Universidade Federal do ABC
Centro de Engenharia, Modelagem e Ciências Sociais Aplicadas
Trabalho de Graduação em Engenharia de Informação

Aplicações de Aprendizado de Máquina para Seleção de ETFs e Construção de Portfólios.

Lucas Eduardo De Mieri

Prof. Dr. Luneque Del Rio de Souza e Silva Junior

Santo André - SP, Maio de 2022

Lucas Eduardo De Mieri

Aplicações de Aprendizado de Máquina para Seleção de ETFs e Construção de Portfólios.

Trabalho de Graduação apresentado ao curso de Graduação em Engenharia de Informação, como parte dos requisitos necessários para a obtenção do Título de bacharel em Engenharia de Informação.

Universidade Federal do ABC – UFABC

Centro de Engenharia, Modelagem e Ciências Sociais Aplicadas

Bacharel em Engenharia de Informação

Orientador: Prof. Dr. Luneque Del Rio de Souza e Silva Junior

Santo André - SP

Maio de 2022

Agradecimentos

Agradeço ao meu orientador, Luneque, por todos os conselhos, pela paciência e ajuda nesse período.

Aos meus amigos Luís Fernando, Vítor Queijo, Enzo Shiraishi, Lucas Pileggi, Juliana Freitas e Juliana Coimbra por toda colaboração, companheirismo e *feedbacks* que recebi ao longo da pesquisa.

Resumo

Exchange-traded funds(*ETFs*) são fundos de investimento negociados em bolsa, que têm seu valor lastreado aos ativos que o compõe, permitindo exposição a estratégias específicas e rebalanceamento automático. A popularização desses fundos fez com que em 2021 a bolsa americana contasse com 2793 *ETFs* ativos, mostrando um avanço significativo frente aos 2204 *ETFs* de 2020. O crescente número de *ETFs* e a diversidade de estratégias tornaram-se um obstáculo na avaliação dos investimentos. Para simplificar o processo de seleção de investimentos, este trabalho propõe o uso de algoritmos de aprendizado de máquina para classificar os *ETFs* mais promissores, auxiliando na composição do portfólio. Os métodos estudados foram: *SVM GRID*, *KPCA SVM GRID* e Árvore de decisão com redução de custo de complexidade, os algoritmos foram treinados com dados de 2010 até 2020, tendo como objetivo classificar os ativos baseado na rentabilidade esperada para 2021. Uma vez selecionados, os *ETFs* com a melhor relação de risco e retorno são combinados em portfólios e comparados ao *benchmark* do S&P 500, fazendo uso do *HRP* como método de otimização e construção dos portfólios. O desenvolvimento foi baseado nas seguintes etapas: (i) Seleção de dados e dos *ETFs* elegíveis; (ii) Seleção de algoritmos; (iii) Otimização dos portfólios; (iv) Avaliação dos portfólios. O *SVM Grid* mostrou-se a melhor abordagem para classificação dos ativos mais rentáveis. Foram construídos dois portfólios com índices de Sharpe de 2.52 e 2.89, superiores ao *benchmark* de 2.34, evidenciando a aplicabilidade da metodologia para construir portfólios e selecionar ativos com uma maior expectativa de retorno, frente a um menor risco.

Palavras-chaves: Aprendizado de Máquina, *Support Vector Machine*, Árvore de Decisão, Otimização de Portfólio.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

ETF	*Exchanged-Traded Fund*
S&P500	*Standard & Poor's 500*
FNDX	*Schawb Fundamental U.S Large Company ETF*
TQQQ	*UltraPro TQQQ*
QQQ	*Invesco QQQ Trust*
MIT	*Massachusetts Institute of Technology*
API	*Application Programming Interface*
XINA11	*Trend ETF MSCI China*
MCHI	*iShares MSCI China ETF*
HRP	*Hierarchical Risk Parity*
FNGO	*MicroSector FANG+ 2X Leveraged ETN*
FNGU	*MicroSector FANG+ 3X Leveraged ETN*
USA	*United States of America*
SVM	*Support Vector Machine*
HRP	*Hierarchical Risk Parity*
KPCA	*Kernel Principal Component Analysis*
TP	*True Positive*
TN	*True Negative*
FP	*False Positive*
FN	*False Negative*

Sumário

1 Introdução

Exchange-traded funds (*ETFs*) são fundos de investimento negociados em bolsa que buscam reproduzir um índice, setor ou estratégia. Como qualquer fundo, os *ETFs* possuem ativos, como títulos de dívida, ações ou *commodities*. Os ativos lastreiam a emissão da cota do fundo, que pode ser negociada em bolsa como uma ação, fazendo com que o dono do *ETF* também seja, indiretamente, o dono dos ativos que o compõem, como detalhado em (PETROVA, 2015).

Segundo (DEVILLE, 2008), *ETFs* são formas eficientes de replicar um *benchmark*, além de oferecerem fácil diversificação, baixas taxas de administração, transparência e eficiência tributária. A maioria dos *ETFs* adotam estratégias fixas, previamente detalhadas no prospecto do fundo, tanto para a composição dos ativos elegíveis a participar da portfólio. Quanto das proporções de cada ativo, permitindo ao investidor um rebalanceamento automático e clareza sobre as regras e critérios que estão sendo aplicados ao portfólio.

Fluxo de capital entre fundos ativos e passivos.

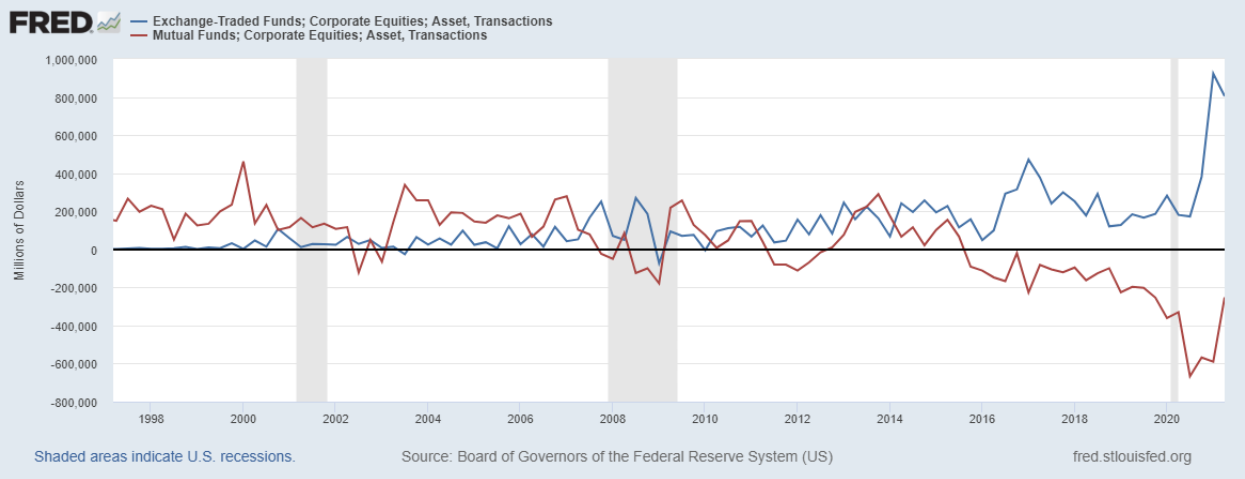

Figura 1 – Redução do fluxo de investimentos de fundos de investimentos de gestão ativa e aumento para fundos de gestão passiva, gerado com dados disponibilizados pelo (FRED, 2021).

Como podemos observar na Figura 1 ao longo dos anos, *ETFs* têm se mostrado alternativas mais atrativas aos investidores do que a alocação ativa do portfólio. Em 2020, foram levantados US$ 507.4 bilhões em novos *ETFs* (ROY, 2021), ultrapassando o antigo recorde de US$ 476.1 bilhões de 2017. Segundo os dados divulgados no relatório trimestral pela (NYSE, 2021), ao final de 2021 existiam 2793 *ETFs* nos *USA*, representando um total de US$ 7.233 trilhões.

Segundo (LAZZARA, 2021), a rentabilidade do S&P 500, que representa a média do mercado americano, foi superior à maioria das ações que compõem o índice nos últimos 20 anos, sendo um forte indício da dificuldade de selecionar ativos individuais, com uma relação vantajosa de risco e retorno, que justificasse a escolha em detrimento do investimento neutro de um *ETF*.

Retorno dos ativos individuais do S&P ao longo de 20 anos.

Figura 2 – Apenas 22% dos ativos tiveram uma rentabilidade superior ao índice. Fonte (LAZZARA, 2021).

Como podemos observar na Figura 2, apenas 22% das 500 maiores ações da bolsa americana tiveram um retorno superior à média neutra do mercado, enquanto a mediana das rentabilidades teve um retorno de 63 % frente aos 322 % do índice, ressaltando que aproximadamente 30% dos ativos tiveram rentabilidade negativa nos últimos 20 anos.

O fato do valor nominal dos ativos ser assimétrico, uma vez que a alta é irrestrita e a baixa é limitada a zero, as perdas de capital são compensadas pelos poucos ativos de alta rentabilidade. A escassez de ativos mais rentáveis que o índice corrobora o que foi observado por (FAMA; FRENCH, 2010), representando uma dificuldade adicional para fundos de investimento ativos de apresentarem retornos acima da média do mercado.

Dados os conceitos apresentados e a relevância do tema, é pertinente estabelecer critérios objetivos, como em (ELTON; GRUBER; BLACKE, 2011), de modo a identificar características em comum entre os *ETFs* com rentabilidade superior à média do mercado, eliminando possíveis vieses. A abordagem consiste na implementação de algoritmos de aprendizado de máquina, como vistos em (BAEK KWAN YONG LEE; OH, 2020), que também destaca o amplo emprego de estratégias de aprendizado de máquina por instituições financeiras como Morgan Stanley e Goldman Sachs.

Também é do interesse que os portfólios construídos pela abordagem apresentada, apresentem um maior qualidade nos retornos do que o *benchmark*, já que não necessariamente é desejável alcançar um alto retorno que esteja diretamente associado a um alto risco.

Esse trabalho adotará como principal métrica de qualidade o índice de Sharpe, detalhado em (SHARPE, 1994), que mede a razão simples entre o retorno real dos ativos e seu desvio padrão. Adicionalmente o índice de Sortino, descrito em (SORTINO, 1994), será considero como métrica secundária para avaliação da qualidade dos retornos, sendo que o índice mede a razão entre o retorno real e o desvio padrão das perdas. A adoção desses índices permite avaliar o impacto do uso de estratégias de maior risco, como (PROSHARES, 2021) que faz uso de alavancagem, para multiplicar os possíveis retornos em troca de um maior risco. Sendo assim nas etapas de avaliação de portfólio o uso de alavancagem será discriminado, afim de verificar o seu impacto.

1.1 Motivação

A preferência por estratégias de gestão passiva tem aumentado, conforme observado na Figura 1, que é coerente com a dificuldade dos fundos de gestão ativa em apresentarem retornos consistentes no longo prazo e das altas taxas associadas. Destaca-se que com *ETFs* é possível compor um portfólio de investimentos diversificado, onde o investidor compra a cota do fundo em bolsa e prontamente já tem sua carteira exposta a uma ampla gama de ações.

Tabela 1 – Fundos de gestão ativa que rentabilizaram abaixo do benchmark de gestão passiva ao longo dos últimos 10 anos, dados de (S&P-SPIVA, 2022).

