UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Previsão de Redes Financeiras Utilizando Aprendizado de Máquina para Gerenciamento de Portfólio

Douglas Donizeti de Castilho Braz

Tese de Doutorado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

ICMC USP
SÃO CARLOS

Douglas Donizeti de Castilho Braz

Previsão de Redes Financeiras Utilizando Aprendizado de Máquina para Gerenciamento de Portfólio

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências – Ciências de Computação e Matemática Computacional. *EXEMPLAR DE DEFESA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. André Carlos Ponce de Leon Ferreira de Carvalho
Coorientador: Prof. Dr. João Manoel Portela da Gama

USP – São Carlos
Novembro de 2021

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

Donizeti de Castilho Braz, Douglas
D352p Previsão de Redes Financeiras Utilizando Aprendizado de Máquina para Gerenciamento de Portfólio / Douglas Donizeti de Castilho Braz; orientador André Carlos Ponce de Leon Ferreira de Carvalho; coorientador João Manoel Portela da Gama. -- São Carlos, 2021.
181 p.

Tese (Doutorado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2021.

1. Finanças Computacionais. 2. Aprendizado de Máquina Supervisionado. 3. Previsão de Formação de Links. 4. Redes de Ações. 5. Gerenciamento de Portfólio. I. Carlos Ponce de Leon Ferreira de Carvalho, André, orient. II. Manoel Portela da Gama, João , coorient. III. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Douglas Donizeti de Castilho Braz

Forecasting Financial Networks Using Machine Learning to Portfolio Management

Thesis submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP – in accordance with the requirements of the Computer and Mathematical Sciences Graduate Program, for the degree of Doctor in Science. *EXAMINATION BOARD PRESENTATION COPY*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. André Carlos Ponce de Leon Ferreira de Carvalho
Co-advisor: Prof. Dr. João Manoel Portela da Gama

USP – São Carlos
November 2021

À Lailiane, Maria Aparecida, Leandro e Dênis.

AGRADECIMENTOS

Primeiramente, agradeço a DEUS, à Nossa Senhora Aparecida, à Nossa Senhora de Fátima e à Nossa Senhora da Cabeça pela interseção e proteção durante toda a minha jornada.

Agradeço à minha esposa, Lailiane, por todo apoio, carinho, atenção, parceria, paciência e compreensão durante todos esses anos em que estive envolvido com esta pesquisa. Você foi parte fundamental para conclusão deste trabalho, me acompanhando nos momentos bons e ruins. Sem dúvida nenhuma, você tornou essa jornada menos turbulenta e me fez acreditar que era possível. Agradeço também aos meus pais, Leandro e Cida, pelo incentivo, pelo suporte e pelo esforço que fizeram para que eu chegasse até aqui. Agradeço ao meu irmão, Dênis, pela amizade e incetivo. Aproveito também para agradecer à toda a minha família, em especial aqueles que estiveram ao meu lado e me dando carinho e atenção.

Agradeço ao meu orientador, prof. Dr. André C. P. L. F. de Carvalho, pela imensa contribuição e excelente orientação que me proporcionou durante todo esse período. Seus conhecimentos, direcionamentos e, acima de tudo, sua empatia, foram fundamentais para que eu alcançasse esse tão importante título e também por ter me ensinado a ser uma pessoa melhor, principalmente pela sua preocupação durante os momentos em que fiquei doente e pela sua condução dos trabalhos durante a pandemia da COVID-19.

Agradeço ao meu coorientador, prof. Dr. João Gama, que me recebeu com muito carinho em Portugal e me acolheu eu seu laboratório. Agradeço pela atenção e pela sua excelente coorientação durante esse período.

Gostaria de agradecer à algumas pessoas que foram fundamentais e marcaram presença durante o meu percurso acadêmico, pessoal e profissional: Wellington, Humberto, Everton, Max, Nathália, Thiago, Kênia, Adriano, Karina, Ricardo, Carla, Rute, Guilhermes, Joel, Thársis, Soong, Leandro e prof. Gautam. Deixo aqui o meu sincero agradecimento.

Agradeço ao Conselho Nacional de Pesquisa (CNPq) pela concessão do auxílio financeiro que me proporcionou o Período Sanduíche na Universidade do Porto. Agradeço também ao Instituto Federal de Educação, Ciência e Tecnologia do Sul de Minas Gerais (IFSULDEMINAS) por todo o apoio durante todo esse período da pesquisa.

Finalmente, agradeço a todos os professores e funcionários do ICMC (USP) que contribuíram direta ou indiretamente para o desenvolvimento desta pesquisa.

“Viver é melhor que sonhar!”
(Belchior)

RESUMO

BRAZ, D. D. C. **Previsão de Redes Financeiras Utilizando Aprendizado de Máquina para Gerenciamento de Portfólio**. 2021. 178 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2021.

A dinâmica comportamental entre ativos do mercado financeiro pode ser analisada sob um perspectiva topológica de redes financeiras, conhecidas como redes ativos ou redes de ações. Essas redes permitem a modelagem e análise do mercado financeiro por meio de uma perspectiva topológica e apresentam uma forma robusta de analisar a interdependência do co-movimento de preços dos ativos por meio de grafos de relacionamentos, cujos vértices representam as ações e as arestas representam o relacionamento entre elas. De uma forma geral, o relacionamento pode ser modelado por meio da correlação entre as séries de preços dessas ações. Este estudo apresenta uma abordagem utilizando aprendizado de máquina supervisionado para resolver dois problemas: *(i)* previsão de formação de links em redes de ações; *(ii)* utilização da previsão de links como suporte ao gerenciamento de portfólios (carteiras). Para resolver o primeiro problema, desenvolvemos um modelo baseado em aprendizado de máquina supervisionado, que utiliza como entrada atributos extraídos das redes de ações, para prever a formação de links em redes futuras. Investigamos a previsão de links em redes modeladas através de três métodos de filtragem baseados em correlação: *Dynamic Asset Graphs* (DAG), *Dynamic Threshold Networks* (DTN) e *Dynamic Minimal Spanning Tree* (DMST). Foram propostos experimentos para avaliar o desempenho do método proposto, comparando-o com quinze algoritmos propostos na literatura, além de experimentos qualitativos para proporcionar uma interpretação dos resultados. Em relação ao segundo problema, propusemos uma abordagem para definição de constantes para otimização de portfólio, utilizando o método clássico conhecido como Análise Média-Variância (AMV), através da utilização de algoritmos de aprendizagem de máquina para previsão de links em redes de ações ponderadas. Além dos resultados da previsão de links ponderados, foram apresentados resultados financeiros relacionados ao gerenciamento de portfólio. Os resultados apresentados sugerem que o método proposto é capaz de melhorar o gerenciamento de risco na maioria dos índices de mercado estudados.

Palavras-chave: Finanças Computacionais, Aprendizado de Máquina Supervisionado, Previsão de Formação de Links, Redes de Ações, Gerenciamento de Portfólio.

ABSTRACT

BRAZ, D. D. C. **Forecasting Financial Networks Using Machine Learning to Portfolio Management**. 2021. 178 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2021.

The dynamics among assets in financial market can be analyzed from a topological perspective of financial networks, known as asset networks or stock networks. These networks allow the modeling and analysis of the financial market through a topological perspective and present a robust approach to analyze the interdependence of the co-movement of asset prices through relationship graphs, where vertices represent the assets and the edges represent the relationship between them. In general, the relationship can be modeled through the correlation between price series of these stocks. This study presents an approach using supervised machine learning to solve two problems: *(i)* link prediction in stock networks; *(ii)* use of link prediction results to support portfolio management. To solve the first problem, we developed a model based on supervised machine learning, which uses derived features extracted from stock networks to predict the link formation in future networks. We investigate link prediction in networks modeled through three correlation-based filtering methods: Dynamic Asset Graphs (DAG), Dynamic Threshold Networks (DTN) and Dynamic Minimal Spanning Tree (DMST). Experiments were proposed to evaluate the performance of the proposed method, comparing it with fifteen benchmark algorithms in the literature, in addition to qualitative experiments in order to understand of results. Regarding the second problem, we proposed an approach to provide input constants of portfolio optimization models, known as Mean-Variance Analysis (MVA), by using supervised machine learning algorithms to predict links in weighted stock networks. In addition to the results of the weighted link prediction, financial results related to portfolio management were presented. The results suggest that the proposed method is able to improve risk management in most of the market indices studied.

Keywords: Computational Finance, Supervised Machine Learning, Link Prediction, Stock Networks, Portfolio Management.

LISTA DE ILUSTRAÇÕES

LISTA DE ALGORITMOS

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

A existência de relações entre flutuações em mercados financeiros tem sido alvo de vários estudos (YANG *et al.*, 2014) (BONANNO *et al.*, 2003) (MANTEGNA; STANLEY, 1999) (BONANNO *et al.*, 2004). Muitos pesquisadores buscam formas de extrair informações de ambientes dinâmicos para analisá-los por meio da perspectiva topológica de redes complexas. Redes de Ações (*Stock Networks*), também conhecidas como Redes Baseadas em Correlação ou Redes Financeiras, se apresentam como uma forma robusta de analisar a interdependência do co-movimento de preços no mercado de ações. Elas permitem analisar a interação existente entre diferentes ativos ou índices financeiros por meio de grafos de relacionamento: os vértices representam os ativos e as arestas representam os relacionamentos entre eles (MANTEGNA, 1999). De uma forma geral, os relacionamentos são obtidos por meio de métricas baseadas em correlação e distância. Ativos com alta correlação entre si indicam a existência de uma distância pequena entre seus comportamentos. Assim, podemos dizer que eles possuem algum tipo de relacionamento, seja ele caracterizado pela forma como o preço das ações ou volume financeiro reagem a diferentes estímulos externos e especulações do mercado, ou até mesmo pela forma como as flutuações nos seus preços vão impactar outras empresas no mesmo setor, por exemplo.

Considerando a crescente disponibilidade de grande volume de dados proveniente dos mercados de ações, é possível notar um considerável aumento em pesquisas relacionadas esta área. Podemos verificar também um crescente aumento da utilização de soluções baseadas em aprendizado de máquina (*Machine Learning - ML*) em aplicações na área de finanças. Em geral, a utilização de algoritmos de aprendizagem de máquina no processo de tomada de decisão em bolsas de valores não é uma tarefa recente (TRIPPI; DESIENO, 1992). Um número crescente de aplicações foi criado utilizando modelos baseados em aprendizado de máquina para prever o comportamento da série temporal de preços (LONG; LU; CUI, 2019), previsão de volatilidade (LIU, 2019), análise de sentimento para investimento (PAGOLU *et al.*, 2016) e geração de regras automáticas de negociação (POTVIN; SORIANO; VALLÉE, 2004). Apesar da grande diversidade, o grande alvo de pesquisas que utilizam aprendizado de máquina é na

previsão de preços e tendências (HENRIQUE; SOBREIRO; KIMURA, 2019).

Nesta pesquisa, propomos um modelo para previsão de estrutura de mercado utilizando aprendizado de máquinas supervisionado. Para tal, a estrutura de mercado é modelada como uma rede dinâmica de ativos, quantificando sincronizadamente o co-movimento dos retornos dos preços das ações das empresas constituintes dos principais índices do mercado global. O problema de previsão da estrutura do mercado financeiro é formulado como um problema de previsão de links, onde estimamos a probabilidade de adicionar ou remover links em redes futuras. Desenvolvemos um modelo baseado em aprendizado de máquina, que utiliza como entrada atributos extraídos das redes financeiras, para prever a formação de links em redes de ações, construídas através de três métodos diferentes de filtragem de rede para estimar a estrutura do mercado: *Dynamic Asset Graph* (DAG), *Dynamic Minimal Spanning Tree* (DMST) e *Dynamic Threshold Networks* (DTN). Resultados experimentais mostraram que o modelo proposto pode prever a estrutura de mercado com alto desempenho preditivo, com até 40% de melhoria em relação a um benchmark invariável no tempo. Os atributos preditivos baseados na topologia da rede mostraram-se importantes em comparação com as atributos tradicionalmente usadas baseados na medida de correlação, considerando todos os mercados estudados, particularmente na previsão de longo prazo da estrutura do mercado de ações. Evidências são fornecidas para ações constituintes dos índices DAX30, EUROSTOXX50, FTSE100, HANGSENG50, NASDAQ100 e NIFTY50. As descobertas podem ser úteis para melhorar a seleção de portfólio e os métodos de gerenciamento de risco, que normalmente dependem de uma matriz de covariância para estimar o risco do portfólio.

Além disso, o gerenciamento de portfólio no mercado de ações tem sido investigado por muitos pesquisadores ao longo de décadas. Esta classe de investimento procura alocar o capital do investidor em um subconjunto de ativos de maneira a manter um bom controle de entre risco e retorno financeiro. Vários algoritmos têm sido propostos para gerenciamento de portfólio. De uma forma geral, grande parte dos algoritmos usam dados de retorno dos ativos e a correlação entre eles para recomendar um subconjunto cujo capital deve alocado. Redes dinâmicas de ações, cujos vértices representam ações e arestas representam a correlação entre elas, também podem ser utilizadas como entrada por esses algoritmos de gerenciamento de portfólio. Nesse sentido, esta pesquisa também tem como objetivo propor uma abordagem para definição de constantes da clássica análise de otimização de portfólio, conhecida como Análise Média-Variância (AMV), através da utilização de algoritmos de aprendizagem de máquina para previsão de links em redes de ações. Para avaliar o desempenho do método proposto, foram realizados experimentos com dados reais de seis índices de mercados europeus, asiáticos e norte-americano. Nesses experimentos, o método proposto foi comparado com a abordagens de Média-Variância clássicas. Os resultados experimentais mostraram que a utilização do resultado da previsão de links ponderados em redes de ações como entrada para os modelos de gerenciamento de portfólio produziram resultados satisfatórios em termos financeiros e de gestão de risco.

O restante desse capítulo está organizado da seguinte forma: a Seção 1.1 apresenta as motivações e justificativas para realização desta pesquisa; a Seção 1.2 apresenta as hipóteses, os objetivos e as contribuições desta pesquisa; por fim, a Seção 1.3 apresenta a estrutura relacionada à organização deste documento.

1.1 Motivação

O trabalho de Marti *et al.* (2021) apresenta uma ampla análise sobre 20 anos de trabalhos sobre o assunto de análise de correlações do mercado utilizando redes complexas. Em seu trabalho, o autor analisa os diferentes aspectos relacionados a este assuntos, que vão desde a análise da topologia de diferentes mercado através de redes baseadas em correlações, até na aplicação destes conceitos, de maneira prática, em diferentes segmentos do mercado. De fato, um ponto que é alvo de críticas é a falta de aplicações práticas destes tipo de análise. Além disso, muitos trabalhos que fazem a análise do passado, mas nenhum apontado pelo autor como sendo a extrapolação dessas redes no futuro. Alguns poucos trabalhos que fazem isso, mas nenhum deles segue a linha de pesquisa proposta neste trabalho - utilização de aprendizado de máquina para previsão dessas estruturas. Além disso, para tentar preencher o vazio existente entre as análises de redes de correlação e aplicações práticas, nós propusemos uma abordagem para sua utilização, envolvendo a utilização da previsão da estrutura do mercado como forma de melhorar o gerenciamento de risco em portfólios de investimento. De uma forma geral, este trabalho apresenta uma aplicação diferente da utilizada na literatura, quando realizamos a previsão de formação de redes de correlação utilizando aprendizado de máquina, assim como a aplicação direta dos resultados de previsão nesses dois segmentos de métodos para investimento. A complexidade deste problema possui vários desafios que abrem espaço para várias pesquisas na área, explorando a forte relação entre Ciência da Computação e a área de Finanças.

Existem poucos trabalhos na literatura que empregam abordagem semelhante. O trabalho de Musmeci, Aste e Matteo (2016) faz a previsão da volatilidade futura através de redes de correlação e métodos de filtragem de rede, porém não utiliza aprendizado de máquina e nem tampouco mostra aplicação direta de sua metodologia. O mesmo acontece com o trabalho de SOUZA e ASTE (2019), que apresenta um método de previsão de estruturas de mercado combinando informações de redes sociais e formação de fechamentos triádicos. Park, Chang e Song (2020) analisam a rede de causalidade de Granger do mercado monetário global e propõem um método de predição de link incorporando o eta quadrado das direções de causalidade entre nós como peso da aresta. Outro trabalho recente (JI-HWAN, 2020), que mais se aproxima desta pesquisa, trata-se de uma tese de doutorado que utiliza a previsão de redes ponderadas para gerenciar risco em portfólio através do método proposto em (POZZI; MATTEO; ASTE, 2008). Vale ressaltar que o trabalho não utiliza aprendizado de máquina e não explora o método de otimização de portfólio de Markowitz. Além disso, o trabalho é posterior ao primeiro trabalho publicado desta tese (CASTILHO *et al.*, 2019), que, pelo melhor de nosso conhecimento, foi o

primeiro trabalho a sugerir a união entre essas a previsão de links em redes de ações e otimização de portfólio.

Apesar de ser um assunto novo abordado na área de computação financeira, a previsão de formação de links é um assunto antigo, muito estudado na literatura e que vem sendo cada vez melhorado. Vale ressaltar que a contribuição deste trabalho não está relacionado com a previsão de links em si - apesar do método proposto ter sobressaído em todos os cenários em comparação com os outros algoritmos referentes à previsão de links e que são amplamente utilizados na literatura. Aqui, nós apenas utilizamos técnicas de previsão de link para conseguir estimar a rede de correlação futura e, a partir dessa previsão, avaliar a melhoria que ela pode resultar nos algoritmos de gerenciamento de portfólio. Além disso, o método aqui proposto pode ser utilizados para previsão de links em redes dinâmicas, além da aplicação em análises de redes complexas cujos relacionamentos entre vértices são mensurados através de similaridade entre séries temporais, como análises envolvendo redes energéticas.

1.2 Hipótese e Objetivos da Pesquisa

O principal objetivo deste trabalho é analisar e desenvolver um método para previsão de formação de links em redes de ações, assim como propor e investigar a utilização dessas previsões no problema de gerenciamento de portfólio. Para definirmos o escopo desta pesquisa, nós propusemos responder à seguinte pergunta:

- Como a previsão de formação de links em redes de ações pode auxiliar na melhoria de investimentos no mercado financeiro?

A resposta dessa pergunta está diretamente relacionada com o desenvolvimento deste trabalho. Conforme descrito anteriormente, ressaltamos a escassez de trabalhos na literatura que abordem a previsão de formação de links em redes de ações, apesar de sua importância. De um modo geral, as pesquisas nessa área utilizam somente dados do passado para criação dessas redes e para realizar suas análises. Assim, Ao responder a pergunta central deste trabalho, estamos apontando uma nova perspectiva para análises envolvendo redes de ações.

Dada a pergunta central desse trabalho, temos a formalização das duas hipóteses investigadas nesta pesquisa:

- Um método de aprendizagem de máquina, utilizando informações de redes de ações previamente conhecidas, é capaz de prever a formação de novos links em redes de ações futuras com desempenho preditivo maior que algoritmos comumente utilizados para previsão de links.
- A utilização da previsão de links em redes de ações, realizada através de algoritmos de aprendizagem de máquina, leva à melhoria de resultados financeiros de modelos de

otimização de portfólio quando comparados a modelos que não utilizam a previsão de links.

A formulação das duas hipóteses supracitadas serve de base para a definição dos objetivos específicos desta pesquisa. A seguir, listamos os desafios enfrentados pela pesquisa, assim como as contribuições gerais obtidas através do desenvolvimento deste trabalho de doutorado:

Revisão da literatura: através de uma ampla e constante revisão da literatura, pudemos identificar os trabalhos que analisaram a estrutura do mercado através de técnicas de redes complexas, as principais conclusões destes trabalhos e a formulação das hipóteses aqui apresentadas;

Proposição de um método para previsão de links em redes de ações: investigação, proposição e análise de um método baseado em aprendizagem de máquina, que utiliza como entrada atributos extraídos das redes de ações, para a previsão de formação de links em redes de ações. O método proposto foi comparado com algoritmos de previsão de links pertencentes ao estado-da-arte, através da utilização de dados de seis mercados diferentes. Um conjunto de análises empíricas foram propostas para entender os resultados experimentais e a eficiência preditiva do método proposto;

Utilização da previsão de links no gerenciamento de portfólios: investigação sobre formas de utilizar a previsão de formação de links em redes de ações como insumo para melhoria do gerenciamento de portfólios. Propusemos e desenvolvemos um método que utiliza os resultados da previsão de links como forma de definir as constantes utilizadas como entrada no clássico modelo de otimização de portfólio de Markowitz, criando uma conexão entre esses dois problemas.

1.3 Organização do Documento

O restante deste documento é organizado da seguinte forma: o Capítulo 2 apresenta o referencial teórico e os conceitos sobre redes complexas e redes de ações, aprendizado de máquina, previsão de formação de links e gerenciamento de portfólio, necessários para o entendimento do trabalho, assim como revisão bibliográfica acerca desses assuntos; o Capítulo 3 apresenta o método proposto para previsão de links em redes de ações, assim um conjunto de resultados experimentais relacionados ao problema abordado; o Capítulo 4 apresenta uma análise detalhada sobre a utilização da previsão de links em redes de ações no gerenciamento de carteiras de investimento; por fim, o Capítulo 5 apresenta as conclusões desse estudo e um conjunto de apontamentos sobre a continuação da pesquisa em futuros trabalhos.

CAPÍTULO

2

REFERENCIAL TEÓRICO

Este capítulo apresenta uma visão geral dos principais conceitos necessários para o entendimento deste trabalho, compreendendo definições sobre séries temporais e mercado de ações. Também são apresentados conceitos sobre análise de mercado utilizando de redes complexas, conhecidas como redes de ações ou redes de ativos. O restante deste capítulo apresenta os três principais problemas abordados neste trabalho, organizados da seguinte forma: a Seção 2.4 apresenta conceitos e técnicas relacionadas ao problema de previsão de formação de links em redes complexas e a Seção 2.5 apresenta uma visão geral sobre o problema de Gerenciamento de Portfólio.

2.1 Mercado de Ações e Séries Temporais

O mercado financeiro sempre despertou interesse das pessoas, principalmente por estar assimilado à ideia de ganhos financeiros reais e acumulo de capitais. Nas últimas décadas, com a crescente popularização de investimentos financeiros em países desenvolvidos e o estreitamento das fronteiras físicas entre países com o advento da internet, o mercado financeiro tem atraído cada vez atenção de investidores e pesquisadores de diversas áreas, como econometria, computação financeira, finança quantitativa, econofísica, dentre outras. Nesse trabalho, focamos nossos esforços parar o mercado de ações. O mercado de ações, como o próprio nome já diz, é o mercado onde são negociados ações de diversas empresas de capital social aberto. Ações representam a menor fração de uma empresa, que podem ser compradas ou vendidas em mercados comumente conhecidos como bolsa de valores. Ações são um tipo de ativo financeiro. Ativos representam qualquer coisa não material que possui valor financeiro agregado à sua condição contratual e que podem ser negociados. Exemplos de ativos compreendem ações, títulos públicos do governo, moedas de países e moedas digitais. Empresas cujas ações são negociadas em bolsa de valores possuem capital social aberto, permitindo que qualquer pessoa possa comprar ou vender suas ações. Quando alguém compra uma ação de uma empresa, ela se torna então acionista (sócio)

daquela empresa. Há uma série de exigências para que as empresas tenham suas ações negociadas na bolsa de valores, mas esse não é nosso foco aqui.

Grandes mercados de ações possuem índices que sintetizam, de uma forma geral, o comportamento das ações que estão sendo negociadas e servem de base para descrever o mercado geral. No Brasil, o mais índice mais conhecido é o *Ibovespa*, que compreende a lista das ações mais negociadas no mercado de ações brasileiro, a bolsa de valores B3 (antiga BM&FBOVESPA). Outros exemplos conhecidos são os índices *Dow Jones Industrial Average* (ou somente índice *Dow Jones*), que contém ações negociadas Bolsa de Valores de Nova York (NYSE), e o S&P500, composto por ações negociadas na NYSE e na NASDAQ (*National Association of Securities Dealers Automated Quotations*, tradução para o português - Associação Nacional de Corretores de Títulos de Cotações Automáticas).

Os preços das ações de uma empresa podem variar ao longo do tempo, trazendo lucro ou prejuízo para os acionistas. As razões por trás dessas variações são inúmeras e muitas vezes desconhecidas, embora sejam frequentemente estudadas por analistas e pesquisadores com objetivo de trazer explicações para tais fenômenos. A possibilidade de ganhos com ações torna atraente esse mercado. Apesar de não haver garantia de lucro, o que qualifica os investimentos no mercado de ações como sendo de retorno financeiro variável, cada mais investidores individuais têm se aventurado nesse mercado, principalmente em países desenvolvidos, onde uma parcela significativa da população investe em ações. O retorno de um investimento pode ser calculado através da diferença entre o preço dos ativos em momentos distintos. Seja $p_i(t)$ o preço de um ativo i no tempo t e $p_i(t+1)$ o seu preço no tempo futuro. O retorno desse ativo pode ser calculado da seguinte forma:

$$R_i(t+1) = \frac{p_i(t+1)}{p_i(t)} - 1 \tag{2.1}$$

Existem outras formas de cálculo do retorno, como o log-retorno, que será apresentado na seção seguinte. O preço das ações que costumamos ver em gráficos geralmente refletem o passado, ou seja, o preço no qual aquela ação foi negociada em algum momento. A cotação do preço de uma ação ao longo de um período de tempo formam uma série temporal de preços (série histórica de preços). Uma série temporal é um conjunto de observações de um evento realizadas em um determinado intervalo de tempo para modelar um fenômeno específico (HAMILTON, 1994). Os eventos observados acerca do fenômeno devem possuir alguma relação de temporalidade, que mostram características das informações modeladas pela série. O maior interesse em análises de séries temporais é extrair e estudar essa relação de dependência entre os eventos. Conseguimos observar a existência de séries temporais em várias áreas, como economia, meteorologia, geografia, sociologia, dentre outras. De uma forma geral, séries temporais costuma ser diferenciadas entre séries determinísticas, quando conseguimos descrever a sua evolução através de funções matemáticas, ou estocásticas, quando a série temporal também possui um componente aleatório na sua descrição (CRYER; KELLET, 1986). Séries temporais relacionadas

mercado de ação são exemplos de séries estocásticas, como a série histórica do preço de ações ou a série do volume financeiro negociado.

Além dessa classificação, séries temporais podem ser categorizadas como estacionárias ou não estacionárias. Resumidamente, séries temporais são estacionárias quando a distribuição dos seus valores possuem média e variância independente do tempo e são não estacionárias quando seus valores não convergem (MADSEN, 2007). Geralmente, dados financeiros possuem características não estacionárias, pois valores de média, variância e covariância podem mudar ao longo do tempo. Análises realizadas utilizando séries não estacionárias podem produzir conclusões erradas, uma vez que estes dados, por conta da fator aleatório, não podem ser modelados ou previstos. Porém, uma prática comum, com intuito de produzir análises mais conclusivas e que é adotada por muitos pesquisadores e analistas, é a conversão de séries de preços em séries de retornos. A série de retornos pode ser facilmente obtida através da diferenciação dos preços originais, como mostrado na Equação 2.1.

Vale ressaltar que essa conversão não garante que a série de retornos seja estacionária (apesar de muitas vezes isso ser assumido), mas sim que os preços das ações tenham uma normalização e que diferentes ativos, com valores e preços distintos, sejam comparados. Em geral, análises envolvendo séries temporais compreendem:

- **Previsão de Séries Temporais**: identificação e previsão de valores de novas observações para um determinado intervalo de tempo, visando a extrapolação de valores ainda não conhecidos. Exemplos: previsão da temperatura, previsão da quantidade de chuva, previsão do retorno financeiro de uma ação;
- **Classificação de Séries Temporais**: classificação do comportamento de uma série temporal baseado em padrões previamente conhecidos. Exemplo: classificação de sinal de voz;

A Figura 1 exemplifica séries temporais de duas ações distintas negociadas no mercado de ações brasileiro, colocadas de forma sobreposta. Apesar de serem ações de duas empresas diferentes, elas possuem comportamento bastante similar.

Outra representação frequentemente utilizada, principalmente para realização de análise de padrões ou análise técnica (gráfica), são gráficos que sintetizam as informações dos preços das ações através de *candlesticks* (ou simplesmente *candle* - vela, tradução para o português). Um *candle* sintetiza informações da variação do preço de um ativo dentro em um intervalo de tempo pré-determinado (SILVA *et al.*, 2014). As informações reprentadas em um *candle* são:

- Abertura - preço da primeira negociação no período;
- Fechamento - preço da última negociação no período;

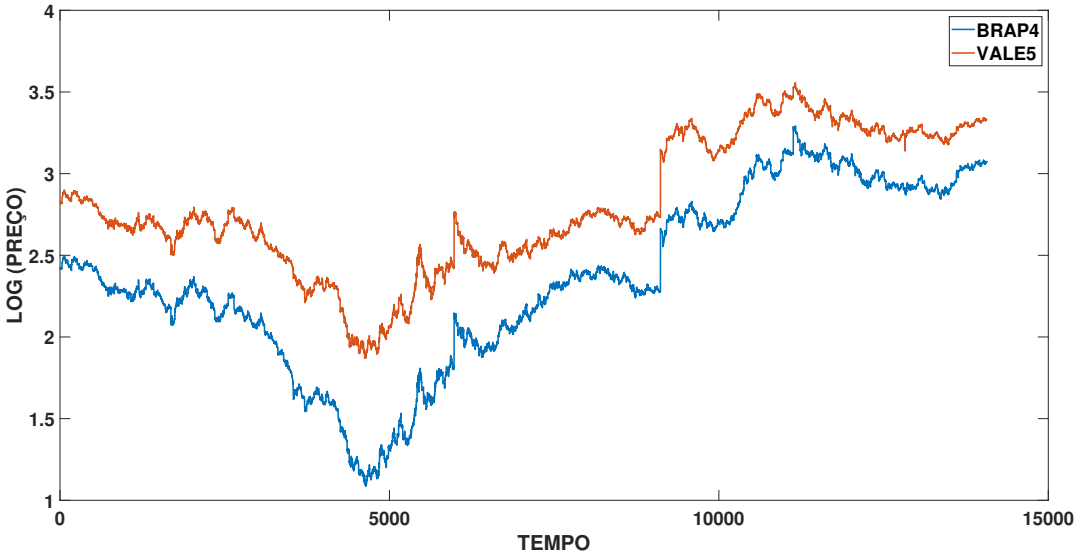


Figura 1 – **Série temporal de duas ações do mercado brasileiro.** Exemplo de duas séries do log do preço negociado de duas ações do mercado brasileiro. Os dados possuem intervalo de 15 minutos entre as medições dos preços e compreendem o período entre 1 de junho de 2015 e 18 de julho de 2017. Apesar de serem ações de duas empresas diferentes, elas possuem comportamento bastante similar. Ambas são empresas do setor de mineração.

- Máximo - maior preço negociado no período;

- Mínimo - menor preço negociado no período;

A Figura 2 apresenta um exemplo de gráfico de *candlesticks* da série temporal dos preços diários do Ibovespa, que é medido através de pontos. Esse gráfico foi extraído da plataforma *Investing*[1] e compreende dados entre entre 1 de agosto de 2021 a 25 de setembro de 2021.

O intervalo de tempo que um *candle* representa depende da análise que se deseja realizar, podendo compreender informações de segundos, minutos ou horas para análises realizadas dentro dos dias (*intraday*), também conhecidas como análises de curto prazo, ou informações de dias, semanas, meses e anos para análises entre dias, conhecidas como análises de médio ou longo prazo. Grande parte dos trabalhos na literatura analisam as séries financeiras formadas pelo preço de fechamento de *candles*. A granularidade da sintetização dos dados também está diretamente relacionada com a frequência de investimento que será abordada. Exemplo: investimentos feitos através de algoritmos que necessitam de uma alta taxa de atualização de informação utilizam dados com menores granularidades (alta frequência); algoritmos que fazem gestão de investimentos diários utilizam informações com maiores granularidades (baixa frequência).

[1] http://www.investing.com.br

Figura 2 – **Exemplo de um gráfico de *candlestick* da variação o Ibovespa.** Exemplo de representação gráfica da variação do preço diário do Índice Bovespa, medido através de pontos. Os dados variam entre 1 de agosto de 2021 a 25 de setembro de 2021. Este exemplo foi extraído da plataforma *Investing*.

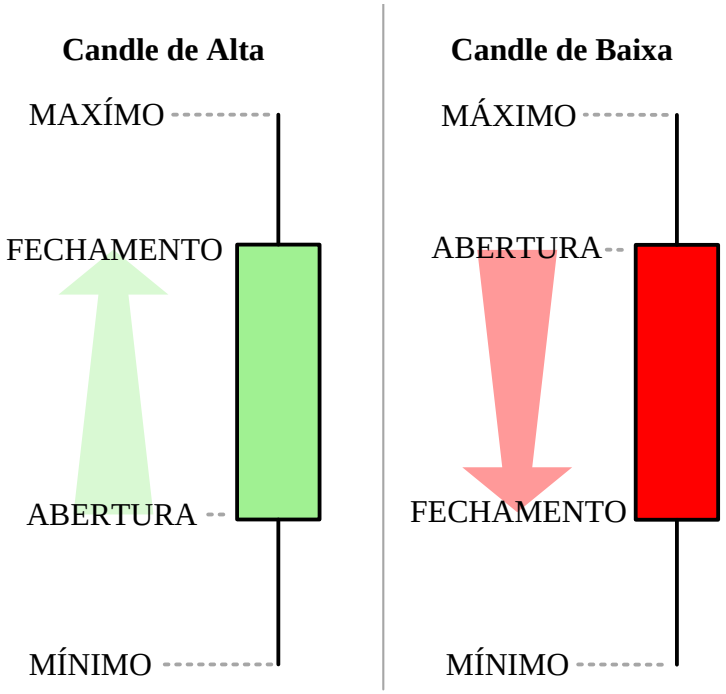


Figura 3 – **Representação de um *candlestick*.** Figura adaptada de Silva *et al.* (2015), mostrando as informações sintetizadas através da representação de um *candle*. As cores utilizadas nessa forma de representação podem variar de acordo com a ferramenta utilizada.

2.2 Redes Complexas e Redes de Ativos

Análises de redes complexas são frequentemente utilizadas em problemas onde é necessário o mapeamento de conexões entre um grande conjunto de variáveis. As redes complexas são como grafos que apresentam propriedades topológicas bastante particulares, não encontradas em

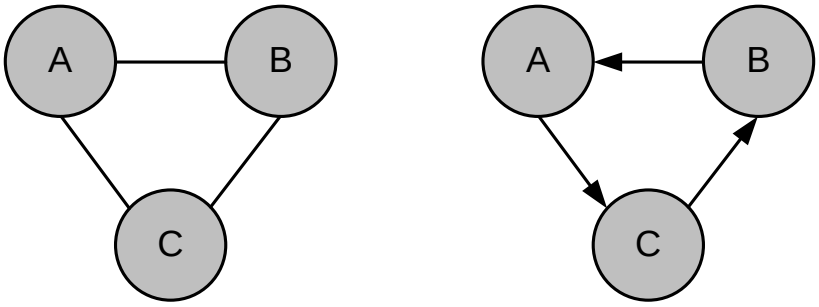


Figura 4 – **Exemplos de grafos.** Exemplo de grafo não direcionado (esquerda) e um grafo direcionado (direita).

grafos mais simples (METZ *et al.*, 2007) (ALBERT; BARABÁSI, 2002). Grafos são estruturas abstratas que representam objetos e as relações existente entre eles. Uma rede ou grafo é denotado como $G = (V, A)$, onde V representa o conjunto não-vazio de objetos ou vértices e A representa um conjunto de pares não ordenados de V, conhecidos como links ou arestas (BONDY; MURTY *et al.*, 1976).

A primeira utilização desta forma de representação é atribuída ao físico e matemático Leonhard Euler, no ano de 1736, no artigo em que é descrito a resolução do "Problema das Pontes de Königsberg". Este tipo de abstração é encontrada em várias áreas para diferentes aplicações que envolvem modelagem topológica de problemas, como circuitos elétricos, estradas, relações sociais, biologia, neurociência, dentre outras (NEWMAN, 2003). Arestas de um grafo podem ser direcionadas ou não-direcionadas. Considere como exemplo $G^D = (V, A)$ como sendo um grafo direcionado. O conjunto A de arestas de G^D possui direcionamento entre os vértices. Seja i e j vértices pertencentes a V. Um aresta $i \rightarrow j$ indica que existe conexão de i para j, mas não necessariamente existe um conexão de j para i. Considere agora um grafo não direcionado $G^N = (V', A')$. O conjunto A' de arestas de G^N não possui direcionamento entre os vértices. Isso significa que se existe uma aresta entre dois vértices i e j pertencentes ao conjunto V', $i \rightarrow j$ e $j \leftarrow i$. A Figura 4 apresenta dois exemplos de representações gráficas de grafos. Além disso, as arestas de um grafo podem assumir valores (pesos) que representem alguma informação dentro da estrutura topológica do grafo, como por exemplo, a distância entre duas cidades em um grafo que represente um mapa. Denominamos um grafo ponderado ou grafo valorado $G^W = (V, A_W)$ quando o conjunto de arestas A_W possuem valores que descrevem as relações entre os nós do conjunto V. Vale ressaltar que grafos ponderados podem ser tanto direcionados quanto não-direcionados.

De maneira simplificada, podemos dizer que as redes complexas são estruturas que não seguem um padrão regular (METZ *et al.*, 2007). Apesar de não haver consenso sobre a definição exata do que seja um padrão regular ou sobre o que são redes complexas, sabemos que elas possuem características que não existem em redes regulares e que, tais características, mostram como os grafos são formados ou como essa formação pode ser explorada em problemas

distintos (BARABÁSI; ALBERT, 1999).

Existem diferentes modelos de redes complexas, descritos por observações de padrões comuns no comportamento das redes. Dois dos principais modelos são:

Redes *Small World*: segundo Watts e Strogatz (1998), várias redes apresentam padrão de vértices com pequenos números de conexões, com grande parte das conexões entre vértices mais próximos, sendo que distância entre quaisquer par de vértices em redes muito grandes não passa de valores pequenos. Esse padrão implica em um alto coeficiente de agrupamento, com pequeno comprimento médio de caminhos entre dois vértices.

Redes Aleatórias: proposta por Erdos e Rényi (1960), é uma fusão de teoria de grafos com probabilidade, sendo o modelo mais simples que uma rede complexa pode assumir. Neste modelo, as arestas são adicionadas no grafo conectando cada par de vértices seguindo uma probabilidade p.

Existem muitos trabalhos na literatura que utilizam redes complexas para modelar a estrutura do mercado financeiro. A existência de interações entre ativos do mercado é assunto amplamente investigado (LEE; DJAUHARI, 2012). Algumas das abordagens comumente utilizadas incluem redes baseadas em correlação e métodos de filtragem de rede (MARTI *et al.*, 2021). Os trabalhos de Mantegna, Kadtke e Bulsara (1997), Mantegna (1999) e Mantegna e Stanley (1999) introduzem uma das formas mais utilizadas para realizar análise topológica de mercados financeiros por meio de grafos não direcionados, onde os vértices representam as ações e as arestas representam os relacionamentos entre elas, obtidos com base em medidas de distância entre séries temporais do log-retorno de ativos. Essa abordagem é utilizada em grande parte dos trabalhos da literatura envolvendo análises da estrutura e hierarquias dentro dos mercados financeiros através de redes, tais como Onnela *et al.* (2003b), Gopikrishnan *et al.* (2000), Bonanno *et al.* (2003), Bonanno *et al.* (2004) e Eom *et al.* (2009). De uma forma geral, as abordagens utilizam a correlação entre o log-retorno dos preços das ações para estabelecer uma relação de distância entre todos os pares de ações. Como base nas distâncias, é aplicado um método de filtragem de rede, responsável pela identificação dos pares de ações que possuam relacionamento que satisfaça algum critério de seleção. Em outras palavras, é utilizado algum filtro para selecionar os pares de ações que possuem um link na rede. Onnela, Kaski e Kertész (2004) apresenta uma análise sobre como selecionar as correlações relevantes de uma matriz de correlações para montar o grafo de relacionamento entre ações e compara os resultados com grafos gerados randomicamente.

Os métodos de filtragem de rede permitem uma análise momentânea ou temporal da estrutura de mercado, explorando *snapshots* de dados do mercado para modelar redes financeiras que representem a topologia e a sua estrutura geral. Usando uma abordagem de janela deslizante, podemos capturar *snapshots* em cada janela de tempo de comprimento arbitrário, permitindo

explorar a análise temporal da evolução do mercado (MUSMECI; ASTE; MATTEO, 2014), também chamadas de redes dinâmicas ou redes temporais. Cada método de filtragem de rede possui características diferentes. Em geral, os métodos definem exclusivamente arestas selecionadas na rede filtrada e nenhuma restrição topológica é imposta na estrutura de mercado (ONNELA *et al.*, 2003a). Além dos métodos de filtragem de rede, outra modificação que diversos autores propõem para a modelagem da topologia do mercado financeiro é na forma calculo da distância entre as ações. Yang *et al.* (2014) apresenta uma análise da construção de redes financeiras utilizando coeficiente de cointegração entre principais índices de mercados financeiros do mundo (GRANGER, 1981) (JOHANSEN; JUSELIUS, 1990). Em Tabak, Serra e Cajueiro (2010) são investigadas propriedades topológicas das redes financeiras do mercado brasileiro utilizando a construção de árvores através do conceito de ultrametricidade, que utiliza matriz de correlação da variação do preço de ativos de vários setores. Billio *et al.* (2012) propõem a utilização de causalidade de Granger como medida de distância entre a série de preços das ações, enquanto Wang *et al.* (2012) propõem a utilização de *Dynamic Time Warping* como métrica.

