sobre o algoritmo K-means", a eficácia do algoritmo k-means é destacada, especialmente em sua capacidade de convergir dados em clusters, aplicando-se a diversos problemas.

Além das abordagens fundamentais e técnicas, a inovação tecnológica, especialmente em inteligência artificial e aprendizado de máquina, está revolucionando a análise de dados no mercado financeiro. A integração dessas tecnologias permite uma análise mais eficiente e profunda de grandes conjuntos de dados financeiros, revelando tendências e padrões ocultos cruciais para previsões de mercado e tomadas de decisão de investimento (LUDERMIR, 2021).

Paralelamente, o contexto econômico e político brasileiro exerce influência significativa no Ibovespa. A análise do impacto de políticas econômicas e eventos políticos no mercado acionário brasileiro oferece uma perspectiva importante, demonstrando como fatores externos afetam o comportamento do mercado (MONTES, 2008). Essa dimensão adiciona uma camada de complexidade à pesquisa, permitindo uma compreensão mais ampla das forças que moldam o mercado de ações no Brasil.

Portanto, a questão orientadora do presente trabalho é: de que forma a técnica *K-means* pode ser utilizada para segmentação de opções listadas na B3 e viabilizar a realização de análises sobre as mesmas no mercado financeiro?

1.2 Objetivos

O presente trabalho de conclusão de curso objetiva aplicar as técnicas de aprendizado de máquina, especificamente o algoritmo *K-means*, na análise e segmentação de ações do mercado financeiro brasileiro, utilizando indicadores financeiros para agrupar empresas listadas na B3.

O objetivo geral desdobra-se nos seguintes objetivos específicos:

1. Selecionar e analisar indicadores financeiros relevantes que podem ser utilizados para a segmentação de empresas, como Margem EBITDA, CAGR de receitas de 5 anos, ROE, EV/EBITDA e Dívida Líquida/EBITDA.

2. Aplicar o algoritmo de clusterização K-means aos dados financeiros das empresas listadas na B3 para identificar padrões e agrupamentos significativos.
3. Analisar os clusters formados para entender as características comuns e diferenciadoras entre as empresas, focando em aspectos como eficiência operacional, crescimento e valorização de mercado.
4. Interpretar os resultados à luz da teoria financeira e das condições do mercado de ações brasileiro, visando compreender as implicações práticas para investidores e analistas.

2 REFERENCIAL TEÓRICO

2.1 Aspectos dos Investimentos em Renda Variável

A diversidade e complexidade do universo de investimentos têm intrigado e fascinado acadêmicos e profissionais por décadas. No contexto da renda variável, em particular, existe uma rica tapeçaria de estudos e pesquisas que oferecem insights e abordagens para navegar neste ambiente dinâmico e muitas vezes volátil.

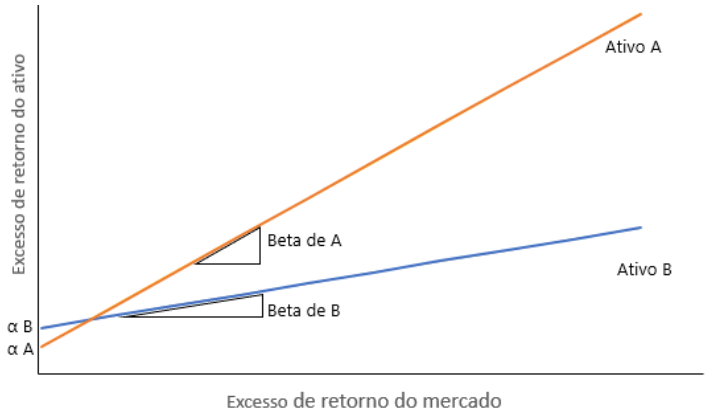
Historicamente, o mercado de renda variável foi visto como uma das formas mais tradicionais de investimento, ao lado dos mercados de renda fixa e commodities. Seus atrativos retornos potenciais, no entanto, vêm acompanhados de riscos consideráveis. Graham e Dodd (1934) em seu seminal "Security Analysis", por exemplo, lançaram as bases para a análise fundamentalista de ações. Eles defendiam que, por meio de uma avaliação cuidadosa das demonstrações financeiras de uma empresa, um investidor poderia determinar seu valor intrínseco e, assim, identificar oportunidades de investimento.

Fama (1970), em seu influente trabalho sobre mercados eficientes, argumentou que os preços das ações refletem sempre todas as informações disponíveis. Isso tornaria a tarefa de superar o mercado consistentemente, através de estratégias de seleção de ações, extremamente desafiadoras. Esse pensamento deu origem ao conceito de "forma eficiente" do mercado e teve um impacto profundo sobre a gestão de investimentos, incentivando uma mudança em direção aos investimentos passivos.

Por outro lado, estudos mais recentes têm identificado anomalias de mercado que desafiam a teoria do mercado eficiente. Por exemplo, Jegadeesh e Titman (1993) identificaram momentum em retornos de ações, sugerindo que ações que tiveram bom desempenho no passado tendem a continuar superando no futuro próximo. Este e outros trabalhos semelhantes abriram as portas para estratégias de investimento quantitativo baseadas em fatores de risco.

Outro tópico central nos estudos sobre renda variável é o conceito de risco e recompensa. A Capital Asset Pricing Model (CAPM) proposta por Sharpe (1964) sugere que o retorno esperado de um ativo é linearmente relacionado ao seu beta de mercado (como pode ser observado na Figura 1). Embora amplamente ensinado e utilizado, o CAPM tem sido objeto de críticas e revisões, com muitos sugerindo que outros fatores, além do beta de mercado, podem ser importantes preditores de retornos futuros.

Figura 1 - Gráfico CAPM



Fonte: Pro Educacional (2008).

Mais recentemente, com a evolução das tecnologias de informação e o advento da era digital, grandes conjuntos de dados (big data) e técnicas de análise avançada, como machine learning, estão sendo aplicados no domínio dos investimentos. Estes novos métodos estão permitindo que os investidores explorem padrões e correlações anteriormente não reconhecidos (KORAJCZYK; MURPHY, 2020).

A crescente globalização dos mercados financeiros também trouxe desafios adicionais e oportunidades para investidores em renda variável. Estudos como o de Bekaert e Harvey (2000) destacam a importância de considerar correlações dinâmicas ao investir internacionalmente, dada a evolução das interdependências de mercado ao longo do tempo.

Em suma, o campo dos investimentos em renda variável é vasto e em constante evolução. As teorias e estratégias que foram relevantes em uma época podem não ser mais aplicáveis hoje. É, portanto, essencial para os investidores manterem-se atualizados com as mais recentes pesquisas e desenvolvimentos acadêmicos, bem como serem adaptáveis e flexíveis em suas abordagens de investimento.

2.2 Aprendizado de Máquina (*Machine Learning*) e Análise de *Clusters*

A evolução da tecnologia levou ao surgimento de sistemas que se ajustam e evoluem com base na informação que recebem. Esta capacidade de se adaptar, de aprender, deu origem ao campo do Aprendizado de Máquina (ML). Originando-se como uma ramificação da Inteligência Artificial, o ML foca na capacidade das máquinas de aprender e se adaptar sem serem explicitamente programadas para uma tarefa específica (SAMUEL, 1959). De detecção

de spam a sistemas de recomendação, o ML tem transformado a maneira como interagimos com a tecnologia (JORDAN & MITCHELL, 2015).

O que diferencia o ML de outras abordagens computacionais é o seu enfoque nos dados. Em vez de um programa seguir um conjunto específico de instruções, ele utiliza dados para criar um modelo. Esse modelo é então usado para fazer previsões ou tomar decisões (HASTIE, TIBSHIRANI & FRIEDMAN, 2009).

O aprendizado de máquina pode ser subdividido em várias categorias, sendo as mais comuns: supervisionado, não supervisionado e por reforço, como podem ser vistos na Figura 2. No aprendizado supervisionado, modelos são treinados usando conjuntos de dados com rótulos pré-definidos. Já no aprendizado não supervisionado, não há rótulos, e o algoritmo tenta identificar estruturas nos dados. É aqui que entra a análise de clusters (MURPHY, 2012).

Figura 2 - Aprendizado de Máquina: Subáreas e Aplicações.

Fonte: Aquarela Analytics (2015).

A clusterização é um tipo de aprendizado não supervisionado onde o objetivo é agrupar objetos semelhantes, ou seja, formar "clusters". Um cluster é essencialmente um conjunto de dados que são similares entre si e distintos de dados em outros clusters. A ideia é que os dados dentro de um cluster compartilhem certas características ou propriedades que os diferenciam dos dados fora desse cluster (PAPENBROCK, 2011; LINDEN, 2009).

Para determinar esses clusters, uma métrica de distância é utilizada. Esta métrica avalia o quão "próximos" ou "distantes" os pontos de dados estão entre si. Várias métricas de

distância, como a distância euclidiana ou a distância de Manhattan, podem ser empregadas dependendo do tipo de dados e do problema (SIBSON, 1973).

Uma vez que uma métrica é escolhida, o próximo passo é a seleção do algoritmo de clusterização. O algoritmo de K-means, por exemplo, é um dos mais populares e envolve a seleção de "k" centros e a alocação de pontos de dados ao centro mais próximo. O processo é repetido até que os centros não se movam significativamente ou atinjam um número máximo de iterações (THORNDIKE, 1953).

A capacidade dos modelos de Aprendizado de Máquina de se adaptar e evoluir com novos dados os torna uma ferramenta valiosa em diversas áreas, desde medicina e biologia até finanças e marketing (GOODFELLOW et al., 2016). Esta adaptabilidade provém da capacidade do modelo de ajustar seus parâmetros internos em resposta aos feedbacks do ambiente externo. Além das categorias de aprendizado já mencionadas, também temos o aprendizado semi-supervisionado e o aprendizado por transferência, ampliando ainda mais as possibilidades de uso do ML.

Ao abordar Análise de Clusters, é essencial destacar que os clusters são, em sua essência, agrupamentos de dados com características semelhantes. O sucesso dessa técnica depende da escolha correta do algoritmo, bem como da métrica de distância, como você mencionou anteriormente. Em contextos mais amplos, a clusterização pode ser usada em áreas como biologia, para agrupar genes com funções similares, ou em marketing, para segmentar clientes com base em comportamentos de compra (TAN et al., 2005).

Enquanto o aprendizado supervisionado busca fazer previsões com base em exemplos rotulados, o aprendizado não supervisionado, do qual a clusterização é um subconjunto, busca identificar estruturas subjacentes nos dados sem a necessidade de rótulos pré-definidos. O K-means, um dos algoritmos mais conhecidos para clusterização, procura minimizar a variação dentro de cada cluster, garantindo que os dados dentro de um cluster sejam o mais homogêneos possível (BISHOP, 2006).

Um ponto notável é que, com o advento de técnicas de aprendizado profundo e redes neurais, a fronteira entre aprendizado supervisionado e não supervisionado tornou-se mais permeável. Por exemplo, autoencoders, uma forma de rede neural, são usados para redução de dimensionalidade e podem ser combinados com técnicas de clusterização para fornecer insights mais profundos sobre os dados (HINTON et al., 2006).

2.3 O Algoritmo *k-means*

Dentro do vasto domínio do aprendizado de máquina, o algoritmo k-means tem sido um pilar central de métodos de clusterização. Desde sua concepção por MacQueen em 1967, este algoritmo não apenas pavimentou o caminho para avanços significativos na área de data science, mas também provou ser uma ferramenta indispensável em uma variedade de aplicações práticas, de análises de mercado a sistemas de recomendação (MACQUEEN, 1967).

A lógica fundamental do k-means é agrupar um conjunto de dados de *n* observações em *k* clusters distintos (Figura 3). Cada observação é atribuída ao cluster cujo centroide (a média dos pontos desse cluster) é o mais próximo.

O algoritmo é fundamentado em princípios matemáticos que visam otimizar a alocação de um conjunto de dados em clusters distintos. Matematicamente, o objetivo principal do *K-means* é minimizar a soma das distâncias quadradas entre os pontos de dados e o centroide de seu respectivo cluster. Esse processo é conhecido como minimização da variação intra-cluster.

Inicialmente, o algoritmo seleciona k centroides de forma aleatória ou com base em uma heurística. Esses centroides são pontos representativos iniciais para cada um dos *k clusters*. Em seguida, cada ponto de dado no conjunto é atribuído ao *cluster* cujo centroide é mais próximo. A proximidade é geralmente determinada usando a distância euclidiana, que é calculada como a raiz quadrada da soma das diferenças quadradas entre as coordenadas de dois pontos.

Após a atribuição inicial de todos os pontos a um cluster, o centroide de cada cluster é recalculado. Isso é feito encontrando a média de todas as coordenadas dos pontos pertencentes a esse cluster. A atualização dos centroides é uma etapa crucial, pois reflete a nova posição central com base nos pontos atribuídos.

O algoritmo então repete as etapas de atribuição e atualização de centroides sucessivamente. Este processo iterativo continua até que uma condição de parada seja satisfeita, que ocorre quando os centroides não se movem significativamente entre as iterações consecutivas ou quando as atribuições de pontos aos clusters permanecem constantes.