Categoria do fundo	***benchmark***	Percentual de fundos que tiveram retornos abaixo do **benchmark**			
		1 Ano (%)	**3 Anos (%)**	**5 Anos (%)**	**10 Anos (%)**
Large-Cap	*S&P 500*	58.20	67.64	72.67	82.51
Mid-Cap	*S&P MidCap 400*	75.52	49,35	59.20	73.09
Small-Cap	*S&P SmallCap 600*	78.02	54.83	66.73	83.51
Multi-Cap	*S&P Composite 1500*	50.55	68.62	69.81	88.58
Large-Cap Growth	*S&P 500 Growth*	64.98	53.14	52.78	81.46
Large-Cap Core	*S&P 500*	62.74	74.8	81.58	93.32
Large-Cap Value	*S&P 500 Value*	48.35	72.56	71.26	87.75
Mid-Cap Growth	*S&P MidCap 400 Growth*	60.15	19.53	30.56	55.43
Mid-Cap Core	*S&P MidCap 400*	72.55	72.55	72.55	72.55
Mid-Cap Value	*S&P MidCap 400 Value*	76.47	79.63	87.93	87.84
Small-Cap Growth	*S&P SmallCap 600 Growth*	77.13	16.48	32.12	66.99
Small-Cap Core	*S&P SmallCap 600*	77.05	69.85	86.06	96.08
Small-Cap Value	*S&P SmallCap 600 Value*	75	79.38	86.84	98.1
média		65.75	60.70	67.26	82.37

Estratégias de baixo viés e atreladas a critérios objetivos, como *ETFs*, têm apresentado retornos significativamente melhores no longo prazo que fundos de gestão ativa, conforme a Tabela 1. Destacando, que a categoria dos fundos é referente ao tipo de ação comprada, descriminadas em três categorias de tamanho: *Large, Mid* e *Small*, além de três categoria de avaliação de ação: *Value, Growth* e *Core*, ambas as categorias detalhadas em (MORNINGSTAR, 2009). Observa-se como a média dos fundos ativos apresentou retornos consistentemente inferiores ao *benchmark*, principalmente em um intervalo de 10 anos, reforçando a necessidade de estabelecer critérios para a seleção das abordagens mais

promissoras e que possam superar o seu respectivo *benchmark*, evidenciando a dificuldade da gestão ativa em apresentar retornos no longo prazo.

Em geral, *ETFs* contam com pouca ou nenhuma decisão direta da gestão na alocação dos seus recursos, se baseando em regras objetivas e de baixo viés. Tomando como exemplo o *ETF FNDX*, descrito no prospecto (SCHWAB, 2021), que seleciona *Large Caps* americanas e leva em conta o tamanho da companhia, receitas, remuneração ao acionista e distribuição de pesos por setor, com rebalanceamentos anuais seguindo as mesmas regras descritas pelo prospecto. Evidenciando como *ETFs* seguem estratégias delimitadas e com objetivo específico de se fazer cumprir as normas propostas e não de gerar ganhos adicionais como nos fundos de gestão ativa, que pode incorrer em riscos desnecessário e aumento de custos, como visto em (FAMA; FRENCH, 2010).

A maioria dos *ETFs* têm uma estratégia algorítmica, com conjuntos claros e replicáveis de regras que foram selecionadas durante a sua construção, de modo a reduzir o grande número de ativos disponíveis no mercado a um grupo cada vez menor que possa compor um portfólio, filtrando os ativos até obter as ações que melhor representam as características buscadas pela estratégia, tendo um funcionamento compatível a algoritmos de aprendizado de máquina, como demonstrado por (LEE, 2019).

Dentre as possíveis estratégias para a seleção de ações, destaca-se o uso de derivativos para alavancar o retorno dos ativos, como usado pelo *ETF TQQQ*, que oferece o triplo da volatilidade do *ETF QQQ*. O alto risco associado ao emprego de alavancagem em *ETFs* não é recomendado para estratégias de longo prazo, como descrito no prospecto (PROSHARES, 2021). Para avaliar se o emprego de alavancagem de fato pode produzir resultados indesejados na volatilidade do portfólio, a análise dos resultados desse trabalho empregará a comparação do portfólio com e sem *ETFs* alavancados.

Sabendo da dificuldade em selecionar investimentos que ter desempenho acima da média do mercado e fazendo uso da característica algorítmica dos *ETFs*, este estudo se propõe a empregar algoritmos de classificação, como usados em (ITO; MURAKAMI; DUTTA, 2021), para selecionar ativos com estratégias com rentabilidades esperadas superiores à média do mercado. Uma vez que os *ETFs* possam ser selecionados, eles são combinados em portfólios, a fim de avaliar o seu desempenho conjunto contra o *benchmark*.

1.2 Objetivos

Fazendo uso da estrutura algorítmica e diversificada dos *ETFs*, este projeto busca analisar o retorno histórico das diversas estratégias, usando técnicas de aprendizado de máquina para selecionar *ETFs* com rentabilidades atrativas e empregar técnicas de otimização para construir diversos portfólios de teste baseados nos *ETFs* selecionados. Para verificar se os métodos foram efetivos em construir portfólios capazes de apresentarem retornos acima da média do mercado, o S&P 500, sendo que tanto para a seleção dos ativos quanto para a otimização dos portfólios serão empregadas técnicas de aprendizado de máquina.

2 Metodologia

A metodologia adotada neste projeto foi desenvolvida integralmente em *Python* e pode ser dividida em quatro tópicos: (i)coleta de dados de cada um dos ETFs disponíveis no mercado americano; (ii) classificação dos ETFs por meio de algoritmos de aprendizado de máquina; (iii) construção de portfólios otimizados a partir dos resultados da etapa anterior; (iv) avaliação qualitativa dos portfólios e comparação direta com o *benchmark*.

A abordagem se propõe a utilizar os métodos de aprendizado de máquina para selecionar *ETFs* com potencial de compor um portfólio, uma vez que os *ETFs* sejam classificados, eles são combinados em portfólios. O sistema é baseado no emprego de três algoritmos para classificação: *SVM* (*Support Vector Machine*) com *grid search*, *KPCA (Kernel Principal Component Analysis) SVM* com *grid search* e Árvores de Decisão com redução de complexidade, detalhados em (BAEK KWAN YONG LEE; OH, 2020), (LI; ROSSI, 2021) e (YANG, 2020). Por fim o portfólio é criado usando o *HRP (Hierarchical Risk Parity)*, visto em (PRADO, 2016), as quantidades de *ETFs* em um portfólio são gradualmente reduzidas removendo o *ETF* de menor índice de *Sharpe*.

Fluxograma do sistema proposto

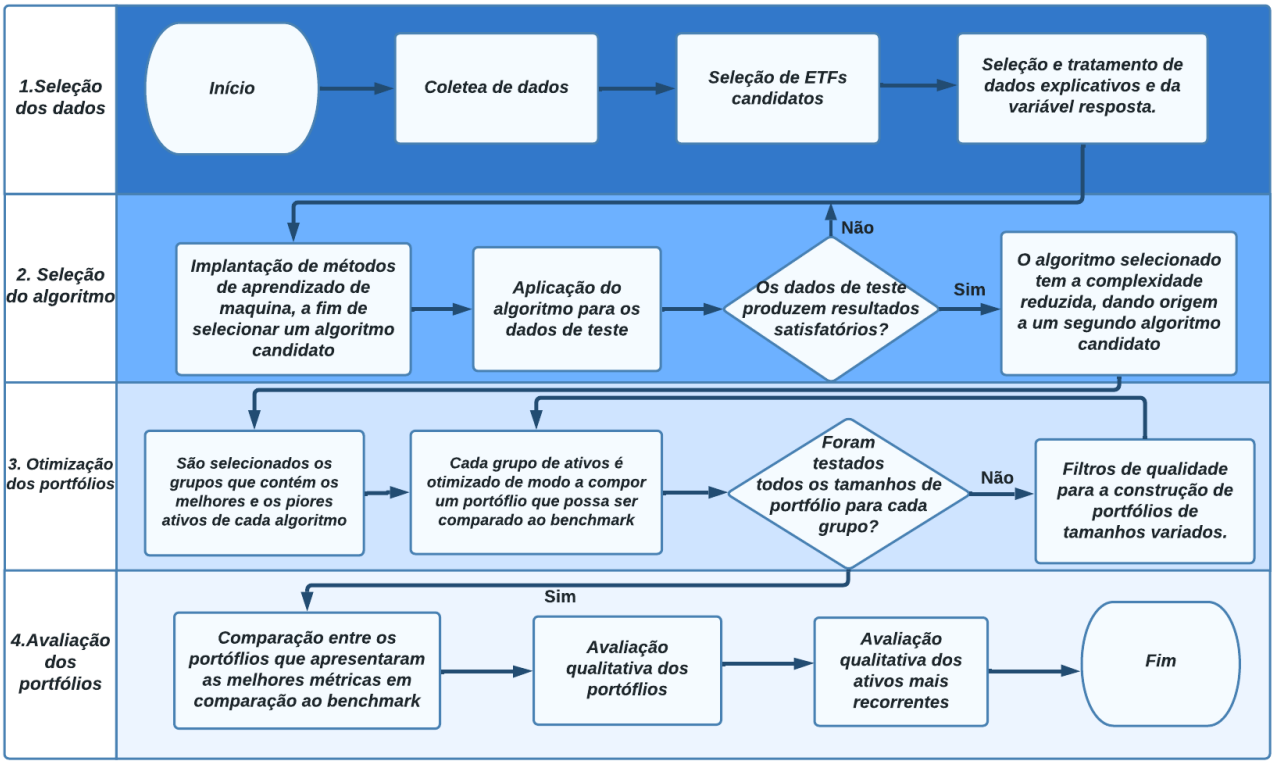

Figura 3 – Fluxograma de todas as etapas do desenvolvimento do sistema proposto, destacado em 4 tópicos: Seleção dos dados, Seleção do algoritmo, Otimização e Avaliação dos portfólios.

O sistema empregado para o desenvolvimento, descrito pela Figura 3, permite um desenvolvimento flexível, com passos claros e bem definidos, fazendo com que o escopo da pesquisa possa ser ampliado para mais algoritmos ou até reaproveitado em outro tipo de abordagem.

2.1 Seleção de dados

As variáveis do problema foram separadas de modo a isolar as variáveis explicativas, que seguem o intervalo 01-01-2010 até 31-12-2020, e a variável resposta, que compreende o retorno esperado de 2021, representado pela série histórica do *ETF*. No total, foram consideradas 82 variáveis explicativas e 2959 *ETFs*.

Foram desconsiderados *ETFs* que não tenham ao menos um ano de vida durante o desenvolvimento do estudo,que não tenham parte predominante do seu patrimônio alocados em ações, *ETFs* que não estejam domiciliados no mercado americano, para evitar variações cambiais e *ETFs* replicados na base, como por exemplo o *ETF XINA11*, domiciliado no Brasil e que replica o ativo americano *MCHI*, o que reduz a amostra de 2959 para 1058 *ETFs*

2.1.1 Coleta de dados

A coleta dados via *Python* se deu com o auxilio das bibliotecas: *Yahoo Query*, *Yahoo Finance* e *Investpy*. As bibliotecas foram selecionadas por possuírem licença MIT e por consultarem sites públicos, garantindo a confiabilidade da informação e permitindo a validação em tempo real, semelhante à coleta de dados descrita por (LIEW; MAYSTER, 2018).

2.1.2 Detalhamento das variáveis

Para atribuir a categoria da variável resposta, foi adotada a média dos retornos dos ativos. Como observado pela Figura 1, *ETFs* têm se tornado alternativas mais procuradas nos últimos 5 anos, sendo assim parte significativa dos ativos tem uma série inferior a uma década, que pode trazer à variável resposta um viés de *momentum*, ou seja, orientada a curto prazo, semelhante a abordagem de (BAEK KWAN YONG LEE; OH, 2020).