De uma forma geral, a abordagem para criação das redes financeiras é realizada através do coeficiente de correlação entre séries temporais sincronizadas (BONANNO *et al.*, 2004). Seja V um conjunto de ativos, tais como ações, moedas, commodities e fundos. Primeiramente, sincronizamos o intervalo de tempo a ser analisado e computamos a correlação entre todos os pares de ativos em V. Para quantificar o grau de similaridade entre a evolução síncrona do preços desses ativos, calculamos o coeficiente de correlação de Pearson da seguinte forma (MANTEGNA, 1999):

$$\rho_{ij} = \frac{< Y_i Y_j > - < Y_i >< Y_J >}{\sqrt{(< Y_i^2 > - < Y_i >^2)(< Y_j^2 > - < Y_j >^2)}} \tag{2.2}$$

onde i e j representam ativos do conjunto V. Em Mantegna, Kadtke e Bulsara (1997), o valor de $Y_i = ln(P_i(t)) - ln(P_i(t-1))$, sendo $P_i(t)$ o preço de fechamento do ativo i no tempo t. Y_i é dado pela diferença entre o logaritmo do preço de fechamento do tempo atual e o logaritmo do preço de fechamento do tempo anterior. Além disso, vale ressaltar que $< Y_i >$ representa a média de Y_i no período t. Alguns trabalhos na literatura, como Chi, Liu e Lau (2010), utilizam outras séries do mercado para construção de redes financeiras, tais como volume financeiro e preços de fechamento.

Seja uma matriz de correlações $M_c = N \times N$, que contém a correlação entre todos os pares de ativos do portfólio V, tal que $N = |V|$. Por definição, ρ_{ij} está contido em um intervalo de -1 a 1. Assim, $\rho_{ij} = 1$ indica correlação total entre os ativos i e j, $\rho_{ij} = -1$ indica anticorrelação total entre os ativos i e j e $\rho_{ij} = 0$ indica que os ativos i e j não possuem correlação. A matriz de correlações M_c é uma matriz simétrica e possui diagonal principal com $\rho_{ij} = 1$. Assim como na metodologia inicial de (MANTEGNA, 1999), para realizarmos análises utilizando variáveis

que representem distância, é necessário que sejam satisfeitos os três axiomas que definem uma medida Euclidiana (TABAK; SERRA; CAJUEIRO, 2010):

(i) $d(i,j) = 0$ se e somente se $i = j$;

(ii) $d(i,j) = d(j,i)$;

(iii) $d(i,j) \leq d(i,k) + d(k,j)$;

A função comumente utilizada para representação da distância entre dois ativos i e j é dada por:

$$d(i,j) = 1 - \rho_{ij}^2 \tag{2.3}$$

Através desta métrica podemos obter a matriz D, que corresponde a distância entre todos os pares de ativos contidos no conjunto V. O primeiro axioma é satisfeito porque $d(i,j) = 0)$ se e somente se a correlação (ou a anticorrelação) é completa ($|\rho| = 1$). O segundo é válido porque a matriz de correlações M_c é simétrica por definição e, consequentemente, a matriz D também. O terceiro axioma pode ser verificado numericamente através dos dados da matriz D.

Dada a matriz de distâncias D, utilizamos um método de filtragem de rede para selecionar os links que irão representar a rede financeira. Duas abordagens amplamente utilizada na literatura para criação de redes financeiras é através da criação de uma Árvore Geradora Mínima (AGM) com base na matriz D, através da execução do Algoritmo de Kruskal (KRUSKAL, 1956). A segunda abordagem para criação de redes financeiras requer a definição de um critério para considerar a existência ou inexistência de arestas: definimos um limiar α e, dada a correlação ρ_{ij}, existe aresta (i,j) se $\rho_{ij} >= \alpha$ (CHI; LIU; LAU, 2010). Vale ressaltar que esta segunda abordagem pode resultar em grafos desconexos. A Figura 5 ilustra uma AGM construída utilizando preços de fechamento diários das ações do mercado brasileiro negociadas no período entre 18 de julho de 2016 e 18 de julho de 2017.

2.3 Aprendizado de Máquina

Mudando o foco, apresentamos nessa seção conceitos relacionado com Aprendizado de Máquina (AM), outro objeto de investigação neste trabalho. Aprendizado de Máquina (do inglês, *Machine Learning*) é uma subárea da Inteligência Artificial (IA), campo de pesquisa centrado na interseção de áreas como Estatística e Ciência da Computação e pode ser visto como sendo uma área de estudos que objetiva a criação de sistemas de computação capazes de realizar tarefas de forma inteligente (CARVALHO *et al.*, 2011). Esse termo foi proposto inicialmente em 1956 pelo pesquisador John McCarthy. Apesar da simplificação do termo aqui proposta, esse é um conceito

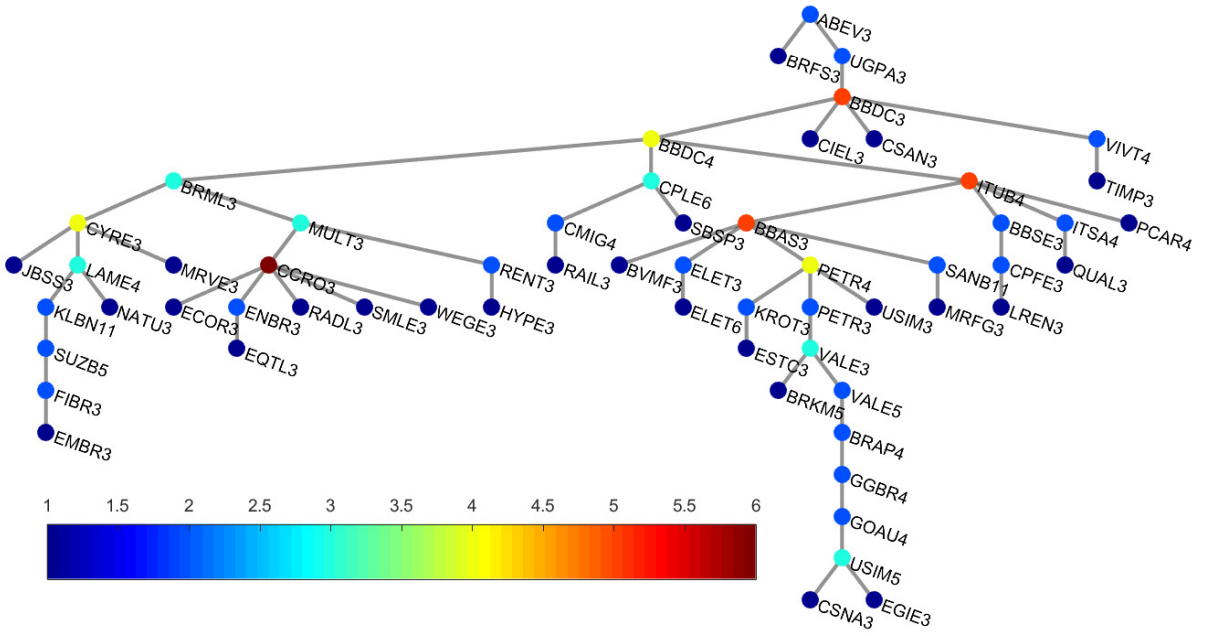


Figura 5 – **Árvore Geradora Mínima de ações do mercado brasileiro.** AGM gerada através da metodologia supracita, proposta por Mantegna (1999), para modelagem e análise da estrutura do mercado brasileiro. As cores dos nós da rede representa a quantidade de arestas que cada nó possui.

difícil de ser definido principalmente porque a área está em constante evolução e o entendimento sobre o que ele significa evolui ao longo do tempo. Outra razão para a dificuldade em definir IA é a natureza interdisciplinar do campo. Antropólogos, biólogos, cientistas da computação, linguistas, filósofos, psicólogos e neurocientistas contribuem para o campo da IA, e cada grupo traz sua própria perspectiva e terminologia (LUCKIN *et al.*, 2016). A discussão se aprofunda ainda mais e se torna filosófica quando tentamos definir o que significa ser "inteligente". Uma boa definição para "ser inteligente" é "ser racional". Assim, um sistema é inteligente e, ao mesmo tempo, racional, se "faz tudo certo" com os dados que tem (RUSSELL; NORVIG, 2004). A inteligência artificial sistematiza e automatiza tarefas intelectuais e, portanto, é potencialmente relevante para qualquer esfera da atividade intelectual humana (GOMES, 2010).

De uma forma geral, algoritmos de Aprendizado de Máquina podem ser vistos como sendo funções que buscam fazer o mapeamento entre um conjunto de características, utilizadas como entrada, para extrair algum tipo aprendizado. Os algoritmos de Aprendizado de Máquina são frequentemente divididos em dois grupos: Aprendizado Supervisionado e Aprendizado Não-Supervisionado. Na primeira classe, os algoritmos que possuem a propriedade de utilizar rótulos previamente conhecidos para induzir funções que relacionem o conjunto de características de entrada com o atributo alvo. Seja X o espaço de entrada e Y o espaço de saída. O objetivo do Aprendizado Supervisado é aprender um função $f : X \rightarrow Y$ (LUXBURG; SCHÖLKOPF, 2011). Na segunda classe, os algoritmos relacionados com Aprendizado Não-Supervisionado

lidam apenas com o espaço de entrada X, uma vez que o rótulos das instâncias não são conhecidos (exemplos não são rotulados). Este tipo de algoritmo é utilizado para clusterização ou agrupamento de dados e redução da dimensão do espaço de entrada.

Ainda sobre Aprendizado Supervisionado, os algoritmos são comumente divididos em duas sub-categorias, de acordo com a natureza do problema em que serão utilizados:

- **Algoritmos de Classificação:** utilizados em problemas onde o atributo alvo pode ser descrito por classes ou assumem valores discretos. O principal exemplo de problema dentro dessa sub-categoria é a classificação binária, onde existem duas classes possíveis para rotulagem das instâncias. Exemplos de algoritmos: Árvores de Decisão e *Support Vector Machines*.

- **Algoritmos de Regressão:** aplicados em problemas onde o atributo alvo é um valor numérico (contínuo). O objetivo dos algoritmos é induzir funções (lineares ou não-lineares) que aproximem ao máximo os atributos de entrada à variável de saída. Um exemplo de problema dentro dessa sub-categoria é a previsão do preço de ações ou previsão de temperatura. Exemplos de algoritmos: Regressão Linear e Regressão Logística.

Não é recente a utilização de técnicas de Aprendizado de Máquinas dentro do processo de tomada de decisões no mercado de capitais (TRIPPI; DESIENO, 1992). Um grande número de aplicações têm sido criadas utilizando Redes Neurais Artificiais (RNA) para a previsão do comportamento de séries temporais (ZHANG; PATUWO; HU, 1998). Neste contexto, também podem ser citados como exemplos Wang *et al.* (2011), Kamijo e Tanigawa (1990), Wang (2009), Kimoto *et al.* (1990), Trippi e Turban (1992), Yoon e Swales (1991) e Hsieh, Hsiao e Yeh (2011). Além da previsão de séries temporais, é possível encontrar na literatura trabalhos que possuem classificadores, na sua base principal, e que identificam instantes interessantes para adquirir ou abandonar uma posição no mercado de ações (TSAI *et al.*, 2011), (SHIHAVUDDIN *et al.*, 2010). Também existem trabalhos da literatura reportam a utilização de heurísticas dentro de seus processos de tomada de decisão (CHEN; HU; ZHOU, 2010; KABOUDAN, 2000).

Um mecanismo muito encontrado na literatura de aprendizado de máquina para o mercado de ações é a geração de regras para escolher o melhor momento para a execução de trades. Nesta linha, podem ser citados os trabalhos de Potvin, Soriano e Vallée (2004) e Allen e Karjalainen (1999), que utilizam técnicas de otimização baseadas em algoritmos evolutivos. Em geral, estes trabalhos criam e aperfeiçoam suas regras de investimento por meio da técnica de Programação Genética, originalmente descrita no trabalho de Koza (1992). Lendasse *et al.* (2000) e Hiemstra e Jones (1994) utilizam métodos estatísticos de regressão para extrapolar e prever o comportamento do preço dos ativos em bolsas de valores. Em Shihavuddin *et al.* (2010), os autores utilizam informações oriundas de notícias para auxiliar seu algoritmo inteligente no processo de tomada de decisão para investimentos no mercado de ações. Esta área tem

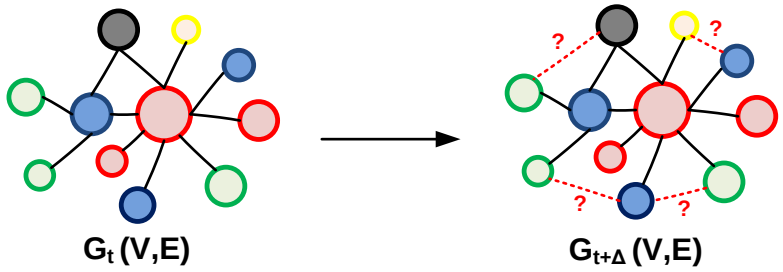


Figura 6 – **Previsão de formação de links**. Exemplificação da formação de links, considerando o grafo $G_t = (V,A)$ como sendo o grafo conhecido no tempo t e $G_{t+\Delta}(V,A)$ o grafo no futuro, cujos possíveis novos links são sugeridos em vermelho.

atraído crescente atenção nos últimos anos (MITRA; MITRA; BARTOLOMEO, 2008), (MITRA; MITRA, 2011), (MITRA *et al.*, 2015).

2.4 Previsão de Formação de Links

Outro assunto relacionado com este trabalho é a previsão da estrutura do mercado de ações. Como discutido anteriormente, a modelagem da estrutura do mercado pode ser feita através de métodos de filtragem de rede, cujo resultado proporciona uma análise topológica do mercado através de redes complexas. Sendo assim, abordamos o problema de previsão de estrutura do mercado como sendo um problema de previsão de links em redes financeiras. Este problema pode ser visto como a utilização de informações de redes previamente conhecidas para localizar links que possam surgir no futuro ou simplesmente para identificar relacionamentos que estão ocultos. Esta tarefa preditiva é investigada em muitos problemas reais, como em redes sociais (MARTÍNEZ; BERZAL; CUBERO, 2017; GROVER; LESKOVEC, 2016), redes biológicas (SHOJAIE, 2013), redes de co-autoria (LIBEN-NOWELL; KLEINBERG, 2007), redes tecnológicas (LÜ; ZHOU, 2010), redes criminais (BERLUSCONI *et al.*, 2016), dentre outros.

O termo "*link prediction*" (previsão de links, em tradução livre) é amplamente utilizado na literatura, sendo que 'link' é o mesmo que uma conexão ou aresta entre dois nós de uma rede. Seja $G_t = (V,A)$ um grafo ou rede identificado no tempo t, onde V é o conjunto de vértices e A é o conjunto de arestas ou pares de vértices que possuem alguma ligação entre si. A previsão de formação de links tem como objetivo identificar a existência de novas arestas (ou arestas ocultas) entre $i, j, \forall i, j \in V$ em $G_{t+\Delta}$, sendo $t+\Delta$ a representação de um tempo no futuro, como exemplificado na Figura 6.

Existem inúmeros métodos e algoritmos para resolver problemas de previsão de links. Alguns autores categorizam o tipo de previsão de link de acordo com o método utilizado (Lü; ZHOU, 2011; MARTÍNEZ; BERZAL; CUBERO, 2017). Os principais métodos encontrados na literatura abrangem as seguintes categorias: *(i)* métodos baseados em similaridade, que buscam

identificar nós similares para sugestão de novos relacionamentos, e *(ii)*) métodos baseados em classificadores, que incluem métodos que utilizam aprendizado de máquina, tratando o problema de previsão de links como uma tarefa de classificação binária. Além dessa categorização, outros autores sugerem a classificação do método pelo tipo de predição, de acordo com a natureza específica do problema, como predição de links dinâmicos ou predição de links ocultos, e também de acordo com a aplicação do método de predição de link, como recomendação em redes sociais ou complementação de redes (WANG *et al.*, 2015). Outro problema envolvendo a previsão de links é a previsão dos pesos da arestas de uma rede ("*weighted link prediction*" - previsão de links ponderados, em tradução livre) (LÜ; ZHOU, 2010).

Em um cenário real de mercado de ações, a relação entre ativos pode não apenas aparecer ao longo do tempo, mas também desaparecer. Uma sequência de interações dinâmicas ao longo do tempo introduz outra dimensão para o desafio de minerar e prever links da rede, chamada de previsão temporal de links (DUNLAVY; KOLDA; ACAR, 2011). Redes temporais são um tipo específico de redes dinâmicas nas quais o tempo pode ser organizado como um tensor de terceira ordem ou matriz multidimensional (WANG *et al.*, 2015). Uma deficiência comum desses métodos é manter os links entre os nós, mesmo quando seu relacionamento não existe mais, ou seja, prever o aparecimento e o desaparecimento de links em redes futuras. O comportamento evolutivo de aprendizagem das redes está diretamente relacionado ao problema de previsão de links, pois a adição ou remoção de novos links ou arestas ao longo do tempo leva à evolução da rede (DIVAKARAN; MOHAN, 2020).

A seguir, serão apresentados conceitos relacionados às duas categorias de métodos comumente utilizados na literatura para previsão de formação de links.

2.4.1 Métodos Baseados em Similaridade

Em métodos baseados em similaridade, o algoritmo atribui um peso $score(i, j)$ a pares de nós $\langle i, j \rangle$, com base no grafo de entrada $G(V, E)$, $\forall i, j \in V$. Em seguida, produz uma lista de classificação em ordem decrescente de $score(i, j)$ (LIBEN-NOWELL; KLEINBERG, 2007). Esses algoritmos podem ser vistos como uma medida de proximidade ou "similaridade" entre os nós i e j.

Os métodos nesta categoria podem ser divididos de acordo com o nível de informações acessadas para medir a similaridade entre pares de nós: *(i)*) métodos de similaridade local; *(ii)* métodos de similaridade semi-local e *(iii)* métodos de similaridade global. Embora os métodos de similaridade local e semi-local sejam mais simples de serem calculados, geralmente, os métodos de similaridade global podem fornecer previsões mais precisas (MUTLU; OGHAZ, 2019). A seguir, são apresentados alguns dos métodos comumente utilizados:

1. **Métodos de Similaridade Local**

***Common Neighbors* (CN):** É um método de predição de link simples e eficaz baseado em vizinhos comuns compartilhados por dois nós. Pares de nós com alto número de vizinhos comuns tendem a estabelecer um link (LIBEN-NOWELL; KLEINBERG, 2007);

***Preferential Attachment* (PA):** Este método define que novos links são formados entre nós com graus mais altos ao invés de nós com graus mais baixos (BARABÁSI; ALBERT, 1999);

***Jaccard Coefficient* (JC):** Este método é baseado no coeficiente de similaridade de Jaccard, levando em consideração o número de vizinhos comuns compartilhados por dois nós, mas normalizado pelo número total de vizinhos de ambos os nós (MUTLU; OGHAZ, 2019);

***Adamic-Adar* (AA):** Este método também se baseia em vizinhos comuns compartilhados por dois nós. Em vez de usar o número bruto de vizinhos comuns, como no método CN, ele é definido usando a soma do inverso do grau logarítmico de cada vizinho compartilhado (ADAMIC; ADAR, 2003) .

2. **Métodos de Similaridade Semi-Local**

***Local Path Index* (LP):** Semelhante ao CN, este método usa informações dos 2 e 3 vizinhos mais próximos, em vez de usar apenas informações dos vizinhos compartilhados por dois nós (ZHOU; LÜ; ZHANG, 2009) .

3. **Métodos de Similaridade Global**

***Random Walk with Restart* (RW):** Com base no *Random Walk*, é um caso especial de caminho, partindo de um determinado nó e alcançando aleatoriamente um vizinho selecionado. O reinício procura a probabilidade de um caminhante aleatório começando do nó i visitar o nó j e voltar ao nó de estado inicial i (BRIN; PAGE, 1998; MUTLU; OGHAZ, 2019).

Estes métodos são frequentemente utilizados em trabalhos relacionados com previsão de links, seja para resolução do problema em si ou para servir de *benchmark* em análises comparativas com outros métodos.

2.4.2 Métodos Baseados em Classificadores

Nessa categoria de métodos para previsão de links, a abordagem baseada em classificadores define o problema como sendo um problema de classificação binária. Aqui, um vetor de característica é extraído para cada par de nós e um rótulo $1/0$ deve ser atribuído com base na existência/inexistência desse link na rede. Qualquer método baseado em similaridade, descritos anteriormente, podem formar o vetor de características necessário para um método de

aprendizado supervisionado (Al Hasan *et al.*, 2006). Posteriormente, qualquer algoritmo de aprendizado de máquina supervisionado convencional pode ser aplicado para treinar um preditor de link. O aprendizado de máquina supervisionado é a busca por algoritmos que raciocinam a partir de instâncias fornecidas externamente para produzir hipóteses gerais, que então fazem previsões sobre instâncias futuras. Em outras palavras, o objetivo da aprendizagem supervisionada é construir um modelo conciso da distribuição de rótulos com base em características preditoras (KOTSIANTIS *et al.*, 2007). Além da previsão da existência/inexistência de links em redes futuras, nesta categoria também se enquadra a previsão de links ponderados (SÁ; PRUDÊNCIO, 2011).

Outra forma abordada na literatura para extração de características das redes são conhecidos como *embedding*. Algoritmos de *embedding* são métodos que mapeiam objetos entre diferentes espaços através de informações implícitas e estruturais. Em nosso contexto, um algoritmo de *embedding* tenta fazer esse mapeamento, preservando a estrutura do grafo no espaço vetorial, mantendo os nós vizinhos mais próximos uns dos outros (GOYAL; FERRARA, 2018). As características extraídas através de algoritmos de *embeddings* podem ser utilizados como preditores na formação de links, assim como na seção anterior, através da similaridade entre os vetores de atributos dos nós, ou como atributos preditores na aprendizagem supervisionada. *Embeddings* podem ser utilizados para aprender características de grafos, arestas e nós.

A seguir, são apresentados alguns dos algoritmos de *embedding* mais utilizados na literatura, categorizados de acordo com a abordagem utilizada para extrair as características de aprendizado dos nós da rede (MARA; LIJFFIJT; BIE, 2020).

1. **Métodos Baseados em Aprendizado Profundo** - como o próprio nome sugere, métodos nessa categoria utilizam a capacidade que *Deep Neural Networks* (Redes Neurais Profundas) possuem de capturar informações de relacionamentos não-lineares para extrair *embeddings* de grafos e nós do grafo.

   ***Structural Deep Network Embedding (SDNE)*:** Método baseado em um modelo semi-supervisionado profundo para extração de características de nós e previsão de links, que possui várias camadas de funções não lineares. Com intuito de preservar a estrutura e esparsidade da rede, são explorados a proximidade de primeira ordem e a proximidade de segunda ordem para caracterizar a estrutura da rede local e global, respectivamente. Ao serem otimizados em conjunto no modelo semi-supervisionado profundo, as representações aprendidas são preservadas pela estrutura local-global (WANG; CUI; ZHU, 2016).

2. **Métodos Baseados em *Random Walks*** - os métodos nessa categoria são baseados na execução de algoritmos de busca em grafos que percorrem os nós através de *random walks* (passeios aleatórios, tradução livre), onde os nós são "visitados" de forma aleatória, levando em consideração a posição nós no grafo e as arestas que interconectam os nós. Eles

se diferenciam principalmente pelas informações acessadas no grafo durante o *random walk* e a estrutura utilizada para realização da busca.

***Node2vec*:** Método utilizado para aprendizado de características de nós e para previsão de links. Desenvolvido para aprender continuamente representações das características de nós, seu objetivo é aprendizagem através de um mapeamento de nós para um espaço de baixa dimensão que maximiza a probabilidade de preservação de vizinhanças dos nós do grafo. É definido com base no conceito de vizinhança de nós, que utiliza método de busca através de *random walk*, explorando de forma eficiente diversas vizinhanças (GROVER; LESKOVEC, 2016). Este método pode ser visto como sendo uma generalização do algoritmo *DeepWalk*.

***DeepWalk*:** Similar ao *Node2vec*, usa informações de *random walks* locais e uniformes com probabilidades de transição fixas para medir semelhanças entre nós. Os *embeddings* são criados através de um modelo que é utilizado para capturar a estrutura semântica e sintática da linguagem humana, mas que analisa os *random walks* gerados ao invés de utilizar frases ou sentenças (PEROZZI; AL-RFOU; SKIENA, 2014)

***Struc2vec*:** Esse método extrai características de nós através de representações que capturam a identidade estrutural dos nós, ou seja, as funções ou papéis desempenhados pelos nós dentro da rede. A similaridade estrutural de pares de nós é acessada através de uma métrica hierárquica definida pela sequência de graus dos nós ordenados, utilizadas para criar um grafo multicamadas ponderado que represente o contexto. Após isso, *random walks* são aplicados ao grafo multicamadas ponderado para criação dos *embeddings* (RIBEIRO; SAVERESE; FIGUEIREDO, 2017).

3. **Métodos Baseados em Fatoração de Matrizes** - os métodos nessa categoria utilizam a forma de representação computacional dos grafos para realização das análises, acessando e extraindo informações oriundas da forma de representação. Em geral, grafos são expressos através de matrizes numéricas.

***Laplacian Eigenmaps*:** Primeiramente, esse método constrói uma representação ponderada da matriz original, alavancando as proximidades de primeira ordem no grafo. A matriz Laplaciana L é calculada usando a matriz ponderada e os *embeddings* são obtidos dos d autovetores correspondentes aos autovalores mais baixos da matriz Laplaciana L (MARA; LIJFFIJT; BIE, 2020; BELKIN; NIYOGI, 2003).

***Locally Linear Embedding*:** Esse método assume que os nós podem ser descritos pela combinação linear de seus vizinhos no espaço de *embedding* (GOYAL; FERRARA, 2018). Assim, o *embedding* de um nó pode ser derivado dos coeficientes lineares que melhor reconstroem o nó a partir dos *embeddings* de seus vizinhos (MARA; LIJFFIJT; BIE, 2020; ROWEIS; SAUL, 2000).

***Graph Factorization*:** Esse método utilizada a matriz de adjacência S como representação da similaridade entre os nós. A fatoração dessa matriz é realizada para minimizar a erro e encontrar a matriz Z contendo os *embeddings*, tal que $Z^T Z$ seja próximo de S (AHMED *et al.*, 2013).

***Higher-Order Proximity preserved Embedding (HOPE)*:** Método baseado na fatoração de matrizes, escalável para preservar proximidades de alta ordem de grafos de grande escala, capaz de capturar a transitividade assimétrica de grafos direcionados. Primeiramente, é utilizada uma formulação geral que cobre várias medições populares de proximidade de alta ordem e, em seguida, é aplicado um algoritmo de *embedding* escalável para aproximar as medições de proximidade de alta ordem com base em sua formulação geral (OU *et al.*, 2016).

Ainda, algumas variações dentro desta segunda categoria de métodos baseados em classificadores, Soares e Prudêncio (2012) propuseram a previsão de links usando previsões de séries temporais sobre métricas de similaridade.(HUANG; LIN, 2009) introduziram o problema de predição de link de série temporal, levando em consideração as evoluções temporais de ocorrências de link para prever probabilidades de ocorrência de link em um determinado momento, e mostraram que modelos de série temporal de ocorrências de link alcançam desempenho de previsão de link comparável com algoritmos de previsão de link baseados em similaridade comumente usados.

2.5 Gerenciamento de Portfólio

Para finalizar este capítulo, esta seção apresenta conceitos relacionados à gestão de portfólio (carteira). Quando um investidor deseja aplicar uma certa quantia de dinheiro em algum investimento financeiro, ele se depara com incerteza de qual seria a melhor forma de alocar essa riqueza. De uma forma resumida, todo investidor gostaria de obter o maior retorno financeiro assumindo o menor risco possível. Contudo, normalmente, quanto maior o risco de um investimento, maior o retorno esperado. Em 1952, Harry Markowitz (MARKOWITZ, 1952) introduziu Teoria Moderna de Portfólios (TMP) sob a ideia de que a diversificação do investimento poderia diminuir a risco de uma carteira de investimento. Nessa teoria, o objetivo de um investidor é distribuir um capital inicial dentre um conjunto de investimentos de forma a minimizar o risco e maximizar o retorno. A diversificação na alocação do capital é proposta através do método conhecido como Análise Média-Variância (AMV), que tem como objetivo a maximização do retorno e a minimização do risco de um investimento. A modelagem desse problema é realizada através da análise do risco de um ativo, representado através da variância do retorno, e do retorno esperado, quantificado através do seu retorno médio de um ativo. Essas duas métricas cunham o nome do método Média-Variância.

Matematicamente, a análise MV é descrita através de um modelo de programação quadrática bi-objetivo utilizado para encontrar a fronteira eficiente de possíveis portfólios ótimos. Seja N o conjunto de ativos disponíveis para investimento de tamanho n, cujo retorno esperado é dado por $\mu \in \mathbb{R}^n$. Seja matriz de correlação entre os retornos dos ativos $\rho_{ij}, \forall\, i, j \in N \mid \rho \in \mathbb{R}^{n \times n}$. A fração (peso) de investimento em cada ativo i no portfólio ótimo é dado por $w_i, \forall i \in N$. O modelo matemático é defino da seguinte forma:

$$
\begin{aligned}
\text{Minimize} \quad & v^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \rho_{ij} \sigma_i \sigma_j \\
\text{Sujeito a:} \quad & \sum_{i=1}^{n} w_i \mu_i = \mathscr{E}, \\
& \sum_{i=1}^{n} w_i = 1, \\
& w_i \geq 0,\ \forall\, i = 1, ..., n.
\end{aligned}
\tag{2.4}
$$

onde $\mathscr{E}$ é o retorno esperado do portfólio, σ_i é o desvio padrão do retorno do ativo i e v^2 é a variância máxima. Vale ressaltar que $\rho_{ij}\sigma_i\sigma_j$ é equivalente à covariância entre i e j. Note que a soma dos pesos é 1, representando que o investimento total não pode ser maior que o capital disponível e que investimentos negativos não são permitidos ($w_i \geq 0$). Um peso negativo nesse tipo de investimento representaria que operações do tipo *short*, conhecidas como "venda a descoberto", poderiam ser realizadas, ou seja, seria possível vender um ativo mesmo sem tê-lo. Essa operação é comum em operações dentro do mesmo dia (*daytrade*), mas aqui não serão abordadas. Para calcular a fronteira eficiente de Markowitz, podemos utilizar o modelo, através da minimização do risco considerando diversos níveis de retorno esperado $\mathscr{E}$.

CAPÍTULO

3

PREVISÃO DE LINKS EM REDES FINANCEIRAS

Neste capítulo, apresentamos um método para previsão de estrutura de mercado utilizando aprendizado de máquinas supervisionado. Para tal, a estrutura de mercado é modelada como uma rede dinâmica de ações, quantificando sincronizadamente o co-movimento dos retornos dos preços das ações das empresas constituintes dos principais índices do mercado global. Fornecemos evidências empíricas usando três métodos diferentes de filtragem de rede para estimar a estrutura do mercado: *Dynamic Asset Graph* (DAG), *Dynamic Minimal Spanning Tree* (DMST) e *Dynamic Threshold Networks* (DTN). O restante desse capítulo é organizado da seguinte forma: na Seção 3.1 apresenta uma introdução e revisão bibliográfica sobre a previsão de links em redes de ações; a Seção 3.2 apresenta a metodologia utilizada para resolver o problema proposto; a Seção 3.3 apresenta os resultados experimentais e a discussão desses resultados; por fim, a Seção 3.6 apresenta as considerações finais deste capítulo.

3.1 Introdução

As análises financeiras de múltiplos ativos, particularmente a seleção de portfólio e gerenciamento de risco, tradicionalmente dependem do uso de uma matriz de covariância representativa da estrutura de mercado, que é comumente assumida como invariante no tempo. Sob essa suposição, no entanto, a não estacionariedade (LIVAN; INOUE; SCALAS, 2012; MORALES; MATTEO; ASTE, 2013) e a memória de longo alcance (CONT, 2005) podem levar à conclusões enganosas e prejudicar a capacidade de explicar a dinâmica futura da estrutura do mercado.

Análises empíricas de redes em finanças têm sido usadas com sucesso para estudar a dinâmica da estrutura de mercado, particularmente para explicar a interconectividade do mercado a partir de dados de alta dimensão (MANTEGNA, 1999; TUMMINELLO; LILLO;

MANTEGNA, 2010; IORI; MANTEGNA, 2018; MARTI *et al.*, 2021). Sob essa abordagem, a estrutura do mercado é modelada como uma rede cujos vértices representam diferentes ativos financeiros e as arestas representam um ou vários tipos de relacionamentos relevantes entre esses ativos. Embora haja muitas pesquisas descrevendo a dinâmica do mercado, há pouca pesquisa sobre a inferência da estrutura do mercado.

Há uma vasta literatura que aplica redes financeiras à análise descritiva do mercado e dinâmica de portfólio, incluindo estabilidade de mercado (MORALES *et al.*, 2012), extração de informações (SONG; ASTE; MATTEO, 2008), alocação de ativos (POZZI; MATTEO; ASTE, 2013; HüTTNER; MAI; MINEO, 2018) e dependência de estrutura de mercado (MANTEGNA, 1999; TUMMINELLO; LILLO; MANTEGNA, 2010; SONG; MATTEO; ASTE, 2012; MUSMECI *et al.*, 2017). No entanto, existem poucas pesquisas sobre a aplicação de redes financeiras na previsão da estrutura futura do mercado. Pesquisas recentes sobre inferência de estrutura de mercado fazem uso de filtragem de informação em redes para produzir uma estimativa robusta da matriz de covariância inversa esparsa global, alcançando resultados computacionalmente eficientes (BARFUSS *et al.*, 2016). SOUZA e ASTE (2019) desenvolveram um método para prever a estrutura de mercado com base em um modelo que usa um princípio de formação de links por fechamento triádico em redes de ações. Spelta (2017) propôs um método para prever mudanças abruptas de mercado, inferindo a dinâmica futura dos preços das ações por meio da previsão de distâncias futuras entre eles, usando uma técnica de decomposição de tensores. Musmeci, Aste e Matteo (2016) propuseram uma nova ferramenta para prever a volatilidade futura do mercado utilizando redes de ações baseadas em correlação, meta-correlação e regressão logística, e mostraram a capacidade de previsão da ferramenta utilizando base de dados de dois mercados diferentes. Park, Chang e Song (2020) analisaram a rede de causalidade de Granger do mercado monetário global e propuseram um método de predição de links incorporando o eta quadrado das direções de causalidade de dois nós como sendo peso da aresta futura. Para construir a rede de causalidade, utilizaram a taxa de câmbio efetiva de 61 países e mostraram que a capacidade de predição de seu modelo supera a de outros métodos estáticos para previsão de links.

Neste trabalho, o problema de previsão da estrutura do mercado financeiro é formulado como um problema de previsão de links, onde estimamos a probabilidade de adicionar ou remover links em redes futuras. Para resolver esse problema, desenvolvemos um modelo baseado em aprendizado de máquina, que utiliza como entrada atributos extraídos das redes financeiras em nível de nós e arestas para a previsão de formação de links. Este é o primeiro trabalho da literatura que utiliza aprendizado de máquina para previsão de formação de links em redes financeiras

Para entender como o desempenho preditivo do modelo é influenciado pelos atributos topológicos extraídos das redes, nós fornecemos um conjunto de experimentos empíricos projetados para responder às seguintes questões:

1. Até que ponto as redes financeiras dinâmicas podem ajudar a prever a estrutura de correlação do mercado de ações?

2. Como os atributos relacionados à topologia da rede financeira se comportam em relação aos dados de correlação, tradicionalmente usados para prever a estrutura do mercado de ações?

3. Como a previsibilidade da estrutura do mercado varia em diferentes índices financeiros para os modelos propostos?

4. As propriedades da rede financeira podem explicar a previsibilidade da estrutura do mercado de ações?

As descobertas podem ser particularmente úteis para melhoria da seleção do portfólio e o gerenciamento de risco, que normalmente dependem de uma matriz de correlação ou covariância para estimar o risco do portfólio.

3.2 Materiais e Métodos

Nesta seção, descrevemos as principais etapas do método proposto para previsão da estrutura de mercado através de atributos de rede financeira utilizando aprendizado de máquina. A Figura 7 apresenta a metodologia.

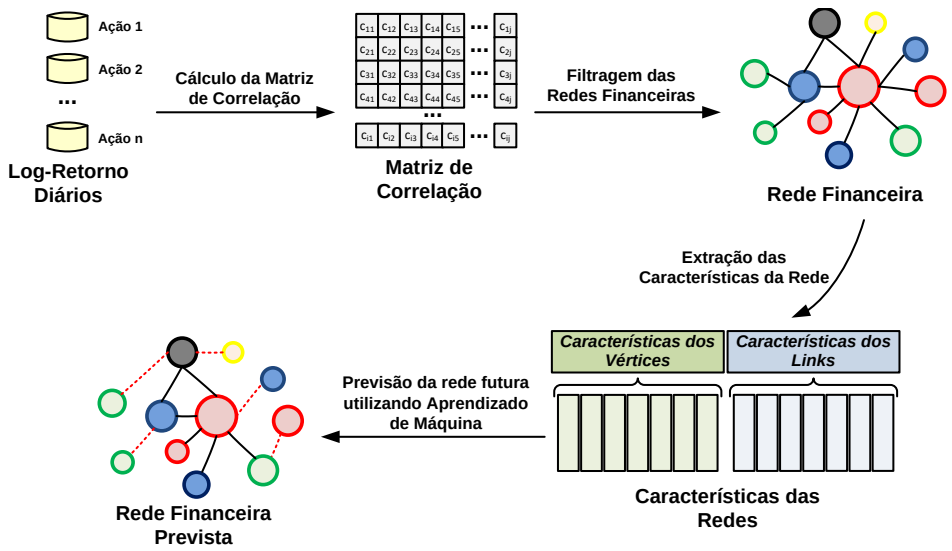


Figura 7 – **Principais etapas da metodologia usada para previsão da estrutura de mercado.** Com base nos preços de fechamento diários das ações constituintes de um índice financeiro, calculamos a matriz de correlação e criamos uma rede financeira através de três algoritmos de filtragem de rede diferentes. Dada a rede financeira, extraímos características derivadas da rede em nível de nós e em nível de links. Essas características são utilizadas como entrada para um algoritmo de aprendizado de máquina para previsão de redes financeiras futuras.

Inicialmente, calculamos a matriz de correlação entre todos pares de ações com base na série diária de preços de fechamento. Dada a matriz de correlação, a estrutura de mercado é modelada como uma rede financeira, calculando a matriz de distância de ativos e aplicando um método de filtragem de rede. Avaliamos três métodos de filtragem de rede diferentes para modelar a estrutura do mercado financeiro, descritos na Seção 3.2.1. Em seguida, extraímos um conjunto de características da rede, usados como atributos de entrada para o modelo de aprendizado de máquina, através da extração de características em nível de nó e em nível de link, conforme descrito na Seção 3.2.3. Finalmente, aplicamos um modelo de aprendizado de máquina, descrito na Seção 3.2.2, para prever redes financeiras usando as próprias informações de rede como entrada.

3.2.1 Redes Financeiras Dinâmicas

Existem diferentes métodos na literatura para modelar a estrutura do mercado financeiro através de redes. Alguns dos métodos mais comumente usados incluem redes baseadas em correlação e métodos de filtragem de rede (MARTI *et al.*, 2021), como descrito anteriormente na seção. Usando uma abordagem de janela deslizante, podemos tirar *snapshots* em cada janela de tempo de comprimento arbitrário, permitindo explorar a análise temporal da evolução do mercado (MUSMECI; ASTE; MATTEO, 2014), também chamadas de redes dinâmicas ou temporais. Alguns exemplos dos métodos mais comuns incluem a abordagem de *Minimal Spanning Tree* (Árvore Geradora Mínima) (MANTEGNA, 1999), *Planar Maximally filtered graph* (gráfico planar maximamente filtrado) (TUMMINELLO *et al.*, 2005), *Directed Bubble Hierarchical Tree* (árvore hierárquica de bolha direcionada) (SONG; MATTEO; ASTE, 2012), *asset graphs* (grafo de ativos) (ONNELA *et al.*, 2003a) e outras abordagens baseadas na filtragem de redes utilizando limiar (ONNELA; KASKI; KERTÉSZ, 2004).