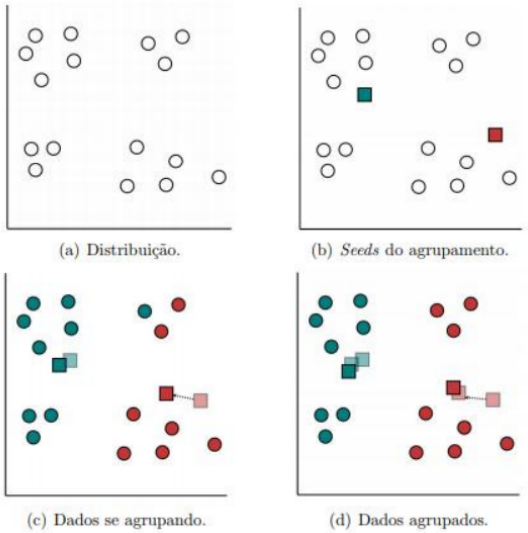
Matematicamente, esse processo iterativo pode ser visto como uma forma de otimização, onde o algoritmo busca encontrar a configuração de clusters que minimiza a soma total das distâncias quadradas de cada ponto ao centroide de seu respectivo cluster. Esta

abordagem é eficaz para identificar agrupamentos naturais dentro de um conjunto de dados, desde que os clusters tendam a ser esfericamente distribuídos e de tamanhos relativamente similares. Formalmente, isso pode ser representado pela equação:

$$d(P, X) = \frac{1}{n} \sum_{i=1}^{n} d(p_i, x)^2 \qquad \textbf{(1)}$$

O algoritmo é executado em etapas iterativas. Inicia-se com uma seleção, que pode ser aleatória ou heurística, dos *k* centroides. O que se segue é um processo contínuo de atribuição de pontos ao centroide mais próximo e de nova determinação dos centroides. Essa iteração é repetida até que os centroides não se movam significativamente entre iterações consecutivas ou até que um limite máximo de iterações seja atingido.

Figura 3 - Dados agrupados utilizando k-means



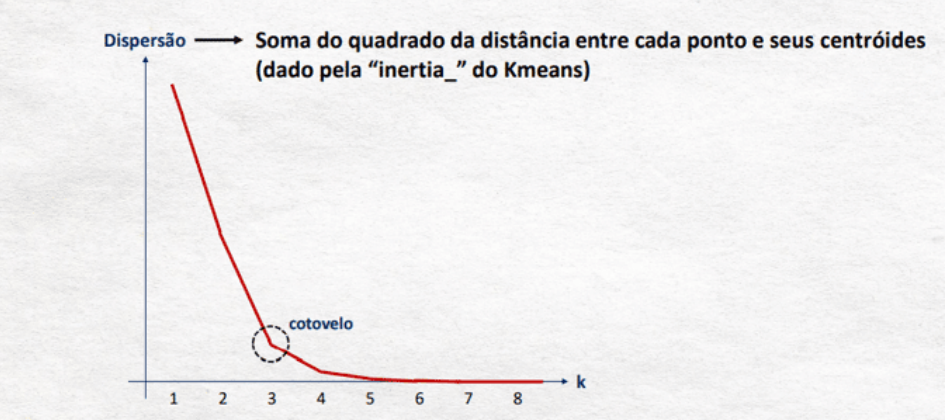
Fonte: Prado (2008).

A determinação do número ideal de clusters, *k*, é talvez uma das etapas mais desafiadoras na aplicação do k-means. A escolha do número de clusters apropriado é vital porque pode influenciar diretamente a qualidade da clusterização. O "método do cotovelo" é uma das abordagens mais populares para determinar o *k* ótimo.

Neste método, a clusterização k-means é executada para uma série de valores de *k* (por exemplo, *k* de 1 a 100), e para cada valor, a inércia total é calculada. A inércia é definida como a soma das distâncias quadradas das amostras ao centroide de cluster mais próximo e é

uma medida da coesão interna dos clusters. Ao plotar os valores de inércia para diferentes números de clusters, geralmente observa-se uma curva (Figura 4). O ponto onde a diminuição da inércia começa a se estabilizar, formando um "cotovelo" na curva, sugere um número ótimo de clusters para o conjunto de dados (KODINARIYA et al., 2013).

Figura 4 - Exemplo Gráfico do Método do Cotovelo



Fonte: *Hashtag* Treinamentos (2017).

Além disso, a entropia, um conceito derivado da teoria da informação, pode ser extremamente útil na validação e interpretação dos clusters identificados pelo k-means. A entropia, como medida da incerteza ou impureza dentro de um conjunto de dados, pode ser aplicada para avaliar a homogeneidade dentro dos clusters. Quando utilizada em aprendizado não-supervisionado, onde os dados não são etiquetados, a entropia ajuda a determinar se os clusters formados são consistentes com as etiquetas conhecidas. Esta aplicação é discutida em detalhes por Hastie et al. (2009) em '*The Elements of Statistical Learning*', onde enfatizam o valor da entropia na avaliação da qualidade dos clusters em cenários de aprendizado não-supervisionados.

Portanto, a combinação do método do cotovelo e da análise de entropia no contexto do k-means oferece uma abordagem informativa para a análise de cluster, particularmente em situações onde o aprendizado não-supervisionado é aplicável. Essas técnicas juntas fornecem uma maneira eficaz de validar a estrutura dos clusters e garantir que eles sejam estatisticamente significativos e relevantes para as categorias predefinidas nos dados.

Em síntese, ao aplicar o k-means, é fundamental considerar cuidadosamente o valor de *k*. O método do cotovelo, empregando inércia e, em certos contextos, entropia, fornece uma abordagem sistemática para determinar um número ótimo de clusters. Estas ferramentas,

quando usadas corretamente, melhoram a qualidade e a interpretabilidade dos resultados da clusterização.

3 MÉTODO DE PESQUISA

O presente estudo adota uma metodologia quantitativa e alicerçada em dados, inspirando-se em técnicas de análise de dados e aprendizado de máquina para examinar indicadores financeiros e realizar a clusterização das empresas líderes do Ibovespa. Esta metodologia incorpora práticas recomendadas e descobertas de pesquisas anteriores na área de análise financeira e aprendizado de máquina.

Neste estudo, busca-se explorar a complexidade e as nuances do mercado financeiro brasileiro utilizando uma abordagem multifacetada que combina análises financeiras com técnicas de aprendizado de máquina. A pesquisa se estrutura em cinco etapas principais, cada uma fundamentada em metodologias e teorias reconhecidas, visando fornecer resultados sobre o mercado de ações brasileiro.

1. **Seleção de Indicadores Financeiros**: A seleção dos indicadores financeiros fundamenta-se no trabalho de Koller, Goedhart e Wessels (2010), privilegiando métricas como Margem EBITDA e CAGR de Receita de 5 anos, reconhecidas por sua capacidade de refletir eficiência operacional e crescimento potencial. Em linha com as recomendações de Damodaran (2012), estes indicadores são usados para compreender o desenvolvimento sustentável das empresas ao longo do tempo.
2. **Coleta de Dados**: A coleta e verificação dos dados financeiros segue as orientações de Brigham e Ehrhardt (2013), assegurando a precisão e confiabilidade dos dados obtidos de fontes públicas e privadas. Esta etapa é crucial para garantir a integridade dos dados que fundamentam a análise.
3. **Implementação do Algoritmo K-means**: A implementação do algoritmo K-means segue a metodologia proposta por MacQueen (1967), com adaptações para o contexto financeiro. Utiliza-se o "Método do Cotovelo" conforme descrito por Kodinariya e Makwana (2013) para determinar o número ideal de clusters, um passo fundamental para garantir agrupamentos significativos.
4. **Análise e Interpretação dos Clusters**: A análise dos clusters formados será realizada com base na teoria financeira contemporânea e nas condições atuais do mercado de ações brasileiro. Serão exploradas as anomalias de mercado e estratégias de investimento quantitativo, seguindo as linhas de pesquisa estabelecidas por Fama (1970) e Jegadeesh e Titman (1993).

5. **Avaliação da Metodologia Adotada**: A eficácia da metodologia será avaliada em relação aos avanços recentes no campo do aprendizado de máquina, particularmente as contribuições de Goodfellow et al. (2016). Este aspecto é fundamental para assegurar que a abordagem adotada permaneça relevante e eficaz no contexto atual do mercado financeiro.

Esta metodologia aprimorada propõe uma combinação entre as práticas convencionais de análise financeira e as técnicas de aprendizado de máquina, proporcionando uma análise das dinâmicas do mercado de ações brasileiro. O objetivo é oferecer uma compreensão ampla e detalhada que seja fundamental para estratégias de investimento informadas e eficazes.

3.1 Escolha de indicadores

Uma etapa crucial neste estudo é a seleção dos indicadores financeiros que serão analisados. Dado que as empresas incluídas no índice Ibovespa pertencem a setores variados, foi imperativo escolher métricas que possam ser aplicáveis de forma generalizada, independentemente do setor. Abaixo estão os indicadores selecionados para o estudo, cada um com suas respectivas justificativas.

A Margem EBITDA é um indicador que mostra a porcentagem do lucro antes de juros, impostos, depreciação e amortização (EBITDA) em relação à receita total da empresa. Este é um indicador-chave da eficiência operacional de uma empresa e é frequentemente utilizado para comparar a rentabilidade entre diferentes empresas (KOLLER et al, 2010). Dada a sua aplicabilidade em vários setores, torna-se uma escolha ideal para este estudo. A margem segue a fórmula:

$$Margem\ EBITDA = \frac{EBITDA}{Receita\ Líquida \times 100} \qquad \textbf{(2)}$$

O próximo indicador utilizado no estudo foi o CAGR (Taxa de Crescimento Anual Composta) de Receita ao longo de 5 anos, este é uma medida do crescimento sustentável de uma empresa ao longo do tempo. Este indicador ajuda os investidores a entenderem a taxa de expansão dos negócios e é particularmente útil para comparar o crescimento entre empresas de setores distintos (DAMODARAN, 2012).

O Retorno sobre o Patrimônio Líquido (ROE) é uma métrica que mede a rentabilidade de uma empresa em relação ao seu patrimônio líquido. É um indicador amplamente utilizado para avaliar a eficiência com que a gestão está usando o capital dos acionistas para gerar lucros (BRIGHAM et al., 2013).

$$ROE = \frac{Lucro\ Líquido}{Patrimônio\ Líquido} \qquad \textbf{(3)}$$

O EV/EBITDA é uma razão de avaliação que compara o valor da empresa (*Enterprise Value* ou EV) ao seu lucro antes de juros, impostos, depreciação e amortização (EBITDA). É uma métrica útil para avaliar empresas com estruturas de capital diferentes e é especialmente útil quando se compara empresas de setores diversos (COPELAND et al., 2000).

A razão Dívida Líquida/EBITDA é uma medida da alavancagem financeira de uma empresa e indica quantos anos seriam necessários para pagar a dívida líquida com o EBITDA gerado. É um indicador crítico do risco financeiro e é frequentemente usado em análises de crédito e de investimento (ALTMAN, 1968).

Cada um desses indicadores foi selecionado com o intuito de fornecer uma visão abrangente do desempenho financeiro das empresas, considerando sua aplicabilidade em diferentes setores. Isso é particularmente importante dado o escopo deste estudo, que visa agrupar empresas do índice Ibovespa, notoriamente diversificado em termos de setores representados.

3.2 Coleta de dados

O processo de coleta de dados para este estudo envolve múltiplas etapas e fontes para garantir a abrangência e a precisão das informações utilizadas. As empresas objeto deste estudo são as 24 que mais têm participação no índice Ibovespa, uma escolha que oferece uma visão representativa da economia brasileira e da performance do mercado de ações.

Para começar, dados financeiros dessas empresas são coletados de bases de dados públicas e privadas, como a Comissão de Valores Mobiliários (CVM), Bloomberg, e relatórios anuais, mais precisamente os demonstrativos do resultado do exercício (DRE), como mostrado na Figura 5. Esses dados são necessários para calcular os indicadores financeiros que serão posteriormente utilizados para a clusterização. Dentre os indicadores coletados estão a Margem EBITDA, CAGR de Receita de 5 anos, ROE, EV/EBITDA e

Dívida Líquida/EBITDA. É fundamental que os dados sejam os mais atuais possíveis, e que sejam consistentes em termos de período de relato para permitir uma comparação justa.

Figura 5 – Exemplo de Demonstrativo do Resultado do Exercício (DRE)

Demonstração do Resultado do Exercício	2019	2020	2021
Receita Operacional	1.000.000,00	2.000.000,00	3.000.000,00
Deduções da Receita	-86.500,00	-173.000,00	-259.500,00
Receita Operacional Líquida	913.500,00	1.827.000,00	2.740.500,00
Custos Operacionais	-235.241,24	-370.750,00	-480.151,00
Lucro Bruto	**678.258,76**	**1.456.250,00**	**2.260.349,00**
% Margem Bruta	74%	80%	82%
Despesas Operacionais	-343.518,99	-649.700,92	-730.265,00
EBITDA	**334.739,77**	**806.549,08**	**1.530.084,00**
% Margem EBITDA	37%	44%	56%
Depreciação/ Amortização	-8.716,63	-10.018,01	-12.709,44
Receitas Financeiras	1.000,00	1.500,00	2.300,00
Despesas Financeiras	-119.534,16	-215.436,50	-193.644,98
Outras Receitas	3.000,00	7.000,00	6.000,00
Lucro Operacional	210.488,98	589.594,57	1.332.029,58
Tributos IRPJ e CSLL	-108.800,00	-217.600,00	-326.400,00
Lucro Líquido	**101.688,98**	**371.994,57**	**1.005.629,58**
% Margem Líquida	11%	20%	37%

% Margem Bruta é Receita Operacional Líquida/ Lucro Bruto
% Margem Ebitda é Receita Operacional Líquida/ EBITDA

Fonte: Suno *Research* (2020).