A variável **Categoria da Morningstar** foi construída para sanar eventuais ausências de categoria encontradas na base criada pela API. Sendo assim, adotou-se o critério descrito pela (MORNINGSTAR, 2016) para a atribuição manual das categorias faltantes, mantendo a uniformidade com as categorias já preenchidas. Para r_i o retorno do investimento, de um ativo de preço x_i, negociado por n dias, r_f sendo o ativo livre de risco e r_m o retorno médio do mercado. Definem-se as variáveis avaliadas, detalhadas em (CATALANO, 2021), como:

Alfa: Mede o retorno adicional do investimento sobre um investimento de abordagem.

$$\alpha = r_i - r_m$$

Beta: Avalia o risco sistemática, faz uma análise de regressão entre o investimento e a média do mercado.

$$\beta = \frac{cov(r_i, r_m)}{var(r_m)}$$

Desvio Padrão: Dispersão entre o valor diário e a média do valor do ativo.

$$\sigma^2 = \frac{\sum_{x_1=1}^{n}(x_i - \overline{x})^2)}{n-1}$$

Sharpe: Avalia a qualidade dos retornos, ponderada pelo risco do desvio padrão, adotando o ativo livre de risco do período R_f como a rentabilidade do título público americano, temos que $R_f \approx 0$.

$$S = \frac{R_p - R_f}{\sigma} \approx \frac{R_p}{\sigma}$$

Treynor: Mede o retorno pela volatilidade, determina o excesso de retorno que remunerou o investidor por unidade de risco

$$T = \frac{ri - rf}{\beta_i}$$

R-Squared: Explica o quanto o movimento do ativo, pode ser explicado pelo movimento do *benchmark*

$$R^2 = 1 - \frac{var(r_i)}{var(r_m)}$$

As variáveis explicativas são distribuídas em cinco grupos: Gerenciais; Categóricas; Retornos; Riscos; Riscos da Categoria e Setores. As variáveis são descrita por:

Gerenciais: ticker do *ETF*, data de consulta da informação.

Categóricas: Categoria e Categoria da Morningstar.

Retornos: Retorno total (2010 até 2020), Média do retorno (3 anos, 5 anos e 10 anos), Média do retorno da categoria (3 anos, 5 anos e 10 anos).

Risco: Alfa(3 anos, 5 anos e 10 anos), Beta (3 anos, 5 anos e 10 anos), Sharpe (3 anos, 5 anos e 10 anos), Treynor (3 anos, 5 anos e 10 anos), R^2(3 anos, 5 anos e 10 anos), Desvio Padrão (3 anos, 5 anos e 10 anos)

Risco da Categoria: Alfa da Categoria(3 anos, 5 anos e 10 anos), Beta da Categoria(3 anos, 5 anos e 10 anos), Sharpe da Categoria(3 anos, 5 anos e 10 anos), Treynor da Categoria(3 anos, 5 anos e 10 anos), R^2 da Categoria(3 anos, 5 anos e 10 anos), Desvio Padrão da Categoria (3 anos, 5 anos e 10 anos)

Setores: Distribuição dos pesos macros setoriais de cada *ETF*, distribuídos em : *Basic Materials, Consumer Cyclical, Financial Services, Real Estate, Consumer Defensive, Healthcare, Utilities, Communication Services, Energy, Industrials, Technology*

Classes: Dados baseados nas rentabilidades descrita pela variável resposta, categorizada em : Perda de Capital para rentabilidade negativa, Baixa Rentabilidade para o primeiro quartil das rentabilidades, Rentabilidade Intermediaria para o segundo quartil de rentabilidades, Rentabilidade Alta para o terceiro quartil das rentabilidades, Rentabilidade Muito Alta para o último quartil das rentabilidades, representando as 25% maiores da base.

2.1.3 Classificação dos *ETFs*

A etapa de classificação de *ETFs* busca treinar um algoritmo de aprendizado de máquina, utilizando cada uma das 5 classes, utilizando o período de 2010 até 2020 e avaliar as classificações frente ao período de 2021. Para verificar a capacidade preditiva do método em selecionar um conjunto de *ETFs* que possam ser candidatos para a composição de um portfólio, as classes são distribuídas entre os quartis da variável resposta.

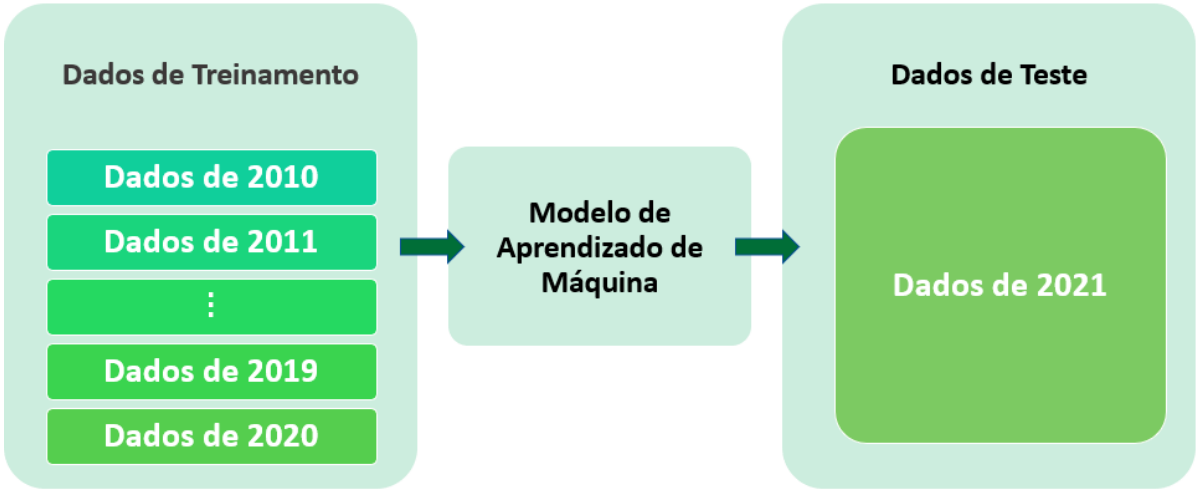

Figura 4 – Estrutura de treino e teste do modelo

Como visto na Figura 4, a estrutura proposta separa os períodos de treinamento e teste, para evitar um possível *overfit* do método e garantindo que a capacidade preditiva possa ser avaliada.

2.1.4 Visão geral dos *ETFs* selecionados

O estudo busca construir uma série de portfólios que possam ter uma expectativa de retorno superior ao S&P 500, suficientemente atrativo para justificar o investimento, conforme o observado em (S&P, 2017), que é composto majoritariamente por companhias na categoria *Large Blend*, descrito em (MORNINGSTAR, 2021).

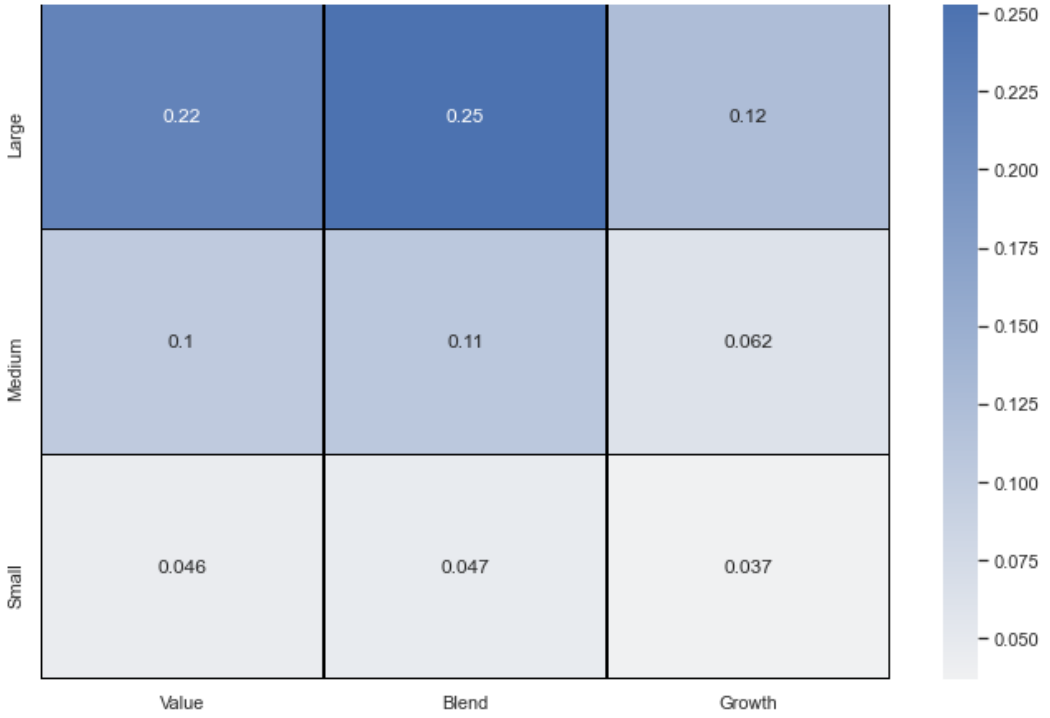

Figura 5 – Distribuição dos ETFs que compõe a base pelas 9 categorias da (MORNINGSTAR, 2016)

Cerca de 59% da base de *ETFs* é composta por ativos do mesmo porte do *benchmark*, conforme observado na Figura 5, com menor representatividade de *ETF* baseados na categoria de *Small*.

2.2 Seleção dos algoritmos

A robustez de métodos de aprendizado de máquina para problemas com alta dimensionalidade faz com que essa abordagem seja cada vez mais necessária para a seleção de ETFs, que comumente é usada em algoritmos gulosos e de busca de hiperparâmetros, como descrito em (FENG; S.GIGLIO; XIU, 2020). Logo, pode-se detalhar cada um dos métodos empregados e avaliar qualitativamente o seu funcionamento para o problema em questão. Cabe destacar que cada um dos algoritmos tem forte evidência literária quanto ao seu funcionamento na classificação de *ETFs*, como (BAEK KWAN YONG LEE; OH, 2020),(YANG, 2020) e (GUPTA et al., 2019).

2.2.1 *Support Vector Machine*

O *SVM* busca realizar separações lineares entre características, limitando uma fronteira de decisão que separa classes entre os dados, separados por uma margem w que é o vetor normal do hiperplano que discrimina as classes. Conforme descrito em (SMOLA; SCHOLKOPF, 2004) e (PEDREGOSA et al., 2011), a técnica de *SVM* é baseada em aprendizado supervisionado, tendo como objetivo identificar hiperplanos ótimos, ou seja, que maximizem a margem da base de treino. A função de custo pode ser descrita como:

$$min\ \frac{1}{2}w^t w + C\sum_{i=1}^{n}\xi_i, \tag{2.1}$$

Sendo que $w \in \mathbb{R}^p$ e $\xi_i \geq 0$,que só apresentará custo 0 se o valor predito e o valor de treino forem os mesmos, o termo C desempenha o papel de regular a força da penalidade aplicada na predição, atuando principalmente na separação das margens, garantindo que o algoritmo tente criar generalizações e não tente maximizar o tamanho das margens indefinidamente.

A versatilidade do método permite o emprego de abordagens como o *KPCA* que, como visto em (PEDREGOSA et al., 2011), tende a ser mais eficiente para classificar classes pouco lineares, uma vez que o método poderá a recorrer a separações: polinomiais, sigmoidais, cosseno, funções de base radial e outras, detalhado em (ZHANG J; MARSZALEK, 2007).