Neste estudo, investigamos três métodos diferentes de filtragem de rede para criar a estrutura do mercado financeiro: *(i) Dynamic Asset Graph* (Grafo Dinâmico de Ativos); *(ii) Dynamic Threshold Networks* (Rede Dinâmicas de Limiares) e *(iii) Dynamic Minimal Spanning Tree* (Árvore Geradora Mínima Dinâmica). Exploramos esses três métodos devido à sua importância para a análise financeira, visto que são usados para capturar diferentes características da topologia do mercado. Esses métodos estimam uma matriz de distância de ativo por meio de métricas de co-movimento de retornos diários. Seja $P(t)$ o preço de fechamento de um ativo no dia t. Consideramos os log-retornos diários dos ativos $R(t) = \log P(t) - \log P(t-1)$ que são calculados no tempo t. Primeiro, calculamos uma matriz de distância que mede o co-movimento dos log-retornos diários (MANTEGNA, 1999), definida como:

$$D_{i,j}(t) = \sqrt{2(1-\rho_t(i,j))}, \tag{3.1}$$

onde $\rho_t(i,j)$ é o coeficiente de correlação de Pearson entre as séries temporais de log-retornos

dos ativos i e j no tempo t, $\forall i, j \in V$, sendo V o conjunto de ações. A matriz de distância é construída dividindo a série temporal de retornos $R(t)$ em janelas contínuas de tamanho L dias de negociação, com δT dias de negociação entre duas janelas consecutivas (intervalo de tempo). A escolha do tamanho da janela L e do intervalo de tempo δT é arbitrária, sendo que existe uma troca entre ter uma análise que é muito dinâmica ou muito suave (TUMMINELLO *et al.*, 2007). Quanto menor for o tamanho da janela L e quanto maiores forem os intervalos de tempo δT, mais dinâmicos serão os dados. Relatamos os resultados para $L \in \{126, 252, 504\}$ e $\delta T = 5$ dias de negociação. Uma rede financeira dinâmica é definida como uma rede temporal, descrita como:

$$W = \langle V, A_1, \ldots, A_T : A_t \subseteq V \times V, \forall\, t \in \{1, \ldots, T\} \rangle, \tag{3.2}$$

onde os vértices $i \in V$ correspondem às ações de interesse. Para cada par $\langle i, j \rangle$ na janela de tempo t, $\forall i, j \in V \mid i \neq j$, existe uma aresta correspondente $A_{i,j}(t) \in A_t$ e cada aresta tem um peso $w_{i,j}(t) = D_{i,j}(t)$. Considerando a matriz de distância $D_{i,j}(t)$ na Equação 3.1 definida anteriormente, podemos aplicar um método de filtragem de rede para criar redes dinâmicas. Os três métodos avaliados neste trabalho são descritos nas próximas seções.

3.2.1.1 Dynamic Asset Graph (DAG)

Dynamic Asset Graph (Grafo Dinâmico de Ativos) é um tipo de rede financeira que pode ser modelada através do ranqueamento das as arestas em ordem crescente de pesos $w_1(t), w_2(t), \ldots, w_{N(N-1)/2}(t)$, onde N é o número de ações em V (ONNELA *et al.*, 2003a). O grafo resultante é obtido selecionando as arestas com as conexões mais fortes. O número de arestas é arbitrário. Nesta análise, selecionamos arestas com pesos no quartil superior, ou seja, $w_1(t), w_2(t), ..., w_{N(N-1)/8}(t)$, conforme proposto por SOUZA e ASTE (2019). A ideia principal deste método é identificar as menores distâncias no mercado de ações.

3.2.1.2 Dynamic Threshold Networks (DTN)

Considerando a matriz de distância $D(t)$ definida na Equação 3.1, criamos uma matriz de adjacência filtrada A para construir a rede financeira usando as seguintes regras (YANG; YANG, 2008; ONNELA; KASKI; KERTÉSZ, 2004; CHI; LIU; LAU, 2010) :

$$A_{i,j}(t) = \begin{cases} 1, & \left|D_{i,j}(t)\right| \geq r_c \\ 0, & \left|D_{i,j}(t)\right| < r_c \end{cases} \tag{3.3}$$

onde os ativos $i, j \in V$ e $\forall \langle i, j \rangle_t \in A_t$. O valor crítico r_c converte a matriz D em um grafo não direcionada, onde $A_{ij}(t) = 1$ e $A_{ij}(t) = 0$ representa a existência e ausência de arestas entre i e j na janela de tempo t, respectivamente. Fixamos o valor de r_c em $0,65$ porque, para $r_c \leq 0,65$ as características da rede estão submersas em grandes flutuações (YANG; YANG, 2008). É importante observar que o método DTN pode produzir gráficos desconexos e que o número de

arestas é dinâmico. Em geral, o objetivo principal deste método é identificar pares de ativos que estão altamente correlacionados e acima do limite r_c. A principal diferença entre este método e o método anterior (DAG) é que, em DAG, embora sejam selecionados os pares de ações que estejam entre os 25% mais similares, pares com valor de correlação inferior a r_c podem ser adicionados à rede.

3.2.1.3 Dynamic Minimal Spanning Tree (DMST)

Criamos uma *Dynamic Minimal Spanning Tree* (Árvore Geradora Mínima Dinâmica) com base na menor distância necessária para interconectar todas as ações da matriz $D(t)$, definida anteriormente na Equação 3.1 (MANTEGNA, 1999). Usamos o algoritmo de Kruskal para identificar o MST no gráfico totalmente conectado D no tempo t (KRUSKAL, 1956). O número de arestas é fixo e pode ser calculado como $N-1$, onde N é o número de ações em V. Este método fornece a menor distância para interconectar o mercado, produzindo a estrutura de mercado mínima necessária para conectar todos as ativos.

3.2.2 Abordagem Utilizando Aprendizado de Máquina

Nesta seção, descrevemos a abordagem proposta utilizando aprendizado de máquina para previsão da estrutura do mercado, considerando um determinado índice de mercado. Neste estudo, abordamos a previsão da estrutura de mercado como sendo um problema de previsão de formação de links em redes dinâmicas. Dados *snapshots* de redes financeiras até o momento t, queremos prever as arestas que estarão presentes na rede em um determinado momento futuro t'. Escolhemos três tempos $t_0 < t < t'$ e criamos um algoritmo que acessa $W[t_0,t] = \langle V, A_{t_0}, \ldots, A_t \rangle$ para estimar a probabilidade das arestas estarem presentes em $W[t']$, onde $t' = t+h$ e $h = \{1,2,\ldots,20\}$ semanas de negociação.

Métodos baseados em similaridade e métodos baseados em classificadores são duas das abordagens mais comuns para previsão de link, como mencionado na Seção 2.4. Neste trabalho, aplicamos um método baseado em classificadores para prever a estrutura do mercado financeiro. Nossa abordagem usa atributos de redes de ações como entrada para um modelo de aprendizado de máquina supervisionado, a fim de criar um método de predição de links, conforme apresentado na Figura 8.

A Figura 8 apresenta o processo utilizado para desenvolver o banco de dados de utilizado pelo modelo de aprendizado de máquina. Assumindo i e j como dois vértices arbitrários em $G(V,A)$ e t como o tempo atual, uma instância do conjunto de dados acessado pelo algoritmo de aprendizado de máquina tem os seguintes atributos preditivos: (a) atributos em nível de nó relacionados ao vértice i ; (b) atributos em nível de nó relacionados ao vértice j ; (c) atributos em nível de link relacionados ao par $\langle i, j \rangle$. Conforme descrito anteriormente, o objetivo do modelo de aprendizado de máquina supervisionado é prever a existência de links em uma rede $G(t+h)$, considerando $h = 1,2,\ pontos, 20$ semanas de negociação. A Figura 8 apresenta uma ilustração

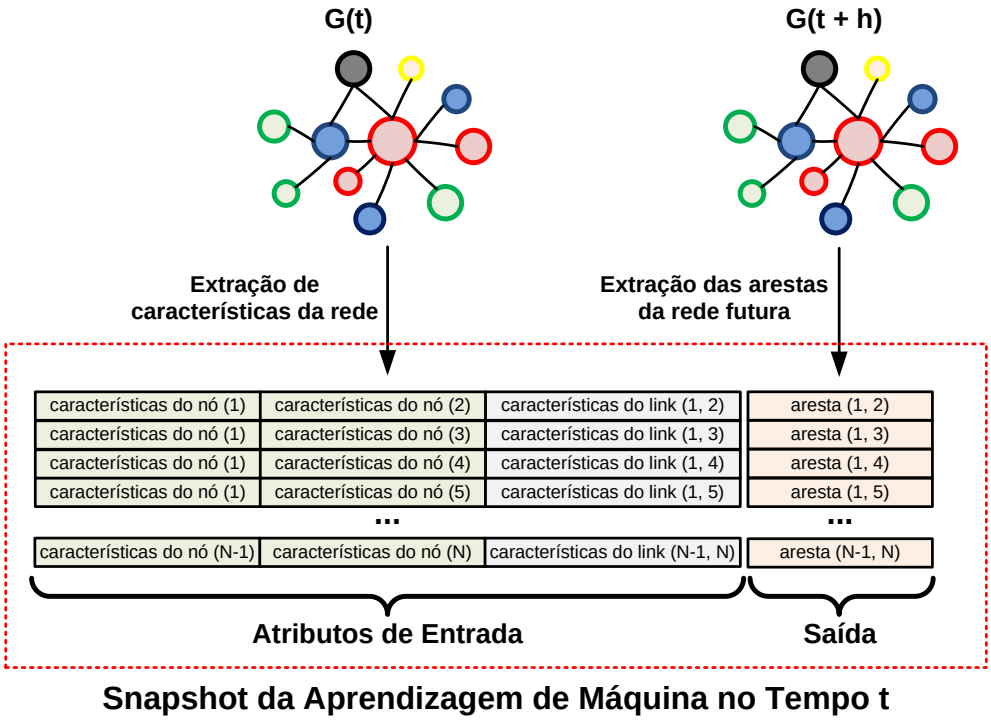


Figura 8 – **Criação das instâncias utilizadas no aprendizado de máquina.** Calculamos atributos para cada nó variando de 1 a N, onde N é o número de ações do grafo $G(V,A)$. Aplicamos uma concatenação entre pares de atributos de nó e de link como variáveis de entrada para a previsão do link, enquanto as arestas da rede no tempo $t+h$ são usadas como a atributo de saída, sendo h o número de semanas de negociação.

de como construímos instâncias para o modelo de aprendizado de máquinas, exemplificado como um *snapshot* no tempo t.

Dividimos conjunto de dados em dois subconjuntos de treinamento e teste levando em consideração a sequência temporal dos dados. O conjunto de treinamento inclui dados produzidos no período de 1 de março de 2005 a 30 maio 2007 e o conjunto de teste tem dados de 30 maio 2007 a 18 de dezembro de 2019. A Figura 9 apresenta uma ilustração que explica como criamos os conjuntos de treinamento e teste. Os modelos de aprendizado de máquina foram treinados e testados usando uma abordagem de janela deslizante. Considerando L como o tamanho da série temporal de log-retornos, t como o tempo atual e $t-k<t<t+h$, criamos o conjunto de treinamento utilizando atributos derivados da rede $G(t-k)$, sendo $k=1,2,\ldots,30$. O conjunto de teste contém dados da rede atual $G(t)$, na qual $G(t+h)$ é a rede alvo, considerando $h=1,2,\ldots,20$. Depois de treinar o modelo de aprendizado de máquina e testá-lo, avançamos na janela deslizante levando em consideração o intervalo de tempo $\delta T=5$ dias de negociação (1 semana de negociação) entre duas execuções consecutivas.

Para avaliar a taxa de informação que um modelo de aprendizado de máquina pode extrair do conjunto de atributos proposto neste trabalho, aplicamos o algoritmo XGboost (CHEN; GUESTRIN, 2016). XGboost é um modelo de aprendizado de máquina rápido, altamente

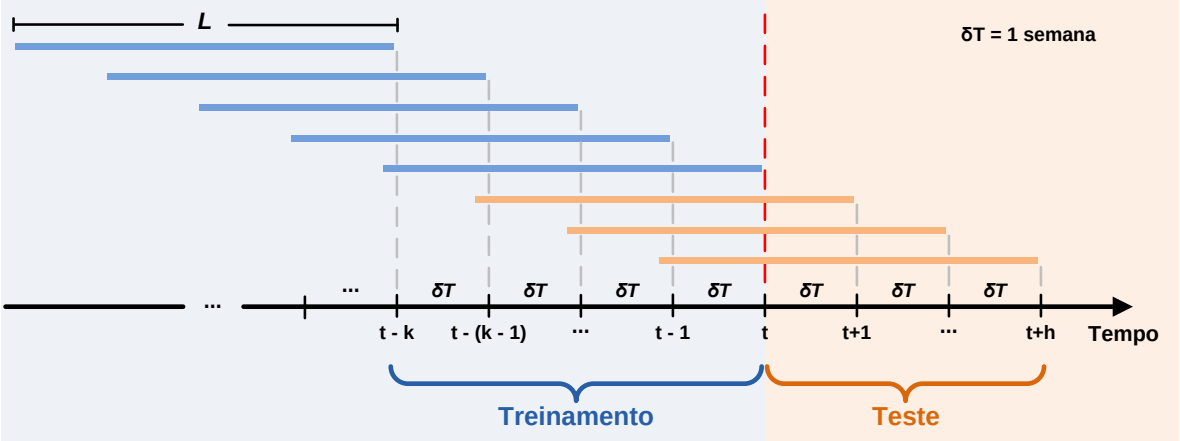


Figura 9 – **Conjuntos de treinamento e teste usados para induzir o modelo de aprendizado de máquina.** Os modelos de aprendizado de máquina foram treinados e testados usando uma abordagem de janela deslizante. Considerando L como o tamanho da série temporal de log-retornos e t como o tempo atual, criamos o conjunto de treinamento usando dados de $t-k$ a $t-1$ e o conjunto de teste usando dados de t. O alvo da aprendizagem supervisionada é a rede $G(t+h)$, onde h é o número de semanas de negociação no futuro. Depois de treinar e testar o modelo de aprendizado de máquina, o intervalo de tempo δT é usado para mover a janela de tempo para frente, a fim de reiniciar o processo e treinar novamente o modelo de aprendizado de máquina. O conjunto de treinamento inclui dados de 1 de março de 2005 a 30 maio 2007 e o conjunto de testes tem dados de 30 maio 2007 a 18 de dezembro de 2019.

eficaz, interpretável e amplamente utilizado em muitas aplicações envolvendo mineração e ciência de dados. Este modelo tem como objetivo a combinação de vários classificadores "fracos", implementados através de algoritmos de Árvores de Decisão, e utiliza o conceito de *Gradient Boosting* (FRIEDMAN, 2001) para criação e evolução de forma sequencial desses novos classificadores. Mais informações sobre a configuração experimental do são descritas no Apêndice A.

A seção seguinte apresenta as características derivadas das redes que foram utilizados para induzir os modelos de aprendizado de máquina para previsão da estrutura de mercado.

3.2.3 Características da Rede

Conforme mencionado anteriormente, nós propusemos uma abordagem para previsão da estrutura de mercado baseada em aprendizado de máquina supervisionado. A fim de fornecer informações para treinar o modelo supervisionado, extraímos um conjunto de características da rede em nível de nó e de link. Esses atributos são usados como entrada para induzir o modelo de aprendizado de máquina. Resumimos as características de rede da seguinte forma:

- **Atributos dos Nós:** avalia a posição de um nó dentro da estrutura geral de um dado grafo $G(V,A)$ (OLIVEIRA; GAMA, 2012). A Tabela 1 apresenta o conjunto de características em nível de nó relacionados ao vértice/ação $i \in V$ usado como entrada para o modelo de

aprendizado de máquina.

- **Atributos dos Links:** examina o conteúdo e os padrões de relacionamentos em um dado grafo $G(V,A)$ e mede as implicações desses relacionamentos (OLIVEIRA; GAMA, 2012). A Tabela 2 apresenta o conjunto de características em nível de link relacionado ao par $\langle i, j\rangle \in A$ usado como entrada para o modelo de aprendizado de máquina.

Em geral, as pesquisas em finanças, especialmente em gestão de portfólio e gestão de risco, utilizam informações de correlação de co-movimento em suas análises. Neste estudo, estamos interessados em analisar como as informações estruturais e topológicas do mercado nos ajudam a prever a própria estrutura do mercado. Por esse motivo, separamos o conjunto de características das redes em dois subconjuntos distintos. Rotulamos os dois subconjuntos de acordo com sua fonte de informação: *(i)* características de correlação de pares, que são atributos baseados em informações de correlação e não derivados de qualquer outra informação de rede, e *(ii)* características topológicas, que são atributos derivados da topologia da rede. Embora os recursos de correlação entre pares sejam tradicionalmente empregados na análise financeira, a importância dos atributos topológicos para prever a estrutura do mercado é uma questão de pesquisa investigada neste trabalho. Assim, podemos comparar seu ganho de informação na previsão da estrutura de mercado. Na Tabela 1, todos os atributos são características topológicas. Na Tabela 2, as características de correlação de pares são marcados com (*).

3.2.4 Avaliação do Método

Calculamos a *Area Under de ROC Curve* (AUC) para avaliar o desempenho preditivo dos métodos de predição de link. Essa métrica é amplamente aplicada em classificação binária e problemas desbalanceados e varia de $0,5$ e 1, onde $0,5$ representa um algoritmo ingênuo aleatório e 1 representa o resultado mais alto. A medida AUC fornece uma métrica resumida para o desempenho geral do algoritmo com diferentes tamanhos de conjuntos de predição, enquanto uma análise detalhada da forma da curva ROC revela o desempenho preditivo do algoritmo em cada tamanho de conjunto de predição (HUANG; LIN, 2009).

Para verificar o desempenho do método proposto, nós o comparamos com os seguintes métodos baseados em similaridade comumente usados na literatura para predição de links, separados em três categorias como segue (MUTLU; OGHAZ, 2019):

Common Neighbors (CN):

$$CN(i,j) = |N_i \cap N_j|, \tag{3.4}$$

onde N_i e N_j representam o conjunto de nós adjacentes aos vértices i e j, respectivamente;

Preferential Attachment (PA):

$$PA(i,j) = |i| * |j|, \tag{3.5}$$

Tabela 1 – **Atributos dos Nós.** As características descritas nesta tabela foram calculadas para o nó i, $\forall i \in V$ para um dado grafo $G(V,A)$. Considere N_i como o conjunto de vértices adjacentes (vizinhança) do nó i. Este conjunto de atributos contém apenas características topológicas.

Nome	Definição
Grau do Nó	$deg(i) = \|i\|$
Grau Ponderado do Nó	$deg_w(i) = \sum_{j \in N_i} w_{<i,j>}$, onde $w_{<i,j>}$ é o peso da aresta $e(i,j)$
Média do Grau dos Vizinhos	$avg(i) = \frac{\sum_{j \in N_i} \|j\|}{\|i\|}$
Propensão de i Aumentar seu Grau	$\gamma(i) = \frac{\|i\|}{deg_w(i)}$
Betweenness do Nó	$b(v) = \sum_{i,j \in V \setminus v} \frac{\sigma_{ij}(v)}{\sigma_{ij}}$, onde $\sigma_{ij}(v)$ é o número de caminhos mínimos entre i e j que possam pelo vértice v e σ_{ij} é o número de caminhos mínimos de i para j, $\forall i, j \in V$
Closeness do Nó	$nc(i) = \frac{n-1}{\sum_{j \in V \setminus i} d(i,j)}$, onde $d(i,j)$ representa a distância entre i e j e n é o número de nós no grafo
Autovetor do Nó	$ne(i) = x_i \frac{1}{\lambda} \sum_{j=1}^{n} d_{ij} x_j$, onde d_{ij} representa uma entrada na matriz de adjacência D (0 ou 1), λ denota o maior autovalor, x_i e x_j representam a centralidade dos nós i e j, respectivamente
Coeficiente de Clusterização do Nó	$cc(i) = \frac{2\left\|e_{jk}\right\|}{\|i\| * (\|i\| - 1)} : j,k \in N_i, e_{jk} \in A$

onde $|i|$ e $|j|$ representam o grau dos vértices i e j, respectivamente;

Jaccard Coefficient(JC):

$$JC(i,j) = \frac{|N_i \cap N_j|}{|N_i \cup N_j|}, \tag{3.6}$$

onde N_i e N_j representam o conjunto de nós adjacentes aos vértices i e j, respectivamente;

Tabela 2 – **Atributos dos Links:** As características descritas nessa tabela foram calculadas entre os nós i e j, $\forall\ text(i,j)\ inA$ para um dado grafo $G(V,A)$. As características de correlação de pares estão marcados com $(^*)$, enquanto o restante são categorizadas como caraterísticas topológicas. Considere N_i e N_j como o conjunto de vértices adjacentes do nó i e j, respectivamente.

Nome	Definição
Existência do Link em $G(t)$ $(^*)$	$E(i,j) = \begin{cases} 1, & \text{existe link,} \\ 0, & \text{não existe link.} \end{cases}$
Valor da Correlação $(^*)$	$C(i,j) = \rho_{ij}$, onde $\rho_{i,j}$ é o coeficiente de Correlação de Pearson entre as séries de log-retornos das ações i e j
Common Neighbors	$CN(i,j) = \lvert N_i \cap N_j \rvert$
Coeficiente de Jaccard	$JC(i,j) = \frac{\lvert N_i \cap N_j \rvert}{\lvert N_i \cup N_j \rvert}$
Coeficiente de Adamic-Adar	$AA(i,j) = \sum_{k \in N_i \cap N_j} \frac{1}{\log \lvert N_k \rvert}$, onde N_k é o conjunto de vértices adjacentes do nó k
Coeficiente de Sorenson-Dice	$SDC(i,j) = \frac{2 * \lvert N_i \cap N_j \rvert}{\lvert i \rvert + \lvert j \rvert}$
Betweenness da Aresta	$B(i,j) = \sum_{i,j \in V} \frac{\sigma_{ij}(e)}{\sigma_{ij}}$, onde $\sigma_{ij}(e)$ é o número de caminhos mínimos entre i e j que cruzam a aresta e e $\sigma_{i,j}$ é o número total de caminhos mínimos de i para j, $\forall i,j \in A$
Mesma Comunidade (BLONDEL *et al.*, 2008)	$SC(i,j) = \begin{cases} 1, & \text{se } i \text{ e } j \in \text{mesma comunidade,} \\ 0, & \text{se } i \text{ e } j \notin \text{mesma comunidade.} \end{cases}$
Preferential Attachment	$PA(i,j) = \lvert i \rvert * \lvert j \rvert$, onde $\lvert i \rvert$ e $\lvert j \rvert$ representam o grau dos vértices i e j

Adamic-Adar (AA):

$$AA(i,j) = \sum_{k \in N_i \cap N_j} \frac{1}{\log |N_k|}, \tag{3.7}$$

onde N_i, N_j e N_k representam o conjunto de nós adjacentes aos vértices i, j e k, respectivamente;

Local Path Index(LP):

$$LP(i,j) = A^2 + \varepsilon A^3 + \varepsilon^2 A4 + \cdots + \varepsilon^{n-2} A^n, \tag{3.8}$$

onde A é a matriz de adjacência, ε um parâmetro livre e $n > 2$ (SRILATHA; MANJULA, 2016). Vale ressaltar que para $\varepsilon = 0$, o algoritmo se torna similar ao CN, acessando A^2, ou seja, o conjunto de nós imediatamente adjacentes.

Random Walk with Restart(RW):

$$RW(i,j) = q_{ij} + q_{ji}, \tag{3.9}$$

onde q_{ij} é o j-ésimo elemento do vetor q_i, definido como sendo a probabilidade que um caminhante aleatório começando pelo nó i visite um nó j qualquer, se movendo com probabilidade c de acessar um vizinho aleatório e retorne ao vértice inicial i com probabilidade $1-c$ (LÜ; ZHOU, 2010).

Além desses métodos, incluímos um algoritmo ingênuo denominado *Time Invariant* (TI - Invariante no Tempo) como outro *benchmark* em nossos experimentos. Este algoritmo usa a ocorrência do link no grafo $G(t)$ como a previsão da ocorrência do link no grafo $G(t+h)$, assumindo que a estrutura de mercado é invariante no tempo. Essa suposição é tradicionalmente usada em algoritmos de gerenciamento de risco, que geralmente contam com uma matriz de covariância retroativa para estimar o risco do portfólio (MARKOWITZ, 1952; SOUZA; ASTE, 2019).

3.2.5 Dados de Mercado

Neste estudo, usamos dados de seis diferentes índices do mercado de ações espalhados pelos mercados americano, europeu e asiático. Os índices de ações foram utilizados para medir o desempenho da abordagem proposta em diferentes cenários, dada a diversidade dos mercados de ações. Além disso, é importante mencionar que eles representam o mercado de ações da região ou país onde estão listados. Consideramos os seguintes índices e países/regiões associados:

- **DAX30** (Alemanha): É um índice de ações que consiste nas 30 maiores e mais líquidas empresas alemãs negociadas na Bolsa de Valores de Frankfurt (*Frankfurt Stock Exchange*).

- **EUROSTOXX50** (Europa): É uma lista das 50 empresas de líderes em seus respectivos setores de onze países da zona do euro, incluindo Áustria, Bélgica, Finlândia, França, Alemanha, Irlanda, Itália, Luxemburgo, Holanda, Portugal e Espanha.
- **FTSE100** (Reino Unido): É um índice listado na Bolsa de Valores de Londres (*London Stock Exchange*). O *Financial Times Stock Exchange Index* (FTSE - Índice da Bolsa de Valores do Financial Times) é o principal indicador de ativos britânico, administrado por uma organização independente e calculado com base nas 100 maiores empresas do Reino Unido.
- **HANGSENG50** (Hong Kong): É um índice listado na Bolsa de Valores de Hong Kong (*Stock Exchange of Hong Kong*). Este índice do mercado de ações tem as 50 empresas constituintes com a maior capitalização de mercado. É o principal indicador do desempenho do mercado em Hong Kong.
- **NASDAQ100** (Estados Unidos da America): É um índice composto pelas 100 maiores empresas não-financeiras listadas na NASDAQ.
- **NIFTY50** (Índia): É um índice do mercado de ações listado na Bolsa de Valores Nacional da Índia (*National Stock Exchange of India*), baseado nas 50 maiores empresas indianas.

Cada índice financeiro tem uma série temporal de preços diários para cada uma de suas ações constituintes. As séries temporais de preços são construídas usando preços de fechamento diários coletados da *Thomson Reuters*. A lista de empresas constituintes de cada índice do mercado de ações não é estática e pode mudar com o tempo. Neste trabalho, consideramos apenas empresas que fizeram parte dos índices subjacentes ao longo de todo o período analisado, como comumente usado em outros trabalhos, quando a previsão de nós está fora do escopo (SOUZA; ASTE, 2019; CASTILHO *et al.*, 2019; MUSMECI; ASTE; MATTEO, 2016). Utilizamos os preços das ações que variam entre 1 de março de 2005 e 18 de dezembro de 2019. O número de ações em cada índice de mercado é: DAX30 = 22; EUROSTOXX50 = 40; FTSE100 = 50; HANGSENG50 = 30; NASDAQ100 = 42; NIFTY50 = 22.

3.3 Resultados e Discussão

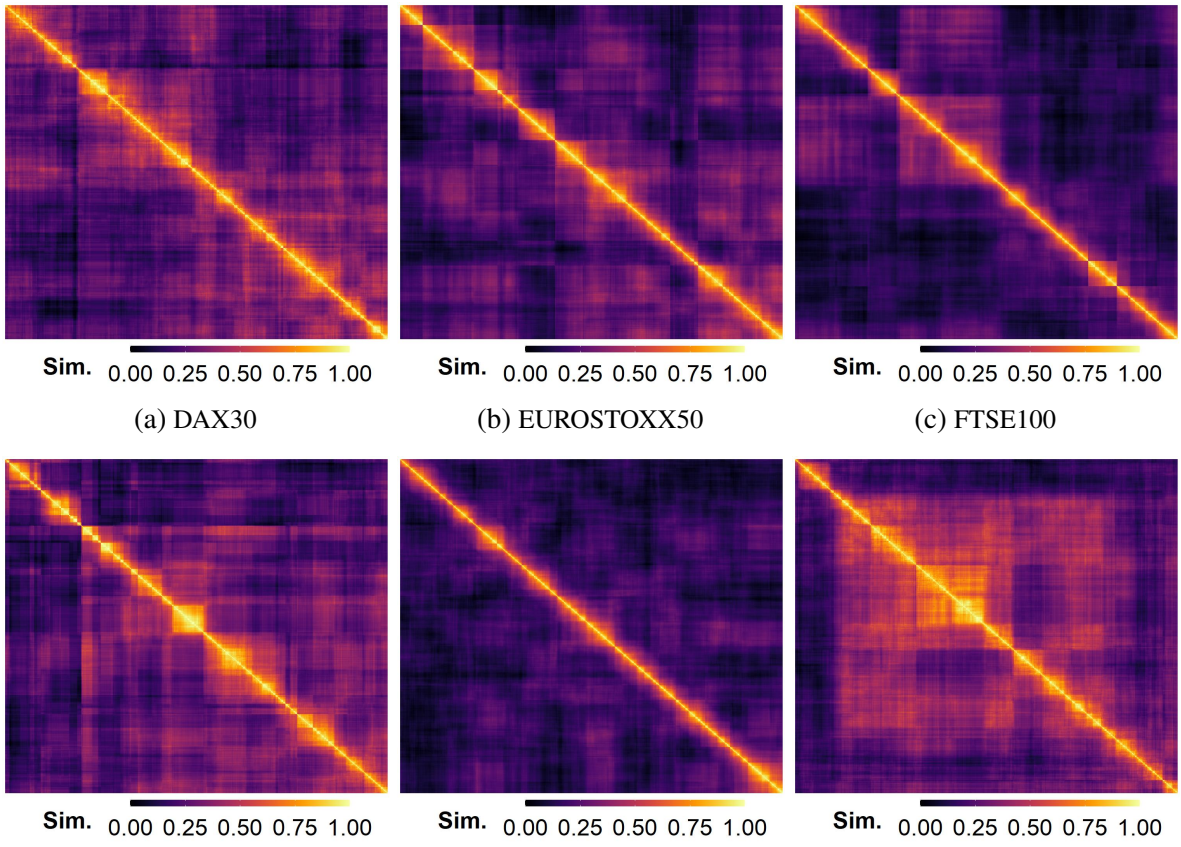
Apresentamos nesta seção um conjunto de análises e resultados experimentais para previsão da estrutura do mercado financeiro. Inicialmente, apresentamos um conjunto de análises descritivas sobre a evolução das redes de ações e uma breve discussão sobre o impacto dos diferentes métodos de filtragem de redes na estrutura do mercado financeiro. Posteriormente, apresentamos um conjunto de análises preditivas relacionadas à abordagem de aprendizado de máquina e aos métodos de *benchmark*. Por fim, apresentamos uma discussão sobre a interpretabilidade dos modelos de aprendizado de máquina.

3.3.1 Análises Descritivas

Apresentamos um conjunto de análises descritivas de redes financeiras criadas em diferentes índices de mercado. Nossa primeira análise descritiva descreve a persistência da rede financeira durante o período de teste, considerando $L = 252$ dias de negociação para criar cada gráfico. Essa análise nos permite medir como as redes financeiras mudam sua estrutura ao longo do tempo. Estimamos a persistência da rede calculando a similaridade entre redes $G(t)$ e $G(t')$ usando a distância de Jaccard, definida como segue:

$$sim(G(t), G(t')) = \frac{|G(t) \cap G(t')|}{|G(t) \cup G(t)|}, \tag{3.10}$$

considerando t e t' variando de 12 de maio de 2006 a 18 de dezembro de 2019.

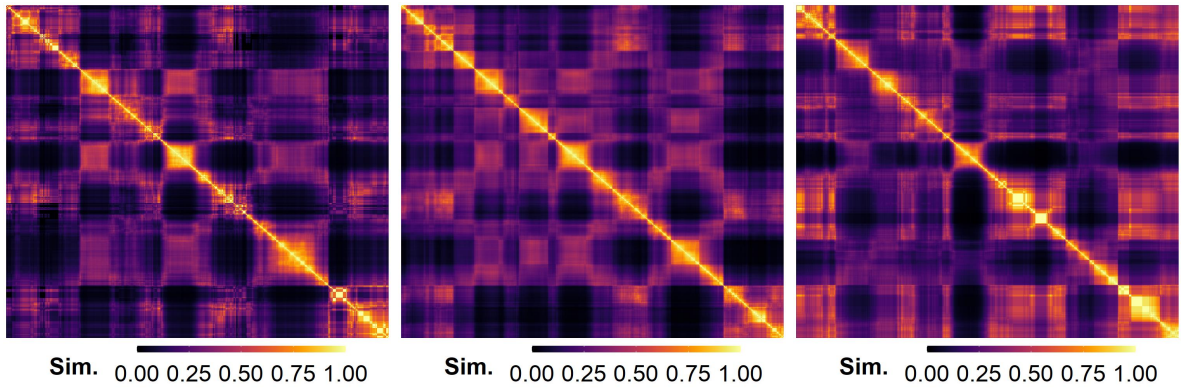


(a) DAX30 (b) EUROSTOXX50 (c) FTSE100

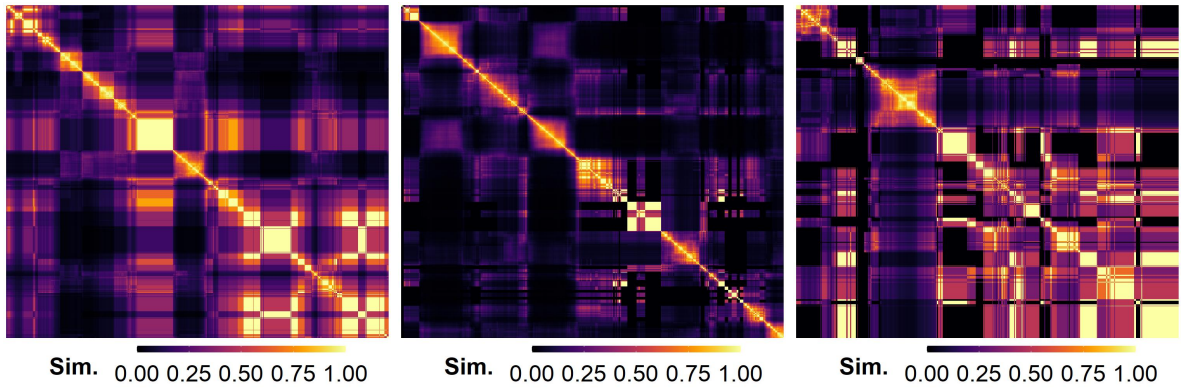
(d) HANGSENG50 (e) NASDAQ100 (f) NIFTY50

Figura 10 – **DAG - Matriz de similaridade cruzada para cada índice de mercado.** Calculamos a persistência da rede usando uma matriz de similaridade. Para criar as matrizes de similaridade, calculamos a similaridade de Jaccard em pares entre todas as redes financeiras $G(t)$ e $G(t')$ durante o período entre de 12 de maio de 2006 e 18 de dezembro de 2019, relacionado a um determinado índice de mercado. Para cada figura de índice de mercado, a primeira rede em 12 de maio de 2006 é representada na parte superior esquerda e a última rede em 18 de dezembro de 2019 na parte inferior direita de cada figura individual.

As Figuras 10, 11 e 12 apresentam a análise de similaridade cruzada utilizando os métodos DAG, DTN e DMST para cada índice do mercado, respectivamente. Na figura individual



(a) DAX30 (b) EUROSTOXX50 (c) FTSE100

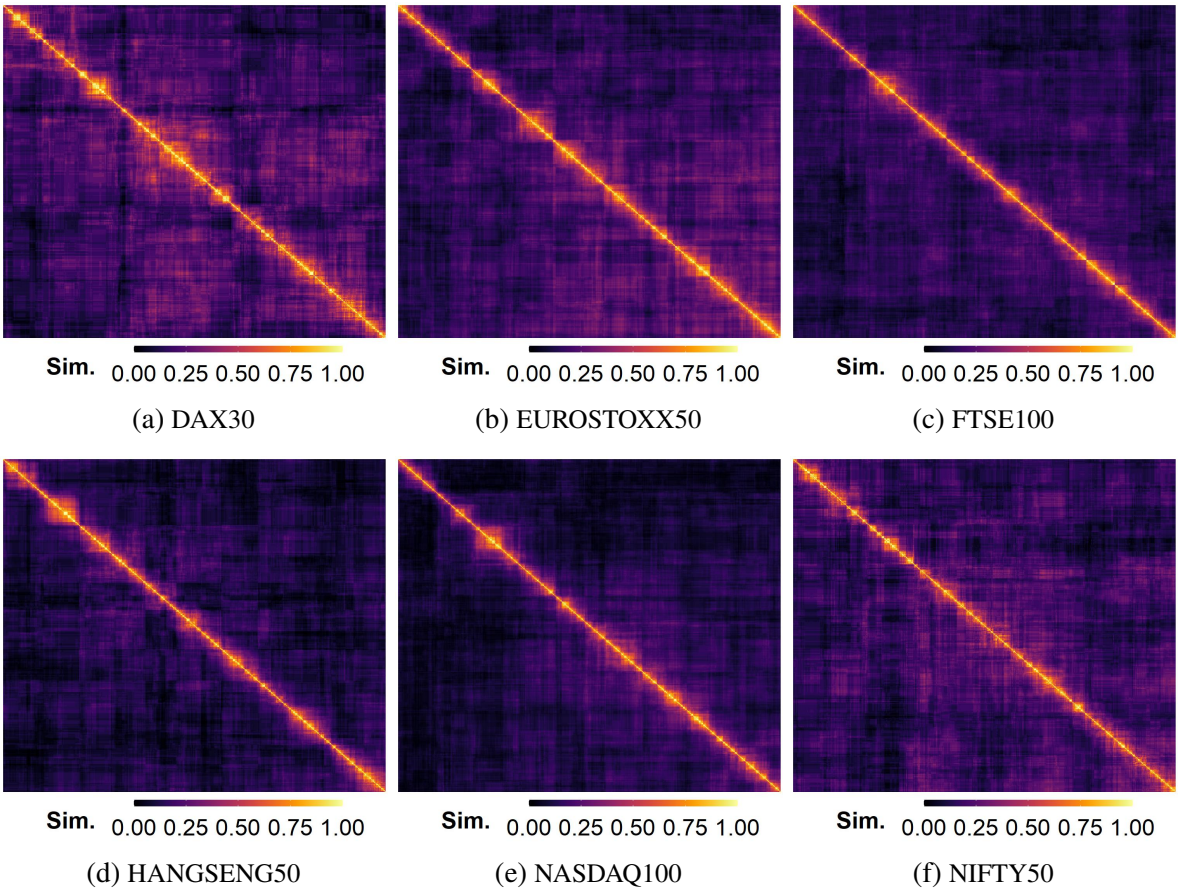


(d) HANGSENG50 (e) NASDAQ100 (f) NIFTY50

Figura 11 – **DTN - Matriz de similaridade cruzada para cada índice de mercado.** Calculamos a persistência da rede usando uma matriz de similaridade. Para criar as matrizes de similaridade, calculamos a similaridade de Jaccard em pares entre todas as redes financeiras $G(t)$ e $G(t')$ durante o período entre de 12 de maio de 2006 e 18 de dezembro de 2019, relacionado a um determinado índice de mercado. Para cada figura de índice de mercado, a primeira rede em 12 de maio de 2006 é representada na parte superior esquerda e a última rede em 18 de dezembro de 2019 na parte inferior direita de cada figura individual.

de cada índice de ações, a primeira rede é representada no canto superior esquerdo e a última rede é representada no canto inferior direito, onde a primeira rede é de 12 de maio de 2006 e a última rede é de 18 de dezembro de 2019. De uma forma geral, podemos observar que a estrutura muda consistentemente ao longo do tempo, o que enfatiza a importância das ferramentas de previsão da estrutura do mercado.

Os resultados de DAG apresentados na Figura 10 mostram que a estrutura da rede muda consideravelmente ao longo do tempo, levando em consideração todos os índices do mercado. A Figura 11 apresenta os resultados do método de filtragem de rede DTN. Podemos observar que a semelhança entre as redes tende a ser mais ruidosa do que o método DAG anterior. Em alguns períodos, a semelhança entre as redes é máxima, enquanto em outros chega a zero, como pode ser verificado em NASDAQ100 e em NIFTY50. O método de filtragem de rede DTN pode produzir gráficos desconexos ou mesmo vazios, o que pode causar essas oscilações de similaridade. Os resultados do DMST são mostrados na Figura 12. Esta figura mostra que há



(a) DAX30 (b) EUROSTOXX50 (c) FTSE100

(d) HANGSENG50 (e) NASDAQ100 (f) NIFTY50

Figura 12 – **DMST - Matriz de similaridade cruzada para cada índice de mercado.** Calculamos a persistência da rede usando uma matriz de similaridade. Para criar as matrizes de similaridade, calculamos a similaridade de Jaccard em pares entre todas as redes financeiras $G(t)$ e $G(t')$ durante o período entre de 12 de maio de 2006 e 18 de dezembro de 2019, relacionado a um determinado índice de mercado. Para cada figura de índice de mercado, a primeira rede em 12 de maio de 2006 é representada na parte superior esquerda e a última rede em 18 de dezembro de 2019 na parte inferior direita de cada figura individual.

baixa similaridade em comparações de longo alcance entre árvores criadas pelo método de filtragem DMST para todos os índices de mercado, sugerindo baixa estabilidade, conforme relatado por outros autores (CARLSSON; MÉMOLI *et al.*, 2010; MARTI *et al.*, 2015).