A qualidade dos dados é garantida por meio de uma revisão rigorosa e técnicas de limpeza de dados. Isso inclui a identificação e correção de outliers, a verificação de consistência temporal e a correção de possíveis erros de entrada ou de registro. Além disso, todos os dados são ajustados aos princípios contábeis geralmente aceitos, para assegurar que as métricas sejam comparáveis entre empresas que podem seguir diferentes normas contábeis.

Para o aspecto computacional e de modelagem, os dados são preparados e transformados em um formato que seja adequado para a análise de cluster utilizando o algoritmo *k-means*. Esse processo envolve a normalização dos dados para que as variáveis tenham o mesmo peso no modelo. Essa etapa é crucial, pois o *k-means* é sensível à escala das variáveis.

Dado o foco deste estudo no uso de métodos quantitativos, todas as etapas da coleta e do tratamento de dados são projetadas para serem replicáveis. Isso não apenas confere maior rigor científico à pesquisa, como também oferece a possibilidade de atualizações futuras e comparações longitudinais, um aspecto especialmente valioso em mercados financeiros em rápida mutação.

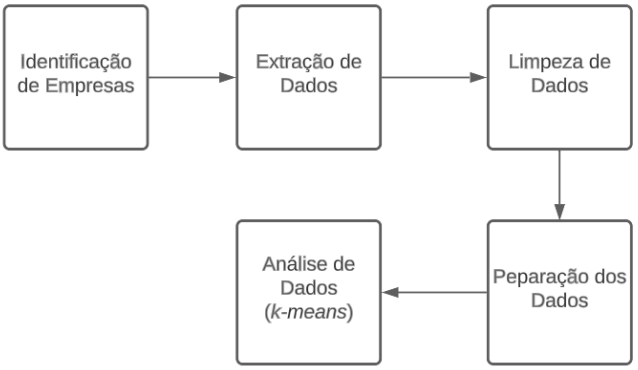
O ambiente de programação escolhido para lidar com a coleta e o tratamento dos dados é o *Python*, devido à sua ampla gama de bibliotecas para manipulação de dados e aprendizado de máquina, como Pandas e *scikit-learn*. Este software permite a automação de

muitos aspectos da coleta de dados, garantindo eficiência e reduzindo a margem de erro humano.

É importante destacar que o processo de coleta de dados também obedece a todas as normas éticas e legais relativas ao uso de informações financeiras. Todos os dados utilizados são de domínio público ou foram adquiridos por meio de licenças que permitem seu uso para fins de pesquisa.

A etapa de coleta de dados é a espinha dorsal deste estudo. Sem um conjunto de dados robusto e bem-curado, as etapas subsequentes de análise seriam comprometidas. É por isso que um investimento significativo de tempo e recursos foi dedicado para garantir que os dados coletados sejam tanto abrangentes quanto precisos. A Figura 6 mostra um fluxograma da coleta de dados.

Figura 6 - Fluxograma do processo da coleta de dados



Fonte: Autoria própria (2023).

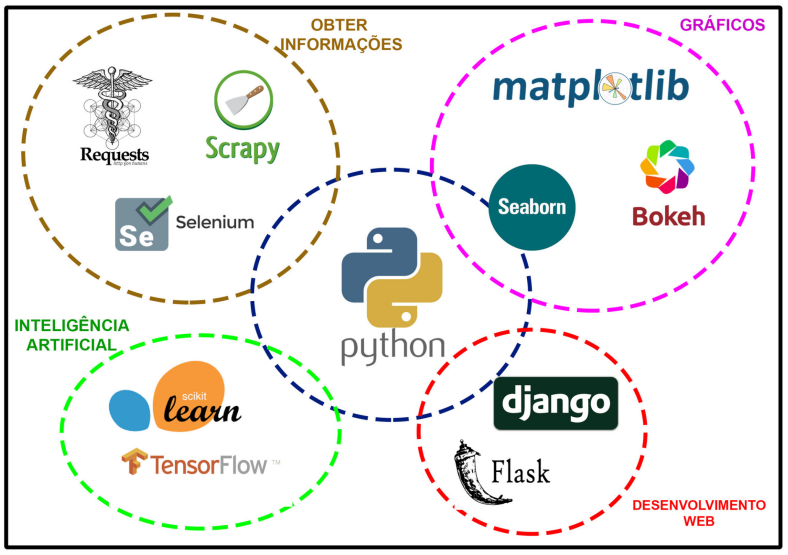
3.3 Implementação e Descrição do Algoritmo de Clusterização

No mundo da programação, existem várias linguagens que facilitam a realização de tarefas específicas. O *Python*, por exemplo, é uma dessas linguagens - amplamente reconhecida por sua simplicidade e eficácia em análise de dados, entre outras funções (PÉREZ et al., 2007). Para começar qualquer análise em *Python*, geralmente importam-se bibliotecas, que são conjuntos de funções e métodos predefinidos. Isso evita que recriemos a roda e nos permite nos concentrar nas questões específicas que estamos tentando resolver.

No código apresentado no Apêndice, as primeiras linhas estão dedicadas à importação de três bibliotecas: "*matplotlib.pyplot*", "*numpy*" e "*sklearn.cluster*". A biblioteca *matplotlib.pyplot* é um padrão na visualização de dados em Python, permitindo a criação de

gráficos de qualidade de publicação (HUNTER, 2007). A *numpy* é uma biblioteca para a linguagem *Python* que suporta *arrays* (incluindo matrizes multidimensionais), além de oferecer uma coleção de funções matemáticas para operar nesses *arrays* (WALT et al., 2011). Finalmente, *sklearn.cluster* faz parte do pacote *scikit-learn*, uma ferramenta simples e eficiente para análise preditiva de dados (PEDREGOSA et al., 2011). A Figura 7 apresenta um diagrama de Venn com as principais bibliotecas utilizadas em *Python*.

Figura 7 - Diagrama representando bibliotecas utilizadas no software Python e suas funcionalidades.



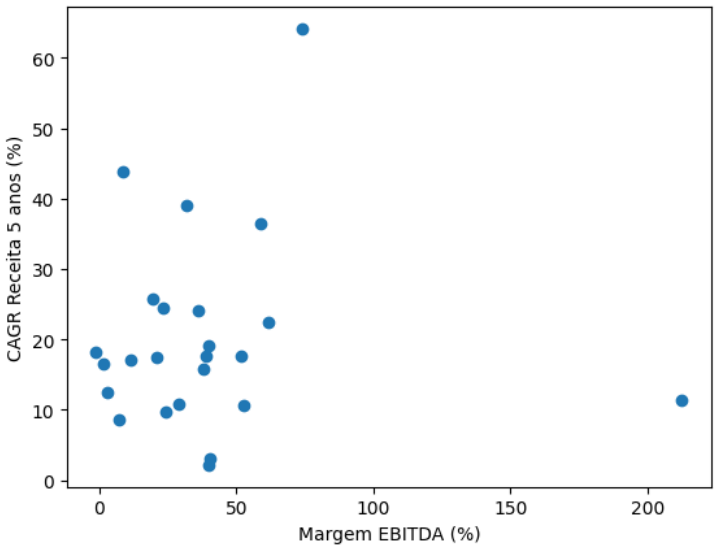
Fonte: LinkedIN (2019).

Posteriormente, o código introduz listas de dados específicas. Estes incluem códigos de ações, margens EBITDA, taxas de crescimento de receita (CAGR) de 5 anos e os outros indicadores financeiros que serão utilizados neste trabalho. Essas listas são, essencialmente, a espinha dorsal da análise subsequente. Ao fornecer essa estrutura de dados, o analista está preparando o terreno para investigar relações, padrões ou anomalias nestes dados.

Uma exibição primária é criada em um gráfico de dispersão ("*scatter plot*") entre os indicadores antes de fazer a clusterização, no caso da Figura 8 o gráfico está mostrando o CAGR Receita (5 anos) no eixo vertical e a Margem EBITDA no horizontal. Gráficos de dispersão são ferramentas valiosas na análise estatística e são frequentemente usados para visualizar a relação entre duas variáveis contínuas. No código, o comando *plt.scatter* é usado para criar este gráfico. Os rótulos dos eixos e o título são definidos para garantir que o leitor

compreenda as dimensões em análise. Finalmente, o método *plt.show*( ) é utilizado para exibir o gráfico para o usuário.

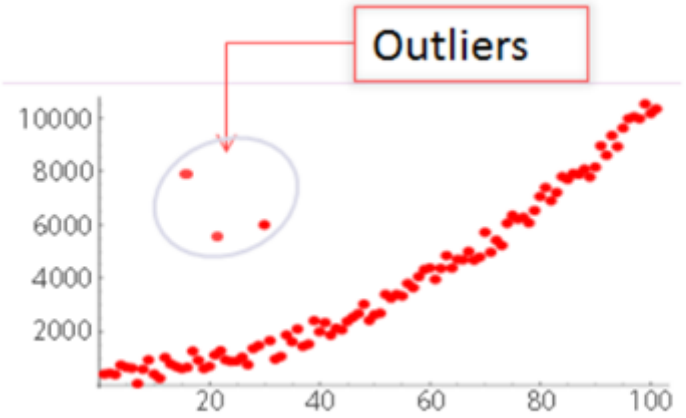
Figura 8 - Gráfico de Dispersão do CAGR Receita 5 anos vs Margem EBITDA



Fonte: Autoria própria e *Python* (2023).

Como pode ser observado na figura acima, existem alguns pontos que possuem valores muito diferentes dos outros, nesse caso é a empresa Itaú Unibanco (ITUB4). Em um estudo detalhado como este, a identificação correta de valores atípicos, frequentemente chamados de *outliers* (Figura 9), é crucial para a integridade da análise. No código fornecido, uma função específica, denominada "*identify_outliers*", foi desenvolvida para abordar essa necessidade. Esta função faz uso de conceitos estatísticos consolidados, particularmente a média e o desvio padrão (HAIR et al., 2010).

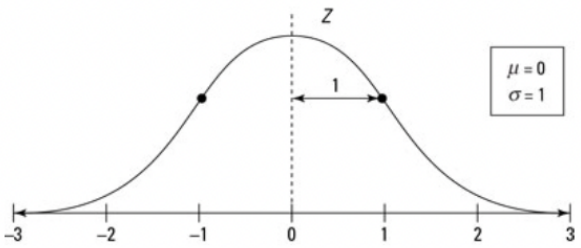
Figura 9 - Exemplo de outliers em um gráfico



Fonte: ElegantJ BI (2013).

O desvio padrão, bem conhecido na literatura estatística, serve como uma medida da variação em um conjunto de números. Em termos mais técnicos, ele quantifica o grau pelo qual cada número em um conjunto varia da média do conjunto. No contexto da análise financeira, o desvio padrão é frequentemente utilizado como uma métrica de volatilidade ou incerteza. A Figura 10 mostra como o desvio padrão (σ) interage com a média (μ) de uma amostra de dados.

Figura 10 - Desvio padrão numa amostra



Fonte: BIOINFO (2016).

No presente trabalho, o critério adotado para identificar um outlier é bastante rigoroso, baseando-se em um múltiplo do desvio padrão. Especificamente, qualquer valor que ultrapasse 1,5 vezes o desvio padrão, seja acima ou abaixo da média do conjunto (Equação 4), é categorizado como um outlier. Esta abordagem não é apenas uma convenção arbitrária, mas é fundamentada em práticas estatísticas bem estabelecidas (IGLEWICZ et al., 1993).

$$\text{Outlier: } se\ x < \mu - 1,5\sigma\ ou\ x > \mu + 1,5\sigma. \qquad \textbf{(4)}$$

A decisão de empregar este critério específico protege a análise de possíveis distorções introduzidas por valores extremos. Além disso, ao utilizar uma métrica objetiva baseada no desvio padrão, assegura-se que a análise é consistente e replicável por outros pesquisadores ou analistas.

Continuando a exploração dos dados e sua análise, um aspecto crucial a ser considerado é a determinação do número adequado de clusters. O método de agrupamento *k-means*, que foi introduzido anteriormente, necessita de um parâmetro para determinar o número de *clusters* a serem formados. Escolher o número correto de *clusters* é fundamental para obter agrupamentos significativos e interpretáveis.