O problema da seleção de *ETFs* elegíveis para compor um portfólio é baseado em determinar estratégias que tenham apresentado resultados atrativos ao longo do período observado. Como observado por (BAEK KWAN YONG LEE; OH, 2020), (LIEW; MAYSTER, 2018) e (GUPTA et al., 2019) técnicas de aprendizado de máquina baseadas em *SVM* tem um amplo poder preditivo quando aplicado a *ETFs*, produzindo evidências significativas sobre a efetividade da técnica e a capacidade de identificar tendências de mercado. O estudo se propõe a treinar o algoritmo conforme a estrutura apresentada na Figura 4.

A abordagem foi empregada para um *Grid SVM*, permitindo que o algoritmo faça iterações entre os parâmetros e selecione aqueles que melhor descrevem uma alta expectativa de retorno. Para construir e avaliar o caráter preditivo das variáveis, o problema foi construído avaliando a capacidade do *SVM* de selecionar *ETFs* com as diferentes faixas de rentabilidade. Foram empregados dois modelos de *SVM*: o *Grid SVM* e o *Grid KPCA SVM* para verificar se algum ganho era obtido ao tentar empregar uma separação não linear entre as classes.

2.2.2 Árvore de decisão com redução do custo de complexidade

Árvores de decisão abordam recursivamente os dados de treino, até que sejam identificadas as etapas que melhor se adéquam as classificações propostas, como descrito em (BREIMAN, 1984). Esta classe de algoritmos tem se mostrado efetiva para a seleção de fundos, como demonstrado em (LI; ROSSI, 2021) e (LIEW; MAYSTER, 2018). Para o problema em questão que conta com um grande número de variáveis, pode-se recorrer ao uso de abordagens que simplifiquem o modelo, de modo a considerar apenas as variáveis de maior explicabilidade. Para reduzir o número de variáveis a árvore sofrerá podas, segundo a abordagem de redução do custo de complexidade.

Entende-se a redução do custo de complexidade ótimo, como a árvore de decisão que obtém a maior acurácia para um menor número de variáveis, sendo assim são penalizadas as árvores que contenham variáveis que acabem por reduzir a acurácia da classificação, o custo de complexidade pode ser descrito por:

$$\alpha_{k+1} = min_{t \in \bar{T}_k^C} \frac{R(t) - R(T_{k,t})}{|\bar{T}_{k,t}| - 1} \tag{2.2}$$

Onde o $R(T)$ descreve o custo de erro da classificação T,α é o custo de complexidade, t o nó interno da árvore, conforme detalhado em (CLARK; FOX; LAPPIN, 2010). Empregando a equação 2.2, podemos podar as variáveis de menor representativa durante o processo de classificação, o que pode representar um ganho frente ao *SVM*, reduzindo vieses da amostra e permitindo uma avaliação qualitativa do comportamento das variáveis.

2.2.3 Métricas

Cada uma das 5 classes, determinadas pelos dois métodos de *SVM* e pela árvore de decisão, serão avaliadas sob a ótica das seguintes métricas:

$$\begin{aligned} precision &= \frac{TP}{TP + FP} \\ recall &= \frac{TP}{TP + FN} \\ support &= TP + TN \\ F1 - Score &= \frac{2 * precision * recall}{precision + recall} \end{aligned}$$

Adotando a principal métrica como o *F1-Score* para a classe Rentabilidades Muito Altas, principalmente por conta da capacidade da métrica de sintetizar as informações tanto do *recall*, quanto *precision*. Ainda que a classe de Rentabilidades Muito Altas seja a principal para a seleção do método, vale destacar que também serão observadas as métricas das demais classes, a fim de identificar uma eventual inconsistência na predição do método.

2.3 Otimização dos portfólios

Os portfólios foram otimizados empregando *HRP*, visto em (PRADO, 2016). A técnica é baseada principalmente em aprendizado de máquina e teoria dos grafos. Técnicas clássicas baseadas em otimização quadrática como (MARKOWITZ, 1952), dependem que a matriz de covariância dos retornos seja invertível, como discutido por (JAIN; JAIN, 2019) estas restrição fazem com que matriz tenha que ter ao menos $\frac{N(N+1)}{2}$ amostras, que tendem a se tornar inviável para um grande número de ativos.

Quando aplicadas a um grande número de ativos, técnicas de otimização quadrática tendem a ser instáveis, concentrada, de baixa performance e difícil implementação, por conta da sua dependência da existência de uma matriz de covariâncias invertível. O conjunto dessas características é conhecido como *Markowitz's Curse*, detalhado em (PRADO, 2016). Uma alternativa mais robusta que os métodos clássicos é recorrer a métodos de aprendizado de máquina como o *HRP*, que não exige que a matriz de covariância dos retornos seja invertível, como evidenciado em (MARTIN, 2021).

O *HRP* consiste em três etapas: (i) árvore de clusters, construindo clusters hierárquicos; (ii) *quasi-diagonalization* onde são reorganizadas as linhas e colunas da matriz de covariância, agrupando elementos similares; (iii) *Recursive Bisection* onde é possível identificar a matriz de variância inversa da matriz quasi-diagonal, como demonstrado por (PRADO, 2016).

Qualitativamente o *HRP*, opera comparando a correlação de cada um dos ativos, para depois ordenar os clusters com possíveis sobreposições e destacar os grupos de ativos de menor correlação, o peso para cada ativo é atribuído na sua distância horizontal aos demais clusters, quanto menor a correlação com os demais, maior é a proporção do ativo no portfólio final.

O *HRP* será implementado para diversos portfólios, onde será possível observar o efeito de portfólios mais e menos concentrados, podendo avaliar qual dos métodos propostos tem um funcionamento melhor para o período de 2021. A poda através da remoções consecutivas dos *ETFs* com menor razão de *Sharpe*, descrito por $S = \frac{R_p - R_f}{\sigma_p}$(SHARPE, 1994), destacando que para 2021 os baixos valores dos juros americano fazem com que $R_f \approx 0$, sendo assim a razão pode ser descrita como $S = \frac{R_p}{\sigma_p}$, fazendo uso apenas do retorno do *ETFs* em relação a sua volatilidade, empregando uma métrica de risco ajustado para a seleção dos *ETFs*, de modo a identificar as estratégias que apresentaram maior retorno pelo menor risco.

2.3.1 Construção de portfólios otimizados

São criados diversos portfólios a partir dos *ETFs* que foram classificados pelo modelo como Rentabilidade Muito Alta. O desempenho dos portfólios é comparado com o *benchmark* do S&P500. São realizadas podas consecutivas do elemento com o menor índice de *Sharpe* a fim de avaliar o desempenho do portfólio ao concentrá-lo nos *ETFs* com o maior retorno pelo menor risco.

2.4 Avaliação dos portfólios

Uma vez que os portfólios de maior *Sharpe* são selecionados, sua rentabilidade é comparada diretamente com a do *benchmark*. Para fins de análise, um dos portfólios será escolhido e terá sua rentabilidade histórica avaliada usando o *Portfolio Visualizer*, disponível em (SCT, 2021), a fim de identificar evidências empíricas a respeito da estabilidade do portfólio.

3 Resultados e Discussões

A primeira etapa consiste na avaliação de qual metodologia de *SVM* apresentou resultados mais consistentes, sobretudo na classificação dos *ETFs* com Rentabilidade Muito Alta. Uma vez que uma das metodologias possa ser selecionada, os resultados são selecionados para o processo de otimização usando o *HRP*, que fará diversas iterações até construir uma série de portfólios otimizados que possam ser comparados com o S&P 500.

3.1 Resultados *GRID SVM* e *GRID KPCA SVM*

Para selecionar o método de *SVM* que melhor se adequá à otimização de portfólio, os resultados de ambos os experimentos são comparados para eleger qual abordagem é mais consistente.

Sobretudo, espera-se que um modelo capaz de selecionar ativos para um portfólio seja capaz de discriminar sobretudo a classe de Rentabilidade Muito Alta, que representará os ativos a serem selecionados para a etapa de otimização.

Tabela 2 – Métricas dos métodos de *SVM*, destacando o resultado superior do método *Grid SVM* em todas as métricas para a atribuição da classe de maior rentabilidade.

Classes e métricas médias		**GRID SVM**			**GRID KPCA SVM**		
		Score	0.69		Score	0.64	
	support	Precision	Recall	f1-score	Precision	Recall	f1-score
Perda de Capital	20	0.56	0.45	0.5	1	0.65	0.79
Baixa Rentabilidade	64	0.65	0.64	0.65	0.67	0.64	0.66
Rentabilidade Intermediaria	86	0.65	0.64	0.64	0.58	0.47	0.52
Rentabilidade Alta	72	0.66	0.69	0.68	0.53	0.64	0.58
Rentabilidade Muito Alta	76	0.83	0.86	0.84	0.73	0.86	0.79
acurácia	318			0.69			0.64
macro avg	318	0.67	0.66	0.66	0.7	0.65	0.67
weighted avg	318	0.69	0.69	0.69	0.65	0.64	0.64

Os resultados apresentados na Tabela 2 apresentam 0.83 de *precision* para o *Grid SVM* e 0.73 para o *Grid KPCA SVM*, na classe de Rentabilidade Muito Alta. Ambos os métodos apresentam *scores* compatíveis. Resta observar se os erros de classificação são dados entre os grupos vizinhos.

Para verificar se ocorrem grandes divergências entre as classificações entre os métodos de *SVM*. As classes foram tratadas como valores numéricos, sendo 1 para a classe de perda de capital e 5 para a classe de rentabilidade muito alta, os valores foram subtraídos, de modo a ser possível verificar se os métodos tendem a cometer erros entre classes distantes.

Distância entre grupos previstos e a variável resposta

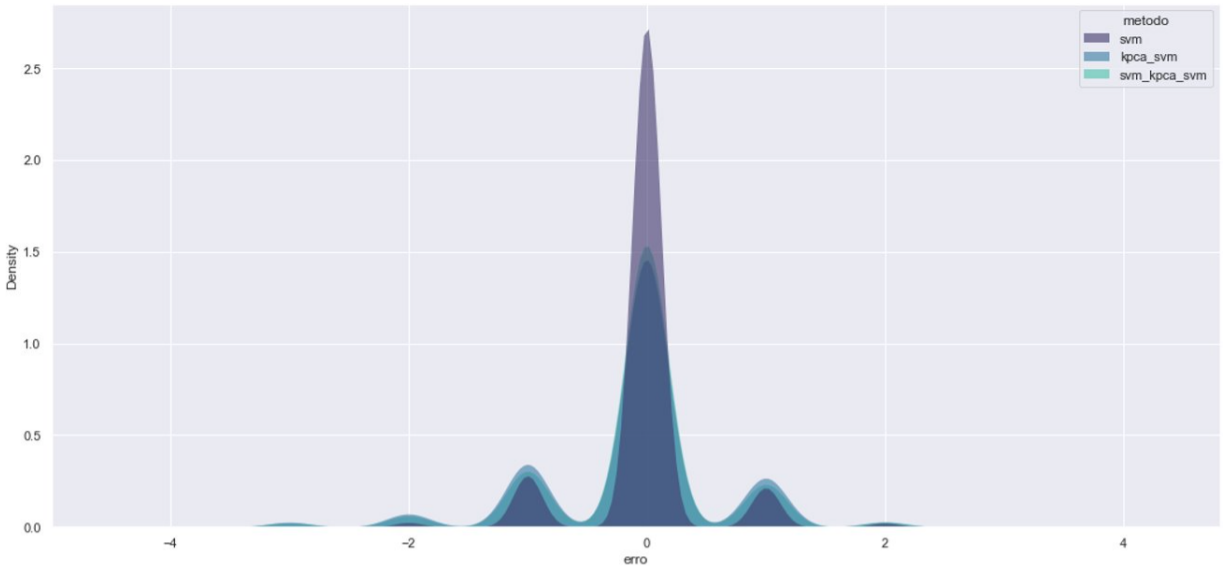

Figura 6 – Distribuição dos erros para cada uma das classificações de ambos os métodos SVM *Score*.