Dadas as matrizes de similaridade de cada mercado, calculamos a distância entre todas as matrizes para medir a similaridade de mercado em termos de evolução da rede. Essa análise permite identificar quais mercados apresentam comportamento semelhante considerando a persistência das redes. Para isso, usamos a similaridade de cosseno, calculada usando a seguinte fórmula:

$$similaridade\_cos(a,b) = \frac{\sqrt{\sum (a-b)^2}}{\sqrt{\sum a^2} * \sqrt{\sum b^2}}, \tag{3.11}$$

onde a e b são dois vetores numéricos diferentes de zero e representam o triângulo superior de duas matrizes de similaridade distintas. Essa métrica varia de 0 a 1 e é definida como a distância

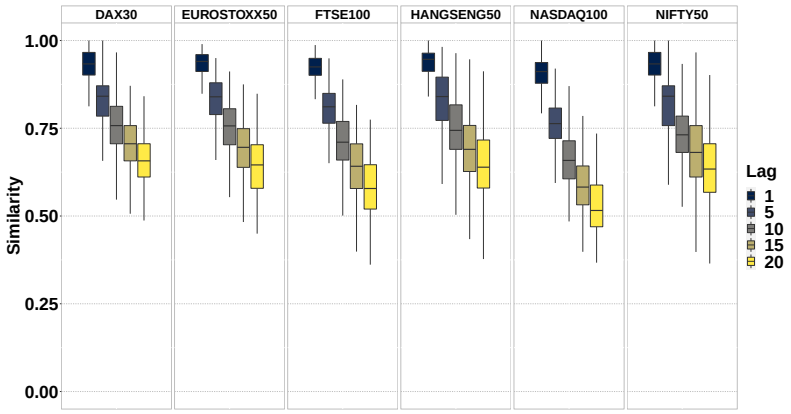
angular entre dois vetores. A Tabela 3 apresenta a similaridade de cosseno par a par para DAG, DTN e DMST. DAX30 e EUROSTOXX50 têm a maior similaridade de cosseno para DAG e DTN. Para DMST, o valor mais alto está entre FTSE100 e EUROSTOXX50. Esta análise demonstra que a persistência da rede entre os mercados da Europa é maior do que os mercados de outras regiões do mundo, dados os três métodos de filtragem de rede.

Tabela 3 – **Similaridade de cosseno entre os resultados de similaridade cruzada.** Calculamos a similaridade do cosseno a partir de matrizes de similaridade cruzada. Usamos o triângulo superior de cada matriz como vetor de entrada. Os mercados europeus apresentam maior similaridade em termos de persistência de rede.

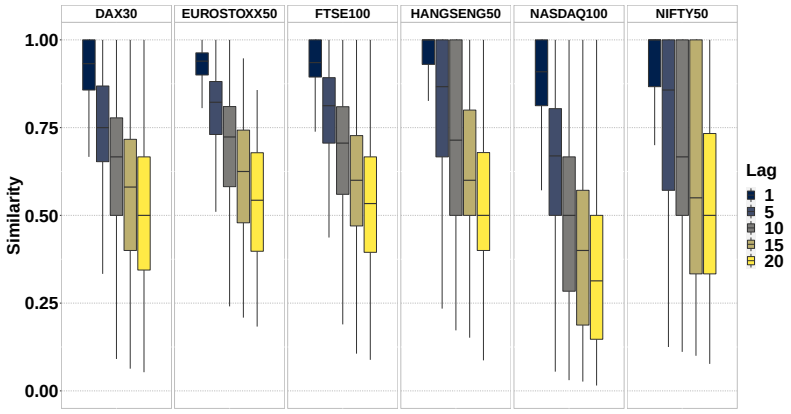
		EUROSTOXX50	***FTSE100***	***HANGSENG50***	***NASDAQ100***	***NIFTY50***
DAG	***DAX30***	**0.9532**	0.9435	0.9472	0.9341	0.9257
	EUROSTOXX50		0.9228	0.9403	0.9420	0.9070
	FTSE100			0.9150	0.9358	0.8978
	HANGSENG50				0.9297	0.9302
	NASDAQ100					0.9137
DTN	***DAX30***	**0.9338**	0.8367	0.7573	0.6209	0.5795
	EUROSTOXX50		0.8755	0.7873	0.6143	0.6000
	FTSE100			0.8331	0.5479	0.5503
	HANGSENG50				0.5892	0.5531
	NASDAQ100					0.4269
DMST	***DAX30***	0.9486	0.9354	0.8967	0.9011	0.9200
	EUROSTOXX50		**0.9500**	0.9058	0.9294	0.9312
	FTSE100			0.9253	0.9400	0.9338
	HANGSENG50				0.9169	0.9080
	NASDAQ100					0.9160

A segunda análise descritiva é uma forma de medir a semelhança entre a rede financeira atual $G(t)$ e a rede futura $G(t+h)$, onde h é o intervalo de tempo, $\forall\ h \in \{1,5,10,15,20\}$ semanas de negociação. Esta análise fornece um ponto de vista preciso sobre como a rede atual muda no futuro próximo - se elas se mantivessem constantes, não precisaríamos fazer uma previsão. Quantificamos as mudanças na estrutura da rede usando a distância de Jaccard entre $G(t)$ e $G(t+h)$, considerando $L = 252$ dias de negociação para criar cada grafo. A Figura 13 apresenta a distribuição de similaridade de redes relacionada aos três métodos de filtragem de rede DAG, DTN e DMST para cada índice do mercado. Os resultados experimentais sugerem uma distribuição de alta similaridade entre as redes considerando $h = 1$ semanas à frente para todos os métodos de filtragem de rede. No entanto, a distribuição de similaridade diminui conforme h aumenta, principalmente no método DMST. Considerando $h = 20$, o DMST apresenta similaridade média inferior a 25% em todos os mercados. Em geral, as redes de ações tendem a ter uma certa margem de similaridade para h baixo, mas à medida que h aumenta, elas se tornam cada vez mais desiguais, justificando, portanto, a importância de prever estruturas de mercado futuras, particularmente em cenários de previsão de longo horizonte. Analisando o método DTN, NIFTY50 e HANGSENG50 apresentam um comportamento diferente para h maiores, onde a distribuição da similaridade se comporta de forma diferente de outros mercados, oscilando entre o valor máximo e quase zero para h maiores, conforme mostrado em $h = 5$, $h = 10$ e $h = 15$. Esta amplitude pode ser explicada pela análise apresentada na Figura 11, que mostra que, para alguns

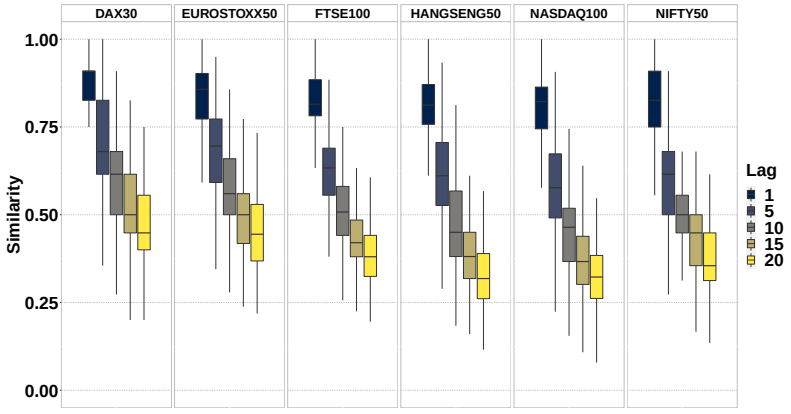
períodos, a similaridade entre as redes é alta, mas também é muito baixa em outros intervalos de tempo. Os menores valores de similaridade são apresentados para o método DMST considerando $L = 20$.



(a) Método de Filtragem de Rede DAG



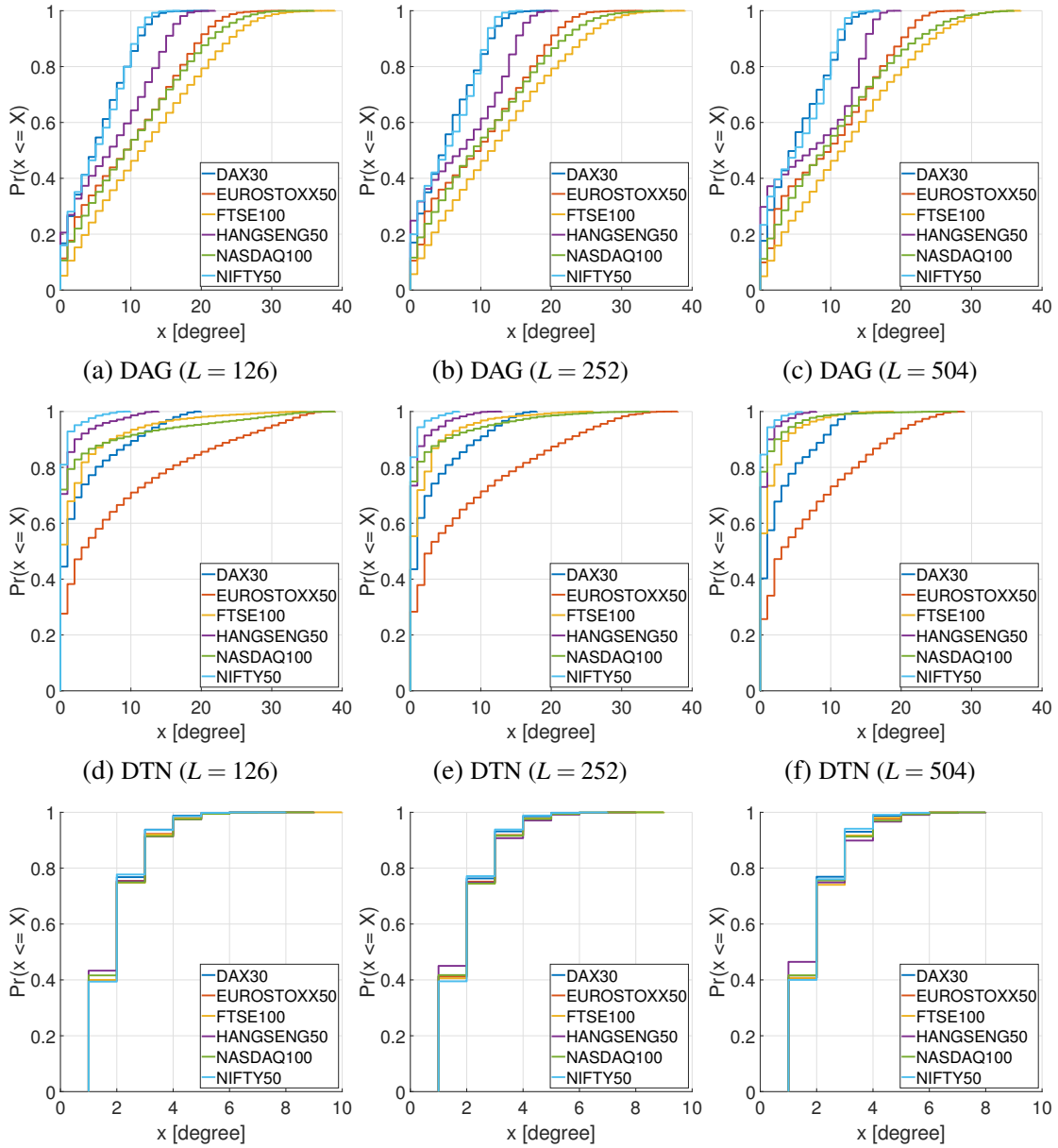
(b) Método de Filtragem de Rede DTN



(c) Método de Filtragem de Rede DMST

Figura 13 – **Similaridade de Redes vs. Intervalo de Previsão.** A figura mostra a distribuição da persistência das redes considerando $h = \ lbrace 1, 5, 10, 15, 20 \ rbrace$ semanas de negociação à frente para os três métodos de filtragem de rede DAG, DTN e DMST. A similaridade de rede é quantificada usando a distância de Jaccard entre os grafos $G(t)$ e $G(t+h)$.

A terceira análise descritiva é apresentada na Figura 14. Apresentamos a Função de Distribuição Cumulativa (*Cumulative Distribution Function* - CDF) do grau dos nós nas redes



(a) DAG ($L = 126$) (b) DAG ($L = 252$) (c) DAG ($L = 504$)

(d) DTN ($L = 126$) (e) DTN ($L = 252$) (f) DTN ($L = 504$)

(g) DMST ($L = 126$) (h) DMST ($L = 252$) (i) DMST ($L = 504$)

Figura 14 – **CDF do grau dos nós em redes usando métodos de filtragem DAG, DTN e DMST.** Calculamos a função de distribuição cumulativa do grau dos nós em todas as redes de ações usando o tamanho da janela deslizante $L = 126, 252$ e 504 dias de negociação. O período dos experimentos varia de 3 março 2007 a 18 de dezembro de 2019. Os índices de mercado com o menor número de ações constituintes apresentam comportamento semelhante considerando o método de filtragem de rede DAG. O método DTN apresenta maior probabilidade de nós sem arestas, principalmente no NIFTY50, NASDAQ100 e HANGSENG50. O EUROSTOXX50 apresenta uma forma distinta em comparação com os restantes índices de mercado em DTN com o menor número de nós sem ligação. Os resultados também sugerem que a distribuição de grau dos índices de mercado são semelhantes para $L = 126, 252e504$ dias de negociação em todos os métodos de filtro de rede.

de cada índice usando os métodos de filtrage de rede DAG, DTN e DMST. Esta análise fornece informações sobre o grau dos nós de acordo com três aspectos principais:*(i)* o impacto do tamanho da série temporal L; *(ii)* método de filtragem da rede e *(iii)* tamanho do índice de mercado, considerando o número de constituintes. Calculamos a distribuição do grau de nós em todas as redes financeiras durante o período entre 3 março 2007 a 18 de dezembro de 2019. São apresentados os resultados usando $L \in \{126, 252, 504\}$ dias de negociação como tamanho da janela deslizante. Observamos na Figura 14 que os índices de mercado com o menor número de constituintes apresentam comportamento semelhante em termos de grau de nós quando usamos o método de filtragem de rede DAG. Além disso, os nós em DAG são propensos a ter maior ocorrência de nós sem conexões. O método DTN também apresenta alta probabilidade de nós sem arestas, principalmente em NIFTY50, NASDAQ100 e HANGSENG50. EUROSTOXX50 apresenta uma forma distinta em comparação com os outros índices de mercado em DTN com o menor número de nós sem conexão - mais de 75% de nós tem grau maior que 1 aresta. Por outro lado, para todos os índices de mercado, pelo menos 50% dos nós têm 4 ou mais conexões no DAG. Considerando o número de ações em cada índice de mercado, também podemos concluir que não há nós conectando-se a todos os outros vértices em nenhum método de filtragem de rede devido ao maior grau de cada índice de mercado. Os resultados também sugerem que a distribuição de grau dos índices de mercado são semelhantes para $L = 126, 252, 504$ dias de negociação em todos os métodos de filtragem de rede, indicando que o tamanho de L não afeta a distribuição de grau das ações em redes de cada índice de mercado.

3.3.2 Análises Preditivas

Nesta seção, apresentamos um conjunto de resultados experimentais relacionados à previsão da estrutura de mercado usando aprendizado de máquina. Inicialmente, investigamos o desempenho preditivo do método proposto em diferentes cenários, comparando-o com os métodos de *benchmark*. Por fim, apresentamos também uma análise qualitativa sobre a interpretabilidade do modelo e suas implicações.

3.3.2.1 Resultados de Desempenho

Neste trabalho, utilizamos uma abordagem de aprendizado de máquina para prever a rede financeira $G(t+h)$, em que h é o número de semanas à frente, considerando $h = 1, 2, \ dots, 20$ semanas de negociação. Discutimos e relatamos os resultados usando o tamanho das janelas deslizantes $L = 252$ dias de negociação para construir as redes financeiras. Os resultados relacionados ao tamanho de janela deslizante $L = \{126, 504\}$ dias de negociação de são relatados no Apêndice A. As Figuras 15, 16 e 17 mostram a medida de AUC do método de aprendizado de máquina proposto em comparação com algoritmos de *benchmark* para os métodos de filtragem de rede DAG, DTN e DMST. Para cada intervalo de tempo h, calculamos a AUC média de cada método e seu respectivo erro padrão durante todo o período de teste, compreendido no período

entre 5 maio 2007 e 18 de dezembro de 2019.

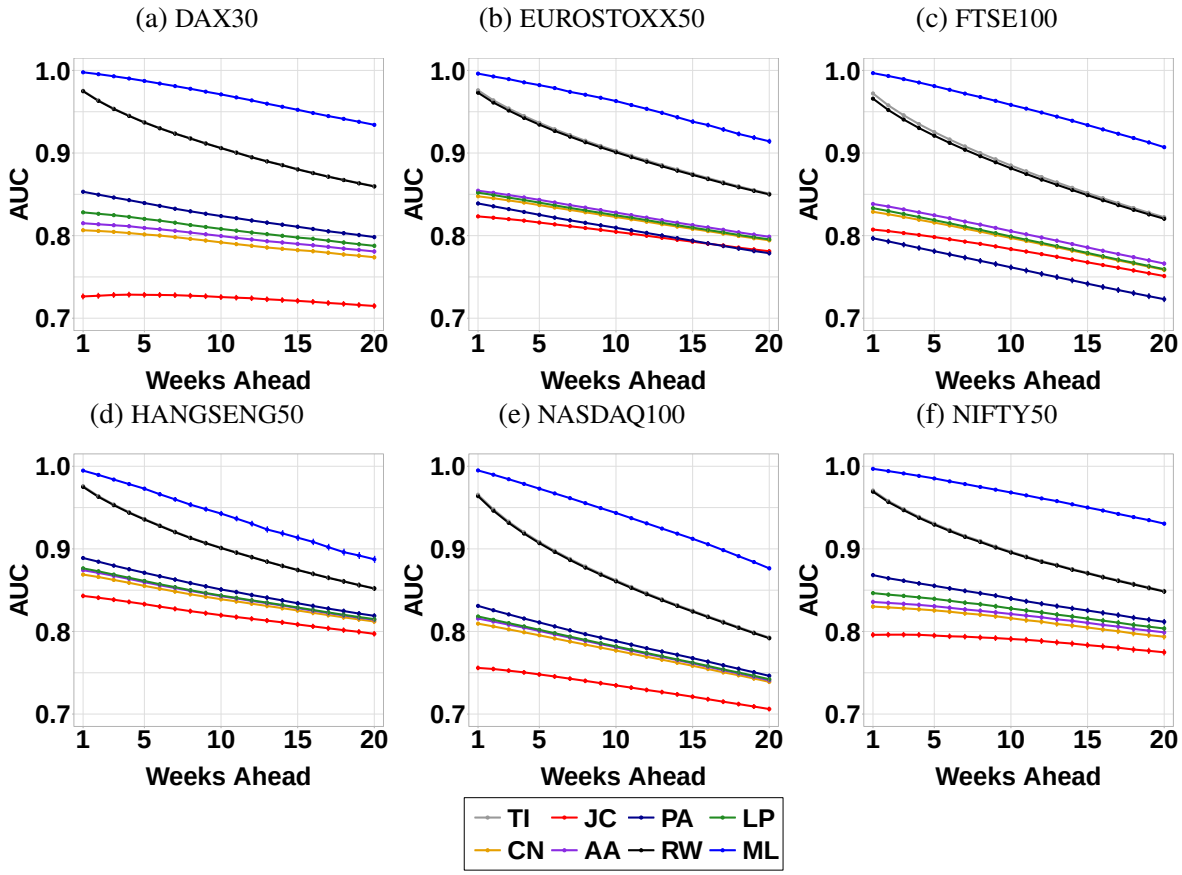


Figura 15 – **DAG - Comparação de desempenho preditiva de todos os métodos.** A figura mostra a AUC do método de aprendizado de máquina em comparação com os métodos base para previsão de links. Para cada intervalo de tempo, calculamos a média AUC de cada método e seu respectivo erro padrão ao longo de todo o período de testes. O método de aprendizado de máquina supera os métodos base em todos os índices de mercado.

Denotado como "ML", o método de aprendizado de máquina supera os métodos base em todos os índices de mercado e todos os métodos de filtragem de rede. Em geral, o desempenho preditivo diminui à medida que o intervalo de tempo h aumenta. Apesar de sua simplicidade, TI é bastante eficaz e apresenta bom desempenho em índices de mercado e métodos de filtragem de rede, semelhante ao algoritmo RW. A Figura 15 apresenta resultados para o método de filtragem de rede DAG, sugerindo que os índices de mercado com um pequeno número de constituintes têm maior AUC do que mercados com grande número. Os resultados também sugerem que o algoritmo RW produz uma classificação de aresta bastante semelhante a TI. O método JC apresenta o pior desempenho preditivo em todos os índices de mercado, exceto para FTSE100 em que PA apresenta valores de AUC mais baixos para o método de filtragem de rede DAG.

A Figura 16 apresenta resultados para o método de filtragem de rede DTN. Os resultados de ML são superiores em todos os mercados e sugerem que o método proposto pode identificar com precisão links com alta correlação, dado o objetivo principal do método DTN. Podemos observar que os algoritmos base apresentam piores resultados em HANGSENG50, NASDAQ100

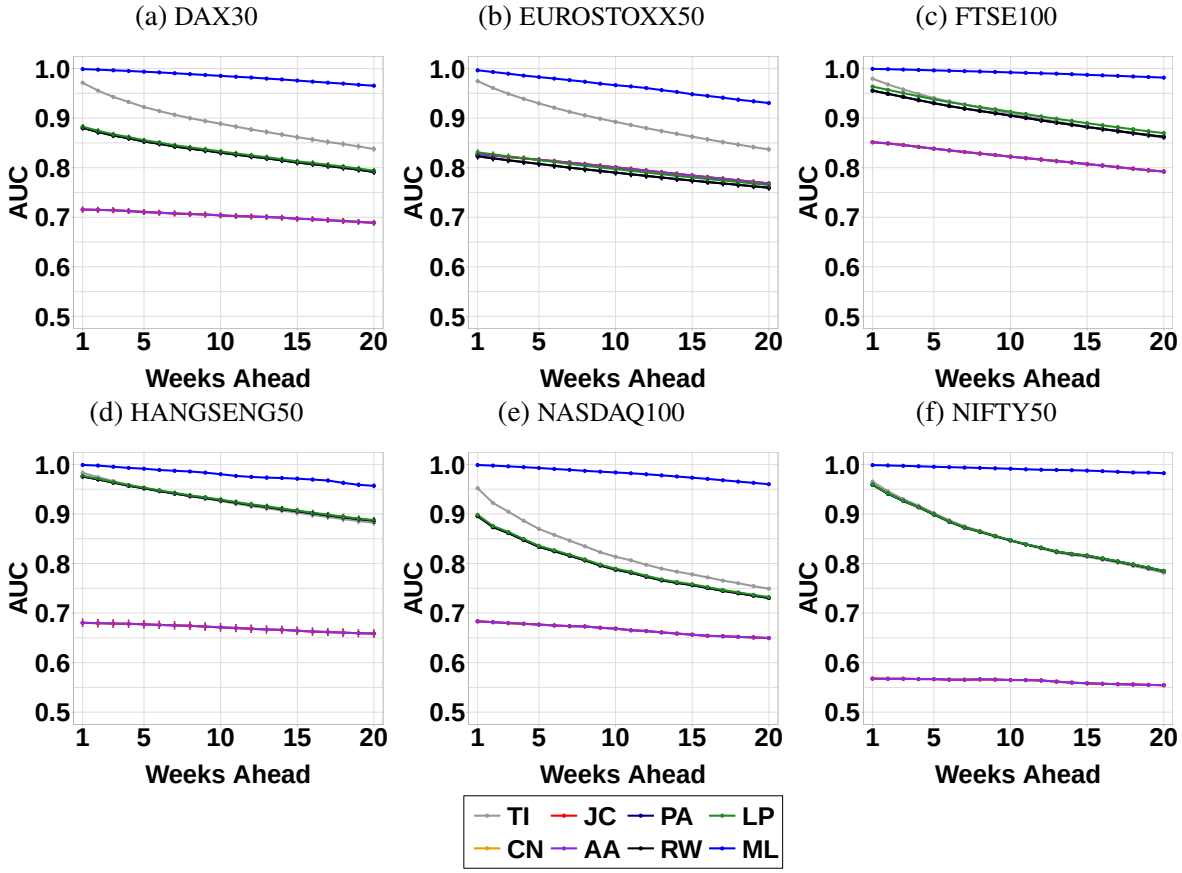


Figura 16 – **DTN - Comparação de desempenho preditiva de todos os métodos.** A figura mostra a AUC do método de aprendizado de máquina em comparação com os métodos base para previsão de links. Para cada intervalo de tempo, calculamos a média AUC de cada método e seu respectivo erro padrão ao longo de todo o período de testes. O método de aprendizado de máquina supera os métodos base em todos os índices de mercado.

e NIFTY50. Conforme apresentado na Figura 14, esses índices de mercado possuem um número expressivo de nós sem conexões. O algoritmo de TI supera os demais algoritmos base em DAX30, EUROSTOXX50 e NASDAQ100. A Figura 17 apresenta resultados relacionados ao método de filtragem de rede DMST. Os métodos base apresentam os piores resultados entre os três métodos de filtragem de rede, com exceção dos algoritmos TI e RW. Novamente, ML supera os métodos base em todos os mercados.

A Figura 18 apresenta o desempenho da AUC do método proposto para previsão de redes, considerando h semanas de negociação à frente ($1 \leq h \leq 20$), utilizando os métodos de filtragem de rede DAG, DTN e DMST. A medida AUC diminui à medida que o intervalo de tempo h aumenta. Também comparamos nossos resultados com o método invariante no tempo TI, onde a rede $G(t)$ é usada como a previsão de $G(t+h)$. Escolhemos o TI para comparar nosso método devido ao seu desempenho superior sobre todos os métodos de benchmark apresentados na análise anterior. Além disso, selecionamos o método TI porque ele é derivado de informações de correlação de pares, conforme descrito na Tabela 2. A melhoria sobre a AUC, representada por AUC*, é calculada da seguinte forma:

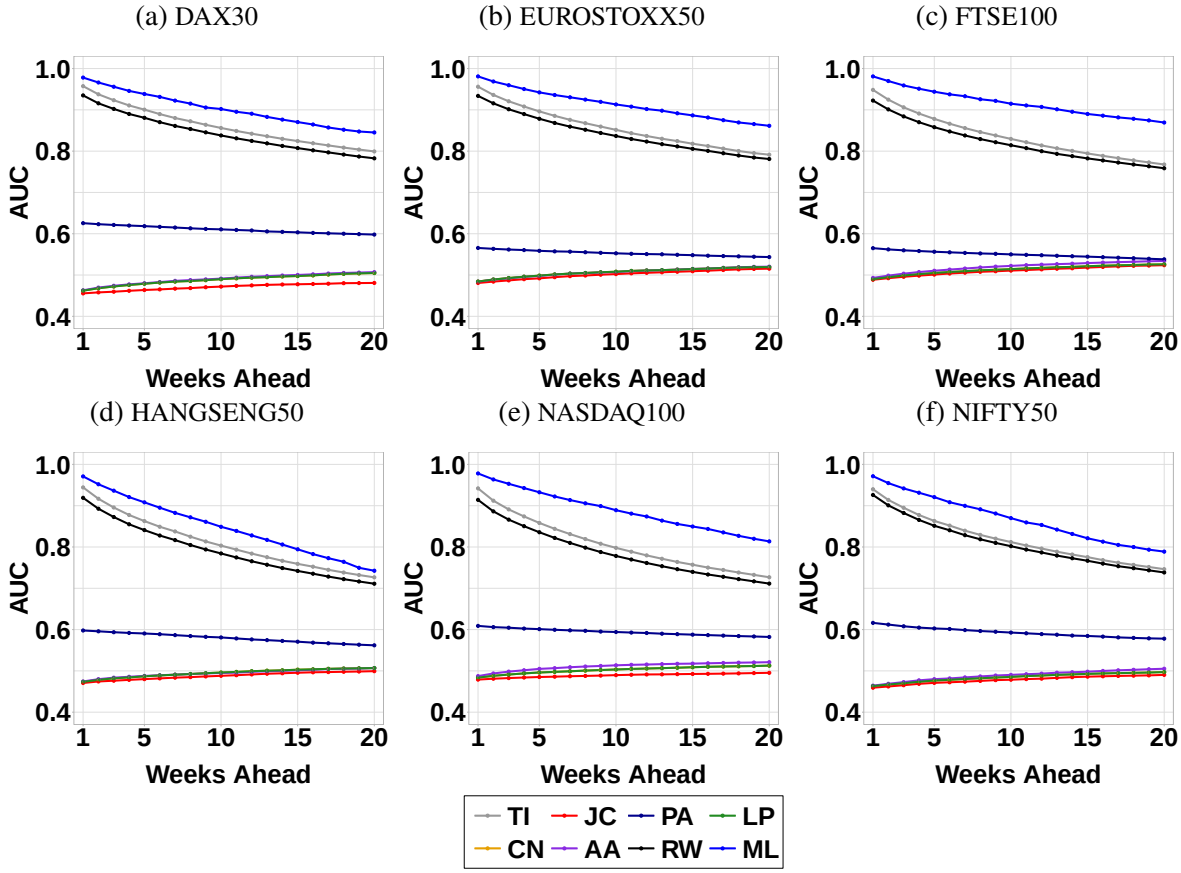
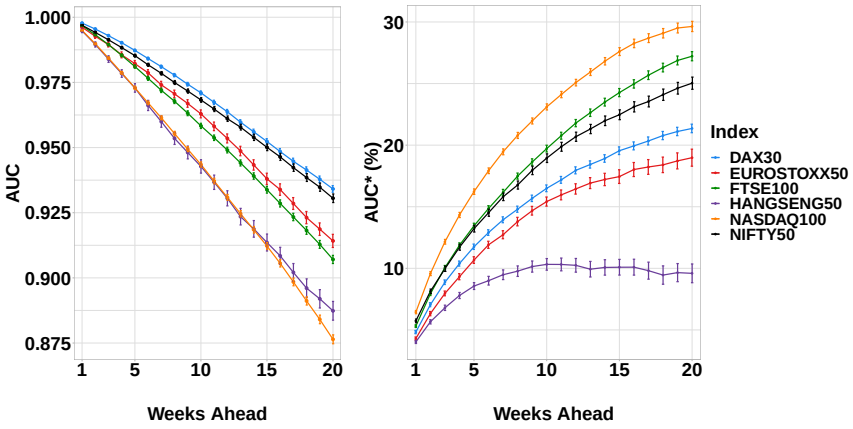


Figura 17 – **DMST - Comparação de desempenho preditiva de todos os métodos.** A figura mostra a AUC do método de aprendizado de máquina em comparação com os métodos base para previsão de links. Para cada intervalo de tempo, calculamos a média AUC de cada método e seu respectivo erro padrão ao longo de todo o período de testes. O método de aprendizado de máquina supera os métodos base em todos os índices de mercado.

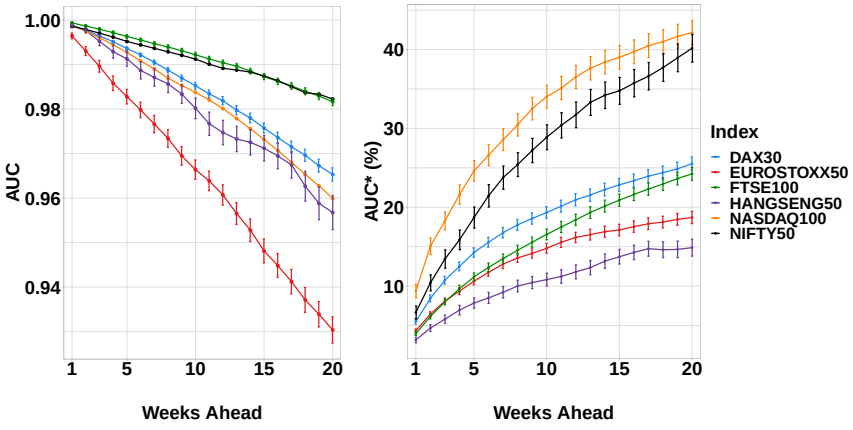
$$AUC^* = (AUC_m - 0.5)/(AUC_b - 0.5) - 1, \tag{3.12}$$

onde AUC_m é a AUC do método de aprendizado de máquina e AUC_b é a AUC do *benchmark*. A Figura 18(b), 18(d) e 18(f) apresentam os resultados de melhoria de AUC * e seu erro padrão para métodos de filtragem de rede DAG, DTN e DMST.

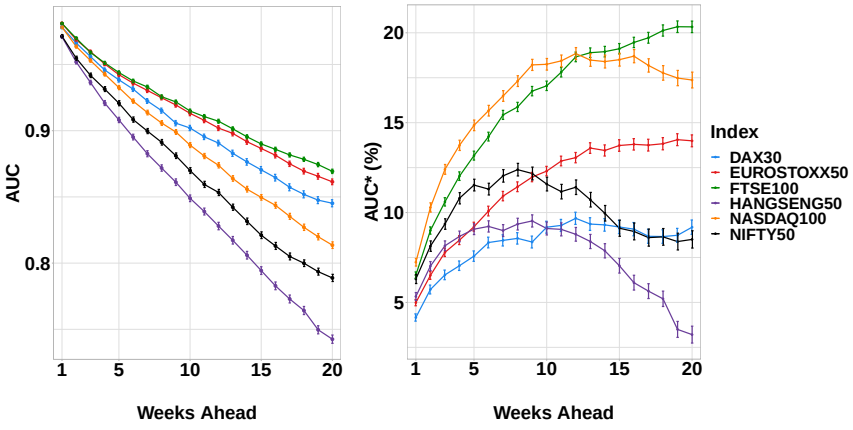
O método proposto apresenta resultados de AUC semelhantes para todos os métodos de filtragem de rede. Os resultados usando DAG, mostrados na Figura 18(a) sugerem que redes com menos constituintes têm melhores resultados de AUC. A Figura 18(b) mostra que a maior melhoria de AUC* acontece em NASDAQ100, chegando a quase 30% para $h = 20$ semanas à frente. Por outro lado, para o método DTN, mostrado na Figura 18(c), os melhores resultados são para FTSE100 e NIFTY50, com EUROSTOXX50 sendo o resultado mais distinto. A maior melhoria de AUC* relacionada ao DTN, mostrado na Figura 18(d), é sobre NASDAQ100 e NIFTY50, chegando a quase 40%. Os resultados mostrados na Figura 18(e) estão relacionados ao método de filtragem de rede DMST e têm decaimento semelhante para todos os mercados, onde DAX30 é o melhor resultado. Curiosamente, a melhoria AUC* mostrada na Figura 18(f) apresenta



(a) DAG (AUC) (b) DAG (AUC*)



(c) DTN (AUC) (d) DTN (AUC*)



(e) DMST (AUC) (f) DMST (AUC*)

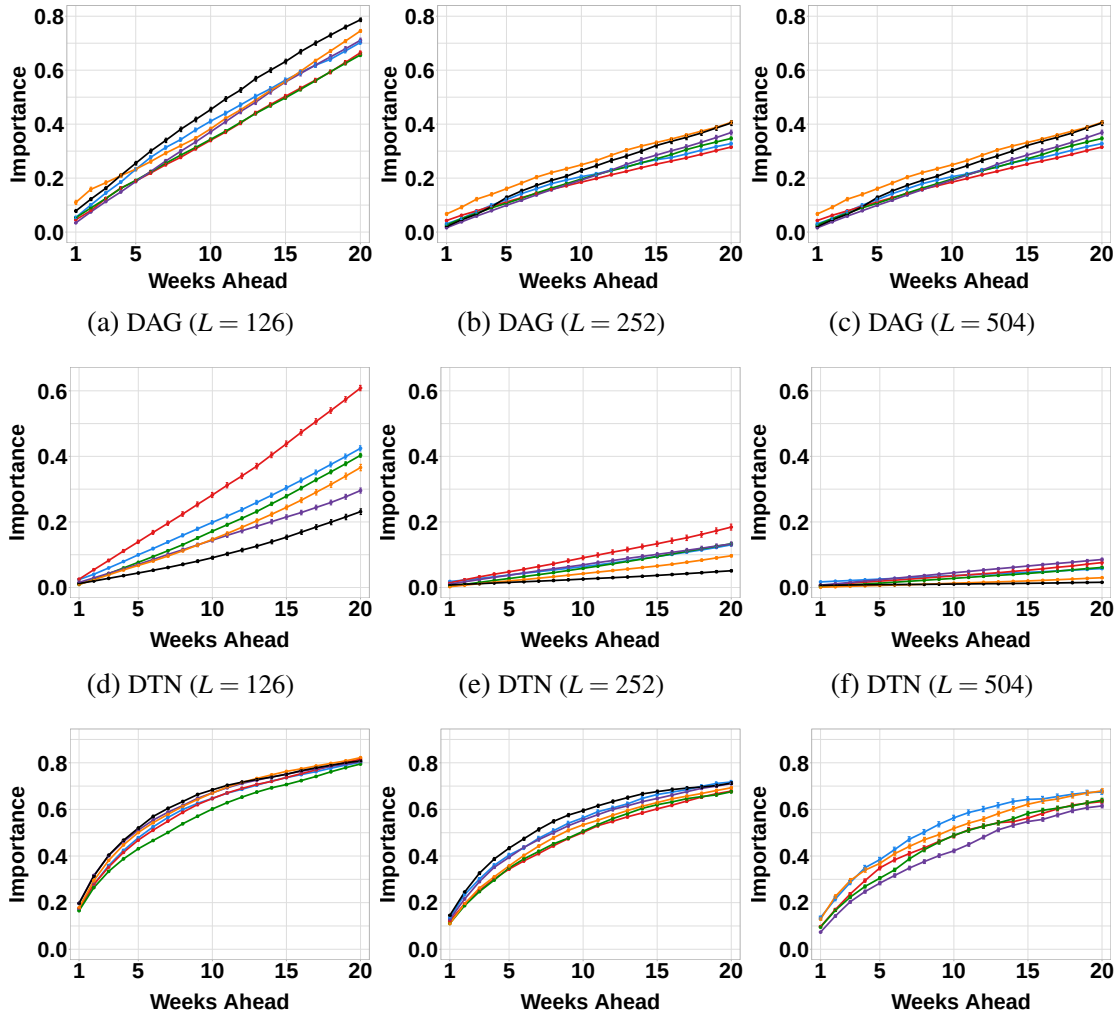
Figura 18 – **AUC e AUC* do algoritmo ML para métodos de filtragem de rede DAG, DTN e DMST.** Os painéis (a), (c) e (e) apresentam a métrica AUC do aprendizado de máquina e seu erro padrão para previsão de h semanas de negociação à frente ($1 \leq h \leq 20$). Os painéis (b), (d) e (f) apresentam a melhoria da AUC em relação ao método invariante no tempo e seu erro padrão. Os resultados apresentados são relacionados a $L = 252$.

curvas semelhantes para os mercados NIFTY50 e HANGSENG50. Os resultados mostram que a melhoria de AUC* para NIFTY50 e HANGSENG50 aumentam até aproximadamente $h = 9$, alcançando quase 13% em NIFTY50. Após este valor máximo, a melhoria de AUC* diminui à medida que h aumenta. FTSE100 apresenta a maior melhoria de AUC*, atingindo quase 21% para $h = 20$ semanas de negociação à frente.

3.3.2.2 Interpretabilidade do Modelo

Em finanças, especialmente em gestão de portfólio, o risco de investimento é calculado usando a correlação e covariância entre os ativos do portfólio, sendo esta uma das principais informações utilizada para estimar o risco. Neste trabalho, queremos medir como a topologia da rede ajuda a prever as próprias redes futuras. Em outras palavras, estamos interessados em avaliar a importância dos atributos topológicos para prever a estrutura do mercado. Conforme descrito na Seção 3.2.3, separamos o conjunto de atributos preditivos em dois subconjuntos: características de correlação em pares e características topológicas. Depois de construir as árvores de decisão no modelo XGBoost, podemos estimar a importância de cada atributo individual. A importância de um atributo está relacionada ao número de vezes que ele é usado para criar decisões de divisão relevantes, ou seja, pontos de divisão que melhoram as métricas de desempenho (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Para cada índice de mercado, calculamos a média e o erro padrão da importância agregada das características de correlação de pares e características topológicas. A Figura 19 apresenta resultados relacionados à importância das características topológicas, considerando os métodos de filtragem de rede DAG, DTN e DMST e $L \in \{126, 252.504\}$ dias de negociação como tamanho da janela deslizante. É importante observar que a importância dos dois subconjuntos de características somam 1.

Os resultados apresentados na Figura 19 mostram que as características topológicas ajudam a prever o mercado futuro utilizando diferentes métodos de filtragem de rede. Observamos que a importância das características topológicas aumenta com h. Além disso, a importância desse subconjunto de atributos muda de acordo com a topologia da rede, isto é, a importância se altera conforme o método de filtragem de rede. Sua importância é evidenciada principalmente para L menores, como $L = 126$, mostrado nas Figuras 19(a), 19(d) e 19(g), onde sua importância para $h = 20$ atinge quase 80% em NIFTY50 através do método DAG, 60% para EUROSTOXX50 usando DTN e quase 90% para todos os mercados usando DMST. Para o método DMST, mostrado nas Figuras 19(g), 19(h) e 19(i), a importância das características topológicas têm forma semelhante para tamanho de janela deslizante $L = 126$, 252 e 504. Os resultados do método DAG são mostrados nas Figuras 19(a), 19(b) e 19(c). Para valores h pequenos, as características topológicas não adicionam muitas informações quando comparados às características de correlação de pares. No entanto, a importância desses atributos aumenta rapidamente com o intervalo de tempo h, sugerindo que esses atributos podem ser mais úteis do que características de correlação de pares para exercícios de previsão de longo prazo, particularmente para tamanhos de janela deslizantes menores. Para $L = 252$ e $L = 504$, as características topológicas têm menos impor-



(a) DAG ($L = 126$) (b) DAG ($L = 252$) (c) DAG ($L = 504$)

(d) DTN ($L = 126$) (e) DTN ($L = 252$) (f) DTN ($L = 504$)

(g) DMST ($L = 126$) (h) DMST ($L = 252$) (i) DMST ($L = 504$)

Figura 19 – **Importância de características topológicas para DAG, DTN e DMST.** A figura mostra a importância agregada de características topológicas utilizando o tamanho da janela deslizante $L = \{126, 252, 504\}$ dias de negociação e métodos de filtragem de rede DAG, DTN e DMST. Os resultados mostram que a importância desses atributos aumentam conforme o intervalo de tempo h aumenta. A importância das características topológicas para $L = 126$ dias de negociação é maior do que $L = 252$ e $L = 504$, considerando todos os métodos de filtro de rede. O crescimento da importância desse subconjunto é consistente em todos os mercados. Um resultado interessante é que a importância das características topológicas mudam de acordo com o método de filtragem de rede.

tância na previsão de redes modeladas através dos métodos de filtragem de rede DAG e DTN. Considerando os resultados do DMST, a importância das características topológicas aumenta rapidamente, mesmo para valores *h* pequanos. Esse comportamento é diferente do apresentado pelos métodos DAG e DTN. Uma possível explicação para isso é a baixa persistência das árvores, conforme mostrado na Figura 12. Assim, os atributos de rede são capazes de adicionar mais informações ao modelo de aprendizado de máquina quando comparados às características de correlação de pares.