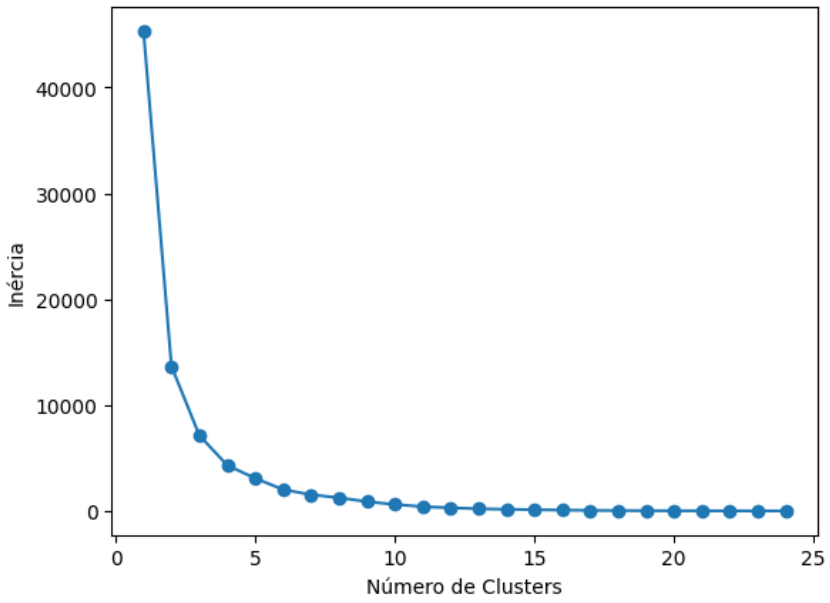
Neste estudo, para determinar o número ótimo de clusters para agrupar as empresas com base em seus indicadores financeiros, empregou-se o "Método do Cotovelo". Esta abordagem tem como princípio identificar um ponto de estabilização na variância explicada à medida que o número de clusters aumenta. Ou seja, é no momento em que a introdução de um cluster adicional não resulta em um aumento significativo da variância explicada. Esse ponto específico é análogo à forma de um cotovelo humano, daí o nome da técnica (KODINARIYA et al., 2013). No código utilizado neste trabalho, este método é implementado da seguinte maneira:

1. **Preparação dos Dados**: Os dados são agrupados em pares usando a função zip. Este passo prepara os dados para a técnica de agrupamento;
2. **Cálculo da Inércia para Diferentes Números de Clusters**: A inércia, também conhecida como soma das distâncias quadradas das amostras para o centro de seu cluster mais próximo, é calculada para diferentes números de clusters. Esta métrica é útil pois fornece uma medida da coesão interna dos clusters. Quanto menor a inércia, mais densos e bem definidos são os clusters;
3. **Visualização do Método do Cotovelo**: A inércia calculada para cada número de clusters é então plotada. O eixo x representa o número de clusters e o eixo y representa a inércia. A visualização auxilia na identificação do "ponto do cotovelo", onde a redução na inércia começa a diminuir, indicando um número adequado de clusters.

Ao aplicar o "Método do Cotovelo", os analistas podem tomar decisões informadas sobre o número de clusters a serem formados, assegurando que os agrupamentos sejam

significativos e otimizados em termos de variância explicada. No caso deste trabalho, um exemplo do método do cotovelo aplicado pode ser visto na Figura 11, novamente com o caso de CAGR Receita 5 anos e Margem EBITDA.

Figura 11 - Método do Cotovelo gerado pelo código



Fonte: Autoria própria e *Python* (2023).

Como pode ser observado na Figura acima, a inércia começa a estabilizar por volta de 5 *clusters*, ou seja, este será o número de clusters utilizados neste trabalho.

Após a decisão sobre o número apropriado de clusters usando o método do cotovelo, o próximo passo lógico é aplicar o algoritmo de agrupamento *k-means* para segmentar os dados em grupos homogêneos. A execução dessa segmentação oferece análises sobre padrões subjacentes nos dados e pode revelar agrupamentos interessantes que seriam difíceis de identificar por meio de uma análise descritiva padrão. No código utilizado, a segmentação é executada e visualizada da seguinte forma:

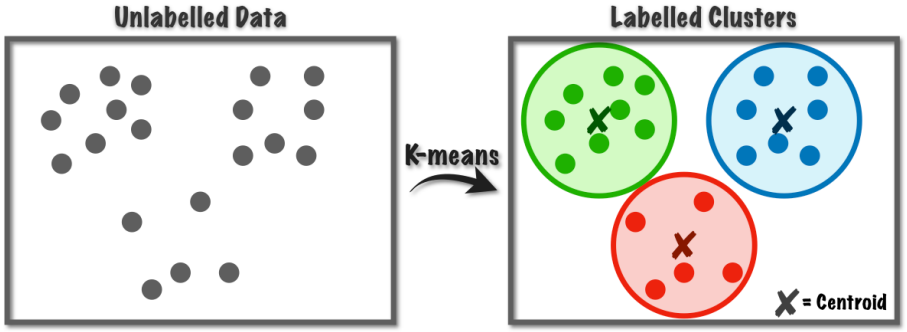
1. **Definindo o Número de *Clusters***: Com base nas informações obtidas do "Método do Cotovelo", um número específico de *clusters* é escolhido. No trabalho, quatro clusters foram selecionados, mas este número pode variar dependendo dos dados e das observações feitas anteriormente;
2. **Implementação do Modelo *k-means***: O algoritmo *k-means* é inicializado com o número escolhido de clusters e é executado com os dados preparados. O

parâmetro "*n_init*" indica o número de execuções com diferentes centroides para escolher a melhor solução em termos de inércia.

3. **Visualização dos *Clusters***: Uma vez que os *clusters* são formados, a visualização é crucial para entender e interpretar os agrupamentos. O código cria um gráfico de dispersão, onde cada ponto representa uma observação e é colorido de acordo com o cluster ao qual pertence.

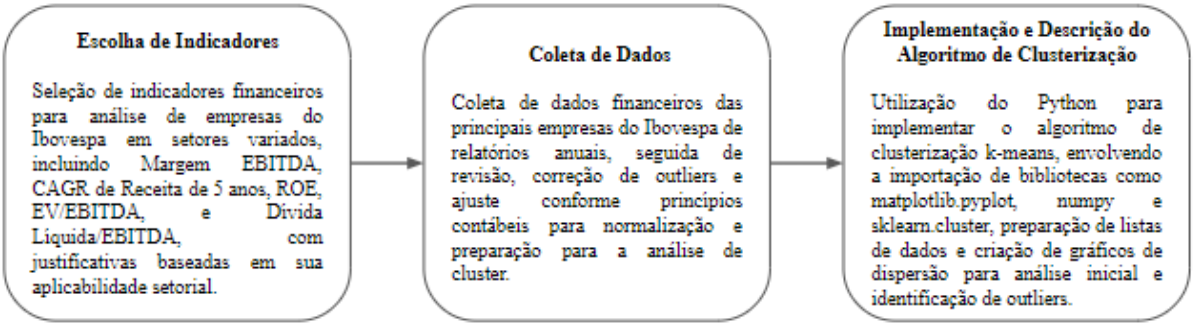
Ao final deste processo, os analistas obtêm uma representação visual clara dos clusters, facilitando a interpretação dos padrões emergentes e proporcionando uma base sólida para análises subsequentes ou tomada de decisão baseada em dados. A Figura 12 mostra um conjunto de dados antes e depois da clusterização pelo método *k-means*. Com isso chegamos no final do código, que pode ser visualizado por completo no Apêndice. Já a Figura 13 resume a metodologia utilizada no trabalho.

Figura 12 - Dados antes e depois da clusterização por *k-means*



Fonte: *Towards Data Science* (2017).

Figura 13 -Representação esquemática do método aplicado na pesquisa



Fonte: Autoria própria (2023).

4 RESULTADOS E DISCUSSÃO

Na sequência do trabalho, esta seção de resultados inicia-se com a apresentação da Tabela 1, que sintetiza os indicadores financeiros fundamentais obtidos em 13 de Setembro de 2023, e que foram extraídos de fontes de domínio público. Estes dados, dispostos de maneira objetiva, proporcionam a base para a análise detalhada que segue, permitindo uma avaliação das variáveis em estudo.

Tabela 1: Tabela com as ações estudadas junto de seus indicadores

Ticker	Ação	Part. (%)	Margem EBITDA (%)	CAGR Receita 5 anos (%)	ROE (%)	EV/EBITDA	Dívida Líquida/EBITDA
VALE3	VALE	12,833	37,97	15,85	30,45	4,2	0,55
PETR4	PETROBRAS	7,243	51,66	17,72	42,19	2,21	0,73
ITUB4	ITAUUNIBANCO	6,631	12,58	13,63	17,65	5,9	0
BBDC4	BRADESCO	3,903	0	8,37	9,78	0	0
B3SA3	B3	3,808	61,93	22,44	20,33	11,65	-0,43
ELET3	ELETROBRAS	3,628	40,21	2,97	1,53	8,84	2,68
BBAS3	BRASIL	3,376	0	9,97	20,01	0	0
ABEV3	AMBEV S/A	3,062	29,22	10,72	16,78	8,87	-0,35
RENT3	LOCALIZA	2,62	36,22	24,03	5,24	10,08	2,91
ITSA4	ITAUSA	2,384	212,51	11,3	17,26	5,86	0,3
WEGE3	WEG	2,252	19,74	25,71	30,74	24,93	-0,36
BPAC11	BTGP BANCO	1,988	21,56	46	18	13,4	0
PRIO3	PETRORIO	1,883	74,08	64,15	34,61	8,44	1,52
SUZB3	SUZA	1,82	58,69	36,49	54,5	4,21	1,9

	NO S.A.						
EQTL3	EQUATORIAL	1,757	23,48	24,5	8,85	8,71	4,19
RADL3	RAIADROGASIL	1,585	11,38	17,08	18,64	13,38	0,56
RDOR3	REDE D OR	1,585	14,34	0	6,02	12,23	0,44
GGBR4	GERDAU	1,374	21,05	17,42	19,52	3,11	0,41
RAIL3	RUMO S.A.	1,361	52,64	10,6	5,11	9,83	2,09
JBSS3	JBS	1,051	-1,07	18,1	10,97	-30,95	-20,39
VBBR3	VIBRA	1,04	1,62	16,5	5,61	11,84	4,02
CSAN3	COSAN	1,038	31,82	38,95	-6,45	5,86	3,34
BBSE3	BBSEGURIDADE	1,021	0	0	80,32	0	0
UGPA3	ULTRAPAR	1,004	2,82	12,42	11,34	7,65	2,23
SBSP3	SABESP	0,983	7,41	8,59	11,18	7,41	2,15
HAPV3	HAPVIDA	0,974	8,62	43,91	-1,59	17,58	2,57
ENEV3	ENEVA	0,956	39,06	17,63	5,06	10,25	4,71
VIVT3	TELEFBRASIL	0,886	39,9	2,14	6,05	4,25	0,7
CMIG4	CEMIG	0,877	24,09	9,68	28,07	4,69	0,94
KLBN11	KLABIN S/A	0,794	39,79	19,06	38,54	5,77	2,49

Fonte: Base de dados B3 e autoria própria (2023).

Após a compilação dos indicadores financeiros na Tabela 1, procedeu-se à sua integração no código Python, detalhado no Apêndice 1. Esta etapa foi essencial para a execução das análises subsequentes, permitindo a manipulação e o tratamento dos dados através do algoritmo *K-means*.

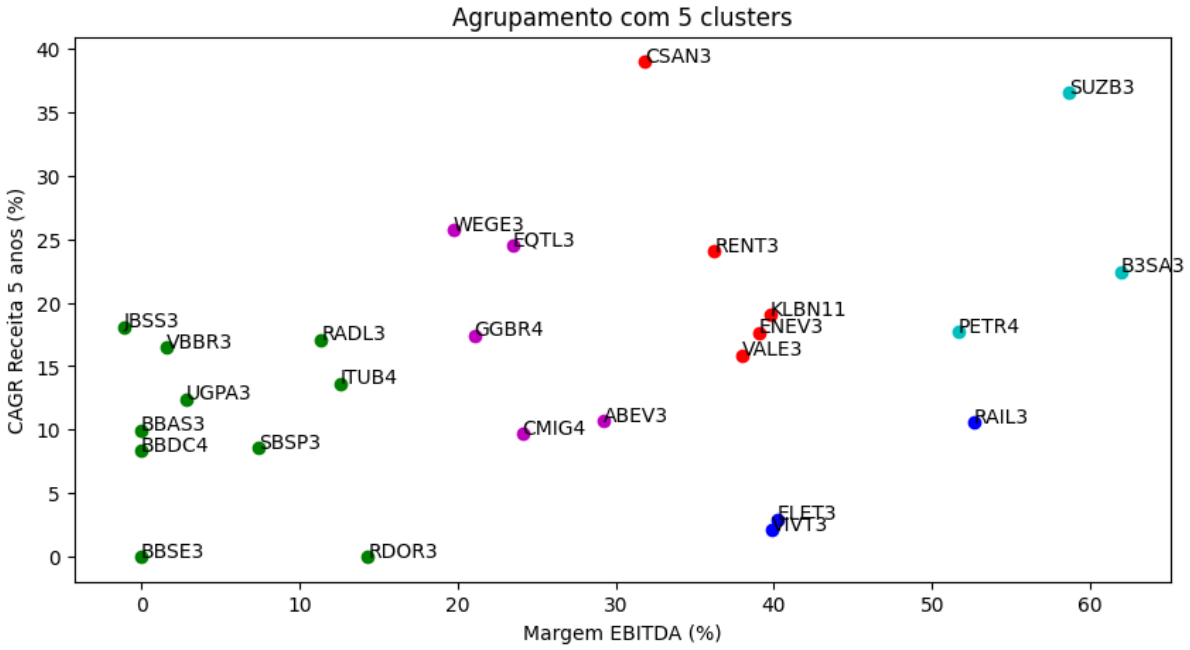
4.1 Eficiência x Crescimento

Nesta seção, avaliamos as empresas listadas com base em dois indicadores chave: margem EBITDA e taxa de crescimento anual composta (CAGR) da receita ao longo de 5 anos. A margem EBITDA foi escolhida para representar a eficiência operacional das empresas, enquanto a CAGR da receita evidencia o crescimento ao longo do período avaliado.

Para garantir a precisão da análise, outliers foram identificados e removidos com base na distância padrão dos dados, eliminando valores que estavam a mais de 1,5 vezes o desvio padrão da média.