A distribuição apresentada na Figura 6, representa a diferença entre os resultados previstos pelo modelo e os resultados de teste, descritos pela Figura 4. Sendo a legenda *SVM _ KPCA _ SVM* referente a comparação dos métodos entre si, verificando as abordagens de *SVM* tendem a ter grandes divergências.

Pela distribuição de erros é possível observar, um baixo risco de propagar um método que seleciona estratégias de baixa rentabilidade para a etapa de otimização de portfólio. Ambos os métodos apresentam uma distribuição de erros equilibrada, com uma concentração dos erros na região central da Figura 6. Apesar dos erros serem compatíveis, é possível observar que o *Grid SVM* é mais assertivo que o *Grid KPCA SVM* e pelo método ter apresentado tanto um maior *Score*, quanto uma *precision* maior na classificação do grupo de Rentabilidade Muito Alta, os resultados do *Grid SVM* serão selecionados para a etapa de otimização.

3.2 Árvore de decisão

Aplicando a técnica de redução de complexidade, espera-se identificar qual valor de custo de complexidade α produz a maior acurácia, permitindo estabelecer o menor número de variáveis que melhor descreve o problema.

Variação da acurácia frente ao aumento do α

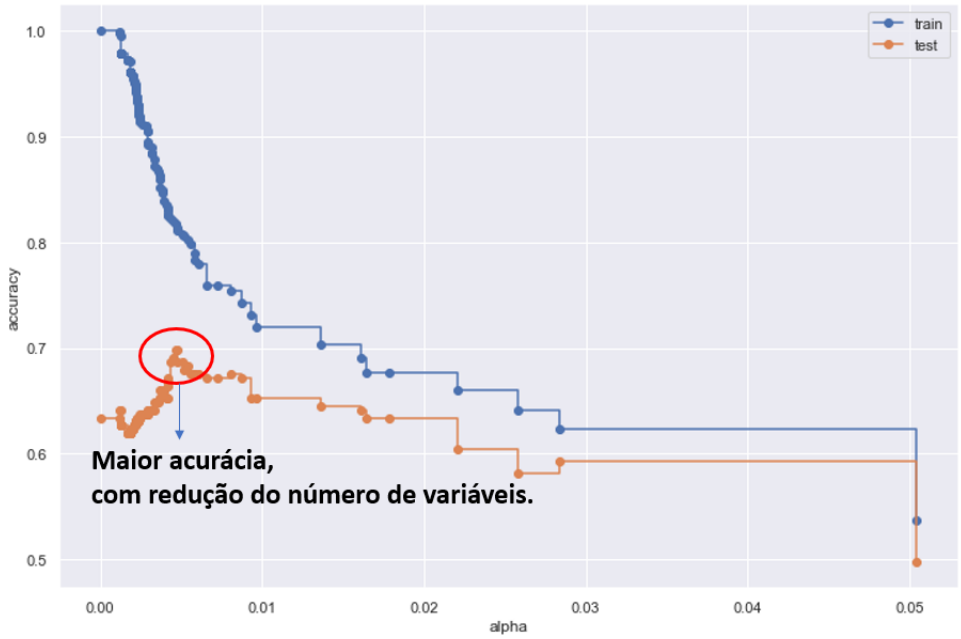

Figura 7 – Evolução dos parâmetros de α frente a acurácia.

Observa-se pela Figura 7 como a acurácia do problema se comporta frente a redução de variáveis, evoluindo significativamente até $\alpha = 0.004761$ e decaindo para valores superiores, representando um total de 50 variáveis frente as 82 variáveis utilizadas nos métodos de *SVM*.

Tabela 3 – Métricas para a Árvore de decisão, destacando o *score* superior em relação aos métodos *Grid SVM* e *KPCA SVM GRID*.

Árvore de decisão	Score	0.72		
Classes e métricas médias	support	Precision	Recall	f1-score
Perda de Capital	20	0.9	0.45	0.6
Baixa Rentabilidade	64	0.68	0.8	0.73
Rentabilidade Intermediaria	86	0.74	0.62	0.67
Rentabilidade Alta	72	0.69	0.71	0.7
Rentabilidade Muito Alta	76	0.77	0.88	0.82
acurácia	318			0.73
macro avg	318	0.76	0.69	0.71
weighted avg	318	0.73	0.73	0.72

O *Score* de 0.72 apresentado pela Tabela 3 é superior ao alcançado pelo uso dos métodos de *SVM* isolados, ainda que as métricas de classificação para a classe de Rentabilidade Muito Alta sejam piores em relação ao *SVM GRID*.

3.3 Comparação entre os métodos

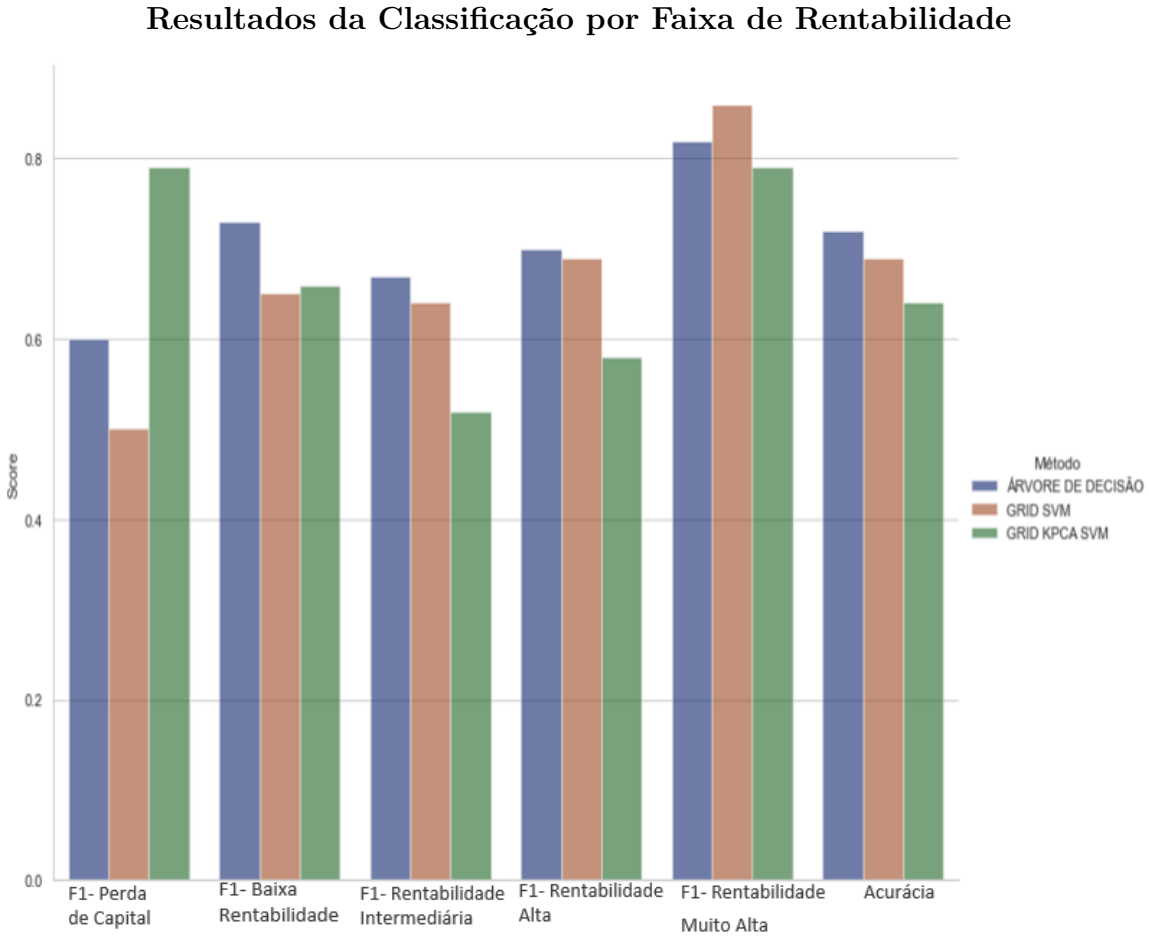

Figura 8 – Comparação direta entre a os modelos de *SVM GRID, KPCA SVM GRID* e Árvore de decisão

Como observado na Figura 8, o *GRID SVM* e Árvore de decisão são significativamente melhores na predição global e na classe de maior interesse. Em média os algoritmos classificaram 277 *ETFs* na classe de maior rentabilidade, posteriormente os *ETFs* da classe de maior rentabilidade serão convertidos em um portfólio que possa ser comparado ao *benchmark*.

3.4 Portfólios pré-otimização

Para avaliar a qualidade dos ativos selecionados e dos métodos empregados, é pertinente compor um portfólio pré-otimização, para cada uma das classes, de modo que seja possível avaliar a qualidade do próprio método no processo de seleção.

Para esta abordagem, será adotado o portfólio pré-otimizado onde cada ativo tem pesos equivalentes, espera-se que essa técnica venha a diminuir a influência de cada ativo individual, minimiza influências da alavancagem e ressalta a qualidade do método em selecionar o maior número de *ETFs* promissores possível.

Rentabilidade dos portfólios de pesos iguais, para cada um dos algoritmos.

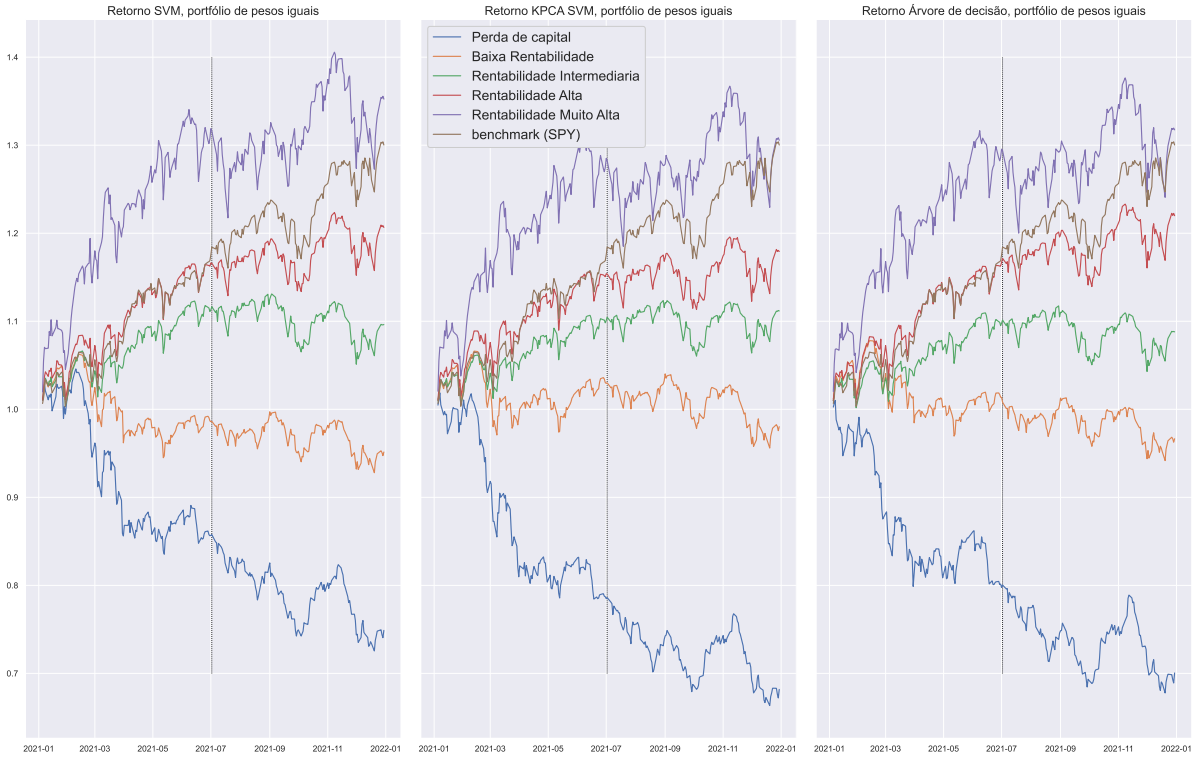

Figura 9 – Rentabilidade anual dos portfólios de pesos iguais, separados por cada um dos algoritmos e discriminados por classes.