3.3.3 O Efeito das Propriedades de Rede na Previsibilidade da Estrutura de Mercado

Nessa seção, nós apresentamos uma análise tentando responder à seguinte questão: o que torna uma rede mais preditiva do que outra? Para responder a essa pergunta, investigamos se as propriedades da rede tornam nosso modelo mais ou menos preditivo. Avaliamos a relevância de seis medidas de rede no desempenho de previsão da estrutura de mercado. O conjunto de métricas de rede usado é o seguinte:

- **Assortatividade** - Correlação do grau dos vértices da rede;
- **Coeficiente de Clusterização** - Uma medida do grau em que os nós em um grafo tendem a se agrupar (WASSERMAN *et al.*, 1994);
- **Diâmetro** - O caminho mais curto máximo entre todos os nós do grafo;
- **Entropia** - Entropia de Shannon dos graus da rede;
- **Modularidade** - Uma medida de quão modular é uma rede em relação a alguma divisão de vértice, conforme definido por by Clauset, Newman e Moore (2004);
- **Rotatividade** - A fração de arestas em comum entre duas amostras consecutivas da rede.

Investigamos essa relação reunindo o *t-value* (valor t) de uma regressão linear múltipla, construída usando o conjunto de variáveis da rede como variáveis explicativas e a métrica AUC como saída do modelo. Isso nos permite verificar quais variáveis independentes podem explicar a variável alvo. Para garantir resultados consistentes, verificamos as premissas de regressão linear subjacentes relacionadas ao modelo e às variáveis explicativas:

- *Suposição 1* - Variáveis explicativas são estacionárias, verificadas usando o teste *Augmented Dickey-Fuller* (ADF);
- *Suposição 2* - Média de resíduos são aproximadamente zero;

- *Suposição 3* - Os resíduos da regressão linear múltipla são normalmente distribuídos, verificados usando o teste de Shapiro-Wilk;

- *Suposição 4* - Variáveis explicativas e resíduos do modelo não estão correlacionados, verificados usando o teste de correlação de Pearson;

- *Suposição 5* - Não há autocorrelação dos resíduos, verificados usando o teste de Durbin-Watson.

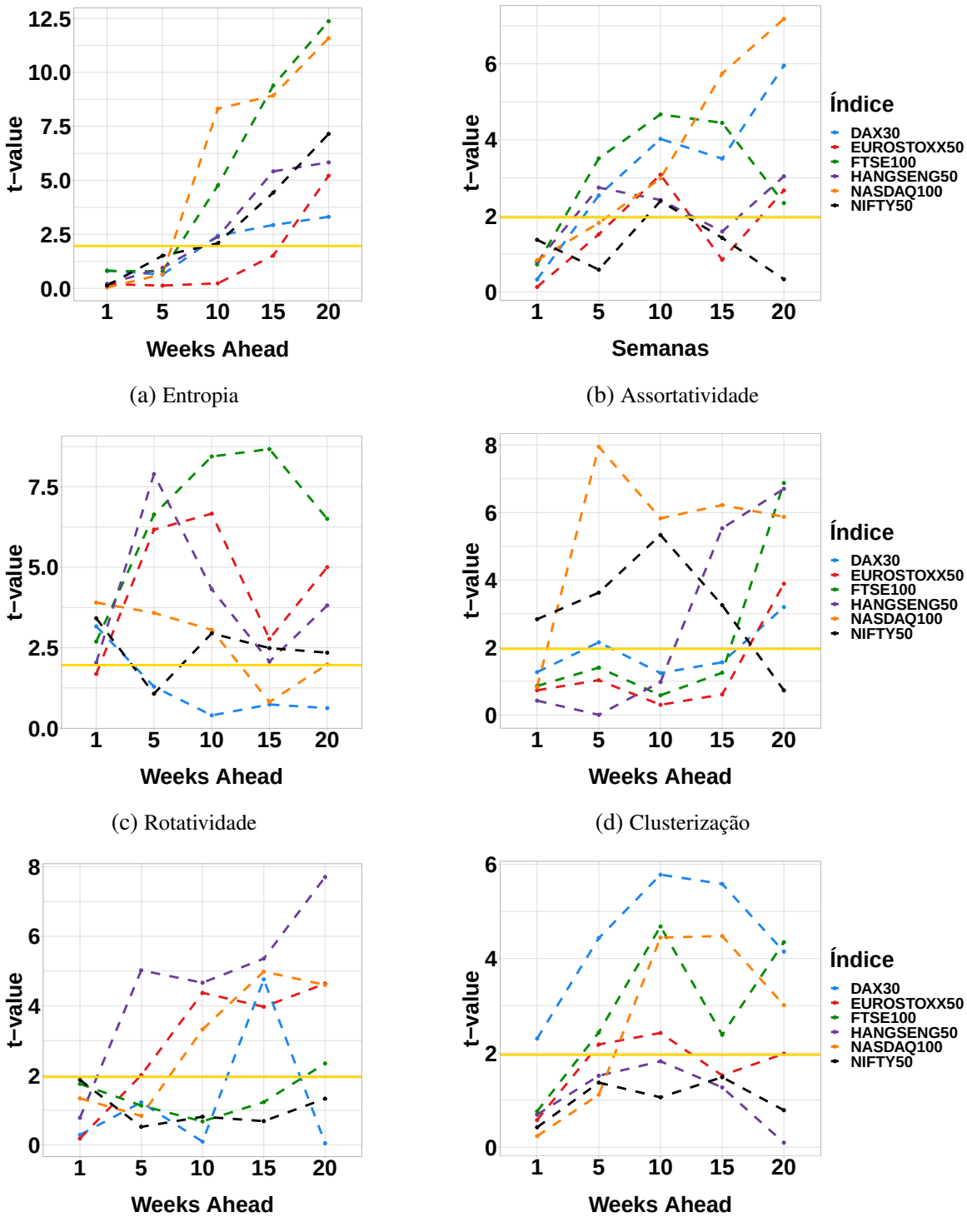
As suposições foram verificadas para cada configuração experimental. A tupla composta pelo índice de mercado, o tamanho da série temporal L, o intervalo de tempo h e o método de filtragem de rede representam uma configuração experimental para realização de tais análises. As Figuras 20, 21 e 22 apresentam resultados para $L = 252$ e os métodos de filtragem de rede DAG, DTN e DMST, respectivamente. Resultados relacionados com $L \in 126, 504$ são mostrados no Apêndice A. Investigamos a relevância estatística das variáveis de rede para inferir a precisão das previsões do modelo de aprendizado de máquina reunindo o valor t da seguinte regressão linear múltipla:

$$AUC = \beta_0 + \beta_1 * a + \beta_2 * c + \beta_3 * d + \beta_4 * e + \beta_5 * m + \beta_6 * t, \tag{3.13}$$

onde a é a assortatividade, c é o coeficiente de agrupamento, d é o diâmetro, e é a entropia, m é a modularidade e t é a rotatividade da rede. Representamos a significância como uma linha amarela tracejada na Figuras 20, 21 e 22. Os resultados acima desta linha são estatisticamente relevantes ($p < 0,05$).

Na Figura 20, observamos que as propriedades da rede são em geral significativamente relacionadas à capacidade preditiva da rede, particularmente na previsão de longo prazo da estrutura do mercado. O nível de imprevisibilidade de informação da rede, medido através da entropia da rede, é apresentado nas Figuras 20(a), 21(a) e 22(a) referentes aos métodos DAG, DTN e DMST, respectivamente. Em DAG, a métrica foi significativa para todos os mercados estudados na previsão de longo prazo ($h = 20$ semanas à frente). Para DTN, a relevância desta métrica para definir o grau de previsibilidade da rede é alta para quase todos os índices de mercado, exceto para NIFTY50. Esse comportamento não é tão evidente para DMST, pois NIFTY50 e HANSENG50 apresentam valores abaixo da linha que define os valores estatisticamente relevantes. No geral, observamos que a significância da entropia aumenta de acordo com h para todos os mercados estudados utilizando os métodos DAG e DTN.

A estrutura da rede, medida por indicadores de assortatividade, agrupamento e modularidade, apresentados nas Figuras 20, 21 e 22, painéis (b), (d) e (e), respectivamente, foram significativas para a maioria dos mercados em uma previsão de $h = 20$ semanas à frente. As alterações nos links da rede, medidas através da rotatividade, apresentadas nas Figuras 20(c), 21(c) e



(a) Entropia

(b) Assortatividade

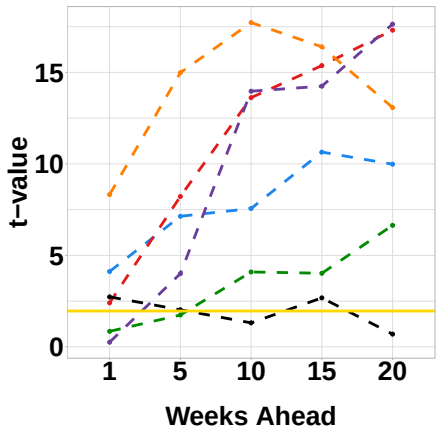
(c) Rotatividade

(d) Clusterização

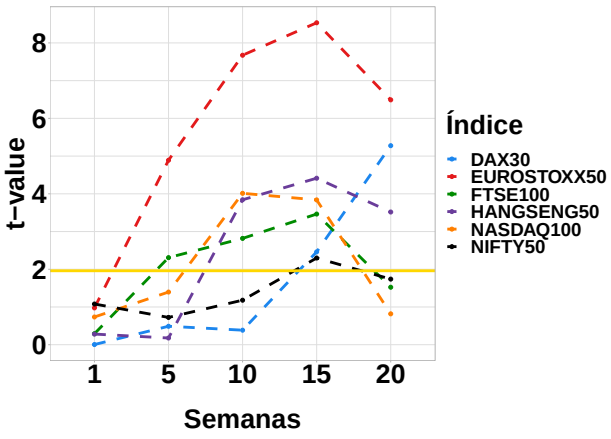
(e) Modularidade

(f) Diâmetro

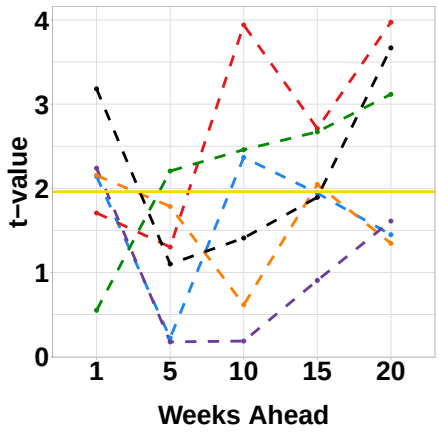
Figura 20 – **DAG - Variáveis de Rede vs. Previsibilidade da Estrutura de Mercado.** A figura apresenta o valor t de uma regressão linear múltipla usando variáveis de rede como variáveis explicativas e a AUC do modelo de aprendizado de máquina como resultado. Os resultados de cada índice de mercado são apresentados individualmente, considerando $L = 252$ e $h = \{1, 5, 10, 15, 20\}$. De acordo com t-student, os valores acima da linha amarela em $1,96$ apresentam relevância estatística.



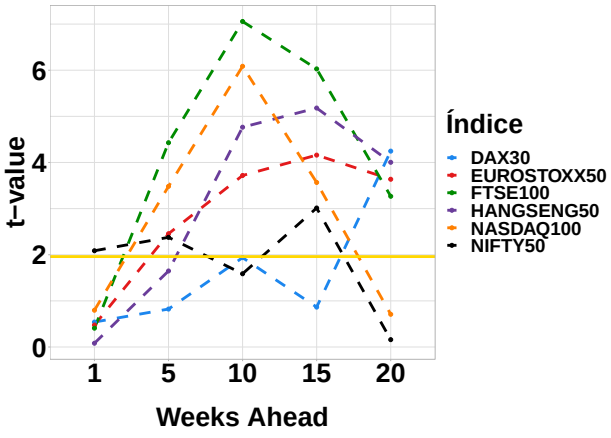
(a) Entropia



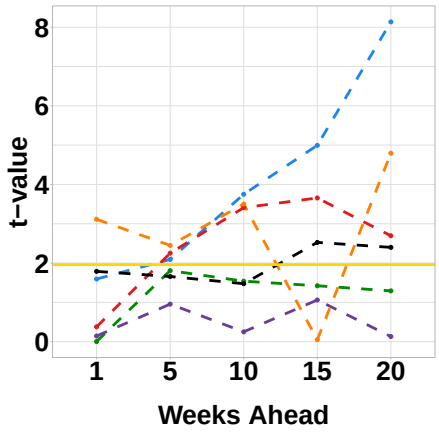
(b) Assortatividade



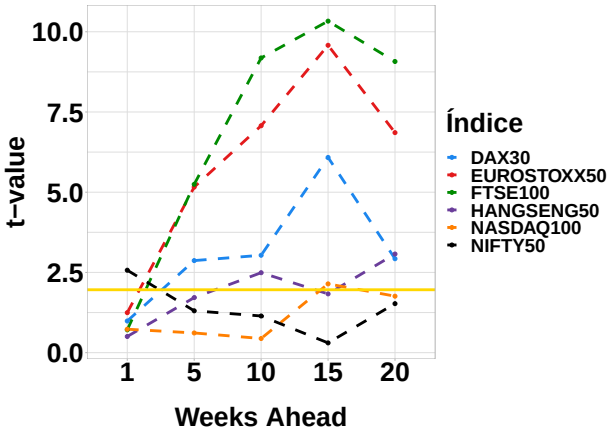
(c) Rotatividade



(d) Clusterização

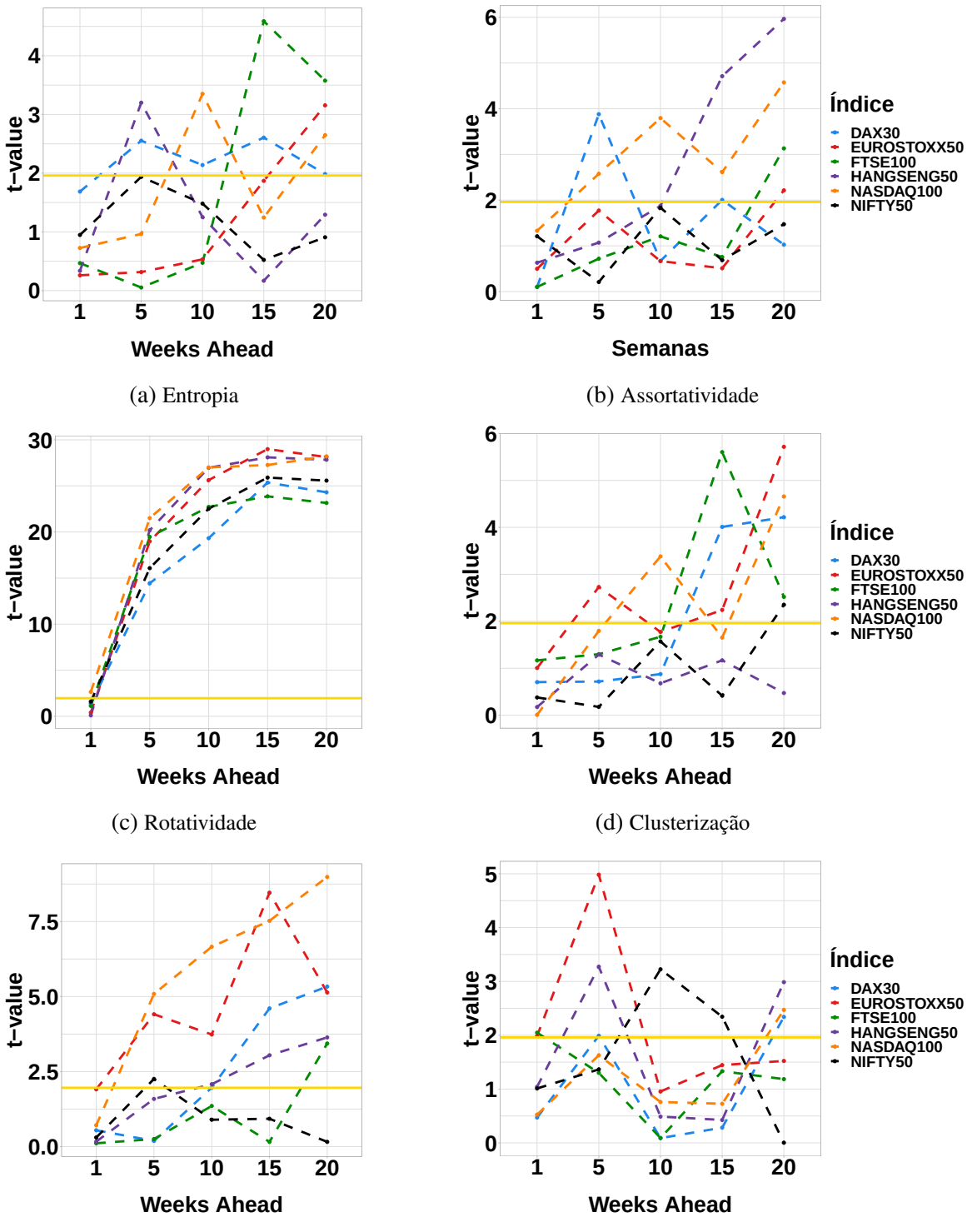


(e) Modularidade



(f) Diâmetro

Figura 21 – **DTN - Variáveis de Rede vs. Previsibilidade da Estrutura de Mercado.** A figura apresenta o valor t de uma regressão linear múltipla usando variáveis de rede como variáveis explicativas e a AUC do modelo de aprendizado de máquina como objetivo. Os resultados de cada índice de mercado são apresentados individualmente, considerando $L = 252$ e $h = \{1, 5, 10, 15, 20\}$. De acordo com *t-student*, os valores acima da linha amarela em $1,96$ têm relevância estatística.



(a) Entropia

(b) Assortatividade

(c) Rotatividade

(d) Clusterização

(e) Modularidade

(f) Diâmetro

Figura 22 – **DMST - Variáveis de Rede vs. Previsibilidade da Estrutura de Mercado.** A figura apresenta o valor t de uma regressão linear múltipla usando variáveis de rede como variáveis explicativas e a AUC do modelo de aprendizado de máquina como resultado. Os resultados de cada índice de mercado são apresentados individualmente, considerando $L = 252$ e $h = \{1, 5, 10, 15, 20\}$. De acordo com *t-student*, os valores acima da linha amarela em $1,96$ têm relevância estatística.

22(c) referentes aos métodos de filtragem de rede DAG, DTN e DMST, mostraram-se significativas para todos os índices de mercado para a maioria dos valores de h, com atenção especial para os resultados do método DMST, cujo comportamento dos *t-values* indicam que essa métrica é altamente relevante para a predição das redes filtradas através deste método.

O diâmetro da rede pode ser considerado uma medida de risco, pois as redes de baixo diâmetro permitem a rápida propagação de choques e, portanto, está relacionado ao risco sistêmico financeiro (DAS; MITCHENER; VOSSMEYER, 2018). Observamos que a relação do diâmetro da rede com o desempenho do modelo foi estatisticamente significativa, conforme apresentado nas Figuras 20(f), 21(f) e 22(f) particularmente em relação aos métodos DAG e DTN e para os mercados dos EUA e do Reino Unido, conforme medido pelos índices DAX30, NASDAQ100 e FTSE100.

Enquanto diferentes propriedades de rede mostraram padrões de relacionamento distintos com o desempenho de previsão do modelo, no geral, as propriedades de rede foram significativamente relacionadas à capacidade de previsão do modelo de aprendizado de máquina. Assim, observamos que a estrutura da rede e suas informações estruturais são importantes na determinação do capacidade de predição de sua estrutura futura.

3.4 Análise Comparativa entre Algoritmos de Aprendizagem de Máquina

Esta seção apresenta uma análise comparativa entre os resultados de previsão da estrutura do mercado utilizando diferentes algoritmos de aprendizagem de máquina. Os resultados anteriores, como descrito na Seção 3.2.2, foram desenvolvidos utilizando o algoritmo XGboost. Com intuito de avaliar como outros algoritmos se comportam quando submetidos ao mesmo problema, propusemos esta análise experimental utilizando os seguintes algoritmos de aprendizagem de máquina supervisionada:

k-Nearest Neighbors (KNN): este método de aprendizagem de máquina é bastante simples e um dos mais comuns utilizados para classificação. O algoritmo é baseado na verificação dos k vizinhos mais próximos à instância de teste e na identificação da classe majoritária dentre estes vizinhos, tal que a classe dos exemplos mais próximos é atribuída à instância de teste. A distância entre os exemplos é medida através de métricas de distância entre dois pontos, como a distância de Manhattan ou a distância Euclidiana, sendo essa a mais comum quando o conjunto de variáveis de entrada são contínuas.

Random Forest (RF): este método, também conhecido como "Florestas Aleatórias", é um algoritmo de aprendizagem de máquina amplamente utilizado, que possui a característica de entregar bons resultados mesmo sem a otimização dos parâmetros de execução. Além disso, é um método que requer pouco tempo para treinamento e que pode ter seus resultados

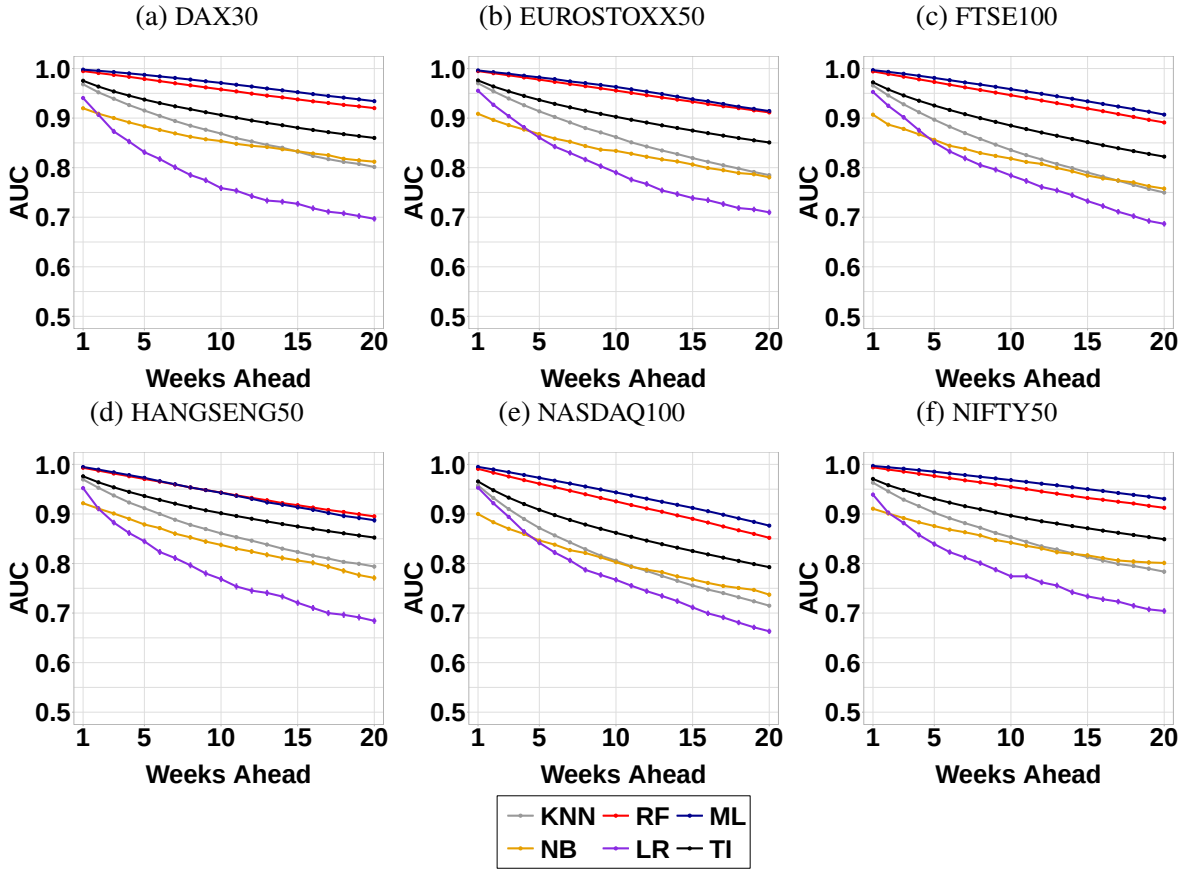


Figura 23 – **DAG - Comparação entre algoritmos de aprendizagem de máquina.** A figura mostra a média da AUC e o erro padrão do algoritmo ML em comparação com outros algoritmos de aprendizagem de máquina. ML e RF superam o algoritmo de *benchmark* TI. LR apresenta o pior resultado preditivo.

interpretados. Essas características o torna atrativo em diversos cenários envolvendo problemas de classificação e regressão. Sua execução se dá através da construção de um grande número de árvores de decisão, geradas de forma aleatória, e apresentando como saída da classificação a classe que representa a maioria das saídas das árvores individuais (BREIMAN, 2001).

Naive Bayes (NB): é um método probabilístico de aprendizagem de máquina utilizado para classificação, desenvolvido com baseado no teorema de Bayes (RISH, 2001). Esse tipo de algoritmo utiliza o conceito de probabilidade posteriori para realizar inferência sobre atributos de entrada que sejam independentes entre si, suposição esse que dificilmente é satisfeita. É um classificador escalável e que não requer parametrização para sua utilização.

Logistic Regression (LR): a regressão logística é um método de classificação que utiliza variáveis de entrada explicativas contínuas para descrever uma variável de saída categórica (binária).

Além dos algoritmos de aprendizagem de máquina supracitados, também incluímos nesta análise comparativa os resultados do algoritmo XGboost (denotado como ML) e o algoritmo

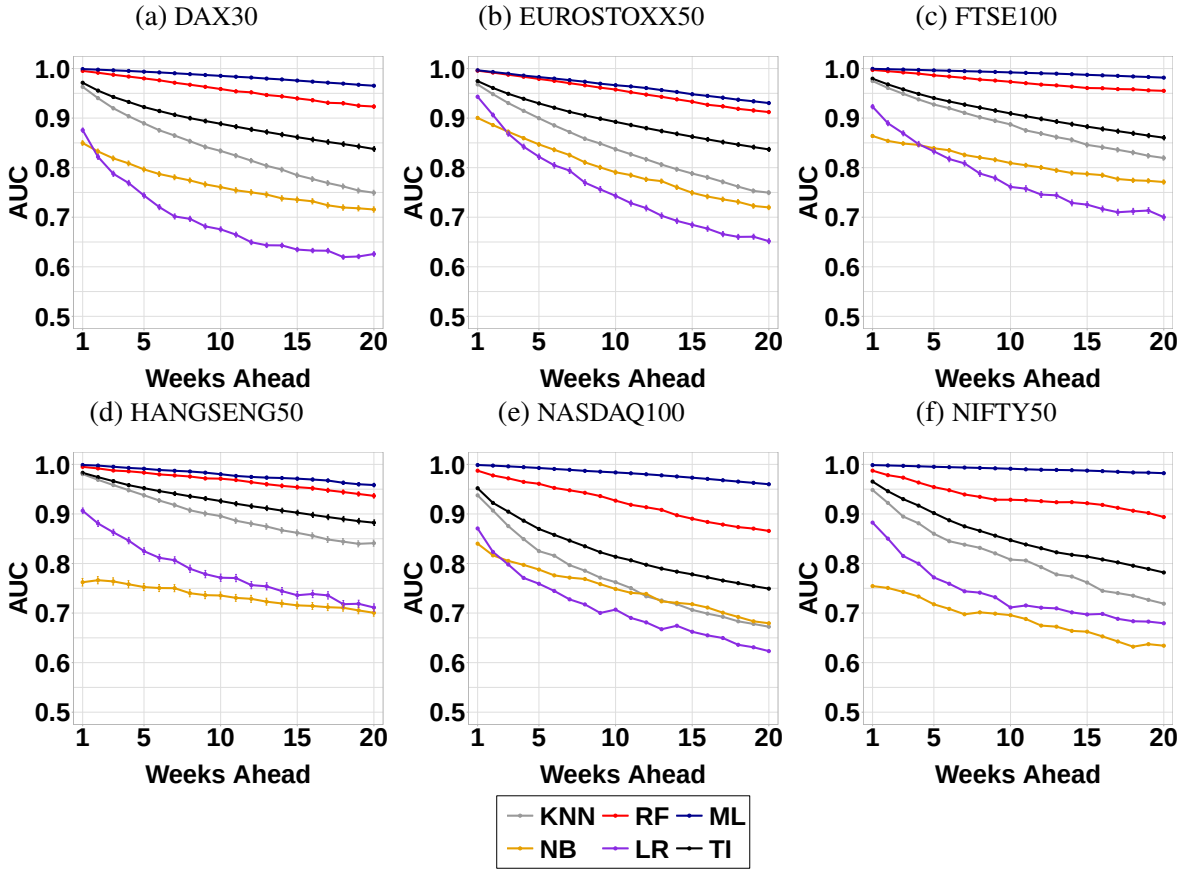


Figura 24 – **DTN - Comparação entre algoritmos de aprendizado de máquina.** A figura mostra a média da AUC e o erro padrão do algoritmo ML em comparação com outros algoritmos de aprendizagem de máquina. ML e RF superam o algoritmo de *benchmark* TI, sendo ML o melhor resultado dentre todos os índices de mercado.

de *benchmark* invariante no tempo (TI). Aplicamos a mesma metodologia descrita nas seções anteriores para prever a rede financeira $G(t+h)$, em que h é o número de semanas à frente, considerando $h = 1, 2, \ dots, 20$ semanas de negociação. Apresentamos aqui os resultados utilizando $L = 252$ dias de negociação para construir as redes financeiras. Os resultados relacionados a $L \in \{126, 504\}$ são apresentados no Apêndice A. Para cada intervalo de tempo h, calculamos a AUC média de cada método e seu respectivo erro padrão durante todo o período de teste, compreendido no período entre 5 maio 2007 e 18 de dezembro de 2019. Os parâmetros utilizados em cada algoritmo são apresentados no Apêndice A.

As Figuras 23, 24 e 25 apresentam as análises comparativas entre os algoritmos de aprendizagem de máquina utilizando os métodos de filtragem de rede DAG, DTN e DMST, respectivamente. Primeiramente, os resultados para o método DAG, apresentados na Figura 23, mostram que os algoritmos ML e RF superam o *benchmark* TI em todos os índices de mercado. Considerando os índices EUROSTOXX50 e HANGSENG50, os algoritmos ML e RF apresentam resultados bastante similares. Para os demais índices, o método ML supera o RF. Além disso, podemos observar que os algoritmos KNN, NB e LR não são capazes de apresentar resultados superiores ao *benchmark* TI. O comportamento destes algoritmos é similar em todos os índices

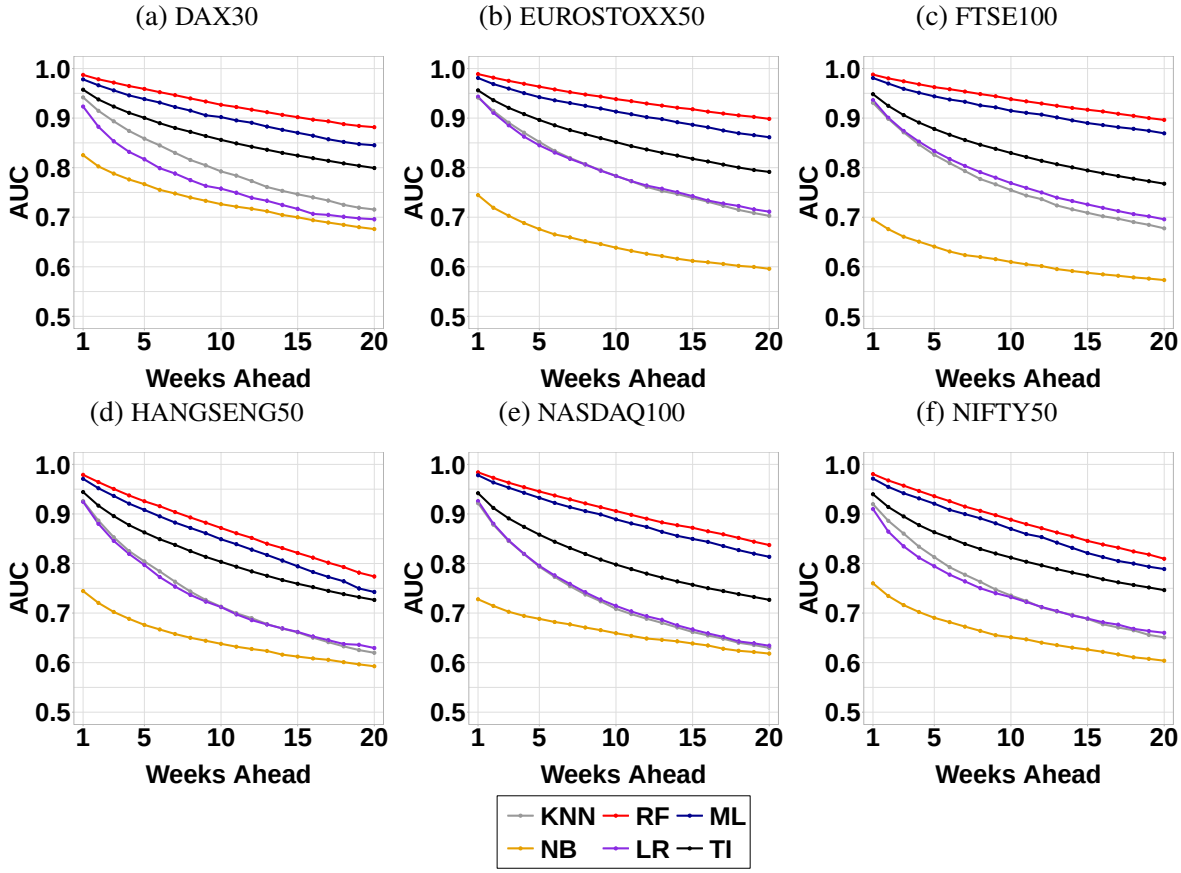


Figura 25 – **DMST - Comparação entre algoritmos de aprendizado de máquina.** A figura mostra a média da AUC e o erro padrão do algoritmo ML em comparação com outros algoritmos de aprendizagem de máquina. ML e RF superam o algoritmo de *benchmark* TI, sendo RF o melhor resultado dentre todos os índices de mercado.

de mercado, sendo LR o pior deles, principalmente para previsões das redes com $h > 5$ semanas à frente.

Em seguida, os resultados relacionados ao método de filtragem DTN são apresentados na Figura 24. Nesse conjunto de experimentos, os algoritmos ML e RF novamente superam o *benchmark*, sendo ML superior ao RF em todos os índices de mercado. Além disso, podemos observar que os algoritmos KNN, NB e LR não superam o *benchmark* TI, sendo NB pior que os demais algoritmos para os índices HANGSENG50 e NIFTY50. Finalmente, a Figura 25 apresenta os resultados relacionados ao método de filtragem de rede DMST. Neste conjunto de experimentos, os algoritmos ML e RF novamente superam o *benchmark* TI, sendo RF melhor que ML em todos os índices de mercado. Podemos observam também que, assim como para DAG e DTN, os algoritmos KNN, NB e LR não superam o algoritmo TI. Porém, neste conjunto de experimentos, o NB apresenta resultado inferior aos algoritmos KNN e LR para todos os índices de mercado.

Vale ressaltar que nosso objetivo aqui não é realizar uma busca por parâmetros ótimos durante a execução envolvendo o conjunto de testes, mas sim mostrar como diferentes algoritmos de aprendizagem de máquina se comportam no mesmo cenário experimental. Estes resultados

sugerem que a escolha do algoritmo de aprendizagem de máquina pode influenciar nos resultados de previsão de estrutura do mercado.

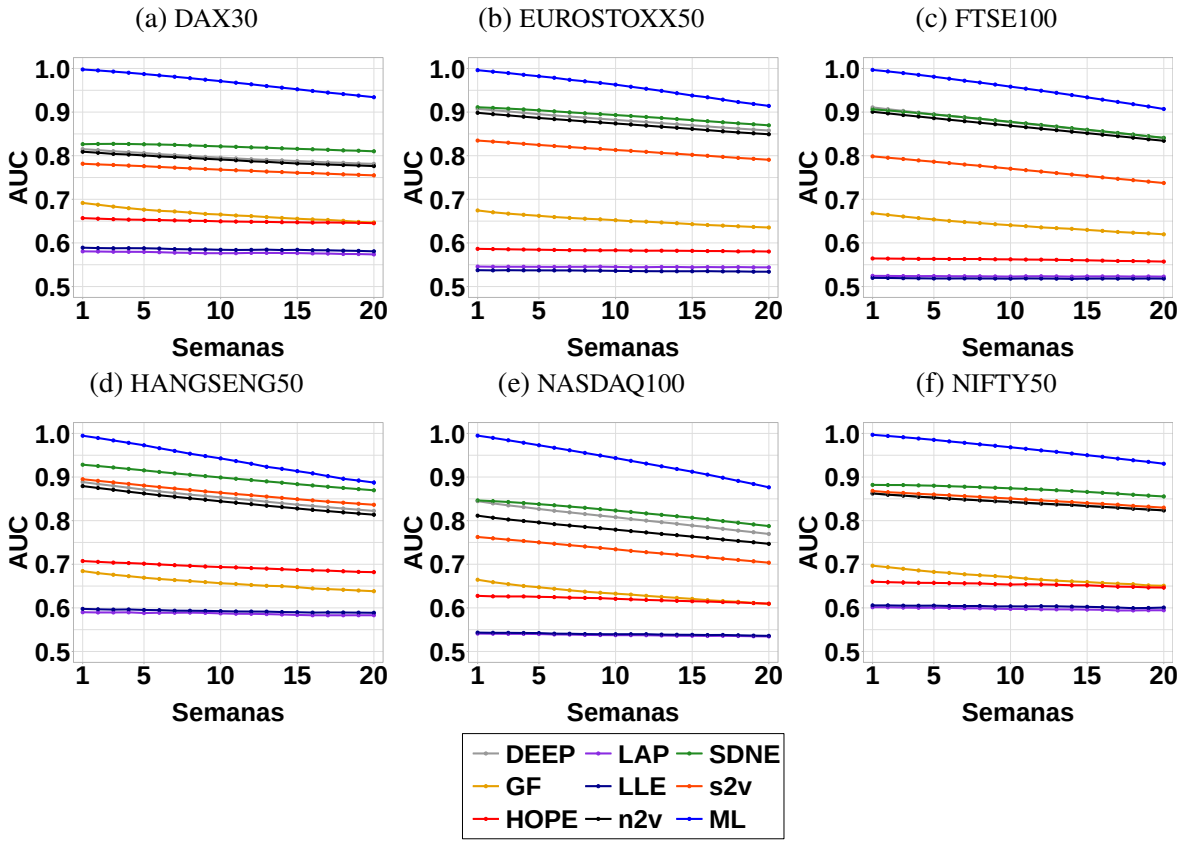


Figura 26 – **DAG - Comparação entre algoritmos de *embedding*.** A figura mostra a média da AUC e o erro padrão do algoritmo ML em comparação com outros algoritmos de *embedding*. ML supera os algoritmos de *embedding* em todos os índices de mercado. Algoritmos baseados em fatoração de matrizes apresentam os piores resultados, enquanto SDNE apresenta os melhores resultados.