Após esta etapa, o algoritmo k-means foi empregado para segmentar as empresas em 5 clusters distintos, permitindo identificar padrões e características comuns entre elas.

Figura 14 - Clusters CAGR Receita 5 anos X Margem EBITDA



Fonte: Autoria própria (2023).

O agrupamento (Verde) inclui RADL3, ITUB4, SBSP3, RDOR3, BBSE3, BBDC4, BBAS3, UGPA3, VBBR3 e JBSS3 representa um grupo diversificado de empresas enfrentando desafios específicos de suas indústrias e econômicos variados. Os bancos brasileiros como ITUB4, BBDC4 e BBAS3, apesar de enfrentarem desafios macroeconômicos e mudanças regulatórias, possuem práticas sólidas de gerenciamento de

risco que sustentam suas operações. O setor de farmácias de varejo, onde a RADL3 opera, tipicamente lida com altos custos operacionais e intensa concorrência, influenciando a rentabilidade. A SBSP3, no setor de saneamento, lida com decisões regulatórias e investimentos em infraestrutura que impactam sua trajetória financeira. A JBSS3 na indústria alimentícia enfrenta condições de mercado e eficiências operacionais que afetam as margens. O desempenho financeiro deste agrupamento reflete uma abordagem de crescimento cautelosa com ênfase em gestão de risco conservadora em meio a um cenário de pressões regulatórias e econômicas.

O segundo *cluster* (roxo), que inclui WEGE3, EQTL3, GGBR4, CMIG4 e ABEV3, é um testemunho de resiliência e crescimento estratégico em setores diversificados. Os ganhos financeiros substanciais da WEGE3 refletem excelência operacional dentro do setor de equipamentos elétricos industriais, indicativo do desempenho robusto do agrupamento. A EQTL3, enraizada no setor de utilidades estáveis, e a CMIG4, do setor de energia regulado, contribuem com confiabilidade e crescimento estável para o perfil do agrupamento. A GGBR4, apesar da volatilidade inerente ao setor siderúrgico, adiciona resiliência através da produção de commodities essenciais. A ABEV3 complementa o agrupamento com a força da marca característica e a durabilidade econômica do setor de bebidas. Coletivamente, essas empresas demonstram um equilíbrio de fluxos de caixa estáveis, demanda consistente e a capacidade de navegar através de desafios específicos do setor, sustentando a saúde geral e o potencial de crescimento do agrupamento.

O terceiro agrupamento (azul), que apresenta ELET3, VIVT3 e RAIL3, representa empresas de setores distintos, mas unificadas por suas robustas margens operacionais e estabilidade de receita. A ELET3 ancora o agrupamento com fortes métricas no setor de eletricidade, ostentando uma margem de EBITDA de 40,21% e um CAGR de receita de 2,97%, sintetizando a estabilidade operacional do agrupamento. A VIVT3, um gigante das telecomunicações, segue este tema com uma margem de EBITDA de 39,9%, indicativa do potencial inerente do setor para rentabilidade sustentada através de investimentos estratégicos. A RAIL3, um ponto crucial logístico no transporte agrícola, completa o agrupamento com uma notável margem de EBITDA de 52,64% e um CAGR de receita de 10,6%, destacando operações eficientes cruciais em indústrias intensivas em ativos. Coletivamente, estas firmas exemplificam uma combinação de saúde financeira duradoura e posicionamento estratégico no mercado, definindo a narrativa compartilhada do agrupamento de resiliência e excelência operacional consistente.

O próximo cluster (vermelho) é composto por VALE3, ENEV3, KLBN11 e CSAN3 e representa uma seleção de empresas com desempenho operacional robusto, marcado por sólidas margens de EBITDA e significativo crescimento de receita. A VALE3 tem se concentrado em prioridades estratégicas em suas soluções de ferro e materiais para transição energética, com um EBITDA ajustado proforma de 2022 de $20,9 bilhões, apesar de uma diminuição em relação ao ano anterior devido à queda dos preços do minério de ferro e ao contínuo investimento em projetos de crescimento. A ENEV3 reportou um recorde de EBITDA, enfatizando a diversificação de receita e a exportação de energia, o que levou a um aumento notável de 224% na receita ano a ano para o primeiro trimestre.

A KLBN11 mostrou um forte desempenho financeiro com um aumento de 44% no lucro líquido e um aumento de 13% no EBITDA ajustado no primeiro trimestre de 2023, alcançando uma margem de EBITDA de 40% através de uma gestão de custos eficaz e ajustes de preço apesar de uma ligeira diminuição no volume de vendas. A CSAN3, embora enfrentando desafios de lucro com uma diminuição de 69,1% no lucro líquido ajustado, gerenciou um aumento de 4,8% no EBITDA e um substancial aumento de 54,2% na receita do primeiro trimestre de 2021, indicando resiliência e potencial de crescimento.

O quinto agrupamento (ciano) inclui PETR4 (Petrobras), B3SA3 (B3 - Brasil Bolsa Balcão) e SUZB3 (Suzano), apresenta uma combinação intrigante de indústrias de energia, serviços financeiros e papel e celulose. Este agrupamento reflete empresas que têm influência significativa dentro de seus respectivos setores e são fundamentais para a economia brasileira.

A Petrobras, um grande *player* no mercado global de energia, historicamente mostrou substanciais fluxos de receita e robustas margens de EBITDA, impulsionadas por suas operações abrangentes de petróleo e gás. A B3 se posiciona como o coração da infraestrutura do mercado financeiro brasileiro, fornecendo serviços de negociação em uma ampla gama de classes de ativos; é uma empresa espinha dorsal que se beneficia da escala das transações financeiras no país. A Suzano, um gigante na indústria de papel e celulose, combina eficiência operacional com práticas sustentáveis para atender à demanda global por papel e bioprodutos.

A análise dos cinco clusters destaca as distintas dinâmicas de eficiência e crescimento no mercado brasileiro. Os bancos como ITUB4 e companhias de setores variados, incluindo RADL3 e SBSP3, formam um cluster caracterizado por uma gestão conservadora frente a desafios macroeconômicos e regulatórios, priorizando a eficiência operacional. Por outro lado, empresas como WEGE3 e ABEV3 compõem um segundo cluster que exibe uma

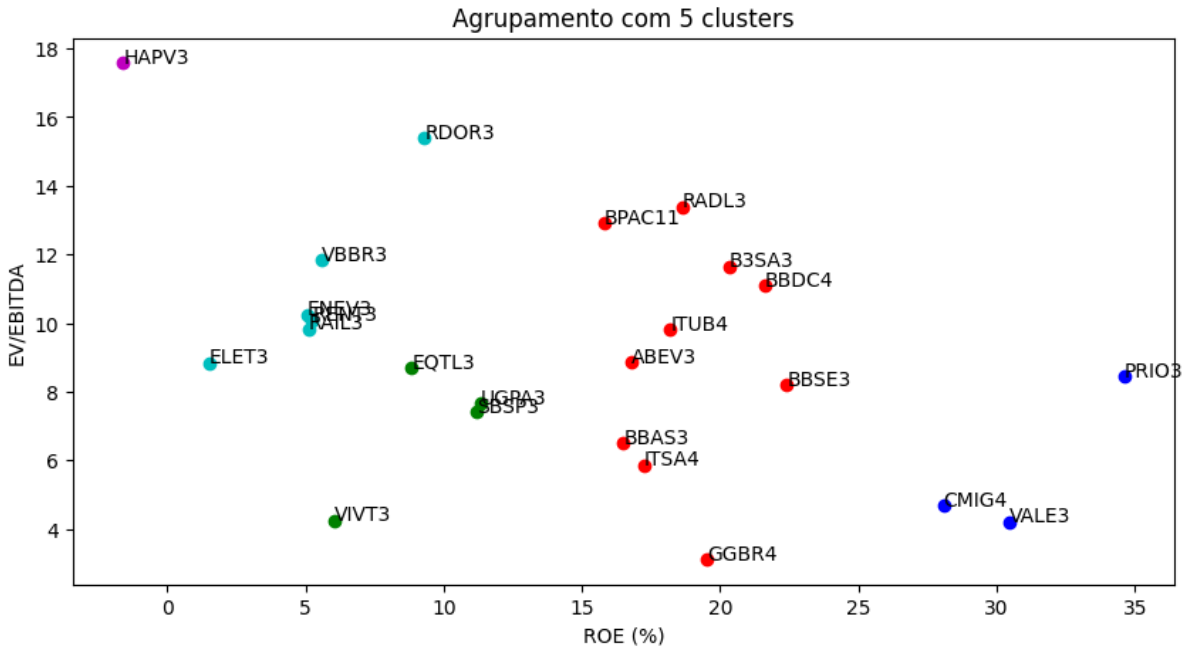
resiliência notável, equilibrando fluxos de caixa estáveis e demanda consistente, apesar das volatilidades inerentes aos seus respectivos setores.

O terceiro grupo, com empresas como ELET3 e VIVT3, destaca-se por margens operacionais robustas e estabilidade de receita, refletindo uma eficiência operacional em setores estratégicos. Já o quarto cluster, que inclui VALE3 e ENEV3, evidencia uma performance operacional forte, com margens de EBITDA sólidas e crescimento de receita significativo. Finalmente, o quinto grupo reúne gigantes como PETR4 e SUZB3, que são fundamentais para a economia brasileira, mostrando a importância estratégica de suas operações e a influência que exercem em tendências econômicas mais amplas e preços globais de commodities.

4.2 Valorização x Rentabilidade

A combinação dos indicadores ROE (Retorno sobre o Patrimônio Líquido) e EV/EBITDA oferece uma visão abrangente da saúde financeira e da valoração de uma empresa. O ROE indica a capacidade de uma empresa de gerar lucro a partir de seu próprio capital, enquanto o EV/EBITDA dá uma ideia da avaliação de uma empresa em relação à sua capacidade de geração de caixa (DAMODARAN, 2012).

Figura 15 - Clusters EV/EBITDA e ROE (%)



Fonte: Autoria própria (2023).

No primeiro agrupamento (verde), que inclui VIVT3, EQTL3, SBSP3 e UGPA3, observamos empresas dos setores de telecomunicações, serviços públicos e distribuição. Essas companhias apresentam índices moderados de Retorno sobre o Patrimônio Líquido (ROE) e de Valor da Empresa sobre EBITDA (EV/EBITDA). A Telefônica Brasil, especificamente, com um ROE de 6,05% e um EV/EBITDA de 4,25, sinaliza uma geração eficiente de lucro em relação ao patrimônio dos acionistas e sugere uma possível subavaliação pelo mercado. Tais características podem refletir uma estabilidade setorial com fluxos de caixa previsíveis e menor susceptibilidade a mudanças rápidas.

As empresas EQTL3 e SBSP3, dos setores de utilidades e saneamento, respectivamente, têm ROEs de 8,85% e 11,18%, e razões EV/EBITDA de 8,71 e 7,41. Esses números indicam uma gestão de capital eficiente e uma sólida posição em seus setores, considerando a natureza regulada de suas receitas e o potencial para crescimento contínuo. A UGPA3, atuante no setor de distribuição, segue uma tendência semelhante com um ROE de 11,34% e um EV/EBITDA de 7,65, apontando para uma capacidade respeitável de converter patrimônio em lucro, alinhada à avaliação de mercado típica do setor.

Este grupo sugere que as empresas são percebidas pelo mercado como estáveis e adequadamente avaliadas. Seus ROEs positivos demonstram a habilidade de reinvestir lucros de forma lucrativa, enquanto os índices EV/EBITDA indicam que não estão superavaliadas nem negligenciadas, tornando-as potencialmente atraentes para investidores que buscam equilíbrio entre valor e estabilidade em seus investimentos.

O segundo *cluster* (ciano) compreende ELET3, RAIL3, VBBR3 e RDOR3, cobrindo os setores de energia elétrica, logística, bancário e saúde, com perfis financeiros variados. A Eletrobras possui um ROE relativamente baixo de 1,53% e um EV/EBITDA de 8,84, o que reflete o impacto de desafios operacionais e regulatórios sobre a rentabilidade, mas uma avaliação de mercado que pode antecipar crescimento ou reestruturação futuros. RAIL3 apresenta um retorno sobre o patrimônio justo de 5,11% e um EV/EBITDA de 9,83, sinalizando a natureza de longo prazo e intensiva em capital dos investimentos em infraestrutura ferroviária.

VBBR3, do setor bancário, exibe um ROE modesto de 5,61% e um EV/EBITDA de 11,84, sugere um crescimento eficiente e conservador no ambiente econômico atual, e uma avaliação de mercado disposta a pagar um prêmio pela estabilidade financeira. Já RDOR3, do setor de saúde, mostra um ROE de 6,02% e um EV/EBITDA de 12,23, representa uma capacidade sólida, embora não excepcional, de geração de lucro a partir do patrimônio, com

uma avaliação de mercado que considera as perspectivas de crescimento da empresa e a demanda estável por serviços de saúde.

Este grupo diversificado indica um espectro de potenciais de crescimento e fatores de estabilidade, onde a saúde financeira de cada empresa e sua avaliação de mercado refletem as características de sua indústria e a natureza essencial dos serviços prestados. Os índices EV/EBITDA moderadamente altos sugerem que o mercado valoriza essas empresas por seus futuros ganhos e fluxos de caixa, tornando o grupo potencialmente atrativo para investidores que buscam variedade e investimentos em indústrias fundamentais.