Comparação da classe de maior rentabilidade de cada algoritmo.

Figura 10 – Classe de Rentabilidade Muito Alta para cada algoritmo, frente ao **benchmark**

Como pode-se observar na Figura 9, apenas as classes de rentabilidade muito alta apresentaram retornos consistentemente superiores ao *benchmark*. A Figura 10 demonstra como o *SVM GRID* tende a ter uma rentabilidade superior aos demais métodos, ainda que os portfólios apresente comportamentos semelhantes.

A abordagem dos portfólios pré-otimizados é equivalente a identificar a esperança dos retornos de cada classe, onde a probabilidade de se escolher um *ETF* que pertença a qualquer classe é proporcional ao número de *ETFs* da própria classe, tomando a forma:

$$E(R_i) = \sum_{i=1}^{n} R_i.p_i = \sum_{i=1}^{n} \frac{R_i}{n}$$

Onde $E(R_i)$ é o valor esperado, caso um *ETF* i, com retorno R_i, seja escolhido aleatoriamente de uma classe com n elementos. Sendo assim é possível observar como os retornos esperados de cada uma das classes tende a manter faixas estáveis, com pouca ou nenhuma troca da posição relativa de cada classe no gráfico, evidenciando uma estabilidade nas classificações do algoritmo.

3.5 Resultados *HRP*

Para que o processo de otimização de portfólios seja viável, se faz necessário que a técnica apresente retornos consistentemente melhores que a solução trivial, observada na Figura 10. Ainda que a otimização usando o *HRP* seja possível para um grande número de ativos, como observado por (PRADO, 2016), espera-se que a técnica apresente melhores resultados à medida em que ocorra uma redução no número de *ETFs* do portfólio.

A abordagem será focada nos métodos que apresentaram os melhores valores de *F1-Score* para a classe de Rentabilidade Muito Alta, *Grid SVM* e a Árvore de decisão, destacando os portfólios que fazem, ou não uso, de alavancagem . Os ativos são sucessivamente removidos, com base no menor índice de Sharpe e as iterações são realizadas até que o portfólio tenha apenas os 0.5% da amostra original, totalizando 20 ativos, com a melhor relação de retornos ponderados pelo risco, atuando como uma métrica da qualidade do investimento, como visto em (SHARPE, 1994).

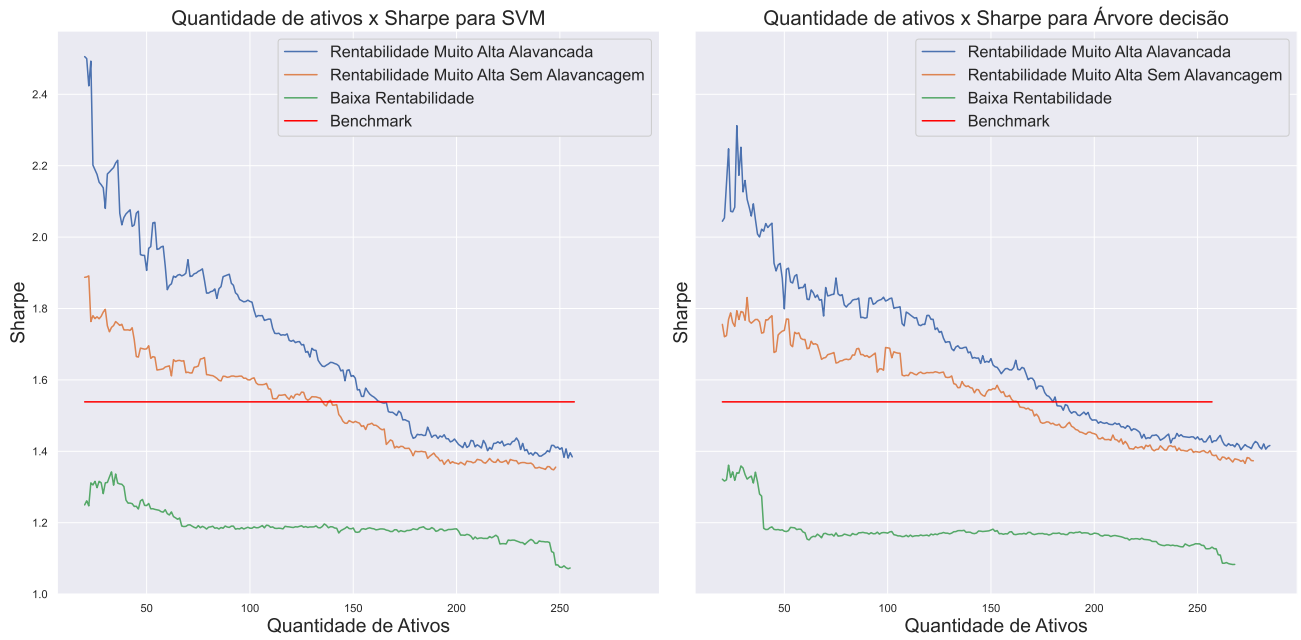

Figura 11 – O aumento da rentabilidade absoluta em contrapartida do aumento de risco e redução de diversificação, considerados os resultados absolutos do portfólio do começo de 2019 até o segundo semestre de 2021.

Como observado na Figura 11, o índice de Sharpe para ambos os métodos é consistentemente maior o *benchmark* para aproximadamente 150 melhores ativos na composição do portfólio, o que representa cerca de 5% da amostra original. Observa-se a distinção clara entre as classes de mais alta e mais baixa rentabilidade, sendo um indício da capacidade dos métodos de discriminar os ativos.

Tabela 4 – Em média, a alavancagem produziu um Sharpe 0.2241 maior para o *SVM GRID* e de 0.1281 para a Árvore de decisão. Destaca-se como a alavancagem, ainda que produzindo maiores Sharpes, apresentou um coeficiente de variação 50% maior para o *SVM GRID* e 30% para a Árvore de decisão.

Sharpe	**SVM alavancado**	**SVM sem alavancagem**	**Árvore alavancada**	**Árvore sem alavancagem**
média	1.7733	1.5492	1.6913	1.5632
mediana	1.6829	1.6360	1.6607	1.6706
desvio padrão	0.3101	0.1303	0.2451	0.1598
máximo	2.5050	1.8873	2.0443	1.7555
mínimo	1.3807	1.3659	1.4042	1.3479
Coeficiente de variação$(\frac{\sigma}{\mu})$	**0.1748**	**0.0841**	**0.1449**	**0.1022**

Os resultados apresentados na Tabela 4 reforçam os ganhos significativos que a alavancagem pode provocar no portfólio. Entretanto, vale ressaltar o ganho de volatilidade na adoção dessa estratégia, provocando grande variabilidade nos resultados do Sharpe, mesmo para pequenas alterações no portfólio.

3.6 Detalhamento e avaliação dos portfólios.

Para avaliar qualitativamente os portfólios apresentados na Figura 11, serão adotados os resultados do *HRP* para 20 ativos, nos casos do *SVM* com e sem alavancagem.

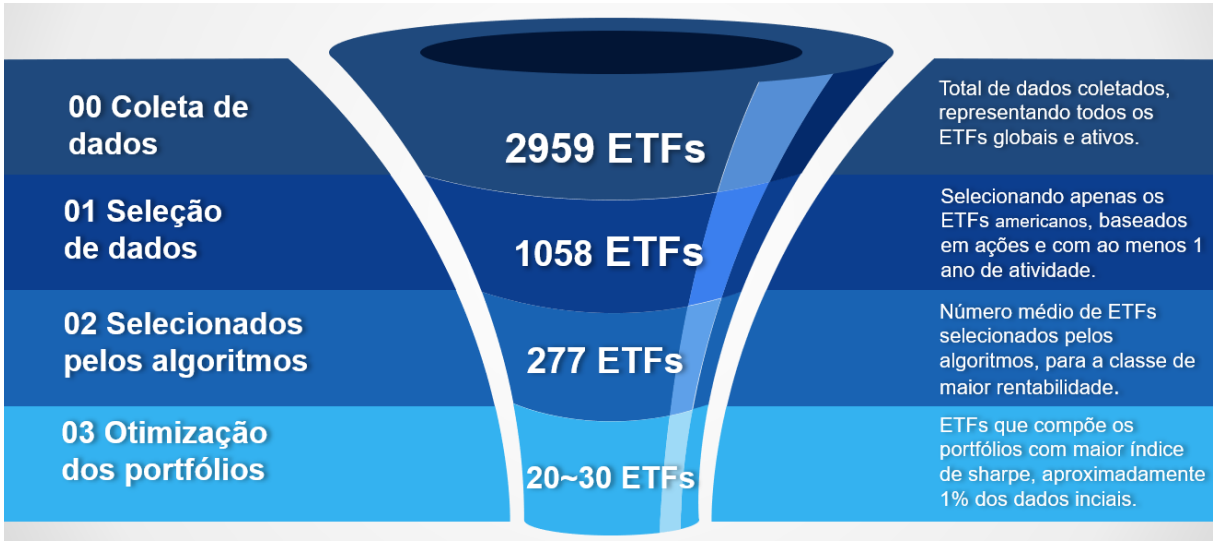

Figura 12 – Filtro detalhando as etapas do processo até a construção dos portfólios, dando enfoque para a quantidade de *ETFs* de cada etapa

A Figura 12 detalha como cada etapa do processo reduz a quantidade de *ETFs*, deixando a otimização de portfólios restrita aos ativos de maior qualidade e de alta expectativa de retorno, de modo a produzir porfólios que sejam mais atrativos do que a solução trivial, vista na Figura 10, que por si só já é mais atrativa que o *benchmark*.

Um dos motivos para se desejar uma alternativa mais atrativa que a solução trivial se dá pela dificuldade na administração de um portfólio tão diversificado, com 258 ativos para o *SVM*. Os pesos iguais implicam em um rebalanceamento recorrente da carteira, que pode resultar em custos adicionais e tornar o portfólio inviável.