3.5 Análise Comparativa entre Algoritmos de *Embedding*

Para finalizar o conjunto de experimentos relacionados à previsão de estruturas de mercado, apresentamos nessa seção uma análise comparativa utilizando algoritmos de *embedding* para previsão de links. Os algoritmos de embedding, assim como descrito na Seção 2.4.2, são métodos utilizados para extração de características das redes. Após criar o mapeamento dos nós da rede em espaços vetoriais n-dimensionais que representem as características dos nós, a previsão de links pode ser realizada através da análise de similaridade entre os vetores de informações dos nós. Esta análise tem como objetivo comparar os resultados do método proposto neste trabalho com algoritmos para previsão de links baseados em *embedding*. Os algoritmos de *embedding* utilizados nesta análise comparativa são: *Structural Deep Network Embedding* (SDNE); *Node2vec* (n2v); *Deep Walk* (DEEP); *Struc2vec* (s2v); *Laplacian Eigenmaps* (LAP); *Locally Linear Embedding* (LLE); *Graph Factorization* (GF) e *High-Order Proximity preserved*

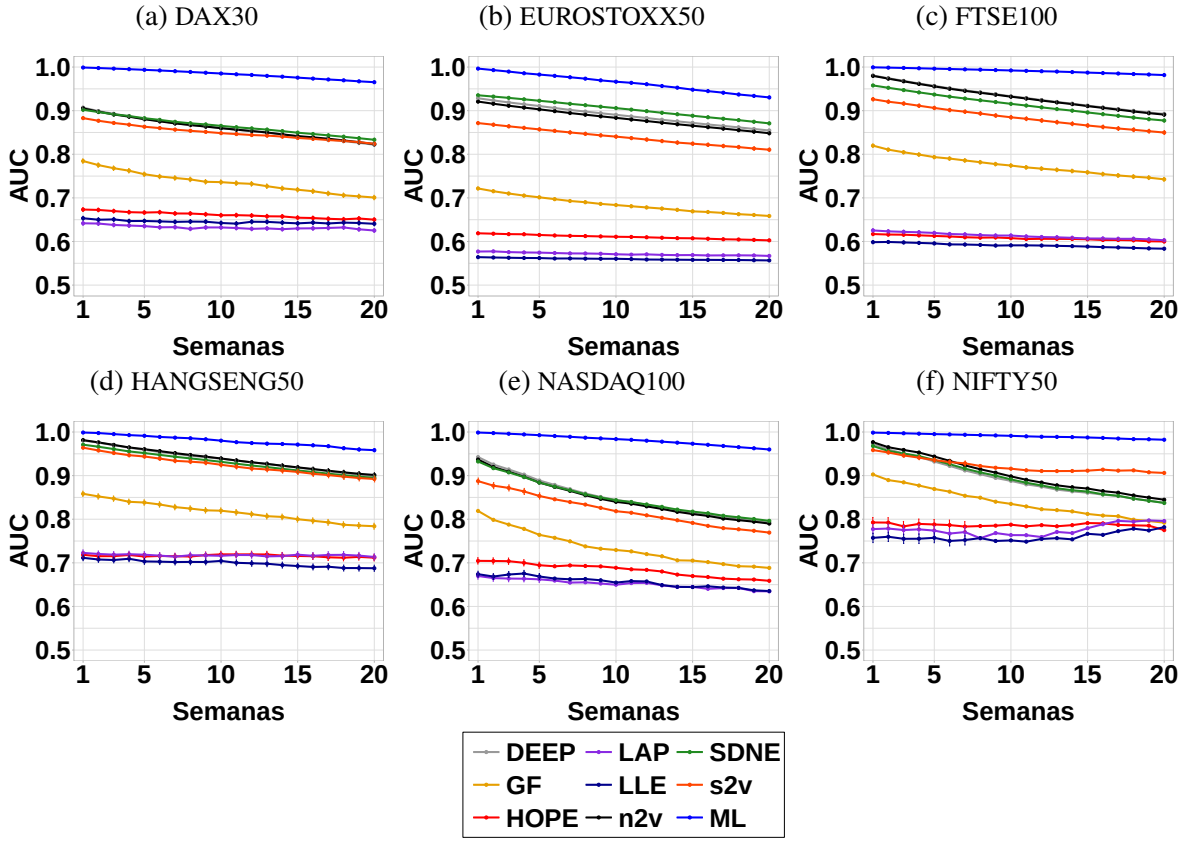


Figura 27 – **DTN - Comparação entre algoritmos de *embedding*.** A figura mostra a média da AUC e o erro padrão do algoritmo ML em comparação com outros algoritmos de *embedding*. ML supera os algoritmos de *embedding* em todos os índices de mercado. Algoritmos baseados em fatoração de matrizes apresentam os piores resultados.

Embedding (HOPE). Para maiores informações sobre estes métodos, veja a Seção 2.4.2. Os parâmetros utilizados em cada algoritmo são apresentados no Apêndice A.

Utilizamos a similaridade de cosseno, definida na Equação 3.11, para calcular a similaridade entre dois vértices i e j. Pares de vértices $\langle i, j\rangle$ com alta similaridade possuem maior probabilidade de estabelecerem uma conexão no futuro. Aplicamos a mesma metodologia descrita nas seções anteriores para prever a rede financeira $G(t+h)$, sendo $h = 1, 2, \ dots, 20$ semanas de negociação e $L = 252$ dias de negociação para construir as redes financeiras. Os resultados relacionados a $L \in \{126, 504\}$ são apresentados no Apêndice A. Calculamos a AUC média de cada método e seu respectivo erro padrão para cada valor de h durante todo o período entre 5 maio 2007 e 18 de dezembro de 2019.

As Figuras 26, 27 e 28 apresentam os resultados comparativos entre os algoritmos de *embedding* e o método de aprendizado de máquina ML proposto neste trabalho, considerando os três métodos de filtragem de rede DAG, DTN e DMST, respectivamente. Primeiramente, os resultados apresentados na Figura 26, relacionados ao método de filtragem DAG, sugerem que ML supera todos os algoritmos de *embedding* avaliados. Em geral, os algoritmos baseados em fatoração de matrizes LAP, LLE, GF e HOPE apresentam os piores resultados para todos

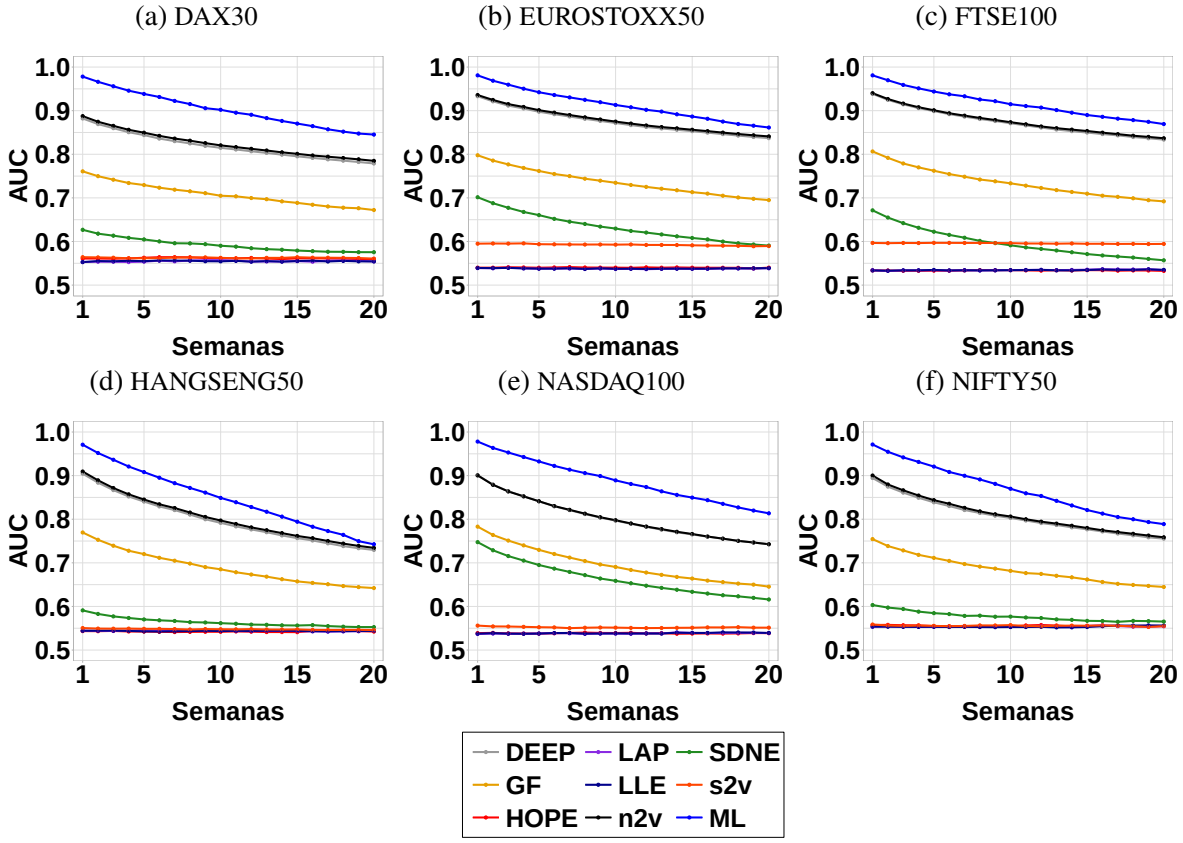


Figura 28 – **DMST - Comparação entre algoritmos de *embedding*.** A figura mostra a média da AUC e o erro padrão do algoritmo ML em comparação com outros algoritmos de *embedding*. ML supera os algoritmos de *embedding* em todos os índices de mercado. O algoritmo SDNE apresenta resultados similar ao piores valores.

os índices de mercado. Em contrapartida, o método SDNE apresente os melhores resultados dentre os algoritmos de *embedding*. Apesar de não superar o algoritmo ML, so algoritmos de *embedding* possuem comportamento quase constante mesmo com aumento do número h de semanas à frente. Em seguida, os resultados para o método de filtragem de rede DTN são apresentados na Figura 27. Novamente, os algoritmos de *embedding* não superam o ML. Os algoritmos baseados em fatoração de matrizes apresentam os piores resultados para todos os índices de mercado, sendo que a variância desses métodos é alta para os índices NIFTY50. Os algoritmos DEEP e n2v apresentam resultados muito similares em todos os índices de mercado. Finalmente, a Figura 28 apresenta os resultados relacionados ao método de filtragem de rede DMST. O ML supera os demais algoritmos em todos os índices de mercado. Diferentemente dos outros resultados, o algoritmo GF não está entre os piores resultados. Nesta configuração, o algoritmo SDNE apresenta resultados similar ao piores valores. Novamente, DEEP e n2v são bastante similares em todos os índices de mercado, inclusive muito próximos ao ML considerando $h = 20$ nos experimentos do índice HANGSENG50.

3.6 Considerações Finais

Neste capítulo, investigamos a previsão da estrutura do mercado de ações de vários mercados financeiros usando redes financeiras, modeladas a partir dos retornos das ações dos principais constituintes dos índices de mercado. A estrutura do mercado de ações foi modelada como redes, onde os nós representam os ativos e as arestas representam o relacionamento entre eles. Três métodos de filtragem baseados em correlação foram utilizados para criar redes de ações: *Dynamic Asset Graphs* (DAG), *Dynamic Threshold Networks* (DTN) e *Dynamic Minimal Spanning Tree* (DMST). Formulamos a previsão da estrutura de mercado como um problema de previsão de formação de links em redes, onde buscamos prever com precisão as arestas que estarão presentes nas redes futuras. Propusemos e avaliamos experimentalmente um modelo de aprendizado de máquina baseado em características das redes financeiras, derivadas de informações de nós e links, para prever a estrutura futura do mercado. Além disso, comparamos o modelo proposto com quinze algoritmos de *benchmark* e apresentamos uma análise qualitativa para explicar os resultados de previsão obtidos. Também utilizamos 4 algoritmos de aprendizagem de máquina distintos para avaliar a capacidade preditiva de outros métodos, diferente daquele utilizado como base.

Existem algumas limitações com relação à análise apresentada. Devemos ressaltar que utilizamos apenas ativos que permaneceram no índice de mercado durante todo o período de testes, o que limita a inserção e retirada de nós nas redes. Além disso, para redes com grande número de nós, o tempo de execução aumentou significativamente, tanto para extração de características derivadas da rede quanto para treinamento dos modelos de aprendizagem de máquina. Para redes com grande número de nós, o método pode se tornar custoso em termos computacionais.

CAPÍTULO

4

GERENCIAMENTO DE PORTFÓLIO ATRAVÉS DE PREVISÃO DE LINKS

O gerenciamento de portfólio no mercado de ações tem sido investigado por muitos pesquisadores ao longo de décadas. Esta classe de investimento procura alocar o capital do investidor em um subconjunto de ativos de maneira a manter um bom controle de entre risco e retorno financeiro. Vários algoritmos têm sido propostos para gerenciamento de portfólio. Neste capítulo, apresentamos um conjunto de experimentos utilizando resultados da previsão de links em redes de ações como entrada para modelos de gerenciamento de portfólio. O capítulo é organizado da seguinte maneira: primeiramente, a Seção 4.1 apresenta uma introdução e contextualização do problema abordado, assim como uma revisão da literatura; em seguida, a Seção 4.2 apresenta a metologia utilizada para resolução do problema proposto; a Seção 4.3 apresenta dois conjuntos de resultados relacionados à previsão de links em redes ponderadas e a otimização de portfólio utilizando previsão de links; finalmente, a Seção 4.4 apresenta as considerações finais deste capítulo.

4.1 Introdução

O gerenciamento de portfólio (carteira) de investimentos no mercados financeiro é o processo de seleção de um subconjunto de ativos que tem como objetivo manter um controle (*trade-off*) entre o risco e o retorno do investimento (MARKOWITZ, 1952). O processo de seleção de carteira no mercado de ações consiste em encontrar, dentre uma grande coleção de empresas, a participação (ou seja, proporção individual) de cada ação que minimize o risco do investimento com base um retorno esperado, ou maximize o retorno esperado da carteira assumindo um determinado risco (FREITAS; SOUZA; ALMEIDA, 2009). Este tópico tem sido investigado por pesquisadores de diversas áreas, como otimização, ciência da computação, estatística e economia.

Em geral, algoritmos de seleção de portfólio usam medidas de retorno e risco de um conjunto de ativos para tomar decisões. Em 1952, Markowitz propôs o primeiro modelo de otimização de portfólio conhecido como Análise Média-Variância (AMV). Neste modelo, o *trade-off* entre risco e retorno é usado para sugerir um subconjunto de ativos que deverão estar na carteira do investidor durante um período futuro. O capital deve ser investido durante um prazo bem definido e então outra sugestão de carteira é feita. Esse processo é conhecido como rebalanceamento do portfólio ou gerenciamento online de portfólio. Métricas comumente utilizadas para dar suporte à essa decisão incluem retorno esperado dos ativos e a variância/covariância dos retornos dos ativos. Além disso, a correlação entre os retornos dos ativos também é métrica importante para o gerenciamento de portfólio (CHRISTOFFERSEN *et al.*, 2011). Assim como apresentado no capítulo anterior, a correlação também é utilizada em vários estudos para criar uma perspectiva topológica do mercado, que caracteriza as estruturas do mercado de ações, também conhecidas como redes de ações ou redes financeiras (MARTI *et al.*, 2021; BONANNO; LILLO; MANTEGNA, 2001). Alguns autores sugerem o uso de informações topológicas derivadas de redes financeiras para gerenciamento de portfólio (PERALTA; ZAREEI, 2016; ZHAO *et al.*, 2018; LI *et al.*, 2019), que sugerem a criação de portfólios aplicando algoritmos de agrupamento ou medidas de centralidade em redes de ações. No entanto, apesar de sua importância, não encontramos trabalhos explorando a predição de links ponderados em redes de ações, nem tampouco sua utilização para melhoria do desempenho em algoritmos de otimização de portfólio - encontramos trabalhos usando previsão de preço e retorno para melhorar os resultados do gerenciamento de portfólio (MISHRA; PANDA; MAJHI, 2016).

Este estudo propõe uma nova abordagem para definir as constantes da clássica Análise Média-Variância (AMV) (*Mean-Variance Analysis*) de Markowitz (1952). As medidas de correlação de ativos são calculadas por meio da previsão de links em redes ponderadas. O método proposto fornece dados de entrada para o modelo matemático de Markowitz, que é utilizado como base para a otimização de portfólio. Para avaliar a eficiência do método proposto, propusemos uma série de experimentos utilizando dados de mercados reais, que foram divididos em duas partes: *(i)* aplicação de algoritmos de aprendizagem de máquina para induzir modelos capazes de prever a formação de links ponderados em redes de ações; *(ii)* aplicação de modelos de otimização de portfólio para calcular o retorno financeiro das carteiras recomendadas. Embora o algoritmo de predição de link ponderado tenha um certo erro em suas previsões, a taxa de acerto é razoável para a criação de portfólios com base nesses dados. Os resultados experimentais mostram que o modelo híbrido sugere melhores carteiras e melhora a qualidade das carteiras recomendadas. A seguir, apresentamos a definição do problema e os trabalhos relacionados à essa pesquisa.

4.1.1 Definição do Problema

A análise do comportamento e da interação entre ativos nos mercados de ações tem sido amplamente estudada na literatura (YANG *et al.*, 2014; BONANNO *et al.*, 2003; BONANNO *et al.*, 2004). Para descrevê-lo formalmente, considere i e j dois ativos distintos pertencentes ao conjunto V. Sejam S_i e S_j séries temporais relacionadas aos ativos i e j, respectivamente. Uma rede de ações ponderadas pode ser representada por um grafo $G^W = (V, A_W)$, onde V é o conjunto de ativos e A_W é o conjunto de pares de ativos $\langle i, j \rangle$ não ordenados, $\forall\, i, j \in V$. A relação entre i e j é medida através da correlação de Pearson, atribuindo um peso w a cada aresta em A. Um conjunto de redes de ações ponderadas ordenadas no tempo representam as redes de ativos dinâmicas (HOLME; SARAMÄKI, 2012).

Neste estudo, queremos responder a seguinte pergunta: dado um conjunto de grafos $G(1), G(2), \ldots, G(t)$ relacionados a uma sequência temporal de redes de ações calculadas até o tempo t e $G(t+h)$ uma rede futura cujos links foram previstos usando algoritmos de aprendizagem de máquina, a utilização de $G(t+h)$ pode melhorar o *trade-off* entre retorno e risco em modelos matemáticos de otimização de portfólio AMV?

Muitos trabalhos na literatura abordam o problema de gestão de portfólio. Neste estudo, estamos propondo uma relação entre a previsão de links em redes de ações e o modelo de otimização AMV proposto por Markowitz. Alguns estudos encontrados na literatura sugerem a relação entre redes de ações e Markowitz, mas não a previsão das redes como entrada do modelo de otimização. Onnela *et al.* (2003b) descobriram empiricamente que os ativos não centrais em uma Árvore Geradora Mínima (*Minimal Spanning Tree* - MST), calculada a partir da matriz de correlação de retorno de ações, são representados de forma proeminente no modelo de otimização AMV correspondente. Em (PERALTA; ZAREEI, 2016), os autores exploram a centralidade em redes complexas para melhorar o processo de seleção de carteiras por meio do direcionamento de um grupo de ações pertencentes a determinada região da rede do mercado de ações. Os autores estudam a relação entre portfólios de Markowitz e a centralidade das redes de ações e fornecem uma conexão entre os dois conceitos. Em (POZZI; MATTEO; ASTE, 2008) e (POZZI; MATTEO; ASTE, 2013), os autores detectaram que o desempenho da carteira é melhorado se os ativos constituintes forem selecionados entre os nós não centrais de uma MST. Giudici, Polinesi e Spelta (2021) combina a teoria da matriz aleatória com redes de ações, incluindo a centralidade da rede explicitamente na função objetivo de Markowitz. Hüttner, Mai e Mineo (2018) analisa a relação entre AMV e MST e mostram que a relação heurística entre a centralidade do grafo e o resultado da AMV não se origina de uma similaridade estrutural entre os dois mecanismos de seleção de portfolio, mas sim devido à características específicas das matrizes de correlação observadas. Li *et al.* (2019) propõe uma otimização de portfólio baseada na topologia de rede, usando correlação cruzada dos retornos de preços diários, para os mercados de ações americano e chinês. Levando em consideração a importância das matrizes de correlação e a possível presença de ruídos nessas matrizes, Pafka e Kondor (2004) apresenta uma

abordagem que permite uma investigação sistemática do efeito das diferentes fontes de ruído nas correlações e no contexto de gestão de risco. O trabalho de Zhao *et al.* (2018) analisa os mercados de ações em evolução no tempo usando a representação de rede temporal. Os autores também propõem uma ferramenta de seleção de portfólio usando centralidade temporal em redes de ações.

Além dessa abordagem envolvendo redes de ações, outros trabalhos usam previsão de preço e retorno para melhorar o gerenciamento de portfólio (MISHRA; PANDA; MAJHI, 2016). Em (FREITAS; SOUZA; ALMEIDA, 2009), uma rede neural é usada para prever o retorno futuro das ações. Os erros de previsão são usados como medida de risco. Bessler e Wolff (2015) também mostra que, para decisões de alocação de ativos, o uso de modelos capazes de prever o retorno é melhor do que usar médias históricas. Mesquita, Valle e Pereira (2020) propõe a utilização de previsões diárias das matrizes de covariância, utilizadas como entrada para algoritmos de otimização de portfólio para alocação de pesos de maneira ideal.

4.2 Materiais e Métodos

Essa seção apresenta as principais etapas para desenvolvimento do método de gestão de portfólio utilizando previsão de links em redes de ações. O método proposto nesta pesquisa é apresentado na Figura 29. O método pode divido nas seguintes etapas: *(i)* utilização de algoritmos de aprendizagem de máquina para prever links ponderados em redes de ações; *(ii)* execução do modelo matemático de otimização de portfólio, usando como entrada as redes de ações previstas anteriormente.

Inicialmente, calculamos a matriz de correlação entre todos pares de ações com base na série diária de preços de fechamento. Em seguida, extraímos um conjunto de características da rede, usados como atributos de entrada para o modelo de aprendizado de máquina, através da extração de características em nível de nó e em nível de link, conforme descrito na Seção 4.2.2.1. Aplicamos um modelo de aprendizado de máquina, descrito na Seção 3.2.2, para prever redes ponderadas. Finalmente, utilizamos o resultado da previsão das redes como entrada no modelo de otimização de portfólio para definir os pesos ideais de cada ativo na composição da carteira.

As traditional portfolio management algorithms, we are interested in reorganize the portfolio every day. Thus, the data were processed to obtain daily price time series. Note that our approach can be easily adapted for weekly, monthly or intraday strategies, according to the main purpose of the investor. It is important to emphasize that these data include an election period with high stock price variations.

4.2.1 Redes de Ativos Ponderadas

Neste trabalho, aplicamos uma versão modificada do método proposto por Mantegna e Stanley (1999) para criar redes financeiras dinâmicas ponderadas. Neste método, os nós da

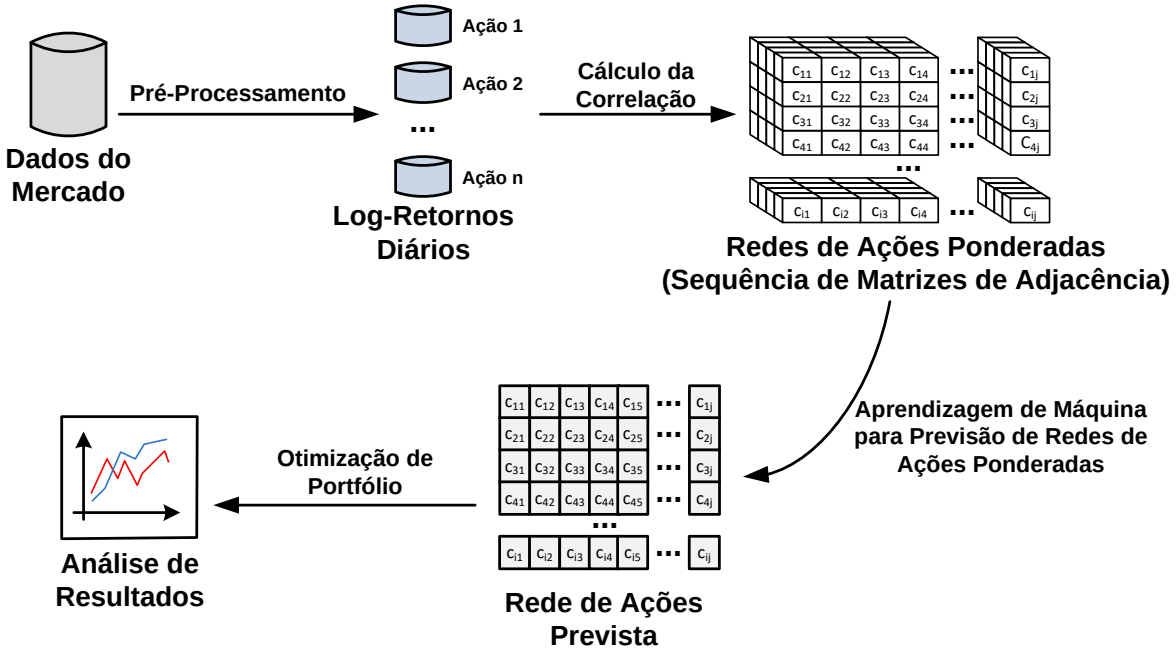


Figura 29 – **Método proposto para otimização de portfólio.** A figura apresenta a metodologia pra gerenciamento de portfólio utilizando previsão de links em redes ponderadas. Primeiramente, calculamos a matriz de correlação entre todas as séries do log-retorno do preço das ações. Em seguida, utilizamos o método baseado em aprendizagem de máquina para previsão de redes de ações futuras. Finalmente, utilizamos esses resultados de previsão como entrada para o modelo de otimização de portfólio.

rede representam ativos e as arestas representam a relação entre eles. Essa relação é baseada na correlação de séries temporais de preços. Sejam S_i e S_j duas séries temporais com comprimento L relacionados a dois ativos distintos i e j, $i, j \in V$, onde V é o conjunto de ações disponíveis. Através dessas séries, obtemos a série de log-retornos usando a seguinte equação:

$$Y_i = ln(P_i(t)) - ln(P_i(t-1)) \tag{4.1}$$

onde $P_i(t) \in S_i$ é o preço de fechamento do ativo i no dia t. Em termos de definições de mercado, a média desse tipo de série temporal tende a ser próxima de zero. Em seguida, utilizamos o coeficiente de Pearson (PEARSON, 1895) para medir a correlação entre as séries temporais de Y_i e Y_j referentes a todos os pares de ações $\langle i, j \rangle$, $\forall\, i, j \in V$:

$$\rho_{ij} = \frac{cov(Y_i, Y_j)}{\sqrt{var(Y_i) \cdot var(Y_j)}} \tag{4.2}$$

A correlação entre os ativos é atribuída aos pesos das aresta A^W no grafo $G^W(V, A^W)$. Usamos a correlação ρ_{ij} entre todos os pares possíveis de ações pertencentes a um determinado

índice de mercado para criar uma matriz de adjacência C. Por definição, os elementos ρ_{ij} estão na faixa de -1 a 1, onde -1 corresponde à anti-correlação perfeita, 1 corresponde à correlação perfeita e 0 corresponde à ausência de correlação. Considere que C representa uma rede de ações não direcionada ponderada completa. Os modelos de otimização AMV utilizam matriz de covariância similar como entrada para sugerir pesos ótimos para investimento, com *trade-off* aceitável entre retorno e risco. A matriz de correlação é construída dividindo a série temporal de log-retornos $S_i(t)$ em janelas contínuas de tamanho L dias de negociação, com δT dias de negociação entre duas janelas consecutivas (intervalo de tempo). Relatamos os resultados para $L \in \{63, 126, 252, 504\}$ e $\delta T = 5$ dias de negociação.

4.2.2 Previsão de Links Ponderados Usando Aprendizado de Máquina

Esta seção descreve o método proposto para previsão de links ponderados em redes de ativos. Neste estudo, ao invés de utilizarmos matriz de correlação conhecida no dia t, propomos utilizar uma estimativa da matriz de correlação $t+h$ como forma de melhorar os resultados do gerenciamento de portfólio. Nosso primeiro problema principal é criar algoritmos de aprendizagem de máquina capazes de induzir modelos que possam prever o peso de todas as arestas em uma futura rede de ações $G^W(t+h)$.

Neste trabalho, abordamos o problema de previsão de links ponderados como uma tarefa de regressão. Para tal, propusemos a utilização de algoritmos de aprendizagem de máquina supervisionados que utilizam como entrada atributos derivados da própria rede. Cada exemplo no conjunto de dados é rotulado com o valor de correlação entre um par de ações i e j para um determinado período futuro. A seguir, apresentamos o conjunto de atributos derivados da rede utilizados como de entrada para treinar os algoritmos de aprendizagem de máquina.

Utilizamos o algoritmo XGboost (CHEN; GUESTRIN, 2016) como modelo principal para previsão de links ponderados. O XGboost é um algoritmo de aprendizagem de máquina rápido, altamente eficaz e amplamente utilizado. Não realizamos uma busca exaustiva pelos parâmetros do modelo porque este não é o nosso principal objetivo aqui. Nossa intenção é mostrar o quão preditivo um algoritmo de aprendizagem de máquina pode ser usando o conjunto de atributos que propomos. O conjunto de parâmetros do modelo utilizados nos experimentos pode ser visto no Apêndice A.

Assim como descrito na Seção 3.2.2, utilizamos uma abordagem de janela deslizante para avaliar os resultados preditivos do método proposto. Separamos os dados em dois subconjuntos de treinamento e teste levando em consideração a sequência temporal dos dados. O conjunto de treinamento inclui dados entre 1 março 2005 a 30 maio 2007 e o conjunto de teste varia entre 30 de maio de 2007 a 18 de dezembro de 2019. Seja L o tamanho das séries temporais dos log-retornos utilizadas para criar as redes de ações e t o tempo atual. Criamos o conjunto de treinamento utilizando atributos derivados das k redes anteriores, sendo $k = 30$. O conjunto de teste contém dados da rede $G(t)$, sendo $G(t+h)$ é a rede alvo. Apresentamos resultados

relacionados a $1 \leq h \leq 20$.

4.2.2.1 Características Derivadas das Redes

As características derivadas das redes são calculados em cada iteração usando métricas de rede complexas. As métricas podem ser divididas de acordo com o nível de análise a ser realizado: no nível do nó, onde os nós representam ativos, e no nível do link (OLIVEIRA; GAMA, 2012). Para criar cada exemplo para o conjunto de dados de aprendizado supervisionado, as métricas relacionadas a i e j são inseridas para ambos os nós. Métricas relacionadas com links são inseridas calculando as medidas entre os nós i e j (NARASIMHAN, 2015). Considere $|i|$ como o grau do nó ou número de arestas. Diferentemente do capítulo anterior, estamos interessados em prever os pesos das arestas. Para isso, adaptamos o conjunto de características da rede previamente definidos para enquadrar com a definição deste problema. As Tabelas 4 e 5 apresentam as características derivadas das redes em nível de nó e em nível de link, respectivamente.

Tabela 4 – **Atributos dos Nós.** As características descritas nesta tabela foram calculadas para o nó i, $\forall\, i \in V$ para um dado grafo $G(V,A)$. Considere N_i como sendo o conjunto de vértices adjacentes (vizinhos) do nó i.

Nome	Definição
Grau Ponderado do Nó	$deg_w(i) = \sum_{j \in N_i} w_{<i,j>}$, onde $w_{<i,j>}$ é o peso da aresta $e(i,j)$
Média do Grau Ponderado dos Vizinhos	$avg_w(i) = \frac{\sum_{j \in N_i} \lvert j \rvert * w_{<i,j>}}{\lvert i \rvert}$
Propensão de i Aumentar seu Grau	$\gamma(i) = \frac{\lvert i \rvert}{deg_w(i)}$
Betweenness do Nó	$b(v) = \sum_{i,j \in V \setminus v} \frac{\sigma_{ij}(v)}{\sigma_{ij}}$, onde $\sigma_{ij}(v)$ é o número de caminhos mínimos entre i e j que possam pelo vértice v e σ_{ij} é o número de caminhos mínimos de i para j, $\forall i, j \in V$
Autovetor do Nó	$ne(i) = x_i \frac{1}{\lambda} \sum_{j=1}^{n} d_{ij} x_j$, onde d_{ij} representa uma entrada na matriz de adjacência C, λ denota o maior autovalor, x_i e x_j representam a centralidade dos nós i e j, respectivamente

Tabela 5 – **Atributos dos Links:** As características descritas nessa tabela foram calculadas entre os nós i e j, $\forall\,(i,j) \in A$ para um dado grafo $G(V,A)$. Considere N_i e N_j como sendo os conjuntos de vértices adjacentes dos nós i e j, respectivamente.

Nome	Definição
Peso da Aresta (*)	$PA(i,j) = \rho_{ij}$, onde $\rho_{i,j}$ é o coeficiente de Correlação de Pearson entre as séries de log-retornos das ações i e j
Common Neighbors Ponderado	$\sum_{z \in N_i \cap N_j} PA(i) + PA(j)$,
Adamic-Adar Ponderado	$AAP(i,j) = \sum_{k \in N_i \cap N_j} \frac{PA(i)+PA(j)}{log[1+s(k)]}$, onde $s(x) = \sum_{z \in N_x} PA(x,z)$ e N_x é o conjunto de vértices adjacentes do nó k
Betweenness da Aresta	$B(i,j) = \sum_{i,j \in V} \frac{\sigma_{ij}(e)}{\sigma_{ij}}$, onde $\sigma_{ij}(e)$ é o número de caminhos mínimos entre i e j que cruzam a aresta e e $\sigma_{i,j}$ é o número total de caminhos mínimos de i para j, $\forall i,j \in A$
Mesma Comunidade	$SC(i,j) = \begin{cases} 1, & \text{se } i \text{ e } j \in \text{ mesma comunidade,} \\ 0, & \text{se } i \text{ e } j \notin \text{ mesma comunidade,} \end{cases}$ sendo o algoritmo de Blondel *et al.* (2008) utilizado para detecção de comunidades.
Preferential Attachment	$PA(i,j) = deg_w(i) * deg_w(j)$, onde $deg_w(i)$ e $deg_w(j)$ representam o grau ponderado dos vértices i e j

4.2.3 Modelo Matemático para Otimização de Portfólio

Em 1952, Markowitz propôs o primeiro modelo de Análise Média-Variância (AMV) que serviu de base para a Teoria Moderna do Portfólio (TMP) para gestão e alocação de recursos (MARKOWITZ, 1952). Nessa teoria, um investidor deseja distribuir uma riqueza inicial em um conjunto de investimentos de forma a minimizar o risco e dado um retorno esperado, ou seja, maximizar o retorno. Esses dois objetivos são conflitantes, pois se houver risco mínimo de investimento e máximo retorno, a decisão é trivial. Normalmente, quanto maior o risco, maior o retorno esperado de um investimento. A grande inovação introduzida pela TMP é a utilização da variância como métrica de risco e sua utilização em um modelo matemático com objetivo único. Para definição matemática do modelo MV, vamos considerar a seguinte notação matemática:

N - conjunto de ativos disponíveis para investimento, de tamanho n;

μ_i - retorno esperado do ativo i, sendo $\mu \in \mathbb{R}^n$;

$\sigma_{i,j}$ - matriz de covariância entre os ativos i e j, sendo $\sigma \in \mathbb{R}^{n \times n}$;

$\mathscr{E}$ - retorno desejado do portfólio;

x_i - fração do ativo i investida no portfólio ótimo, sendo $x \in \mathbb{R}^n$;

No modelo Média-Variância clássico, definido na Equação 4.3, o objetivo é minimizar o risco (variância) sujeito a um nível de retorno desejado ($\mathscr{E}$):

$$\begin{aligned} \text{Minimize} \quad & \sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j \sigma_{ij} \\ \text{Sujeito a:} \quad & \sum_{i=1}^{n} x_i \mu_i = \mathscr{E}, \\ & \sum_{i=1}^{n} x_i = 1, \\ & x_i \geq 0, \ \forall\, i = 1, ..., n. \end{aligned} \tag{4.3}$$

onde a soma dos pesos dos ativos x_i é 1, $\forall i \in N$, representando que o capital todo deve ser alocado, sendo $x_i \geq 0$, representando que o fração/peso investido em cada ativo deve ser positiva.

Esse mesmo modelo pode ser descrito através da sua formal dual, definido da seguinte forma:

$$\begin{aligned} \text{Maximize} \quad & \sum_{i=1}^{n} x_i \mu_i \\ \text{Sujeito a:} \quad & \sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j \sigma_{ij} = \sigma^2, \\ & \sum_{i=1}^{n} x_i = 1, \\ & x_i \geq 0, \ \forall\, i = 1, ..., n. \end{aligned} \tag{4.4}$$

sendo σ^2 a variância máxima do portfólio. No modelo apresentado na Equação 4.4, o objetivo é maximizar o retorno $\mathscr{E}$ sujeito a um nível de risco máximo. A fronteira eficiente de portfólios, que representa a máxima performance obtida por portfólios eficientes e o *trade-off* entre risco e retorno esperado, pode ser calculada minimizando o risco para um vários níveis de retorno esperado na Equação 4.3 ou maximizando o retorno esperado para vários níveis de risco, definido na Equação 4.4.

Apesar das formulações matemáticas 4.3 e 4.4 serem intuitivas, no mundo real pode ser difícil utilizar essas abordagens com objetivo único, uma vez que ambas as formulações possuem

interdependência e os objetivos são integrados. Para isso, pesquisadores desenvolveram formas de lidar com modelos multi-objetivos, ao invés do objetivos únicos. Uma forma comumente utilizada para resolver problemas multi-objetivos é combinar os vários objetivos em apenas um através de soma ponderada, onde o peso associado a cada objetivo indica sua prioridade para a solução final (KALAYCI; ERTENLICE; AKBAY, 2019). citeonlineKallberg83 examina uma formulação alternativa para o problema usando medidas de aversão ao risco absoluto e relativo. Considere que u é uma função de utilidade de Von Neumann-Morgenstern. A aversão absoluta ao risco é definida por $R_a = \frac{u''(w)}{u'(w)}$, onde w é a valorização da carteira. O resultado da formulação é apresentado a seguir:

$$\begin{aligned} \text{Maximize} \quad & \sum_{i=1}^{n} x_i \mu_i - \frac{R_a}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} x_i x_j \sigma_{ij} \\ \text{Sujeito a:} \quad & \sum_{i=1}^{n} x_i = 1 \\ & x_i \geq 0, \; \forall \, i = 1, ..., n. \end{aligned} \tag{4.5}$$

Para o modelo 4.5, a função bi-objetivo tem como finalidade maximizar o retorno esperado do portfólio e minimizar a variância da carteira, que é ponderada pela aversão ao risco $R_a/2$. De acordo com (KALLBERG; ZIEMBA, 1983), as funções de utilidade (R_a) negativas geram carteiras muito avessas ao risco. Assim, a fronteira eficiente pode ser obtida para valores de $R_a > 0$. Resultados empíricos indicam que: as carteiras de risco possuem valores de $R_a \leq 2$; carteiras de risco moderado têm $2 \leq R_a \leq 4$; a carteira avessa ao risco tem $R_a \geq 4$.

Chang *et al.* (2000) reformula o modelo de Markowitz multiobjetivo através de um método de soma ponderada de objetivo único, definido da seguinte forma (KALAYCI; ERTENLICE; AKBAY, 2019):

$$\begin{aligned} \text{Minimize} \quad & \lambda \left[\sum_{i=1}^{n} \sum_{j=1}^{n} x_i x_j \sigma_{ij} \right] - (1 - \lambda) \left[\sum_{i=1}^{n} x_i \mu_i \right] \\ \text{Sujeito a:} \quad & \sum_{i=1}^{n} x_i = 1 \\ & x_i \geq 0, \; \forall \, i = 1, ..., n. \end{aligned} \tag{4.6}$$

sendo λ uma variável para controla o *trade-off* entre retorno e risco, $0 \leq \lambda \leq 1$. Importante observar que $\lambda = 0$ produz um modelo que busca a maximização do retorno com risco desprezado e $\lambda = 1$ transforma o objetivo da função somente na minimização do risco, conhecido como Portfólio de Variância Mínima Global (PVMG).

O modelo apresentado na Equação 4.6 será utilizado nos experimentos envolvendo otimização de portfólio, dado sua característica genérica para abordar risco e retorno. Vale

ressaltar que este modelo não é capaz de produzir todas as soluções ótimas de Pareto com superfícies não convexas (ZITZLER, 1999; KALAYCI; ERTENLICE; AKBAY, 2019)

4.2.4 Otimização de Portfólio Utilizando Previsão de Links

Conforme descrito anteriormente, este estudo tem como objetivo a utilização da previsão de links em redes de ações como forma de melhorar resultados de otimização de portfólio usando o modelo AMV. Para isso, o resultado da previsão de links ponderados usando algoritmos de aprendizagem de máquina é utilizado como entrada para o modelo AMV, descrito na Equação 4.6. Como o resultado da previsão de links é uma matriz de correlação C e o modelo AMV utiliza a matriz de covariância $\sigma_{i,j}$ como entrada, nós realizamos a conversão da previsão de links através da seguinte equação:

$$\sigma_{i,j} = c_{i,j} * \sigma_i * \sigma_j \tag{4.7}$$

onde $c_{i,j} \in C$ é o peso do link previsto (correlação) e σ_i e σ_j o desvio padrão dos ativos i e j. O valor de σ_i é obtido com base no log-retorno esperado μ_i, que é calculado através da média dos L log-retornos do ativo i. Convencionamos a utilização do tamanho L tanto para criação das redes de ações quando para cálculo do retorno estimado. As informações da matriz de covariância e o retorno estimado são usadas como entrada para o algoritmo de otimização de portfólio, descrito na Equação 4.6. O modelo de otimização foi resolvido através de programação quadrática usando o software CPLEX da IBM, amplamente utilizado para resolver problemas de otimização.

Para realização dos experimentos, utilizamos o conjunto de dados apresentado na Seção 3.2.5. Foram utilizados dados reais, compreendidos no período entre 12 maio 2006 e 18 de dezembro de 2019, relacionados aos índices de mercado DAX30, EUROSTOXX50, FTSE100, HANGSENG50, NASDAQ100 e NIFTY50. Para cada índice de mercado, foram selecionadas as ações que permaneceram no índice durante todo o período de testes, para que não haja a inserção de novos nós nas redes. Caso essa restrição não seja aplicada, além da previsão dos links, seria necessário a prever quando um nó entraria na rede e esse não é nosso foco.

Para avaliação dos resultados relacionados à aprendizagem de máquina, utilizamos as seguintes métricas:

Root Mean Squared Error (RMSE): Raiz Quadrada do Erro-Médio

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2} \tag{4.8}$$

sendo y_i o valor previsto e $\widehat{y_i}$ o valor real esperado.

Mean Absolute Error (MAE): Erro Absoluto Médio

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y}_i| \tag{4.9}$$

sendo y_i o valor previsto e $\widehat{y}_i$ o valor real esperado.