O terceiro agrupamento (roxo) é particularmente interessante, pois conta com apenas uma empresa: HAPV3. Com um ROE negativo de -2,10% e um EV/EBITDA alto de 21,01, destaca-se significativamente dos outros grupos. Isso pode indicar desafios recentes, uma reestruturação ou a realização de investimentos estratégicos que o mercado espera que tragam retornos lucrativos no futuro. O EV/EBITDA elevado sugere uma expectativa de crescimento e eficiência operacional significativos, apesar dos desafios atuais.

No quarto *cluster* (vermelho), encontra-se ITUB4 (Itaú Unibanco), ABEV3 e GGBR4, que representam os setores financeiro, bebidas e siderurgia. ITUB4 mostra um ROE alto de 18,77% e um EV/EBITDA de 7,38, indicando uma habilidade excepcional de gerar lucro a partir do patrimônio e uma avaliação de mercado consistente com o potencial da instituição financeira. A Ambev, sob o ticker ABEV3, exibe um ROE sólido de 17,28% e um EV/EBITDA mais baixo de 9,84, que reflete sua posição de mercado consolidada e demanda estável por bebidas.

GGBR4, do setor de siderurgia, tem um ROE forte de 14,58% e um EV/EBITDA de 2,64, que sinaliza uma subavaliação pelo mercado ou uma expectativa conservadora de crescimento futuro. Este *cluster* de empresas têm finanças sólidas e uma presença marcante no mercado. A variação nos índices EV/EBITDA reflete perspectivas diferentes de crescimento e eficiência operacional dentro dos respectivos setores, apresentando-se como uma opção para investidores que buscam companhias com histórico comprovado de rentabilidade e com diferentes avaliações de mercado.

Por fim, o quinto *cluster* (azul) inclui CMIG4 (Cemig), VALE3 e PRIO3, que são empresas dos setores de energia, mineração e petróleo e gás, respectivamente. Essas empresas apresentam ROEs elevados: CMIG4 com 17,14%, VALE3 com 23,59% e PRIO3 com 27,70%. Os índices de EV/EBITDA, no entanto, variam significativamente, com CMIG4 em 2,79, VALE3 em 3,47 e PRIO3 em 1,65. Isso mostra que, embora todas as três empresas tenham uma forte capacidade de gerar lucro, o mercado atribui diferentes valores a elas.

CMIG4 e VALE3 parecem particularmente subvalorizadas, dada a combinação de altos ROEs e baixos múltiplos de EV/EBITDA.

Este grupo representa companhias que não apenas têm demonstrado uma habilidade excepcional de gerar lucro, mas também operam em setores essenciais com demanda constante. Eles podem ser atrativos para investidores que buscam empresas robustas, com potencial de subavaliação e capacidade de manter uma posição dominante em suas indústrias. As variações nos índices EV/EBITDA também sugerem oportunidades de investimento que poderiam ser reconsideradas pelo mercado, oferecendo potencial de valorização.

No contexto do mercado financeiro brasileiro, foram identificados cinco clusters representativos de diferentes perfis de investimento com base nos índices ROE e EV/EBITDA. O primeiro *cluster* inclui empresas de telecomunicações, serviços públicos e distribuição (VIVT3, EQTL3, SBSP3 e UGPA3), as quais apresentam um perfil de estabilidade e eficiência com indicadores que sugerem uma adequada avaliação de mercado, revelando-se como opções atraentes para investidores focados em um equilíbrio entre valor e segurança. O segundo *cluster*, formado por companhias dos setores de energia elétrica, logística, bancário e saúde (ELET3, RAIL3, VBBR3 e RDOR3), exibe uma diversidade de potenciais de crescimento e estabilidade, sugerindo uma valorização do mercado orientada para a previsibilidade dos fluxos de caixa e o essencial dos serviços prestados.

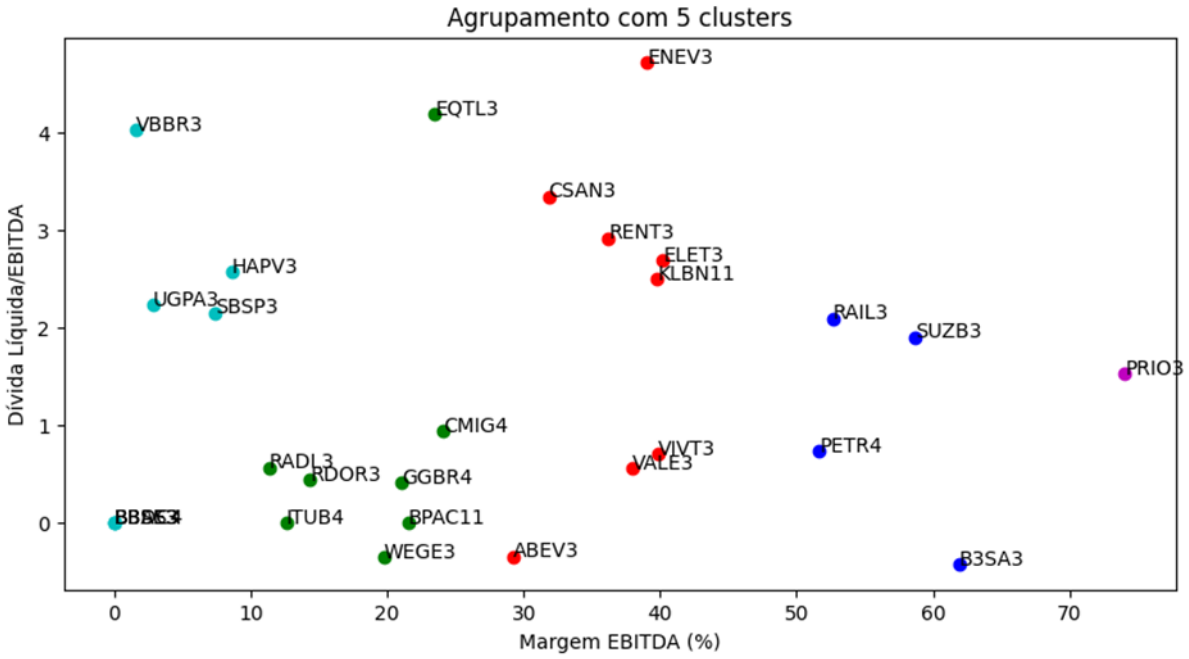
O terceiro agrupamento é apenas composto pela Hapvida (HAPV3), com indicadores financeiros que sinalizam desafios recentes e expectativas de crescimento futuro, destacando-se dos demais grupos pelo seu alto EV/EBITDA em contraste com um ROE negativo. Os clusters quatro e cinco, constituídos respectivamente por Itaú Unibanco, Ambev e Gerdau (ITUB4, ABEV3 e GGBR4), e Cemig, Vale e PetroRio (CMIG4, VALE3 e PRIO3), são caracterizados por empresas com sólida performance financeira e posição dominante em setores fundamentais como o financeiro, bebidas, siderurgia, energia e mineração. Esses clusters sugerem uma subavaliação de mercado, oferecendo um atraente potencial de valorização para investidores dispostos a apostar em companhias com histórico de robustez e eficiência operacional.

4.3 Eficiência x Endividamento

Ao examinar os clusters baseados na Margem EBITDA e na relação Dívida Líquida/EBITDA, pode-se antecipar a identificação de grupos distintos que representam diferentes eficiências operacionais e alavancagens financeiras entre as empresas. Podem

surgir clusters que destacam negócios com lucros operacionais robustos e níveis de dívida conservadores, indicando forte geração de fluxo de caixa e saúde financeira. Alternativamente, *clusters* podem revelar empresas com margens operacionais mais baixas e altas taxas de endividamento, possivelmente refletindo estratégias de crescimento agressivas ou indústrias com maiores despesas de capital. Esse agrupamento oferece percepções de como as empresas equilibram a rentabilidade com o uso de dívidas, revelando diferentes estratégias financeiras e perfis de risco dentro do mercado.

Figura 16 - Clusters Dívida Líquida/EBITDA e Margem EBITDA



Fonte: Autoria própria (2023).

O *Cluster* 1 (ciano) é composto por VBBR3, HAPV3, UGPA3 e SBSP3, um conjunto de companhias de setores distintos, mas conectadas pelos seus indicadores financeiros de Margem EBITDA e razão Dívida Líquida/EBITDA. VBBR3 exibe uma Margem EBITDA menor, de 1,62%, com uma razão Dívida Líquida/EBITDA comparativamente alta de 4,02, o que reflete uma entidade financeira onde a alavancagem é uma ferramenta estratégica e não um fardo. HAPV3 apresenta uma Margem EBITDA de 8,62% com uma Dívida Líquida/EBITDA de 2,57, que indica um prestador de serviços de saúde com sólido desempenho operacional e um nível moderado de endividamento. UGPA3 mostra uma margem menor de 2,82% e uma Dívida Líquida/EBITDA de 2,23, o que sugere uma utilidade ou conglomerado com eficiência operacional estável, porém limitada. SBSP3, com uma

Margem EBITDA mais saudável de 7,41% e uma Dívida Líquida/EBITDA de 2,15, representa uma empresa de saneamento com operações eficientes e uma abordagem equilibrada para alavancagem.

Este *cluster* sugere um padrão de empresas que podem estar utilizando dívidas para investimentos estratégicos ou operações, estabelecidas contra um pano de fundo de rentabilidade operacional variada. Os dados financeiros indicam indústrias onde os investimentos de capital são significativos, com a expectativa de que receitas estáveis e previsíveis servirão para atender ao endividamento associado. Revela-se uma paisagem onde a eficiência operacional e a gestão da dívida são alavancas críticas nas estratégias financeiras das empresas.

O segundo *cluster* (verde) é composto por RADL3, RDOR3, ITUB4, WEGE3, BPAC11, GGBR4, CMIG4 e EQTL3, uma variedade de firmas com Margens EBITDA geralmente robustas e razões Dívida Líquida/EBITDA relativamente baixas, indicativo de operações eficientes e alavancagem financeira prudente. RADL3 e RDOR3, nos setores de varejo e saúde respectivamente, demonstram sólida rentabilidade operacional com margens EBITDA acima de 11% e baixa alavancagem, significando uma estratégia financeira prudente. ITUB4, uma entidade bancária, mostra margens de lucro fortes sem o peso da dívida líquida, um testemunho de sua força financeira e gestão de capital.

A WEGE3, uma empresa de manufatura, destaca-se com uma impressionante margem EBITDA e um indicador Dívida Líquida/EBITDA negativo, sugerindo uma posição financeira robusta e excelência operacional. Paralelamente, BPAC11, atuante no setor de serviços financeiros, apresenta uma rentabilidade substancial e ausência de dívida líquida. O histórico industrial da GGBR4 apoia uma boa margem de lucro e níveis conservadores de dívida, enfatizando a gestão de custos e a cautela financeira. As empresas de utilidades públicas CMIG4 e EQTL3 exibem uma rentabilidade significativa com suas Margens EBITDA se aproximando ou excedendo 23%, embora difiram em alavancagem – EQTL3 apresenta uma razão de dívida mais elevada, o que reflete suas estratégias de investimento ou necessidades de despesas de capital.

De modo geral, o *cluster* 2 é caracterizado por empresas que equilibram operações rentáveis com uma abordagem prudente em relação à dívida, sugerindo um ênfase coletiva na manutenção da estabilidade financeira ao mesmo tempo em que asseguram excelência operacional. Essas firmas, apesar de suas diferenças setoriais, compartilham um ethos financeiro comum que prioriza o crescimento sustentável e o uso eficiente do capital.

O terceiro agrupamento (vermelho) inclui um leque diversificado de empresas, tais como ABEV3, VALE3, VIVT3, KLBN11, ELET3, RENT3, CSAN3 e ENEV3, cada uma demonstrando Margens EBITDA robustas, indicativas de forte eficiência operacional em setores como bebidas, mineração, telecomunicações, papel e celulose, eletricidade e energia. Notavelmente, ABEV3 mostra uma forte rentabilidade com significativas reservas de caixa em contraste com a dívida, enquanto VALE3 combina sua alta Margem EBITDA com baixos níveis de dívida, ilustrando uma combinação potente de geração de caixa e alavancagem financeira conservadora. VIVT3 e KLBN11, ambos mantêm altas margens operacionais e razões de dívida moderadas, sinalizando um controle de custos eficiente e gestão estratégica de dívida.

As empresas ELET3 e RENT3 apresentam altas margens, mas possuem razões de dívida um pouco mais elevadas, o que reflete investimentos substanciais em ativos e infraestrutura. CSAN3 e ENEV3, ambas do setor de energia, exibem margens de lucro saudáveis; contudo, ENEV3 destaca-se com uma razão de dívida superior, na qual sugere uma abordagem de investimento ou crescimento mais agressiva. Em conjunto, o agrupamento retrata empresas operacionalmente fortes com filosofias variadas sobre alavancagem de dívida, de estratégias conservadoras a mais agressivas, sublinhando distintas saúdes financeiras e escolhas estratégicas dentro de suas respectivas indústrias.

O *cluster* 4 (azul) é composto por RAIL3, SUZB3, PETR4 e B3SA3, um conglomerado de empresas atuando nos setores de logística, papel e celulose, petróleo e gás e infraestrutura do mercado financeiro, respectivamente. Este agrupamento é caracterizado por companhias com altas margens EBITDA e uma gama de razões Dívida Líquida/EBITDA, refletindo uma forte capacidade de geração de fluxo de caixa operacional aliada a diversos graus de alavancagem financeira.