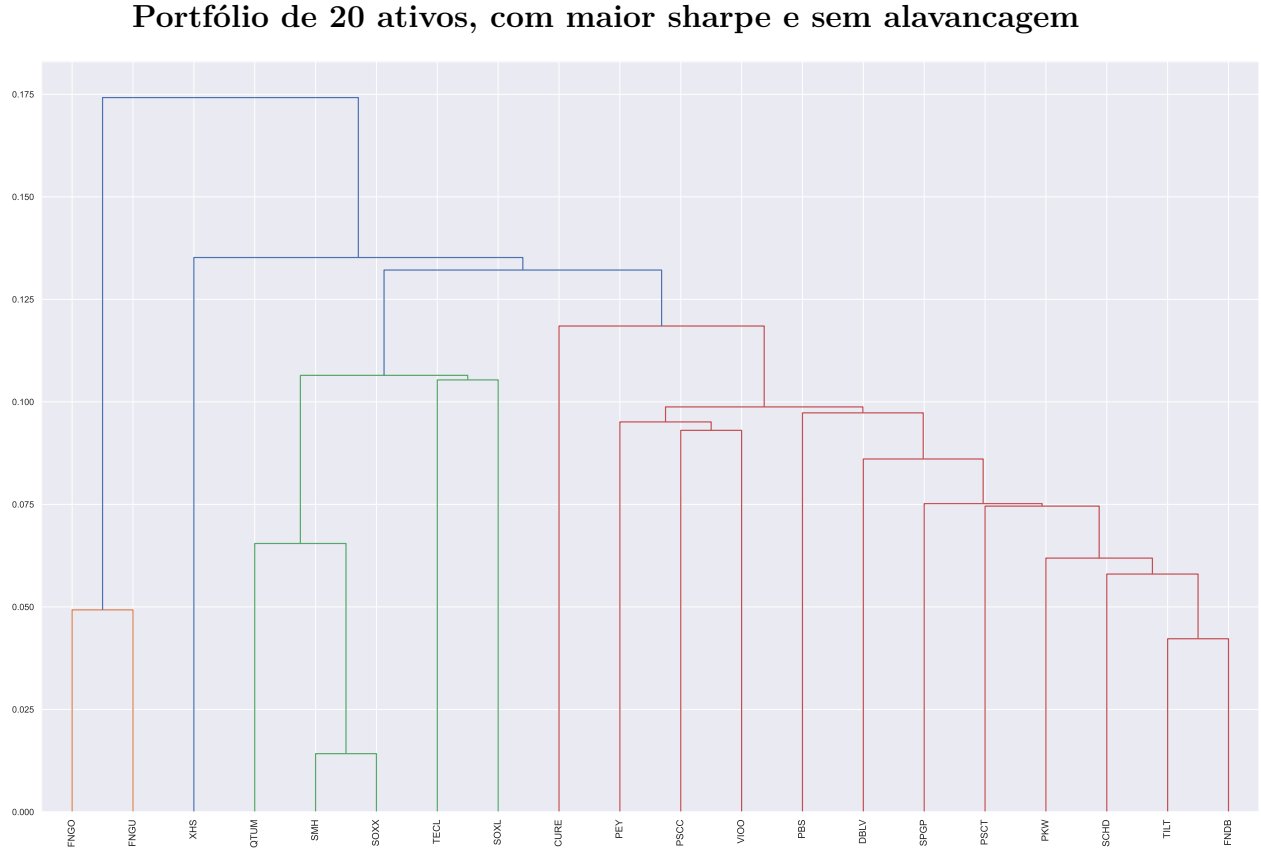

Figura 13 – Destaque nos ativos, considerando alavancagem, que compõe o portfólio de melhor rentabilidade, enfatizando os *clusters* formados pelo *HRP*, sendo o eixo horizontal cada um dos *ETFs* e o eixo vertical a distância entre os clusters, formando um portfólio de 20 ativos.

A Figura 13 mostra a clusterização dos ativos, divididos em quatro ramos. Os dendrogramas fornecem uma ideia prévia sobre a atribuição dos pesos do *HRP*, com ativos com menor distância entre os clusters recebendo um peso menor que os ativos mais distantes ou pouco correlacionados, criando uma estrutura diversificada que tende a favorecer investimentos que forneçam exposições diferentes dos já existentes no portfólio.

Observa-se que o cluster formado pelo *FNGU* e *FNGO* são próximo, compondo um cluster isolado a esquerda do dendrograma, essa proximidade é evidenciada pelo fato dos *ETFs* compartilharem a mesma estratégia de se expor a poucos ativos de grandes empresas de tecnologia americanas, com diferenças apenas no nível de alavancagem, como pode ser verificado em (BMO, 2022b) e (BMO, 2022a).

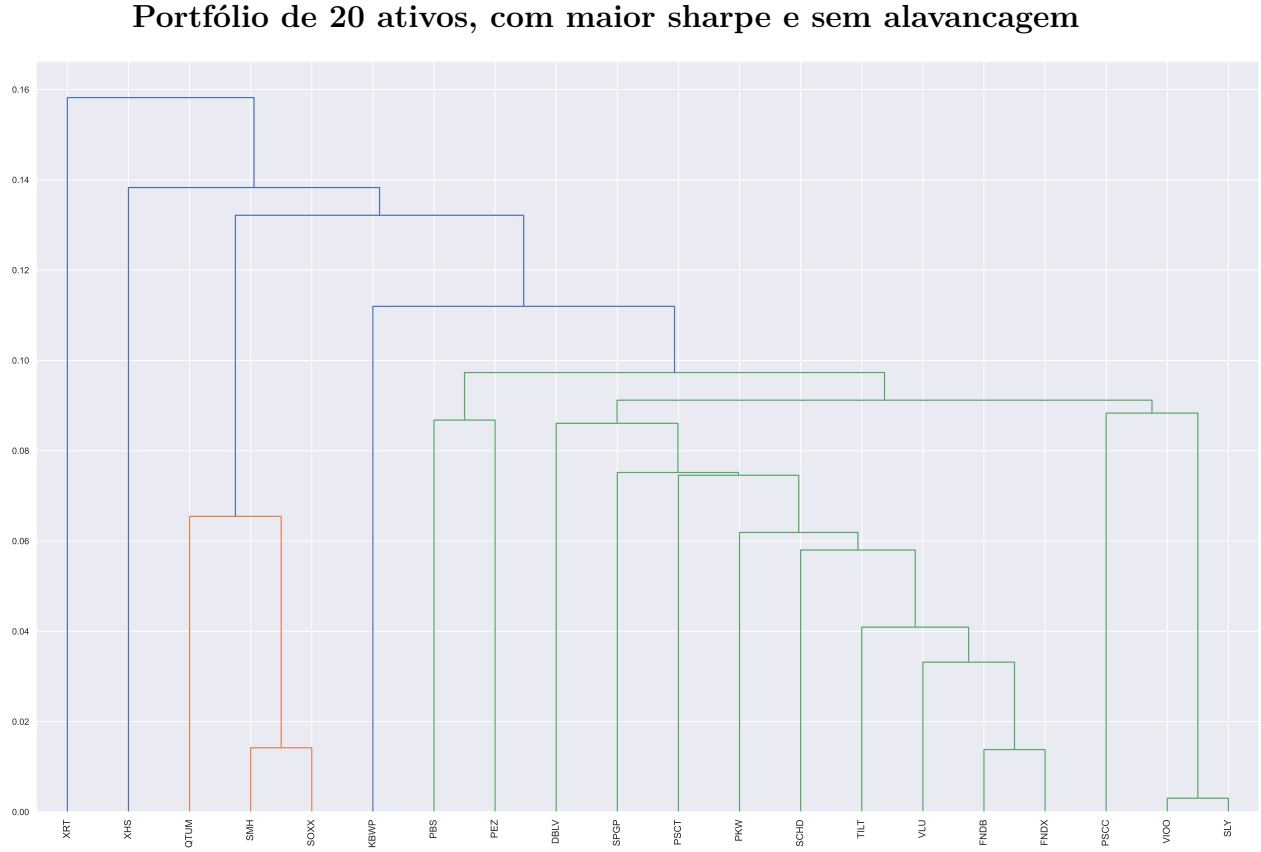

Figura 14 – Destaque nos ativos que compõe o portfólio de melhor rentabilidade, sem o uso de alavancagem, enfatizando os *clusters* formados pelo *HRP*, sendo o eixo horizontal cada um dos *ETFs* e o eixo vertical o a distância entre os clusters, formando um portfólio de 20 ativos.

A Figura 14 apresenta três grandes clusters. Foram observados três *ETFs* em ramos isolados, o que pode indicar uma melhor discriminação dos clusters para os ativos alavancados. Adicionalmente, observa-se como a abordagem de redução dos ativos, baseada puramente no índice de Sharpe, pode levar a seleção de *ETFs* que adotam estratégias similares e altamente correlacionadas, como observado no ramo composto pelo *VIOO* e *SLY*.

Pesos do *SVM* sem alavancagem

SVM sem alavancagem

Ticker	Name	Allocation
SMH	VanEck Semiconductor ETF	0.33%
QTUM	Defiance Quantum ETF	30.33%
SOXX	iShares Semiconductor ETF	0.07%
SCHD	Schwab US Dividend Equity ETF	3.72%
DBLV	AdvisorShares DoubleLine Value Eq ETF	1.14%
TILT	FlexShares Mstar US Mkt Factors Tilt ETF	0.40%
XHS	SPDR S&P Health Care Services ETF	2.76%
PSCC	Invesco S&P SmallCap Consumer Stapl ETF	1.65%
SPGP	Invesco S&P 500 GARP ETF	1.13%
PSCT	Invesco S&P SmallCap Info Tech ETF	0.87%
PKW	Invesco BuyBack Achievers ETF	2.37%
PBS	Invesco Dynamic Media ETF	7.76%
FNDB	Schwab Fundamental US Broad Market ETF	17.11%
VIOO	Vanguard S&P Small-Cap 600 ETF	0.31%
PEZ	Invesco DWA Consumer Cyclicals Mom ETF	2.76%
XRT	SPDR S&P Retail ETF	6.62%
FNDX	Schwab Fundamental US Large Company ETF	16.69%
VLU	SPDR S&P 1500 Value Tilt ETF	0.75%
KBWP	Invesco KBW Property&Casualty Ins ETF	1.98%
SLY	SPDR S&P 600 Small Cap ETF	1.25%

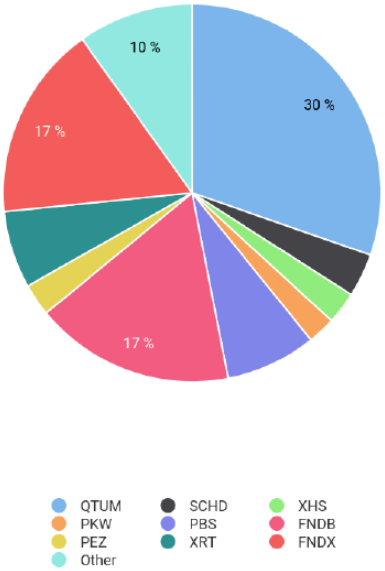

Figura 15 – Detalhamento das estrutura dos pesos e dos ativos que compõe o portfólio *SVM sem alavancagem*, feito em <https://www.portfoliovisualizer.com>.

Pesos do *SVM* alavancado

SVM alavancado

Ticker	Name	Allocation
FNGO	MicroSectors FANG+ 2X Leveraged ETN	4.58%
FNGU	MicroSectors FANG+ 3X Leveraged ETN	8.99%
SMH	VanEck Semiconductor ETF	0.22%
QTUM	Defiance Quantum ETF	15.29%
SOXX	iShares Semiconductor ETF	0.11%
SCHD	Schwab US Dividend Equity ETF	3.31%
DBLV	AdvisorShares DoubleLine Value Eq ETF	0.85%
TECL	Direxion Daily Technology Bull 3X ETF	9.44%
TILT	FlexShares Mstar US Mkt Factors Tilt ETF	0.49%
XHS	SPDR S&P Health Care Services ETF	2.05%
PSCC	Invesco S&P SmallCap Consumer Stapl ETF	1.00%
CURE	Direxion Daily Healthcare Bull 3X ETF	0.82%
SOXL	Direxion Daily Semicondct Bull 3X ETF	13.68%
SPGP	Invesco S&P 500 GARP ETF	0.84%
PSCT	Invesco S&P SmallCap Info Tech ETF	0.61%
PKW	Invesco BuyBack Achievers ETF	1.66%
PBS	Invesco Dynamic Media ETF	5.22%
PEY	Invesco High Yield Eq Div Achiev ETF	21.81%
FNDB	Schwab Fundamental US Broad Market ETF	8.70%
VIOO	Vanguard S&P Small-Cap 600 ETF	0.33%

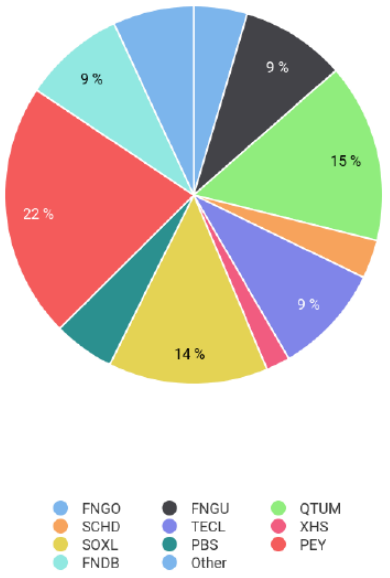

Figura 16 – Detalhamento das estrutura dos pesos e dos ativos que compõe o portfólio *SVM alavancado*, feito em <https://www.portfoliovisualizer.com>.