Para avaliação dos resultados relacionados ao gerenciamento de portfólio, utilizamos métricas de retorno e risco comumente utilizadas para descrever o comportamento de estratégias no mercado financeiro:

Retorno Acumulado (RA): representa o retorno financeiro acumulado pela estratégia durante um período de investimento de tamanho T,

$$\prod_{i=0}^{T}(1+r_t) \tag{4.10}$$

onde r_t é o retorno do portfólio no tempo t.

Sharpe Ratio (SR):

$$Sharpe\ Ratio = \frac{R_P - R_f}{\sigma_P} \tag{4.11}$$

onde R_P é o retorno do portfólio, R_f é a taxa de retorno sem risco e σ_P é o desvio padrão dos retornos do portfólio.

Máximo *DrawDown* (MaxDD): representa o percentual de perda máxima atingido em um intervalo de tempo consecutivo durante o período de investimento da estratégia.

Para avaliar o desempenho do algoritmo de aprendizagem de máquina na previsão dos links ponderados, comparamos o método proposto com o algoritmo de *benchmark* invariante no tempo (*Time Invariant* - TI), que assume que a formação dos pesos dos links é invariante no tempo. Assim como descrito no capítulo anterior, esse método é tradicionalmente utilizado em análises de gerenciamento de risco, que geralmente usam uma matriz de covariância retroativa para estimar o risco do portfólio. Este algoritmo foi escolhido principalmente por conta de seus resultados satisfatórios apresentados no capítulo anterior, que explorou em profundidade a previsão de links em redes de ações.

Para avaliar o desempenho do método proposto para gerenciamento de portfólio, utilizamos dois algoritmos como base de comparação. O primeiro, denominado Igualmente Distribuído (ID), representa os resultados de investimento utilizando um carteira que distribui os pesos uniformemente entre os n ativos do índice, ou seja, a fração do portfólio que será investido em cada ação é dado por $1/n$. O segundo, denominado MV, utiliza a matriz de covariância calculada com dados do passado como entrada para o método AMV.

Algumas suposições sobre o gerenciamento de portfólio devem ser consideradas para garantir a integridade dos resultados experimentais (LI; HOI; GOPALKRISHNAN, 2011):

- A liquidez do mercado é grande o suficiente para que as operações de compra e venda não causem impacto ou qualquer variação no preço das ações;
- A liquidez do mercado é grande o suficiente para que o preço de compra e venda das ações seja o mesmo do preço do último negócio do dia (preço de fechamento);
- Cada ação pode ser dividida em frações proporcionais à quantidade necessária para compra e venda, de acordo com a porção sugerida em cada realocação da carteira;
- Não há custos operacionais e transacionais para executar as ordens de compra e venda.

4.3 Resultados e Discussão

Nesta seção, apresentamos os resultados experimentais agrupados em dois conjuntos: *(i)* resultados relacionados à predição de links ponderados em redes de ações através de algoritmos de aprendizagem de máquina e *(ii)* resultados financeiros relacionados aos modelos de gerenciamento de portfólio.

4.3.1 Previsão de Links Ponderados

Nesta seção, apresentamos um conjunto de resultados experimentais relacionados à previsão da links ponderados usando aprendizagem de máquina. Investigamos o desempenho preditivo do método proposto em diferentes cenários, comparando-o com o algoritmo de *benchmark* TI. Utilizamos uma abordagem de aprendizado de máquina para prever a rede financeira ponderada $G(t+h)$, em que h é o número de semanas à frente, considerando $h = 1, 2, \ dots, 20$ semanas de negociação. Apresentamos resultados utilizando $L = 252$ dias de negociação como sendo o tamanho das janelas deslizantes para construir as redes financeiras. Os resultados para $L = \{63, 126, 504\}$ dias de negociação de são relatados no Apêndice B.

As Figuras 30 e 31 mostram as medidas MAE e RMSE, respectivamente, do método de aprendizado de máquina proposto em comparação com o algoritmo de *benchmark* TI para todos os índices de mercado. Para cada intervalo de tempo h, calculamos a média das métricas MAE e RMSE cada método e seu respectivo erro padrão durante todo o período entre 5 de maio de 2007 e 18 de dezembro de 2019.

As Figuras 30 e 31 apresentam os resultados das métricas MAE e RMSE para todos os índices de mercado. Assim como na previsão de links, apresentado no capítulo anterior, os resultados da previsão dos pesos dos links utilizando aprendizagem de máquina apresenta resultados superiores ao algoritmo de *benchmark* TI. Vale lembrar que esse algoritmo, mesmo sendo excessivamente simples, apresenta resultados significativos - premissa que garante a sua utilização em métodos de gerenciamento de portfólio como uma estimativa da variância dos ativos. Em geral, os resultados sugerem que as correlações dos retornos, representadas através dos pesos

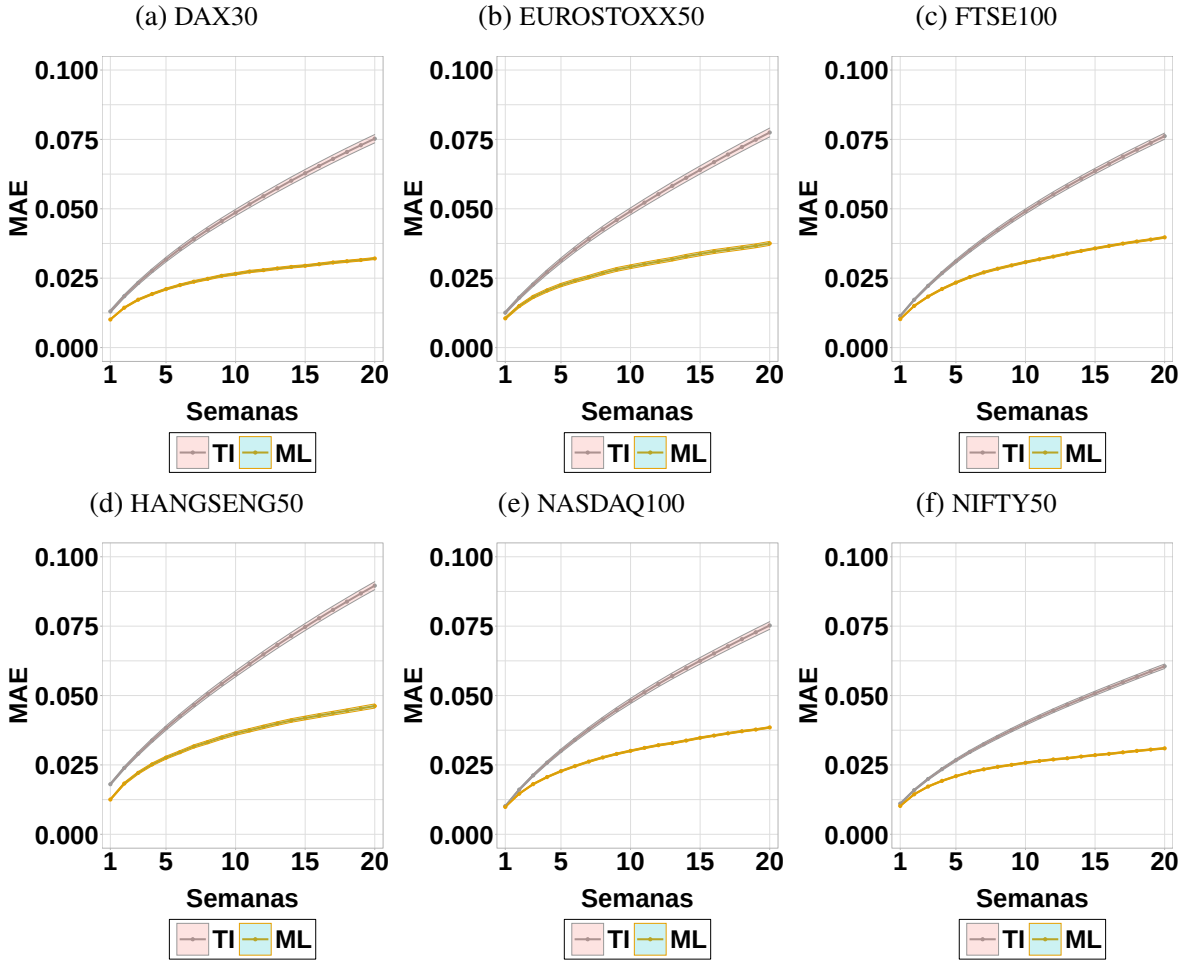


Figura 30 – **MAE - Comparação do desempenho preditivo dos métodos TI e ML.** A figura mostra a métrica MAE dos métodos ML e TI relacionada à previsão de links ponderados. Para cada intervalo de tempo, calculamos a média da métrica MAE de cada método e seu respectivo erro padrão ao longo de todo o período de testes. O método de aprendizado de máquina ML supera o método de *benchmark* TI em todos os índices de mercado.

das arestas nessa abordagem do problema utilizando redes ponderadas, podem ser estimadas como resultados superiores ao algoritmo de *benchmark*. Na seção seguinte, serão apresentados resultados financeiros utilizando os valores dessas previsões para definir as constantes de entrada dos modelos MV para alocação de carteira.

4.3.2 Otimização de Portfólio

Esta seção apresenta um conjunto de resultados experimentais relacionado ao gerenciamento de portfólio utilizando previsão de links ponderados. Os resultados financeiros foram obtidos através da simulação de estratégias de investimento que utilizam as saídas dos modelos MVA como insumo para tomada de decisão. Como descrito anteriormente na Seção 4.2.3, o objetivo do AMV é propor o percentual da carteira que deverá ser alocada para cada ação disponível para investimento. Assim, a fração de cada ação sugerida deverá ser comprada (operação de compra) para que este ativo faça parte do portfólio do investidor. Utilizamos o δT,

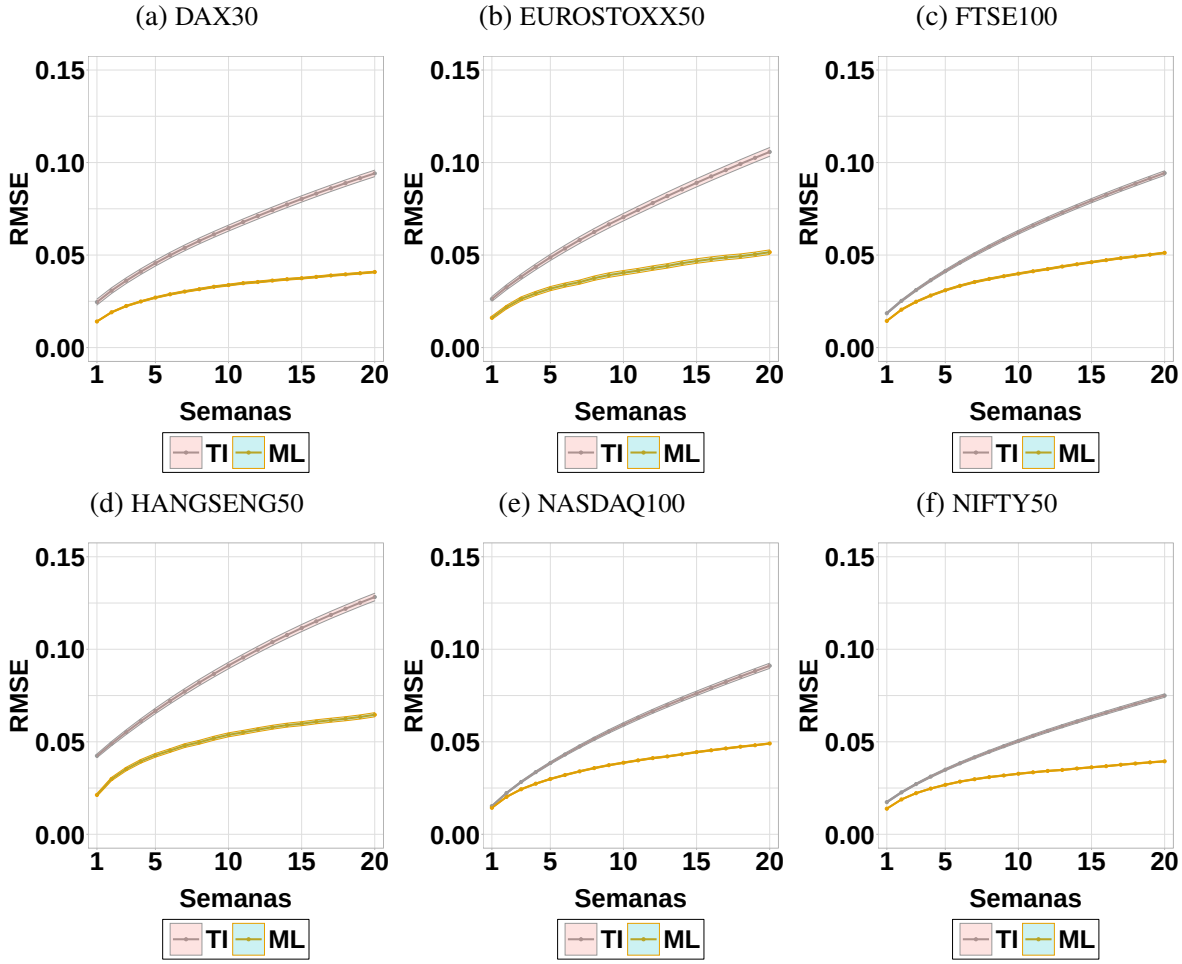


Figura 31 – **RMSE - Comparação do desempenho preditivo dos métodos TI e ML.** A figura mostra a métrica RMSE dos métodos ML e TI relacionada à previsão de links ponderados. Para cada intervalo de tempo, calculamos a média da RMSE de cada método e seu respectivo erro padrão ao longo de todo o período de testes. O método de aprendizado de máquina ML supera o método de *benchmark* TI em todos os índices de mercado.

definido anteriormente como sendo o intervalo de tempo entre duas previsões, como base para o rebalanceamento do portfólio, ou seja, o intervalo em que o modelo AMV será executado para definição e troca das ações da carteira. Em cada operação de rebalanceamento, todas as ações deverão ser vendidas (operação de venda) para que o capital fique disponível para ser alocado novamente. Note que, se uma determinada ação estiver em dois portfólios seguidos, não é necessário que essa seja vendida para ser comprada posteriormente, uma vez que isso gera taxas oriundas das operações de compra e venda - para fins experimentais, iremos assumir que não há custos transacionais nas operações de compra e venda de ações. Outro ponto importante é que os experimentos de previsão de links utilizam preços de fechamento diário. Porém, para a execução da estratégia, é necessário que as operações de compra e venda sejam realizadas durante o período de pregão, ou seja, antes de sabermos o preço de fechamento diário. Nesse sentido, vamos assumir que as operações serão realizadas em um tempo próximo ao encerramento do pregão e que esse tempo seja o mesmo para todas as operações, não havendo impacto de sincronismo entre as operações de compra e venda. Em termos práticos, isso representa a

captura de um *snapshot* do mercado em um tempo próximo ao encerramento do pregão, que servirá de base para todas as tomadas de decisão.

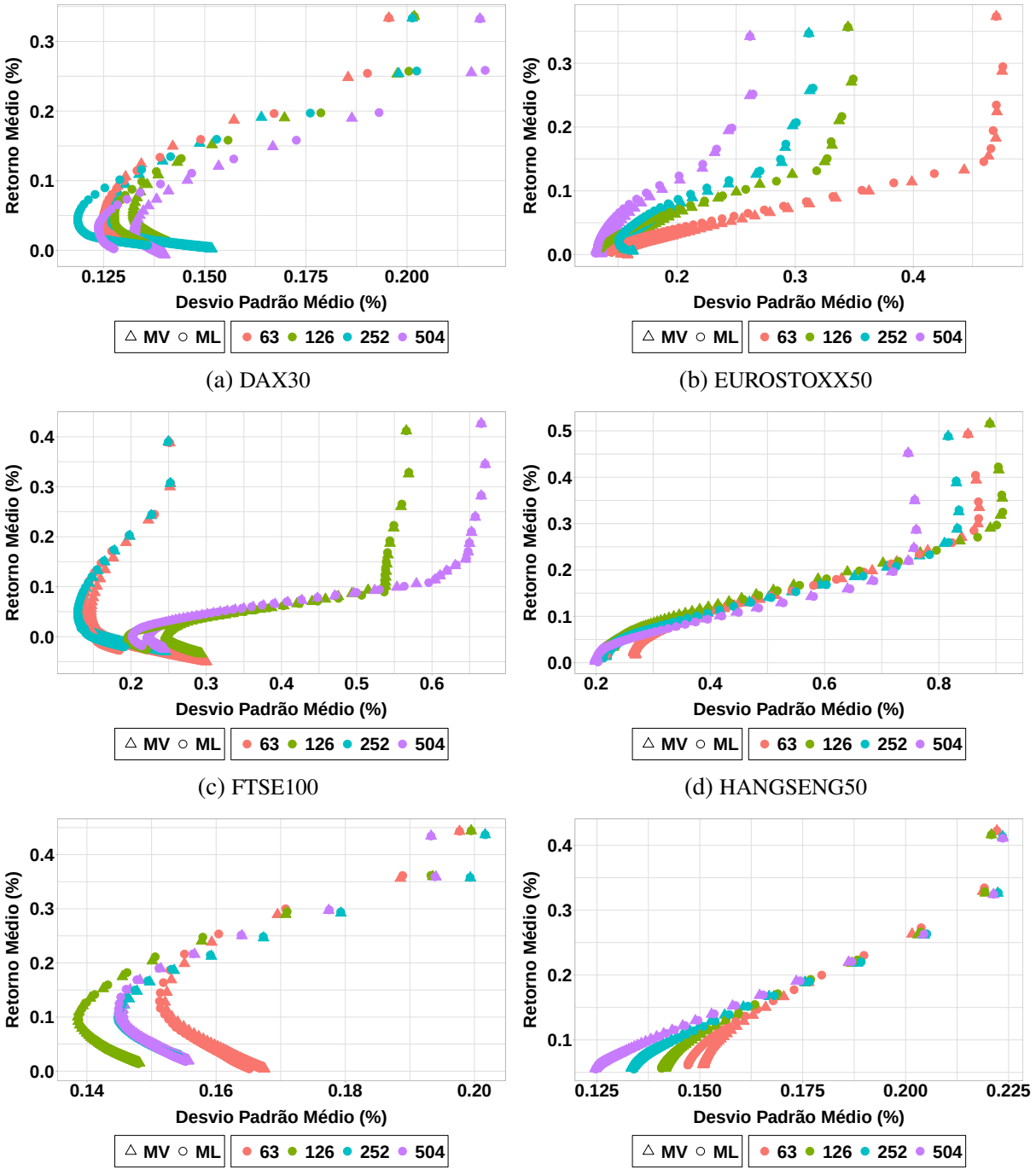
Investigamos o desempenho financeiro do método ML proposto, que usa o resultado da previsão de links ponderados para definir a matriz de covariância, utilizada como entrada para o modelo AMV, comparando-o com os dois algoritmos de *benchmark* ID e MV. Utilizamos uma abordagem de aprendizado de máquina para prever a rede financeira ponderada $G(t+h)$, em que h é o número de semanas à frente, considerando $h = 1, 2, \ dots, 20$ semanas de negociação.

Primeiramente, apresentamos uma análise comparativa da fronteira de portfólios obtidas pelos métodos MV e ML. A fronteira de portfólios foi calculada através da variação do parâmetro λ, definido na Equação 4.6, considerando $0 \leq \lambda \leq 1$. Para cada configuração de *lambda*, executamos o rebalanceamento do portfólio a cada δT dias de negociação em um intervalo de tempo entre 5 de maio de 2007 e 18 de dezembro de 2019. Em cada execução, calculamos o retorno esperado do portfólio ótimo. Assim, cada configuração de *lambda* representa um conjunto de portfólios ótimos encontrados durante todo o período de testes.

As Figuras 32 e 33 apresentam os resultados comparativo da fronteira de portfólios de MV e ML utilizando previsão de links ponderados para $h = 1$ e $h = 10$, respectivamente. Vale ressaltar que essa não é a fronteira eficiente de cada portfólio, uma vez que cada ponto representa um conjunto de execuções que abrange todo o período de experimentos. O eixo x apresenta o desvio padrão médio e o eixo y o retorno médio de cada execução λ. Essa é uma análise comum para identificar o nível de risco que o investidor deve assumir para obter um determinado retorno esperado. Os resultados de MV e ML são calculados para os valores de $L \in \{63, 126, 252, 504\}$. Através da Figura 32, que apresenta resultados para $h = 1$, podemos observar que para DAX30 e FTSE100 a previsão de 1 semana à frente é capaz de melhorar a variância do modelo para todas as configurações de L. Em contrapartida, as fronteiras de portfólio dos demais índices de mercado não apresentam diferenças significativas entre MV e ML. Os resultados de ML mostrados na Figura 33 para $h = 10$ mostram que as fronteiras dos portfólios apresentam, em geral, menor variância quando comparados com MV. Esses resultados sugerem que a previsão de links ponderados para $G(t+10)$ conseguem diminuir o risco de esperado dos portfólios que utilizam esses dados como entrada.

Além da análise do retorno e risco esperado, estamos interessados em entender como a previsão de links afeta os resultados financeiros do gerenciamento de portfólio. Para isso, analisamos o retorno financeiro simulado obtido pelos portfólios sugeridos nas fronteiras das Figuras 32 e 33. Utilizamos o retorno real de cada ativo entre os tempos t e $t + \delta T$ para calcular o retorno simulado de cada portfólio. Em termos práticos, calculamos o retorno semanal obtido pela alocação do portfólio no tempo t. O rebalanceamento do portfólio é feito a cada δT dias de negociação, assim como a previsão dos links ponderados.

As Figuras 34 e 35 apresentam o resultado financeiro dos portfólios ótimos calculados através da simulação do rebalanceamento semanal, comparando os resultados de MV e ML



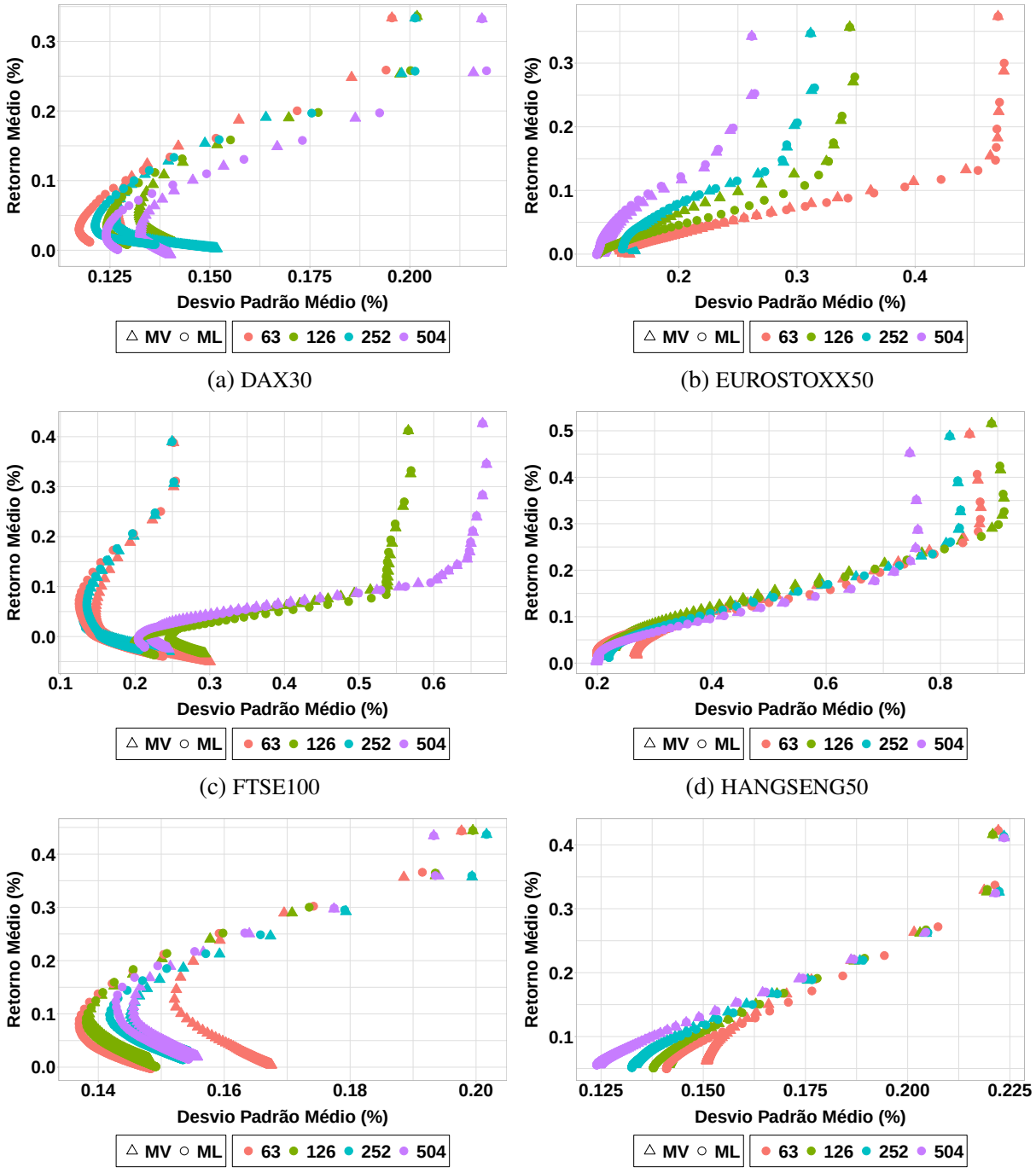
(a) DAX30 (b) EUROSTOXX50

(c) FTSE100 (d) HANGSENG50

(e) NASDAQ100 (f) NIFTY50

Figura 32 – **Fronteira de portfólios para $h = 1$.** Comparação entre os resultados de portfólios ótimos obtidos por MV e ML, que utiliza previsão de $h = 1$ semanas à frente. Cada ponto representa uma configuração λ na Equação 4.6 e contém resultados relacionados ao rebalanceamento do portfólio a cada δT dias de negociação.

utilizando previsão de links ponderados para $h = 1$ e $h = 10$, respectivamente. Calculamos a média e o desvio padrão do retorno simulado para cada configuração de λ. Os resultados de MV e ML são calculados para os valores de $L \in \{63, 126, 252, 504\}$. Nesta análises, quanto mais à esquerda um ponto está, menor o risco de sua configuração. A Figura 34 mostra que a previsão de $h = 1$ traz diminuição do risco (desvio padrão) em algumas configurações de L, considerando



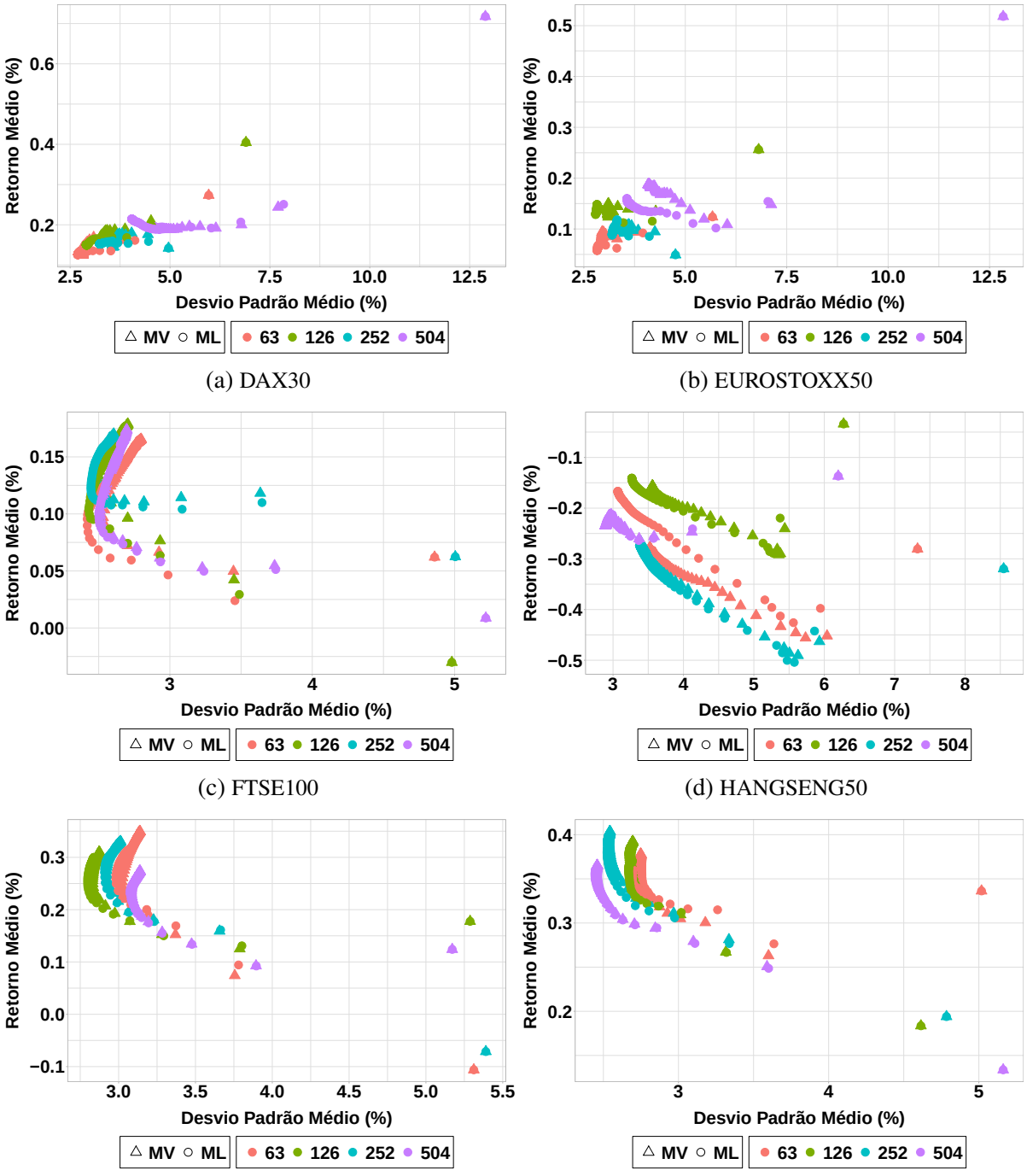
(a) DAX30 (b) EUROSTOXX50

(c) FTSE100 (d) HANGSENG50

(e) NASDAQ100 (f) NIFTY50

Figura 33 – **Fronteira de portfólios para** $h = 10$**.** Comparação entre os resultados de portfólios ótimos obtidos por MV e ML, que utiliza previsão de $h = 10$ semanas à frente. Cada ponto representa uma configuração λ na Equação 4.6 e contém resultados relacionados ao rebalanceamento do portfólio a cada δT dias de negociação.

alguns índices de mercado, como $L = 63$ no índice DAX30, $L = 126$ no índice EUROSTOXX50, $L = 63$ em HANGSENG50, $L = 63$ em FTSE100 e $L = 126$ em NASDAQ100. A Figura 35 sugere que a previsão de $h = 10$ traz certa diminuição do risco em algumas configurações de λ e L para alguns índices de mercado, como $L = 63$ no índice DAX30, $L = 252$ e $L = 504$ no índice EUROSTOXX50, $L = 126$ em FTSE100, $L = 126$ em NASDAQ100 e $L = 63$ em NIFTY50.



(a) DAX30 (b) EUROSTOXX50

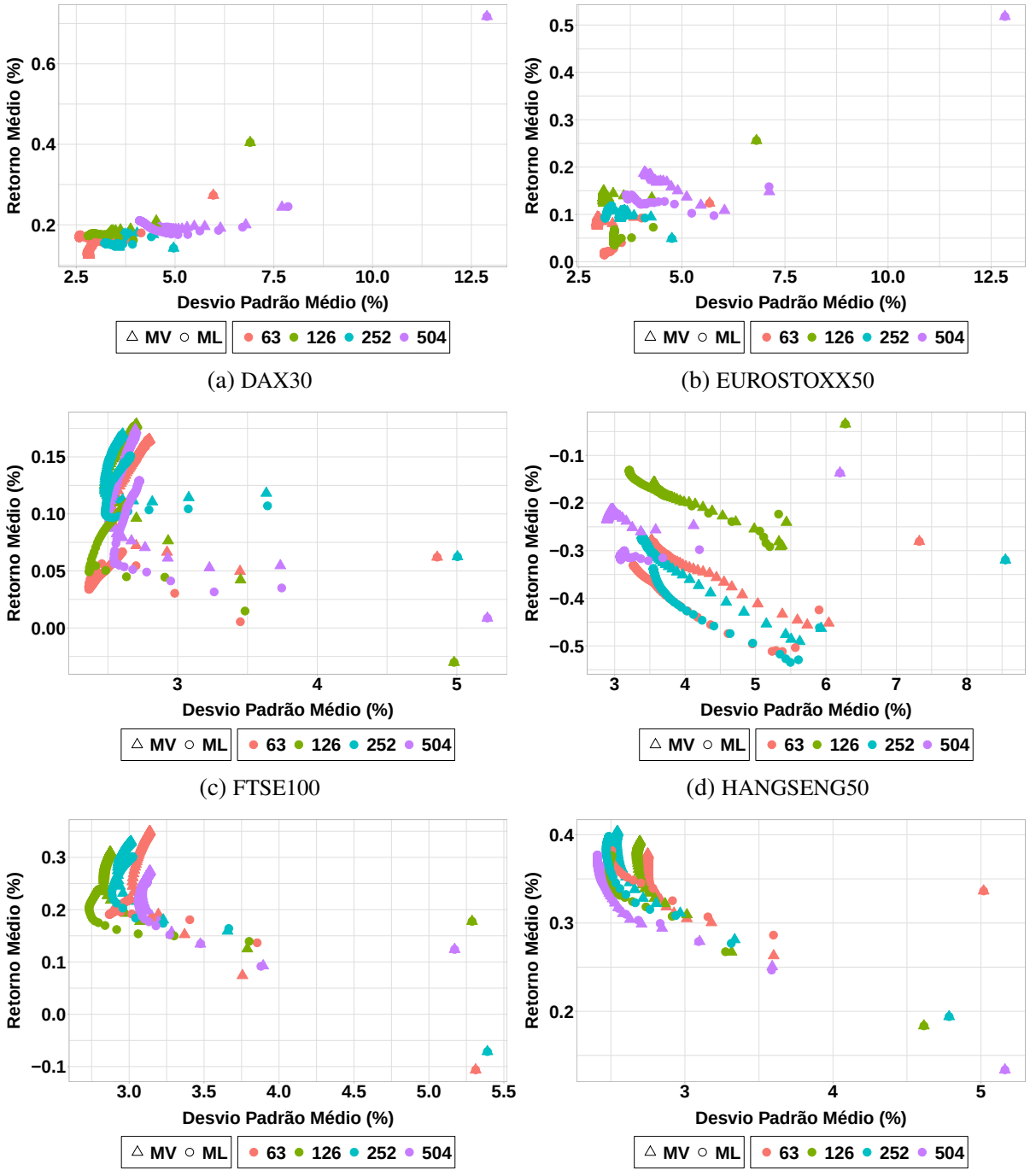
(c) FTSE100 (d) HANGSENG50

(e) NASDAQ100 (f) NIFTY50

Figura 34 – **Retorno simulado de gerenciamento de portfólio utilizando** $h = 1$**.** O eixo x apresenta o desvio padrão médio e o eixo y o retorno médio de cada execução λ.

Importante notar que existe uma diferença significativa entre o risco esperado no portfólio ótimo, apresentado nos resultados das Figuras 32 e 33, e os resultados simulados com dados de retornos reais dos ativos.

Estas análises servem para ilustrar a relação entre o resultado esperado e o retorno real obtido pelo portfólio. Como fazemos a estimativa somente da matriz de covariância futura através da previsão de links ponderados em redes de ações, o retorno estimado μ_i de um ativo i tem



(a) DAX30

(b) EUROSTOXX50

(c) FTSE100

(d) HANGSENG50

(e) NASDAQ100

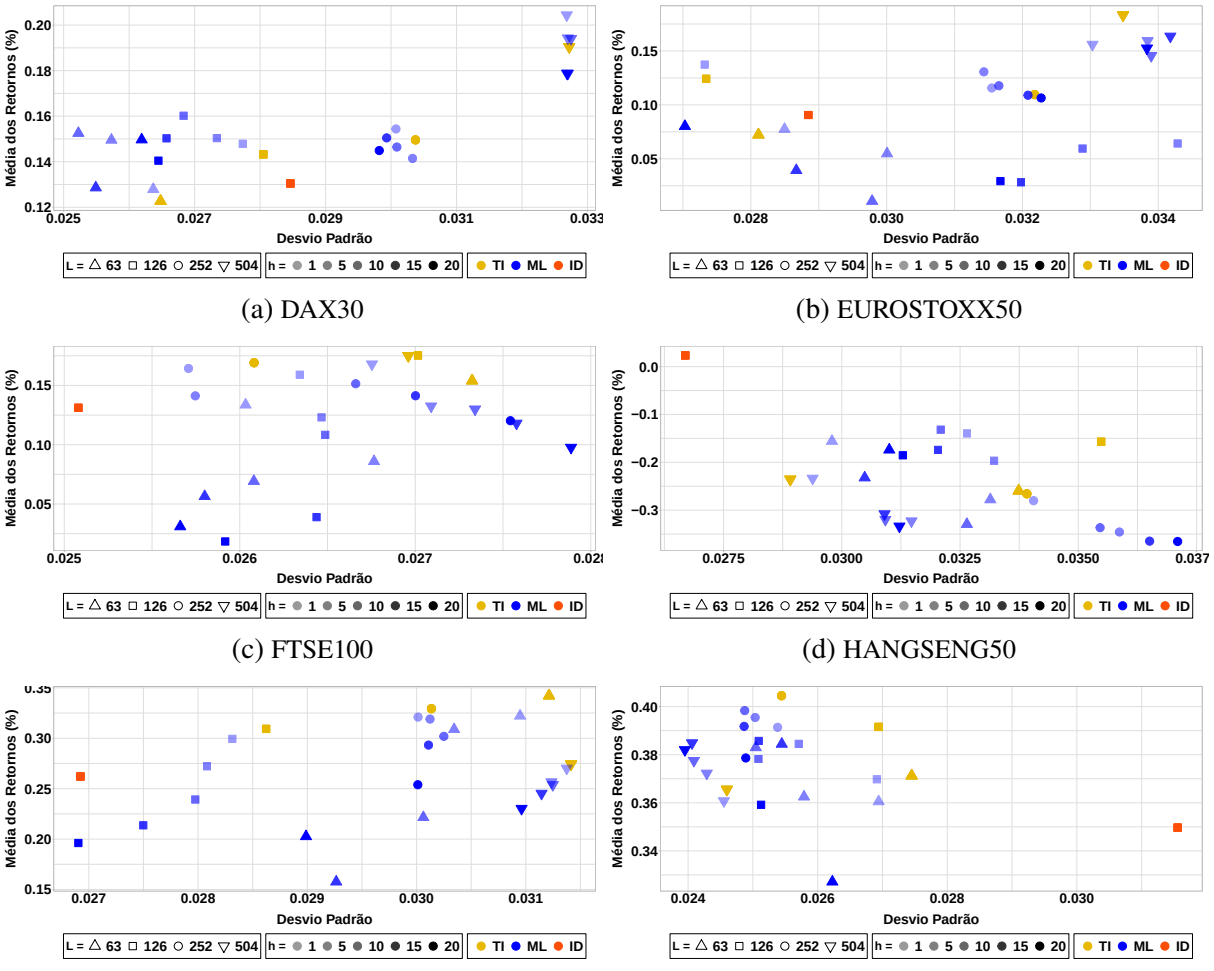
(f) NIFTY50

Figura 35 – **Retorno simulado de gerenciamento de portfólio utilizando** $h = 10$**.** O eixo x apresenta o desvio padrão médio e o eixo y o retorno médio de cada execução λ.

o mesmo impacto na resolução do modelo matemático. Para considerar somente a influência da previsão de links no gerenciamento de portfólio, apresentamos na próxima seção análises experimentais utilizando o modelo de PVMG, quando $\lambda = 1$ na Equação 4.6.

4.3.2.1 Portfólio de Variância Mínima Global

Os resultados apresentados nessa seção utilizam o modelo de Portfólio de Variância Mínima Global (PVMG), que busca o portfólio com menor risco (menor variância) sem considerar o retorno esperado. Esse modelo é derivado da Equação 4.6, considerando $\lambda = 1$. Utilizamos o modelo de otimização PVMG para avaliar o método ML proposto, que utiliza abordagem de aprendizado de máquina para prever a rede financeira ponderada $G(t+h)$, em que h é o número de semanas à frente, considerando $h = 1, 2, \ dots, 20$ semanas de negociação. O método ML usa o resultado da previsão de links ponderados para definir a matriz de covariância do PVMG. Utilizamos dois algoritmos avaliar os resultados do método proposto: *(i)* algoritmo MV, que utiliza a matriz de covariância conhecida como entrada para o modelo de otimização PVMG. e *(ii)* algoritmo ID, que atribui o mesmo peso a todos os ativos disponíveis para alocar o portfólio.



(a) DAX30 (b) EUROSTOXX50

(c) FTSE100 (d) HANGSENG50

(e) NASDAQ100 (f) NIFTY50

Figura 36 – **PVMG - Retorno vs. Risco.** A figura apresenta resultados de simulação de retorno e risco dos portfólios utilizando os métodos ML, MV e ID. As formas geométricas indicam o tamanho de L, a cor representa o método utilizado e a intensidade da forma representa o número de semanas da previsão de links h. Cada ponto representa uma execução durante o período de testes e o resultado financeiro simulado obtida pelo rebalanceamento do portfólio a cada δT dias.