A RAIL3, uma empresa de logística e ferroviária, ostenta uma alta Margem EBITDA de 52,64%, indicativo de eficiência operacional substancial, mas com uma razão Dívida Líquida/EBITDA de 2,09, sugerindo um nível moderado de endividamento em relação aos ganhos. A SUZB3, da indústria de papel e celulose, possui uma impressionante Margem EBITDA de 58,69% e uma razão Dívida Líquida/EBITDA de 1,9, sinalizando um desempenho operacional robusto e uma abordagem equilibrada de gestão de dívida.

A PETR4, grande ator no setor de petróleo, demonstra uma significativa Margem EBITDA de 51,66% com uma razão Dívida Líquida/EBITDA relativamente baixa de 0,73, refletindo operações eficientes e um perfil de dívida conservador para uma indústria de capital intensivo. A B3SA3, associada a serviços financeiros ou operações de bolsa de

valores, tem a maior Margem EBITDA deste *cluster*, com 61,93%, e uma razão Dívida Líquida/EBITDA negativa de -0,43, o que implica que a empresa possui mais caixa do que dívida, uma posição vantajosa para qualquer firma.

Em essência, o *cluster* 4 destaca as empresas com rentabilidade operacional superior, evidenciando a capacidade delas de gerar caixa efetivamente a partir de suas atividades comerciais principais. Os diversos graus de níveis de dívida também revelam diferentes estratégias financeiras e apetites por risco, variando de conservadoras a moderadamente agressivas, mas todas no contexto de margens de lucro fortes. Este agrupamento indica uma combinação de excelência operacional e perspicácia financeira, com cada empresa navegando em sua estrutura de capital para otimizar o crescimento e o valor ao acionista dentro da dinâmica de sua indústria.

O *cluster* 5 (roxo) é único por sua singularidade, consistindo apenas da empresa PRIO3, uma companhia do setor de óleo e gás. Esta empresa se destaca sozinha em seu cluster devido a seus indicadores financeiros distintos: uma Margem EBITDA excepcionalmente alta de 74,08% combinada com uma razão moderada de Dívida Líquida/EBITDA de 1,52. A alta Margem EBITDA indica que a PRIO3 é extremamente eficiente na conversão de receita em lucro operacional, o que é particularmente notável em uma indústria caracterizada pela sua natureza capital-intensiva e volatilidade dos preços das commodities.

O fato de a PRIO3 ser a única empresa neste cluster sugere que possui um perfil financeiro significativamente diferente dos seus pares no conjunto de dados. Sua inclusão indica que não foi considerada um ponto fora da curva no contexto da análise de clusterização, passando pelo teste de outliers que envolve um limiar de 1,5 vezes o desvio padrão. Este teste garante que os pontos de dados da empresa não sejam tão extremos a ponto de não representarem qualquer padrão ou relação dentro do conjunto de dados.

A posição da PRIO3 neste cluster é devido a uma combinação única de eficiência operacional, condições de mercado e gestão financeira estratégica que a diferencia das outras empresas. Isso indica que a PRIO3 encontrou um nicho ou possui uma vantagem competitiva que permite margens de lucro excepcionais enquanto mantém um nível de dívida gerenciável, um equilíbrio que outras empresas no mercado mais amplo podem não demonstrar. Essa distinção poderia despertar o interesse de investidores ou analistas à procura de empresas com fundamentos sólidos e capacidade de gerar fluxo de caixa enquanto mantêm a dívida sob controle.

Os *clusters* 1 e 2 apresentam empresas com diferentes estratégias financeiras e operacionais baseadas em suas margens EBITDA e razões Dívida Líquida/EBITDA. O primeiro agrupamento é composto por empresas de setores diversos, que apesar das diferenças, têm em comum o uso de dívidas para fins estratégicos contra um fundo de rentabilidade operacional variável. Por outro lado, o segundo *cluster* inclui empresas com operações eficientes e uma gestão de dívida prudente, o que sugere um enfoque em estabilidade financeira e excelência operacional.

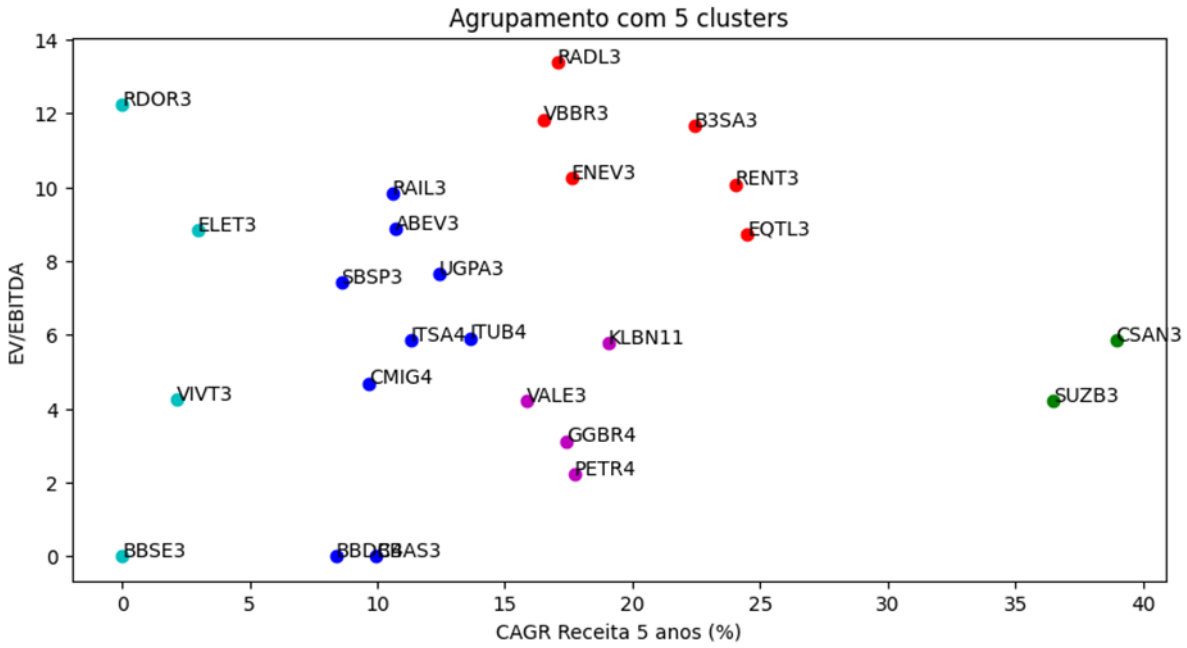
O *cluster* 3 destaca empresas de vários setores com margens EBITDA fortes, ilustrando eficiência operacional e variadas abordagens de alavancagem financeira, desde conservadoras a mais agressivas. Enquanto isso, o quarto agrupamento caracteriza-se por empresas com elevadas margens EBITDA e diferentes graus de dívida, refletindo uma combinação de excelência operacional e astúcia financeira. O último *cluster* é único, destacando a PRIO3, que tem uma margem EBITDA extremamente alta e uma gestão de dívida equilibrada, sugerindo uma posição operacional e estratégica distinta das demais empresas analisadas.

4.4 Crescimento x Valorização

Na análise de *clusters* entre o EV/EBITDA e o CAGR de receitas de 5 anos, objetiva-se desvendar padrões que possam indicar diferentes perfis de investimento no mercado. Companhias que apresentam uma relação EV/EBITDA baixa em conjunto com um CAGR alto apontam uma oportunidades de crescimento que ainda não foram totalmente reconhecidas pelo mercado. Inversamente, aquelas com uma relação EV/EBITDA elevada e um CAGR baixo poderiam sugerir que as expectativas futuras do mercado podem estar excessivamente otimistas, não sendo sustentadas pelo crescimento histórico de receitas.

Adicionalmente, clusters que apresentam tanto EV/EBITDA quanto CAGR elevados podem destacar empresas que já são recompensadas pelo mercado devido à sua comprovada trajetória de crescimento. Esse contraste entre os clusters pode auxiliar na identificação de empresas que, apesar de demonstrarem um sólido crescimento de receita, podem ainda não estar sendo avaliadas de maneira condizente pelo mercado, possivelmente devido a riscos percebidos ou incertezas.

Figura 17 - Clusters EV/EBITDA e CAGR Receita 5 Anos



Fonte: Autoria própria (2023).

O *cluster* 1 (ciano) apresenta BBSE3, VIVT3, ELET3 e RDOR3, um conjunto diversificado de empresas ao se considerar suas relações EV/EBITDA e o CAGR de receitas de 5 anos. O EV/EBITDA de 0 para BBSE3 indica uma avaliação de mercado inusual que pode ser devida a circunstâncias financeiras atípicas ou dados fora do padrão, que também pode ser atribuída por ser uma empresa seguradora. A relação de 4,25 de VIVT3, em conjunto com um modesto crescimento de receitas (CAGR) de 2,14%, sugere uma empresa que é potencialmente avaliada de maneira razoável pelo mercado por seu crescimento estável. ELET3 mostra uma relação EV/EBITDA ligeiramente superior, de 8,84, e uma taxa de crescimento similar à de VIVT3, indicando uma avaliação de mercado que leva em conta expectativas de crescimento modestas. RDOR3, com a relação EV/EBITDA mais alta deste cluster, de 12,23, e um crescimento de receita de 0%, apresenta um caso contrastante, onde o mercado pode estar precificando perspectivas de crescimento futuro ou melhorias de rentabilidade que não estão evidentes nos dados históricos de receitas.

No geral, este *cluster* apresenta um espectro de empresas com crescimento histórico de receitas de baixo a moderado e avaliações de mercado variadas. A gama sugere que essas empresas oferecerem estabilidade com o potencial para crescimento futuro que pode estar atualmente subvalorizado pelo mercado, tornando-as possíveis candidatas para investidores que buscam tanto estabilidade quanto oportunidades de crescimento.

O segundo agrupamento (azul) é composto por BBAS3, CMIG4, ITSA4, ITUB4, UGPA3, SBSP3, ABEV3 e RAIL3, cada um exibindo uma gama de relações EV/EBITDA e índices CAGR de receita de 5 anos. ITSA4 e ITUB4, ambas instituições financeiras, apresentam índices EV/EBITDA semelhantes de aproximadamente 5,9, juntamente com crescimento respeitável de receitas, indicando uma visão de mercado equilibrada de seu crescimento estável de lucros. UGPA3 e SBSP3 compartilham relações EV/EBITDA próximas a 7, com sólidos CAGRs de receitas, apresentando-se como crescimentos estáveis. ABEV3 está avaliada de forma um pouco mais rica, com um EV/EBITDA mais alto de 8,87, correspondendo ao seu maior CAGR de receita de 10,72%. CMIG4 oferece um quadro contrastante, com uma avaliação mais baixa em relação aos seus lucros, apesar de uma robusta taxa de crescimento, enquanto o maior índice de RAIL3 pode refletir um sentimento de mercado positivo em relação ao seu desempenho futuro. No geral, este *cluster* poderia atrair investidores em busca de empresas com desempenho consistente e avaliações justificadas contra seus perfis de crescimento.

O terceiro agrupamento (roxo) é composto por VALE3, GGBR4, PETR4 e KLBN11, empresas enraizadas nos setores de commodities e recursos naturais. Este agrupamento apresenta dinâmicas de avaliação intrigantes, com VALE3 e PETR4 exibindo robustos CAGRs de receita de 5 anos de 15,85% e 17,72%, respectivamente, acompanhados de baixas relações EV/EBITDA de 4,2 e 2,21, sugerindo que o mercado pode estar subestimando o potencial de crescimento de seus lucros. O EV/EBITDA um pouco mais alto de GGBR4 de 3,11 contra uma taxa de crescimento de receita comparável indica um mercado cauteloso com os riscos cíclicos da indústria. KLBN11, com um EV/EBITDA mais alto de 5,77 e o maior crescimento de receita em 19,06%, reflete uma postura cautelosa dos investidores em relação à sustentabilidade do crescimento na indústria de papel e celulose. No geral, este *cluster* representa oportunidades subvalorizadas para investidores dispostos a apostar na demanda contínua por commodities, apostando que o mercado ainda não reconheceu plenamente suas perspectivas de crescimento.

O quarto *cluster* (roxo) é uma variedade de empresas de diversos setores, incluindo RADL3 da saúde, VBBR3 do setor bancário, ENEV3 da energia, B3SA3 que opera uma bolsa de valores, RENT3 de serviços de aluguel de carros e EQTL3 também do setor de energia. RADL3, com um CAGR de receita de 5 anos de 17,08% e um EV/EBITDA de 13,38, e VBBR3, com um CAGR de 16,5% e um EV/EBITDA de 11,84, indicam uma forte avaliação de mercado em relação ao seu potencial de lucro e crescimento. ENEV3 apresenta

uma sólida taxa de crescimento de 17,63% com um EV/EBITDA correspondente de 10,25, apresentando uma visão equilibrada de seu crescimento e valorização.

B3SA3 se destaca com um alto EV/EBITDA de 11,65, justificado por seu impressionante crescimento de receita de 22,44%, sugerindo uma alta expectativa de sua posição estratégica nos mercados financeiros. RENT3, embora tenha o maior CAGR de receita de 24,03%, possui um EV/EBITDA relativamente alto de 10,08, o que reflete suas estratégias de crescimento agressivas e a confiança do mercado em seu modelo de negócio. EQTL3, com um CAGR de receita de 24,5% e um EV/EBITDA de 8,71, completa o *cluster* com um perfil de crescimento forte e uma avaliação moderadamente alta, sugestivo do otimismo dos investidores sobre suas perspectivas futuras. Portanto, este cluster representa um conjunto de empresas de alto desempenho com significativa avaliação de mercado, potencialmente atraente para investidores que buscam investimentos orientados para o crescimento com um plano de fundo de avaliação razoável.