Destaca-se como os portfólios expostos tanto na Figura 16, quanto na Figura 15, apresentam poucas posições de alta concentração de capital. Posições que já são previamente destacadas pelos dendrogramas da Figura 13 e Figura 14, destacando o funcionamento do método, que tende a favorecer ativos de pouca correlação com o restante do portfólio.

Pode-se observar como ativos altamente correlacionados com os demais tendem a ter pouca significância, recebendo pesos que podem tornar desprezível a adição do ativo à composição total do portfólio.

Retorno dos portfólios em 2021

Figura 17 – Resultado histórico do portfólio de melhor desempenho, produzindo evidências empíricas, tanta das rentabilidades acima do *benchmark*, bem como do ganho da otimização frente ao caso trivial, mostrando como os portfólios com apenas 20 ativos podem ter uma rentabilidade superior ao portfólio de 258 ativos do caso do *SVM pré-otimizado*

Pela Figura 17, observa-se como os portfólios otimizados, compostos por 20 ativos, tiveram retornos superiores ao portfólio pré-otimizado que continha 258 ativos, ainda que existam períodos em que o *SVM pré-otimizado* tenha apresentado uma melhor rentabilidade que as demais estratégias.

Tabela 5 – Métricas dos retornos dos portfólios, ajustadas pela inflação, criado usando o *Portfolio Visualizer* <https://www.portfoliovisualizer.com>.

Portfólio (Jan 2021 - Dec 2021)- Ajustado pela inflação (7.04%)			
Métrica	**SVM alavancado**	**SVM sem alavancagem**	***benchmark***
Valor inicial	$ 10,000.00	$ 10,000.00	$ 10,000.00
Valor Final	$ 14,527.00	$ 13,104.72	$ 12,028.07
Retorno anualizado	45.20%	31.05%	20.28%
Desvio Padrão	16.65%	9.59%	11.15%
Máxima perda	-7.05%	-3.08%	-4.66%
Sharpe	2.52	2.89	2.34
Sortino	5.92	8.22	5.38

A Tabela 5 evidencia como os portfólios otimizados apresentaram valores de Sharpe superiores ao *benchmark* e ao portfólio *SVM pré-otimizado*, que apresentou um Sharpe de 2.19. Destaca-se como o *SVM alavancado*, ainda que tenha apresentado um retorno anualizado superior, teve um índices de Sharpe e de Sortino, detalhado em (SORTINO, 1994), inferiores ao *SVM sem alavancagem*.

O processo de otimização utilizando o *HRP* reduz significativamente o custo da manutenção de um portfólio grande, como visto usando *SVM pré-otimizado*, produzindo uma solução mais simples, de fácil alteração, aplicável para situações reais e com retornos significativamente superiores aos apresentados pelo *benchmark*.

Dentre as métricas, destaca-se como o *SVM alavancado* apresentou o maior valor de perda, superior ao dobro do *SVM sem alavancagem*. Os altos valores de perda da estratégia alavancada impactam adversamente os valores do Sortino, fazendo com que o *SVM sem alavancagem* se mostre mais seguro e consistente, ainda que com menor retorno anualizado.

Conclusões e Trabalhos Futuros

Foi possível verificar que, ao empregar e combinar diferentes abordagens de aprendizado de máquina, é possível obter portfólios diversificados, com expectativas de retornos estáveis e consistentes ao ponto de justificarem o investimento no lugar do *benchmark*. Os portfólios otimizados apresentaram um índice de Sharpe de 2.52 para o *SVM alavancado* e de 2.89 para o *SVM sem alavancagem*, frente aos 2.34 do *benchmark*, o que evidencia a capacidade da metodologia em produzir uma melhor relação de risco e retorno que a média do mercado.

Trabalhos Futuros

Fica a cargo de trabalhos futuros a simplificação e aperfeiçoamento da metodologia, de modo a produzir um resultado compatível com o apresentado utilizando um menor número de variáveis, realizar testes históricos para avaliar a consistência da estratégia ao longo do tempo e refinar a abordagem, tornando-a mais seletiva na classificação dos ativos.

Referências

BAEK KWAN YONG LEE, M. U. S.; OH, S. H. Robo-Advisors: Machine Learning in Trend-Following ETF Invests. 2020. Citado 5 vezes nas páginas 2, 6, 7, 11 e 12.

BMO. Microsectors fang+ index 2x leveraged etns. 2022. Disponível em: <https://www.bmoetns.com/ETN/FNGO.P/>. Acesso em: 15.01.2022. Citado na página 25.

BMO. *MicroSectors FANG+ Index 3X Leveraged ETNs*. 2022. Disponível em: <https://www.bmoetn.com/ETN/FNGU.P/>. Acesso em: 15.01.2022. Citado na página 25.

BREIMAN, L. Classification and regression trees. *Taylor Francis*, 1984. Citado na página 13.

CATALANO, C. B. T. J. 2021. Disponível em: <https://www.investopedia.com/investing/measure-mutual-fund-risk/>. Acesso em: 03.02.2022. Citado na página 8.

CLARK, A.; FOX, C.; LAPPIN, S. The handbook of computational linguistics and natural language processing. 2010. Citado na página 13.

DEVILLE, L. Exchange traded funds: History, trading and research. *Handbook of Financial Engineering, Springer, pp.1-37*, 2008. Citado na página 1.

ELTON, E. J.; GRUBER, M. J.; BLACKE, C. R. Holdings data, security returns, and the selection of superior mutual funds. journal of financial and quantitative analysis. 2011. Citado na página 2.

FAMA, E. F.; FRENCH, K. R. Luck versus skill in the cross-section of mutual fund returns. the journal of finance. 2010. Citado 2 vezes nas páginas 2 e 4.

FENG, G.; S.GIGLIO; XIU, D. Taming the factor zoo: A test of new factors. 2020. Citado na página 11.

FRED. *St. Louis Federal Reserve Equity: Diversity and Inclusion*. 2021. Disponível em: <https://fred.stlouisfed.org>. Acesso em: 03.08.2021. Citado 2 vezes nas páginas 4 e 1.

GUPTA, D. et al. Financial time series forecasting using twin support vector regression. 2019. Citado 2 vezes nas páginas 11 e 12.

ITO, H.; MURAKAMI, A.; DUTTA, N. Clustering of etf data for portfolio selection during early period of corona virus outbreak. 2021. Citado na página 4.

JAIN, P.; JAIN, S. Can machine learning based portfolios outperform traditional risk-based portfolios? the need to account for covariance misspecification. 2019. Citado na página 14.

LAZZARA, C. *Equity: Diversity and Inclusion*. 2021. Disponível em: <https://www.spglobal.com/en/research-insights/articles/equity-diversity-and-inclusion>. Acesso em: 03.08.2021. Citado 2 vezes nas páginas 4 e 2.

LEE, J. New revolution in fund management: Etf/index design by machines. 2019. Citado na página 4.

LI, B.; ROSSI, A. Selecting mutual funds from the stocks they hold: a machine learning approach. 2021. Citado 2 vezes nas páginas 6 e 13.

LIEW, J.; MAYSTER, B. Forecasting etfs with machine learning algorithms. 2018. Citado 3 vezes nas páginas 7, 12 e 13.

MARKOWITZ, H. Portfolio selection. *The Journal of Finance*, v. 7, n. 1, p. 77–91, mar. 1952. Disponível em: <https://www.jstor.org/stable/2975974>. Citado na página 14.

MARTIN, R. A. Pyportfolioopt: portfolio optimization in python. *Journal of Open Source Software*, The Open Journal, v. 6, n. 61, p. 3066, 2021. Disponível em: <https://doi.org/10.21105/joss.03066>. Citado na página 14.

MORNINGSTAR. *Morningstar Style Box*. 2009. Disponível em: <https://www.morningstar.com/content/dam/marketing/apac/au/pdfs/Legal/Stylebox_Factsheet.pdf>. Citado na página 3.

MORNINGSTAR. *The Morningstar Category Classifications*. 2016. Disponível em: <http://morningstardirect.morningstar.com/clientcomm/morningstar_categories_us_april_2016.pdf>. Acesso em: 03.08.2021. Citado 3 vezes nas páginas 4, 8 e 11.

MORNINGSTAR. *SPDR S&P 500 ETF Trust SPY*. 2021. Disponível em: <https://www.morningstar.com/etfs/arcx/spy/quote>. Acesso em: 03.08.2021. Citado na página 11.

NYSE. *New York Stock Exchange Arca Q4 2021 Quarterly ETF Report*. 2021. Disponível em: <https://www.nyse.com/etf/exchange-traded-funds-quarterly-report>. Acesso em: 01.02.2022. Citado na página 1.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Citado na página 12.

PETROVA, E. A brief overview of the types of etfs. 2015. Citado na página 1.

PRADO, M. L. de. Building diversified portfolios that outperform out of sample. 2016. Citado 3 vezes nas páginas 6, 14 e 22.

PROSHARES. Summary prospectus tqqq. 2021. Disponível em: <https://www.proshares.com/our-etfs/leveraged-and-inverse/tqqq>. Citado 2 vezes nas páginas 3 e 4.

ROY, S. *Record ETF Assets Growth In 2020*. 2021. Disponível em: <https://www.etf.com/sections/monthly-etf-flows/etf-monthly-fund-flows-december-2020?nopaging=1>. Acesso em: 03.08.2021. Citado na página 1.

SCHWAB, C. *Summary Prospectus Schwab Fundamental U.S. Large Company Index ETF*. 2021. Disponível em: <http://connect.rightprospectus.com/Schwab/TADF/808524771/SP?site=Fund>. Acesso em: 03.08.2021. Citado na página 4.

SCT. *Portfolio Visualizer by Silicon Cloud Technologies LLC*. 2021. Disponível em: <https://www.portfoliovisualizer.com>. Acesso em: 01.02.2022. Citado na página 15.

SHARPE, W. F. The sharpe ratio. 1994. Citado 3 vezes nas páginas 3, 14 e 22.

SMOLA, A. J.; SCHOLKOPF, B. A tutorial on support vector regression, statistics and computing archive volume 14. 2004. Citado na página 12.

SORTINO, F. Performance measurement in a downside risk framework. journal of investing. 1994. Citado 2 vezes nas páginas 3 e 29.

S&P. Understanding the sp managed risk indices. 2017. Citado na página 11.

S&P-SPIVA. *SPDR S&P 500- SPIVA data Results by Region*. 2022. Disponível em: <https://www.spglobal.com/spdji/pt/research-insights/spiva/>. Acesso em: 06.02.2022. Citado 2 vezes nas páginas 6 e 3.

YANG, K. C. Apply k-means techinique and decision tree analysis to predict taiwan etf performance. 2020. Citado 2 vezes nas páginas 6 e 11.

ZHANG J; MARSZALEK, M. L. S. S. C. Local features and kernels for classification of texture and object categories: A comprehensive study international journal of computer vision. 2007. Citado na página 12.