O *cluster* 5 (verde) é composto apenas por SUZB3 e CSAN3, representando, respectivamente, a indústria de papel e celulose e o setor de energia. SUZB3, com um CAGR de receita de 5 anos de 36,49% e um EV/EBITDA de 4,21, exibe uma empresa com crescimento substancial que pode não estar sendo totalmente valorizado pelo mercado, dado o baixo índice EV/EBITDA. Isso indica uma potencial subvalorização ou cautela dos investidores quanto à sustentabilidade de tais taxas de crescimento. Por outro lado, CSAN3 apresenta um cenário contrastante, com um notável CAGR de receita de 38,95% emparelhado com um EV/EBITDA de 5,86, o que sugere uma valorização de mercado mais alta, mas ainda assim razoável, considerando seu ritmo de crescimento. A presença de apenas duas empresas neste cluster aponta para posições de mercado únicas e nichos operacionais especializados que essas empresas ocupam, tornando-as menos diretamente comparáveis a outras. Essa dupla pode atrair investidores interessados em capitalizar o potencial de crescimento que pode ainda não estar totalmente reconhecido em suas avaliações atuais.

No *cluster* 1, as diferenças entre as métricas financeiras das empresas BBSE3, VIVT3, ELET3 e RDOR3 revelam uma diversidade de posições de mercado. VIVT3 está avaliada de forma equilibrada com um EV/EBITDA de 4,25 e crescimento modesto, enquanto ELET3 tem um EV/EBITDA um pouco mais alto, refletindo expectativas moderadas de crescimento. RDOR3 destoa com seu EV/EBITDA elevado e crescimento nulo, indicando expectativas de melhoria futura. Este *cluster* apresenta um mix de estabilidade e potencial de crescimento, possivelmente subvalorizado pelo mercado.

O segundo e terceiro *clusters* representam, respectivamente, empresas com avaliações de mercado equilibradas e desempenho consistente, e empresas de commodities com fortes CAGRs de receita e EV/EBITDAs baixos, indicando potencial subvalorização pelo mercado. O quarto agrupamento contém empresas de alto desempenho com sólidas taxas de crescimento e avaliações de mercado significativas, e o quinto *cluster*, com apenas SUZB3 e CSAN3, sugere um forte potencial de crescimento que pode não estar completamente reconhecido pelas avaliações atuais do mercado.

CONCLUSÕES

O presente estudo forneceu uma análise dos indicadores financeiros das principais empresas listadas no Ibovespa, demonstrando o valor significativo da aplicação de técnicas de aprendizado de máquina, especificamente o algoritmo *K-means*, na análise financeira. Através da seleção cuidadosa de indicadores críticos, como Margem *EBITDA*, *CAGR* de Receita de 5 anos, *ROE*, *EV/EBITDA*, e Dívida Líquida/*EBITDA*, a pesquisa não apenas ilustrou as variações de desempenho entre empresas de setores diversificados, mas também forneceu conhecimentos sobre as estratégias operacionais e financeiras que diferenciam as empresas no índice.

A coleta e análise de dados, realizadas com rigor metodológico, sublinham a importância de fontes de dados confiáveis e técnicas de análise de dados robustas. O uso do *Python* como ferramenta de programação foi crucial para lidar com a complexidade dos dados, e as bibliotecas utilizadas como *matplotlib.pyplot*, *numpy* e *sklearn.cluster* demonstraram ser ferramentas indispensáveis para a visualização e processamento de dados complexos. A identificação de outliers e a normalização de dados garantiram a integridade e a qualidade da análise de clusterização, permitindo uma interpretação precisa dos clusters formados.

Os resultados obtidos revelam informações sobre o mercado de ações brasileiro. Os clusters identificados destacam as características únicas das empresas, variando de gestão conservadora a estratégias de crescimento agressivo, oferecendo um panorama dos diferentes estilos de gestão e abordagens ao risco no mercado. Esta segmentação das empresas permite uma compreensão mais matizada das tendências do mercado, possibilitando aos investidores e analistas fazerem escolhas mais informadas e estratégicas.

Adicionalmente, este estudo amplia o entendimento da aplicabilidade de técnicas de aprendizado de máquina no campo da análise financeira. Demonstrando como a combinação de métodos quantitativos com intuição financeira pode desbloquear novas perspectivas e facilitar uma compreensão mais profunda dos mercados financeiros. Ao destacar as sinergias entre tecnologia e finanças, a pesquisa fornece um modelo valioso para futuros trabalhos na interseção dessas disciplinas.

Em síntese, a pesquisa não só cumpriu seu objetivo de avaliar a aplicabilidade do *K-means* com base em indicadores das empresas do Ibovespa, mas também forneceu uma metodologia replicável e escalável para análises futuras. Os resultados gerados são de grande

relevância para uma variedade de *stakeholders*, incluindo acadêmicos, profissionais de mercado e investidores, reforçando a importância da inovação tecnológica e da análise de dados no desenvolvimento de estratégias de investimento eficazes e bem fundamentadas.

REFERÊNCIAS

[1] ALTMAN, Edward I. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. Journal of Finance, v. 23, p. 589-609, 1968.

[2] BISHOP, Christopher. Pattern Recognition and Machine Learning. Springer, janeiro de 2006.

[3] BRIGHAM, Eugene F.; EHRHARDT, Michael C. Financial Management: Theory and Practice. 14. ed. Cengage Learning, 2013.

[4] COPELAND, Tom; KOLLER, Tim; MURRIN, Jack. Avaliação de empresas – Valuation: Calculando e gerenciando o valor das empresas. São Paulo: Makron Books, 2000.

[5] DAMODARAN, A. Investment Valuation: Tools and Techniques for Determining the Value of Any Asset. 3. ed. ISBN: 978-1-118-01152-2, abril 2012.

[6] Fama, E. F. "The Behavior of Stock Market Prices". Journal of Business, v. 38, n. 1, p. 34-105, jan. 1965.

[7] GRAHAM, B.; DODD, D. L. Security Analysis. Nova York: Whittlesey House, McGraw-Hill Book Co., 1934. 725 p.

[8] HAIR, J. F., et al. So many ways for assessing outliers: What really works and does it matter?. 2010.

[9] HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. "Unsupervised Learning." In: The Elements of Statistical Learning. Springer Series in Statistics. Springer, New York, NY. DOI: 10.1007/978-0-387-84858-7_14, 2009.

[10] HINTON, G.E.; SALAKHUTDINOV, R.R. Reducing the Dimensionality of Data with Neural Networks. Science, v. 313, n. 5786, p. 504–507, 2006.

[11] HUNTER, J. D. Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, v. 9, n. 3, p. 90-95, 2007.

[12] IGLEWICZ, B.; HOAGLIN, D. C. How to Detect and Handle Outliers. Milwaukee: ASQC Quality Press, 1993.

[13] JEGADEESH, N.; TITMAN, S. "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency". The Journal of Finance, v. 48, n. 1, mar. 1993.

[14] JORDAN, M. I.; MITCHELL, T. M. "Machine learning: Trends, perspectives, and prospects." Science, v. 349, n. 6245, p. 255–260, 2015. Disponível em: https://doi.org/10.1126/science.aaa8415.

[15] KODINARIYA, T.M.; MAKWANA, P.R. Review on Determining of Cluster in K-means Clustering. 2013. Disponível em: [link]. Acesso em: 9 nov. 2023.

[16] KOLLER, T.; GOEDHART, M.; WESSELS, D. Valuation: Measuring and Managing the Value of Companies. 5. ed. Hoboken, N.J.: John Wiley & Sons, ©2010.

[17] KORAJCZYK, R. A.; MURPHY, D. "Do High-Frequency Traders Improve your Implementation Shortfall?". Journal of Investment Management, v. 18, n. 1, p. 18-33, First Quarter 2020.

[18] LINDEN, R. Técnicas de agrupamento. Revista de Sistemas de Informação da FSMA, n. 4, p. 18-36, jul./dez. 2009.

[19] MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: LeCam, L.M.; Neyman, J. (Eds.), Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, Califórnia: University of California Press, 1967. v. 1, p. 281-297.

[20] MURPHY, Kevin P. Machine learning - a probabilistic perspective. Adaptive computation and machine learning series, 2012. Disponível em: https://api.semanticscholar.org/CorpusID:17793133.

[21] PAPENBROCK, Jochen. Asset Clusters and Asset Networks in Financial Risk Management and Portfolio Optimization. Karlsruher Institut für Technologie (KIT), 2011.

[22] PEDREGOSA, F., et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, v. 12, p. 2825-2830, 2011.

[23] PÉREZ, F.; GRANGER, B. E. IPython: A System for Interactive Scientific Computing. Computing in Science & Engineering, v. 9, p. 21-29, 2007.

[24] SAMUEL, Arthur L. "Some Studies in Machine Learning Using the Game of Checkers." IBM Journal of Research and Development, v. 3, n. 3, p. 210-29, 1959.

[25] SHARPE, W. F. "Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk". Journal of Finance, v. 19, n. 3, p. 425-442, set. 1964. DOI: 10.1111/j.1540-6261.1964.tb02865.x.

[26] SIBSON, R. SLINK: An Optimally Efficient Algorithm for the Single-linkage Cluster Method. The Computer Journal, v. 16, n. 1, p. 30-34, 1973.

[27] TAN, Pang-Ning et al. Introduction to Data Mining. Michigan State University; University of Minnesota.

[28] VAN DER WALT, S.; COLBERT, S. C.; VAROQUAUX, G. The NumPy Array: A Structure for Efficient Numerical Computation. Computing in Science and Engineering, v. 13, n. 2, p. 22-30, 2011. DOI: 10.1109/MCSE.2011.37.

APÊNDICE

Apêndice 1: Código Completo Utilizado na Geração de Cluster CAGR Receita 5 Anos vs Margem EBITDA.

```
import matplotlib.pyplot as plt
import numpy as np
from sklearn.cluster import KMeans

def identify_outliers(data_list):
    threshold = 1.5
    mean = np.mean(data_list)
    std = np.std(data_list)

    lower_bound = mean - threshold * std
    upper_bound = mean + threshold * std

    outliers = []
    for i in data_list:
        if i < lower_bound or i > upper_bound:
            outliers.append(i)

    return outliers

# Dados
Stock_Code_list = ['ABEV3', 'B3SA3', 'CMIG4', 'CSAN3', 'ELET3', 'ENEV3', 'EQTL3',
'GGBR4', 'HAPV3', 'ITSA4', 'JBSS3', 'KLBN11', 'PETR4', 'PRIO3', 'RADL3', 'RAIL3',
'RENT3', 'SBSP3', 'SUZB3', 'UGPA3', 'VALE3', 'VBBR3', 'VIVT3', 'WEGE3']
EBITDA_Margin_list = [29.22, 61.93, 24.09, 31.82, 40.21, 39.06, 23.48, 21.05, 8.62, 212.51,
-1.07, 39.79, 51.66, 74.08, 11.38, 52.64, 36.22, 7.41, 58.69, 2.82, 37.97, 1.62, 39.9, 19.74]
CAGR_Revenue_5Y_list = [10.72, 22.44, 9.68, 38.95, 2.97, 17.63, 24.5, 17.42, 43.91, 11.3,
18.1, 19.06, 17.72, 64.15, 17.08, 10.6, 24.03, 8.59, 36.49, 12.42, 15.85, 16.5, 2.14, 25.71]
```

```
# Removendo outliers
outliers_EBITDA = identify_outliers(EBITDA_Margin_list)
outliers_CAGR = identify_outliers(CAGR_Revenue_5Y_list)

for outlier in outliers_EBITDA:
    index = EBITDA_Margin_list.index(outlier)
    del Stock_Code_list[index]
    del EBITDA_Margin_list[index]
    del CAGR_Revenue_5Y_list[index]

for outlier in outliers_CAGR:
    index = CAGR_Revenue_5Y_list.index(outlier)
    del Stock_Code_list[index]
    del EBITDA_Margin_list[index]
    del CAGR_Revenue_5Y_list[index]

data = list(zip(EBITDA_Margin_list, CAGR_Revenue_5Y_list))

# Implementação KMeans
n_clusters = 4
kmeans = KMeans(n_clusters=n_clusters, n_init=10)
kmeans.fit(data)

# Scatter plot dos clusters
colors = ['b', 'g', 'r', 'c', 'm', 'y', 'k']
plt.figure(figsize=(10,5))
for i in range(n_clusters):
    plt.scatter([data[j][0] for j, x in enumerate(kmeans.labels_) if x == i],
                [data[j][1] for j, x in enumerate(kmeans.labels_) if x == i],
                color=colors[i], label=f'Cluster {i+1}')

for i, txt in enumerate(Stock_Code_list):
    plt.annotate(txt, (EBITDA_Margin_list[i], CAGR_Revenue_5Y_list[i]))
```

```
plt.xlabel('Margem EBITDA (%)')
plt.ylabel('CAGR Receita 5 anos (%)')
plt.title(f'Agrupamento com {n_clusters} clusters')
plt.show()
```

Fonte: *Python* e autoria própria.