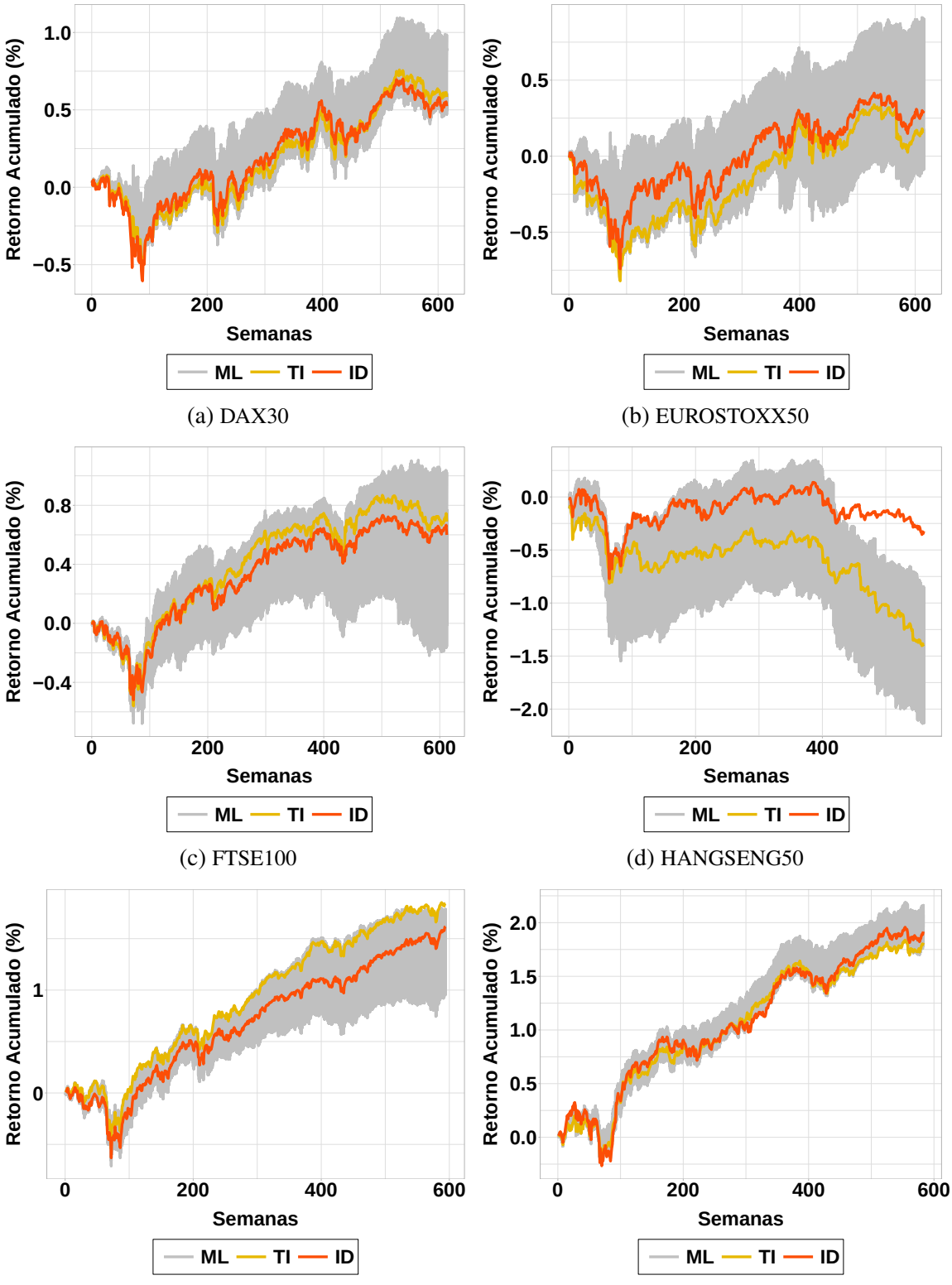
A Figura 36 apresenta uma análise detalhada dos resultados simulados obtidos através da execução do modelo de otimização PVMG utilizando os algoritmos ML, MV e ID. Para cada execução, calculamos a média e o desvio padrão do retorno simulado dos algoritmos. O método MV utiliza resultados da previsão de links da rede ponderada $G(t+h)$ para $h = \{1, 5, 10, 15, 20\}$. Os experimentos foram executados utilizando as configurações de $L \in \{63, 126, 252, 504\}$. Algumas considerações sobre os resultados:

- Para o índice DAX30, o método ML apresenta resultados superiores aos algoritmos MV e ID com relação à gestão de risco e retorno financeiro. Os portfólios com risco mais baixo utilizam $L = 63$ e com maior retorno $L = 504$. É possível notar que os resultados de $L = 252$ se concentram em uma região de *trade-off* entre retorno e risco próxima;
- Os resultados do índice EUROSTOXX50 sugerem que o método ML com $L = 63$ apresenta menor risco dentre todas as configurações. Porém, o maior retorno é referente ao método MV para $L = 504$;
- Para FTSE100, o algoritmo ID apresenta o menor risco, enquanto o método ML apresenta retorno próximo ao máximo encontrado pelo método MV, porém com menor risco, para a configuração utilizando $L = 252$;
- Os resultados do índice HANGSENG50 apresenta retornos negativos para todas as configurações experimentais, exceto para o método ID.
- NASDAQ100 também apresenta o algoritmo ID como sendo o método com menor risco, porém, assim como FTSE100, o método ML apresentado retorno próximo ao máximo encontrado pelo algoritmo MV, mas com menor risco ($L = 126$).
- Para o índice NIFTY50, o método ML apresenta resultados de gestão de risco superiores aos algoritmos MV e ID.

De uma forma geral, podemos ver a influência da previsão de links no gerenciamento do portfólio. Em ativos com menor número de ativos, como NIFTY50 e DAX30, o gerenciamento de risco é melhor que os demais algoritmos, enquanto em índices com maior número de ativos, como FTSE100, NASDAQ100 e EUROSTOXX50, o método proposto apresenta retornos próximos aos máximos encontrados pelos algoritmos de *benchmark*, porém com menor risco.

A Figura 37 apresenta uma projeção do retorno acumulado dos métodos ML, MV e ID. A figura ilustra os limites superior e inferior do retorno acumulado obtido pelo método ML, comparado com os retornos de MV e ID.

A análise apresentada pela Figura 37 é útil para estabelecer um comparativo da evolução do capital obtido pelos métodos, assim como identificar as bordas superior e inferior do retorno acumulado obtido através do método ML utilizando previsão de links para $1 \leq h \leq 20$. Essas bordas auxiliam na distinção entre o comportamento do ML quando comparado aos algoritmos MV



(a) DAX30 (b) EUROSTOXX50

(c) FTSE100 (d) HANGSENG50

(e) NASDAQ100 (f) NIFTY50

Figura 37 – **Projeções de retorno acumulado simulado.** A figura mostra a projeção dos retornos do método ML utilizando previsão de links para $1 \leq h \leq 20$ e $L = \{63, 126, 252, 504\}$. A área cinza de cada figura representa o intervalo dos retornos acumulados dos portfólios sugeridos através do método ML.

Tabela 6 – **Resultado financeiro do PVMG utilizando o método ML.** A tabela apresenta métricas relacionadas ao resultado financeiro do método ML, considerando $L = \{63, 126, 252, 504\}$, $h = \{1, 5, 10, 20\}$ e rebalanceamento de portfólio a cada δT dias.

	L	h	RA	MaxDD	SR	h	RA	MaxDD	SR
DAX30	*63*	1	0.9004	-3.5008	0.0485	10	1.0738	-3.4457	0.0605
		5	1.0529	-3.7729	0.0581	20	1.0537	-3.2065	0.0571
	126	1	1.0219	-2.6569	0.0533	10	1.1070	-2.7224	0.0597
		5	1.0391	-2.5615	0.0550	20	0.9706	-2.4835	0.0531
	252	1	1.0282	-3.9883	0.0513	10	0.9754	-3.5029	0.0487
		5	0.9420	-3.8832	0.0466	20	0.9652	-3.4275	0.0486
	504	1	1.2571	-3.1965	0.0625	10	1.1937	-3.1809	0.0593
		5	1.1948	-3.1586	0.0594	20	1.0996	-3.0961	0.0547
EUROSTOXX50	*63*	1	0.5447	-5.2599	0.0272	10	0.0753	-6.7391	0.0036
		5	0.3851	-8.0951	0.0183	20	0.5638	-4.1186	0.0297
	126	1	0.9492	-3.4437	0.0503	10	0.4109	-3.3169	0.0181
		5	0.4435	-3.9503	0.0187	20	0.2023	-3.2871	0.0092
	252	1	0.7704	-6.9878	0.0367	10	0.7841	-7.3739	0.0372
		5	0.8703	-6.3283	0.0416	20	0.7088	-7.0778	0.0330
	504	1	0.9595	-3.6899	0.0472	10	0.8955	-3.8009	0.0430
		5	0.9807	-3.9774	0.0471	20	0.9382	-3.6570	0.0451
FTSE100	*63*	1	0.9370	-5.1075	0.0513	10	0.4855	-5.2844	0.0266
		5	0.6023	-9.6947	0.0321	20	0.2171	-21.6452	0.0121
	126	1	1.0941	-2.8009	0.0604	10	0.7453	-3.9636	0.0409
		5	0.8468	-4.2086	0.0465	20	0.1268	-3.1325	0.0071
	252	1	1.0895	-2.6228	0.0639	10	1.0040	-2.8284	0.0568
		5	0.9363	-2.7026	0.0549	20	0.7973	-2.9227	0.0437
	504	1	1.0278	-4.3778	0.0628	10	0.7954	-4.2924	0.0475
		5	0.8109	-4.3844	0.0489	20	0.5977	-3.7444	0.0350
HANGSENG50	*63*	1	-1.0085	-4.6416	-0.0522	10	-2.1345	-5.5320	-0.1009
		5	-1.8008	-5.0965	-0.0839	20	-1.1268	-3.5572	-0.0561
	126	1	-0.8865	-3.1823	-0.0428	10	-0.8366	-3.0096	-0.0411
		5	-1.2490	-3.5193	-0.0592	20	-1.1768	-3.2937	-0.0592
	252	1	-1.7094	-3.4505	-0.0823	10	-2.0535	-3.3372	-0.0949
		5	-2.1086	-3.3140	-0.0964	20	-2.2292	-3.1544	-0.0985
	504	1	-1.3082	-4.6878	-0.0795	10	-1.7899	-4.6300	-0.1034
		5	-1.8081	-4.4732	-0.1026	20	-1.8686	-4.3496	-0.1069
NASDAQ100	*63*	1	2.1972	-3.4104	0.1041	10	1.5102	-3.8881	0.0737
		5	2.1054	-3.2227	0.1017	20	1.3819	-3.0628	0.0699
	126	1	2.0028	-3.0796	0.1057	10	1.6005	-3.2468	0.0855
		5	1.8211	-3.1827	0.0969	20	1.3119	-3.7306	0.0729
	252	1	2.0671	-4.2494	0.1069	10	1.9442	-5.1049	0.0998
		5	2.0544	-4.7529	0.1059	20	1.6351	-6.0774	0.0846
	504	1	1.6047	-3.0654	0.0861	10	1.5078	-3.0808	0.0812
		5	1.5240	-2.9950	0.0821	20	1.3676	-3.2111	0.0744
NIFTY50	*63*	1	2.4190	-2.8925	0.1338	10	2.5699	-2.6056	0.1529
		5	2.4323	-2.5781	0.1406	20	2.1951	-2.3844	0.1247
	126	1	2.4368	-3.0960	0.1374	10	2.4923	-2.9234	0.1508
		5	2.5338	-3.1602	0.1496	20	2.3669	-2.5831	0.1429
	252	1	2.4774	-2.5547	0.1542	10	2.5215	-2.6348	0.1601
		5	2.5034	-2.5473	0.1580	20	2.3968	-2.7387	0.1521
	504	1	2.1032	-3.2819	0.1469	10	2.2011	-3.1355	0.1567
		5	2.1702	-3.2682	0.1533	20	2.2271	-3.0502	0.1595

e ID. A área cinza de cada figura representa o intervalo dos retornos acumulados dos portfólios sugeridos através do método ML. A Tabela 6 apresenta um resumo das métricas relacionadas ao resultado financeiro do método ML para $L = \{63, 126, 252, 504\}$ e $h = \{1, 5, 10, 20\}$.

4.4 Considerações Finais

Neste capítulo, investigamos a utilização da previsão de links como suporte ao gerenciamento de portfólio. Apresentamos um conjunto de experimentos utilizando resultados da previsão de formação de links em redes de ações como entrada para modelos de otimização de portfólio. Para isso, propusemos que a definição das constantes da clássica Análise Média-Variância (AMV), proposta por Markowitz em 1952, seja feita através da previsão da matriz de correlação futura. A previsão da correlação futura entre os preços das ações foi modelada como sendo um problema de previsão de links ponderados em redes de ações. Rede de ações são estruturas baseadas em grafos utilizadas para analisar a topologia do mercado financeiro, onde os nós representam os ativos e as arestas representam o relacionamento entre eles. Utilizamos uma modelagem utilizando grafos ponderados, onde o peso de cada aresta corresponde ao coeficiente de correlação de Pearson. Para realizar a previsão de links, propusemos a utilização de um método utilizando aprendizado supervisionado, cujos atributos de entrada são características derivadas da rede e o atributo alvo é o peso das arestas (correlação de Pearson). Propusemos também um conjunto de experimentos para verificar a capacidade preditiva do algoritmo de aprendizagem de máquina e os resultados financeiros do modelo de otimização. Usamos dados de empresas constituintes de seis índices dos mercados dos Estados Unidos da América, Reino Unido, Índia, Europa, Alemanha e Hong Kong, compreendidos em um período entre 1 de março de 2005 e 18 de dezembro de 2019.

Vale ressaltar que existem algumas limitações acerca do estudo apresentado. Primeiramente, destacamos que as ações foram selecionados através de uma análise a posteriori, o que pode influenciar o retorno de cada índice, uma vez que, em certos casos, a rentabilidade positiva é um dos fatores utilizados para realizar a seleção dos ativos que farão parte do índice. Nesse caso, as análises de retorno podem sofrer a influência desta seleção previa. Porém, como todos os algoritmos foram colocados nas mesmas condições experimentais, as conclusões empíricas extraídas da análise dos resultados são confiáveis, íntegras e seguem o método científico. Além disso, destacamos que utilizamos apenas ativos que permaneceram no índice de mercado durante todo o período de testes, o que limita a inserção e retirada de nós nas redes.

CAPÍTULO

5

CONCLUSÕES

Neste estudo, propusemos a análise de dois problemas envolvendo redes financeiras, modeladas a partir dos retornos das ações. As redes de ações (ou redes de ativos) compreendem um tipo de redes financeiras, que permitem a modelagem e análise das informações do mercado de ações através de uma perspectiva topológica. Nesse tipo de rede, os nós representam as ações e as arestas representam tipo de relacionamento entre elas.

As análises foram propostas para responder ao seguinte questionamento: como a previsão de formação de links em redes de ações pode auxiliar na melhoria de investimentos no mercado financeiro? Através dessa pergunta, propusemos a abordagem de dois problemas: *(i)* previsão de links em redes de ações; *(ii)* utilização da previsão de links como suporte ao gerenciamento de portfólios (carteiras).

O primeiro problema, envolvendo a previsão de links em redes de ações, foi apresentado no Capítulo 3. Nele, investigamos a previsão de links em redes modeladas através de três métodos de filtragem baseados em correlação: *Dynamic Asset Graphs* (DAG), *Dynamic Threshold Networks* (DTN) e *Dynamic Minimal Spanning Tree* (DMST). Formulamos o problema de previsão de formação de links em redes de ações, onde buscamos prever com as arestas que estarão presentes nas redes futuras. Propusemos e avaliamos um modelo de aprendizado de máquina supervisionado baseado em atributos derivados das redes financeiras para prever links em redes futuras. Usamos dados de empresas constituintes de seis índices dos mercados dos Estados Unidos da América, Reino Unido, Índia, Europa, Alemanha e Hong Kong (NASDAQ100, FTSE100, NIFTY50, EUROSTOXX50, DAX30 e HANGSENG50, respectivamente), compreendidos em um período entre de 1 de março de 2005 a 18 de dezembro de 2019. Para avaliar o desempenho preditivo do modelo, nós o comparamos com quinze algoritmos de *benchmark* de predição de links amplamente utilizados na literatura. Resultados experimentais mostraram que o modelo proposto foi capaz de prever a estrutura de mercado com desempenho superior a todos os métodos de *benchmark* e para todos os índices de mercado, independentemente do método de

filtragem de rede. Também medimos a melhoria em relação ao algoritmo *Time Invariant* (TI), que assume que a rede não muda com o tempo. Os resultados experimentais mostraram uma melhoria maior em relação ao TI em redes criadas usando o método de filtragem DTN, chegando a quase 40% de melhoria para NASDAQ100. Nossos resultados experimentais também sugeriram que as características topológicas da rede são úteis na previsão da estrutura do mercado de ações em comparação com as características de correlação de pares, particularmente para previsões de longo prazo. Apresentamos também uma análise qualitativa para entender o que pode tornar as redes mais ou menos previsíveis, buscando a explicação em variáveis que descrevem as redes de ações. Além disso, apresentamos uma análise comparativa utilizando quatro algoritmos de aprendizagem de máquina para avaliar o impacto deste nos resultados.

O segundo problema, envolvendo a utilização da previsão de links em redes de ações como suporte ao gerenciamento de portfólio, foi apresentado no Capítulo 5. A previsão da correlação futura entre os preços das ações foi modelada como sendo um problema de previsão de links ponderados em redes de ações. Para analisar os resultados da previsão de links, comparamos os resultados do método proposto com o algoritmo *Time Invariant* (TI), que assume que a rede é invariante no tempo tempo. Este algoritmo foi escolhido principalmente por conta de seus resultados satisfatórios apresentados no capítulo anterior, que explorou em profundidade a previsão de links em redes de ações. Para analisar os resultados de gerenciamento de portfólio, comparamos o resultado do método proposto com dois algoritmos: Igualmente Distribuído (ID), que divide o capital em frações iguais para cada ação disponível; e Média-Variância (MV), que faz a otimização do portfólio através do modelo AMV utilizando dados conhecidos de covariância e retorno, sem estimar nenhuma informação futura. Os resultados experimentais sobre a previsão de links ponderados sugerem que o método proposto consegue estimar a correlação futura melhor que o algoritmo de *benchmark* TI. Através das métricas propostas foi possível identificar que a abordagem proposta consegue realizar uma boa estimativa dos pesos dos links em redes de ações ponderadas. Os resultados experimentais sobre otimização de portfólio sugerem que o método proposto, que utiliza a matriz de correlação prevista como entrada para o modelo AMV, é capaz de apresentar uma diminuição no risco de investimento em certos cenários, levando em consideração as configurações experimentais. Os índices DAX30, EUROSTOXX50 e NIFTY50 apresentaram melhoria na gestão do risco através das configurações experimentais realizadas. Em alguns mercados, como DAX30, EUROSTOXX50, FTSE100 e NIFTY50, o modelo proposto foi capaz de apresentar configurações com maiores retornos que os algoritmos de *benchmark*. De uma forma geral, os resultados mostraram que a previsão de redes com intervalos de tempo futuro maiores, mesmo que tenham uma taxa de acerto menor, apresenta resultados de gerenciamento de risco melhores.

O restante desse capítulo é organizado da seguinte forma: a Seção 5.1 apresenta as principais contribuições científicas dessa tese de doutorado e a Seção 5.2 descreve as perspectivas de trabalhos futuros e desdobramentos da pesquisa.

5.1 Contribuições da Pesquisa

A pesquisa de doutorado apresentada possui contribuições interdisciplinares, envolvendo as áreas de análise de redes complexas, aprendizado de máquina, redes financeiras, gerenciamento e otimização de portfólio. A seguir, apontamos as contribuições obtidas como resultado da realização deste estudo:

- **Previsão de links em redes dinâmicas utilizando aprendizado de máquina:** apresentamos um método com base em algoritmos de aprendizado de máquina para previsão de redes dinâmicas, onde atributos derivados da rede são utilizados como entrada para previsão de redes dinâmicas e efêmeras.

- **Previsão de links em redes baseadas em similaridade de séries temporais:** a principal características das redes de ações é a análise topológica do mercado através de redes complexas, utilizando similaridade entre séries temporais como filtro para modelagem da rede. Assim como no mercado financeiro, outros domínios que possuem tal característica podem utilizar a mesma metodologia aqui proposta para previsão das estruturas futuras, como no caso do setor energético, através da modelagem do *stream* de dados relacionados à demanda de energia ou à produção de energia em unidades distintas, como produção de energia eólica ou solar.

- **Arcabouço para previsão de redes financeiras:** análise comparativa entre métodos de filtragem de rede, utilizando dados de mercados distintos e através de configurações diferentes. Mostramos que as características topológicas da rede são tão importantes quanto informações comumente utilizadas em análises no mercado financeiro para previsão de redes futuras. O arcabouço proposto pode ser utilizado para prever quaisquer estrutura de mercado, tais como mercado de moedas digitais, mercado de energia, mercado de commodities, mercado de câmbio, dentre outros.

- **Análise comparativa entre diversos métodos de previsão de links:** propusemos uma extensa análise comparativo utilizando algoritmos para previsão de links amplamente utilizados na literatura. O método proposto, que utiliza aprendizado de máquina e atributos derivados da rede, apresentou melhores resultados, considerando o domínio do mercado financeiro no qual eles foram utilizados;

- **Metodologia para otimização de portfólio utilizando previsão de links:** propusemos uma metodologia conectando a previsão de links em redes de ações e a Teoria Moderna do Portfólio. Para isso, utilizamos o resultado da previsão de links ponderados em redes de ações como forma de definir as constantes de entrada do método Análise Média-Variância, modelo utilizado para otimização de portfólio.

A seguir, listamos as publicações que foram desenvolvidas durante esta pesquisa de doutorado:

Conferência Internacional

- Castilho D., Gama J., Mundim L.R., Carvalho A.C.P.L.F. (2019) **Improving Portfolio Optimization Using Weighted Link Prediction in Dynamic Stock Networks**. *International Conference on Computational Science - ICCS*.

Periódico Internacional

- Castilho, D., Souza, T.T.P., Kang, S.M., Gama, J., Carvalho, A.C.P.L.F. (2021). **Forecasting Financial Market Structure from Network Features using Machine Learning**. Submetido ao periódico *Scientific Reports*.

5.2 Trabalhos Futuros

Os resultados desta pesquisa abrem um leque de possíveis desdobramentos e trabalhos futuros. Dentre esses trabalhos, podemos citar:

- Utilização do método de previsão de links em problemas que utilizem redes baseadas em séries temporais ou *stream* de dados, como no setor energético;
- Utilização do método de previsão de links em redes ações para melhoria no gerenciamento de risco utilizando clusterização em redes de ações, através da seleção de único ativo por *cluster* (DURANTE *et al.*, 2013);
- Utilização do método de previsão de links em redes de ações para investimento utilizando as folhas das redes previstas (POZZI; MATTEO; ASTE, 2013);
- Aplicação do método de previsão de links em problemas envolvendo arbitragem estatística e *pair-trading* (negociação em pares), através da identificação de pares e gatilhos para operações *intraday* e entre dias.
- Utilização do método de previsão de links em redes de ações como entrada para o método de gerenciamento de portólio *Maximum Decorrelation* (Máxima Decorrelação) (CHRISTOFFERSEN *et al.*, 2011).
- Aplicação do método proposto em outros mercados, como mercado de energia elétrica, moedas digitais, commodities, dentre outros.
- Aplicação do método proposto para seleção de portfólio online em estratégias de investimento *intraday*;

REFERÊNCIAS

ADAMIC, L. A.; ADAR, E. Friends and neighbors on the web. **Social networks**, Elsevier, v. 25, n. 3, p. 211–230, 2003. Citado na página 54.

AHMED, A.; SHERVASHIDZE, N.; NARAYANAMURTHY, S.; JOSIFOVSKI, V.; SMOLA, A. J. Distributed large-scale natural graph factorization. In: **Proceedings of the 22nd international conference on World Wide Web**. [S.l.: s.n.], 2013. p. 37–48. Citado na página 57.

Al Hasan, M.; CHAOJI, V.; SALEM, S.; ZAKI, M. Link prediction using supervised learning. **SDM06: workshop on link ...**, 2006. ISSN 10987576. Citado na página 55.

ALBERT, R.; BARABÁSI, A.-L. Statistical mechanics of complex networks. **Reviews of modern physics**, APS, v. 74, n. 1, p. 47, 2002. Citado na página 46.

ALLEN, F.; KARJALAINEN, R. Using genetic algorithms to find technical trading rules. **Journal of financial Economics**, Elsevier, v. 51, n. 2, p. 245–271, 1999. Citado na página 51.

BARABÁSI, A.-L.; ALBERT, R. Emergence of scaling in random networks. **science**, American Association for the Advancement of Science, v. 286, n. 5439, p. 509–512, 1999. Citado nas páginas 47 e 54.

BARFUSS, W.; MASSARA, G. P.; MATTEO, T. D.; ASTE, T. Parsimonious modeling with information filtering networks. **Phys. Rev. E**, American Physical Society, v. 94, p. 062306, Dec 2016. Disponível em: <https://link.aps.org/doi/10.1103/PhysRevE.94.062306>. Citado na página 60.

BELKIN, M.; NIYOGI, P. Laplacian eigenmaps for dimensionality reduction and data representation. **Neural computation**, MIT Press, v. 15, n. 6, p. 1373–1396, 2003. Citado na página 56.

BERLUSCONI, G.; CALDERONI, F.; PAROLINI, N.; VERANI, M.; PICCARDI, C. Link prediction in criminal networks: A tool for criminal intelligence analysis. **PloS one**, Public Library of Science San Francisco, CA USA, v. 11, n. 4, p. e0154244, 2016. Citado na página 52.

BESSLER, W.; WOLFF, D. Portfolio optimization with return prediction models evidence for industry portfolios. In: **World Finance and Banking Symposium**. [S.l.: s.n.], 2015. Citado na página 102.

BILLIO, M.; GETMANSKY, M.; LO, A. W.; PELIZZON, L. Econometric measures of connectedness and systemic risk in the finance and insurance sectors. **Journal of financial economics**, Elsevier, v. 104, n. 3, p. 535–559, 2012. Citado na página 48.

BLONDEL, V. D.; GUILLAUME, J.-L.; LAMBIOTTE, R.; LEFEBVRE, E. Fast unfolding of communities in large networks. **Journal of statistical mechanics: theory and experiment**, IOP Publishing, v. 2008, n. 10, p. P10008, 2008. Citado nas páginas 69 e 106.

BONANNO, G.; CALDARELLI, G.; LILLO, F.; MANTEGNA, R. N. Topology of correlation-based minimal spanning trees in real and model markets. **Physical Review E**, APS, v. 68, n. 4, p. 046130, 2003. Citado nas páginas 35, 47 e 101.

BONANNO, G.; CALDARELLI, G.; LILLO, F.; MICCICHE, S.; VANDEWALLE, N.; MANTEGNA, R. N. Networks of equities in financial markets. **The European Physical Journal B-Condensed Matter and Complex Systems**, Springer, v. 38, n. 2, p. 363–371, 2004. Citado nas páginas 35, 47, 48 e 101.

BONANNO, G.; LILLO, F.; MANTEGNA, R. N. High-frequency cross-correlation in a set of stocks. Taylor & Francis, 2001. Citado na página 100.

BONDY, J. A.; MURTY, U. S. R. *et al.* **Graph theory with applications**. [S.l.]: Citeseer, 1976. v. 290. Citado na página 46.

BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, n. 1, p. 5–32, 2001. Citado na página 91.

BRIN, S.; PAGE, L. The anatomy of a large-scale hypertextual web search engine. **Computer networks and ISDN systems**, Elsevier, v. 30, n. 1-7, p. 107–117, 1998. Citado na página 54.

CARLSSON, G. E.; MÉMOLI, F. *et al.* Characterization, stability and convergence of hierarchical clustering methods. **J. Mach. Learn. Res.**, v. 11, n. Apr, p. 1425–1470, 2010. Citado na página 74.

CARVALHO, A.; FACELI, K.; LORENA, A.; GAMA, J. Inteligência artificial–uma abordagem de aprendizado de máquina. **Rio de Janeiro: LTC**, p. 45, 2011. Citado na página 49.

CASTILHO, D.; GAMA, J.; MUNDIM, L. R.; CARVALHO, A. C. P. L. F. de. Improving portfolio optimization using weighted link prediction in dynamic stock networks. In: **International Conference on Computational Science - ICCS 2019**. [S.l.]: Springer International Publishing, 2019. p. 340–353. ISBN 978-3-030-22744-9. Citado nas páginas 37 e 71.

CHANG, T.-J.; MEADE, N.; BEASLEY, J. E.; SHARAIHA, Y. M. Heuristics for cardinality constrained portfolio optimisation. **Computers & Operations Research**, Elsevier, v. 27, n. 13, p. 1271–1302, 2000. Citado na página 108.

CHEN, S.; HU, C.; ZHOU, Y. Order book simulator and optimal liquidation strategies. 2010. Citado na página 51.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: ACM. **Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining**. [S.l.], 2016. p. 785–794. Citado nas páginas 65 e 104.

CHI, K. T.; LIU, J.; LAU, F. C. A network perspective of the stock market. **Journal of Empirical Finance**, Elsevier, v. 17, n. 4, p. 659–667, 2010. Citado nas páginas 48, 49 e 63.

CHRISTOFFERSEN, P.; ERRUNZA, V. R.; JACOBS, K.; HUGUES, L. Is the potential for international diversification disappearing? **Available at SSRN 1783960**, 2011. Citado nas páginas 100 e 128.

CLAUSET, A.; NEWMAN, M. E. J.; MOORE, C. Finding community structure in very large networks. **Phys. Rev. E**, American Physical Society, v. 70, p. 066111, Dec 2004. Disponível em: <https://link.aps.org/doi/10.1103/PhysRevE.70.066111>. Citado na página 85.

CONT, R. Long range dependence in financial markets. In: ____. [S.l.]: Springer London, 2005. p. 159–179. Citado na página 59.

CRYER, J. D.; KELLET, N. **Time series analysis**. [S.l.]: Springer, 1986. v. 101. Citado na página 42.

DAS, S. R.; MITCHENER, K. J.; VOSSMEYER, A. **Systemic Risk and the Great Depression**. [S.l.], 2018. (Working Paper Series, 25405). Disponível em: <http://www.nber.org/papers/w25405>. Citado na página 90.

DIVAKARAN, A.; MOHAN, A. Temporal link prediction: A survey. **New Generation Computing**, Springer, v. 38, n. 1, p. 213–258, 2020. Citado na página 53.

DUNLAVY, D. M.; KOLDA, T. G.; ACAR, E. Temporal link prediction using matrix and tensor factorizations. **ACM Transactions on Knowledge Discovery from Data (TKDD)**, ACM, v. 5, n. 2, p. 10, 2011. Citado na página 53.

DURANTE, F.; FOSCOLO, E.; PAPPADA, R.; WANG, H. A portfolio diversification strategy via tail dependence measures. In: . [S.l.: s.n.], 2013. Citado na página 128.

EOM, C.; OH, G.; JUNG, W.-S.; JEONG, H.; KIM, S. Topological properties of stock networks based on minimal spanning tree and random matrix theory in financial time series. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 388, n. 6, 2009. Citado na página 47.

ERDOS, P.; RÉNYI, A. On the evolution of random graphs. **Publ. Math. Inst. Hung. Acad. Sci**, v. 5, n. 1, p. 17–60, 1960. Citado na página 47.

FREITAS, F. D.; SOUZA, A. F. D.; ALMEIDA, A. R. de. Prediction-based portfolio optimization model using neural networks. **Neurocomputing**, Elsevier, v. 72, n. 10-12, 2009. Citado nas páginas 99 e 102.

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. **Annals of statistics**, JSTOR, p. 1189–1232, 2001. Citado na página 66.

GIUDICI, P.; POLINESI, G.; SPELTA, A. Network models to improve robot advisory portfolios. **Annals of Operations Research**, Springer, p. 1–25, 2021. Citado na página 101.

GOMES, D. d. S. Inteligência artificial: conceitos e aplicações. **Olhar Científico. v1**, n. 2, p. 234–246, 2010. Citado na página 50.

GOPIKRISHNAN, P.; PLEROU, V.; LIU, Y.; AMARAL, L. N.; GABAIX, X.; STANLEY, H. E. Scaling and correlation in financial time series. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 287, n. 3-4, p. 362–373, 2000. Citado na página 47.

GOYAL, P.; FERRARA, E. Graph embedding techniques, applications, and performance: A survey. **Knowledge-Based Systems**, Elsevier, v. 151, p. 78–94, 2018. Citado nas páginas 55 e 56.

GRANGER, C. W. Some properties of time series data and their use in econometric model specification. **Journal of econometrics**, Elsevier, v. 16, n. 1, p. 121–130, 1981. Citado na página 48.

GROVER, A.; LESKOVEC, J. node2vec: Scalable feature learning for networks. In: ACM. **Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining**. [S.l.], 2016. p. 855–864. Citado nas páginas 52 e 56.

HAMILTON, J. D. **Time series analysis**. [S.l.]: Princeton university press Princeton, 1994. v. 2. Citado na página 42.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning: data mining, inference, and prediction**. [S.l.]: Springer Science & Business Media, 2009. Citado na página 83.

HENRIQUE, B. M.; SOBREIRO, V. A.; KIMURA, H. Literature review: Machine learning techniques applied to financial market prediction. **Expert Systems with Applications**, Elsevier, v. 124, p. 226–251, 2019. Citado na página 36.

HIEMSTRA, C.; JONES, J. D. Testing for linear and nonlinear granger causality in the stock price-volume relation. **The Journal of Finance**, Wiley Online Library, v. 49, n. 5, p. 1639–1664, 1994. Citado na página 51.

HOLME, P.; SARAMÄKI, J. Temporal networks. **Physics reports**, Elsevier, v. 519, n. 3, p. 97–125, 2012. Citado na página 101.

HSIEH, T.-J.; HSIAO, H.-F.; YEH, W.-C. Forecasting stock markets using wavelet transforms and recurrent neural networks: An integrated system based on artificial bee colony algorithm. **Applied soft computing**, Elsevier, v. 11, n. 2, p. 2510–2525, 2011. Citado na página 51.

HüTTNER, A.; MAI, J.-F.; MINEO, S. Portfolio selection based on graphs: Does it align with markowitz-optimal portfolios? **Dependence Modeling**, De Gruyter, Berlin, Boston, v. 6, n. 1, p. 63 – 87, 2018. Disponível em: <https://www.degruyter.com/view/journals/demo/6/1/article-p63.xml>. Citado nas páginas 60 e 101.

HUANG, Z.; LIN, D. K. The time-series link prediction problem with applications in communication surveillance. **INFORMS Journal on Computing**, INFORMS, v. 21, n. 2, p. 286–303, 2009. Citado nas páginas 57 e 67.

IORI, G.; MANTEGNA, R. N. Chapter 11 - empirical analyses of networks in finance. In: HOMMES, C.; LEBARON, B. (Ed.). **Handbook of Computational Economics**. Elsevier, 2018, (Handbook of Computational Economics, v. 4). p. 637 – 685. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1574002118300054>. Citado nas páginas 59 e 60.

JI-HWAN, P. **Complex Network Analysis of Financial Market using Link Prediction**. Tese (Doutorado), 2020. Citado na página 37.

JOHANSEN, S.; JUSELIUS, K. Maximum likelihood estimation and inference on cointegration with applications to the demand for money. **Oxford Bulletin of Economics and statistics**, Wiley Online Library, v. 52, n. 2, p. 169–210, 1990. Citado na página 48.

KABOUDAN, M. A. Genetic programming prediction of stock prices. **Computational Economics**, Springer, v. 16, n. 3, p. 207–236, 2000. Citado na página 51.

KALAYCI, C. B.; ERTENLICE, O.; AKBAY, M. A. A comprehensive review of deterministic models and applications for mean-variance portfolio optimization. **Expert Systems with Applications**, Elsevier, v. 125, p. 345–368, 2019. Citado nas páginas 108 e 109.

KALLBERG, J. G.; ZIEMBA, W. T. Comparison of alternate utility functions in portfolio selection problems. **Management Science**, v. 29, p. 1257–1276, 1983. Citado na página 108.

KAMIJO, K.-i.; TANIGAWA, T. Stock price pattern recognition-a recurrent neural network approach. In: IEEE. **Neural Networks, 1990., 1990 IJCNN International Joint Conference on**. [S.l.], 1990. p. 215–221. Citado na página 51.

KIMOTO, T.; ASAKAWA, K.; YODA, M.; TAKEOKA, M. Stock market prediction system with modular neural networks. In: IEEE. **Neural Networks, 1990., 1990 IJCNN International Joint Conference on**. [S.l.], 1990. p. 1–6. Citado na página 51.

KOTSIANTIS, S. B.; ZAHARAKIS, I.; PINTELAS, P. *et al.* Supervised machine learning: A review of classification techniques. **Emerging artificial intelligence applications in computer engineering**, Amsterdam, v. 160, n. 1, p. 3–24, 2007. Citado na página 55.

KOZA, J. R. **Genetic programming: on the programming of computers by means of natural selection**. [S.l.]: MIT press, 1992. v. 1. Citado na página 51.

KRUSKAL, J. B. On the shortest spanning subtree of a graph and the traveling salesman problem. **Proceedings of the American Mathematical society**, JSTOR, v. 7, n. 1, p. 48–50, 1956. Citado nas páginas 49 e 64.

Lü, L.; ZHOU, T. Link prediction in complex networks: A survey. **Physica A: Statistical Mechanics and its Applications**, v. 390, n. 6, p. 1150 – 1170, 2011. ISSN 0378-4371. Disponível em: <http://www.sciencedirect.com/science/article/pii/S037843711000991X>. Citado na página 52.

LEE, G. S.; DJAUHARI, M. A. An overall centrality measure: The case of us stock market. **International Journal of Electrical & Computer Sciences**, v. 12, n. 6, 2012. Citado na página 47.

LENDASSE, A.; BODT, E. de; WERTZ, V.; VERLEYSEN, M. Non-linear financial time series forecasting-application to the bel 20 stock market index. **European Journal of Economic and Social Systems**, EDP Sciences, v. 14, n. 1, p. 81–91, 2000. Citado na página 51.

LI, B.; HOI, S. C.; GOPALKRISHNAN, V. Corn: Correlation-driven nonparametric learning approach for portfolio selection. **ACM Transactions on Intelligent Systems and Technology (TIST)**, ACM New York, NY, USA, v. 2, n. 3, p. 1–29, 2011. Citado na página 110.

LI, Y.; JIANG, X.-F.; TIAN, Y.; LI, S.-P.; ZHENG, B. Portfolio optimization based on network topology. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 515, p. 671–681, 2019. Citado nas páginas 100 e 101.

LIBEN-NOWELL, D.; KLEINBERG, J. The link-prediction problem for social networks. **Journal of the American society for information science and technology**, Wiley Online Library, v. 58, n. 7, p. 1019–1031, 2007. Citado nas páginas 52, 53 e 54.

LIU, Y. Novel volatility forecasting using deep learning–long short term memory recurrent neural networks. **Expert Systems with Applications**, Elsevier, v. 132, p. 99–109, 2019. Citado na página 35.

LIVAN, G.; INOUE, J. ichi; SCALAS, E. On the non-stationarity of financial time series: impact on optimal portfolio selection. **Journal of Statistical Mechanics: Theory and Experiment**, v. 2012, n. 07, p. P07025, 2012. Citado na página 59.

LONG, W.; LU, Z.; CUI, L. Deep learning-based feature engineering for stock price movement prediction. **Knowledge-Based Systems**, Elsevier, v. 164, p. 163–173, 2019. Citado na página 35.

LÜ, L.; ZHOU, T. Link prediction in weighted networks: The role of weak ties. **EPL (Europhysics Letters)**, IOP Publishing, v. 89, n. 1, p. 18001, 2010. Citado nas páginas 52, 53 e 70.

LUCKIN, R.; HOLMES, W.; GRIFFITHS, M.; FORCIER, L. B. Intelligence unleashed: An argument for ai in education. Pearson Education, 2016. Citado na página 50.

LUXBURG, U. V.; SCHÖLKOPF, B. Statistical learning theory: Models, concepts, and results. In: **Handbook of the History of Logic**. [S.l.]: Elsevier, 2011. v. 10, p. 651–706. Citado na página 50.

MADSEN, H. **Time series analysis**. [S.l.]: CRC Press, 2007. Citado na página 43.

MANTEGNA, R.; KADTKE, J. B.; BULSARA, A. Degree of correlation inside a financial market. In: AIP. **AIP Conference Proceedings**. [S.l.], 1997. v. 411, n. 1, p. 197–202. Citado nas páginas 47 e 48.

MANTEGNA, R. N. Hierarchical structure in financial markets. **The European Physical Journal B - Condensed Matter and Complex Systems**, v. 11, n. 1, p. 193–197, 1999. ISSN 1434-6036. Disponível em: <http://dx.doi.org/10.1007/s100510050929>. Citado nas páginas 15, 35, 47, 48, 50, 59, 60, 62 e 64.

MANTEGNA, R. N.; STANLEY, H. E. **Introduction to econophysics: correlations and complexity in finance**. [S.l.]: Cambridge university press, 1999. Citado nas páginas 35, 47 e 102.

MARA, A. C.; LIJFFIJT, J.; BIE, T. D. Benchmarking network embedding models for link prediction: Are we making progress? In: IEEE. **2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)**. [S.l.], 2020. p. 138–147. Citado nas páginas 55 e 56.

MARKOWITZ, H. Portfolio selection. **The Journal of Finance**, [American Finance Association, Wiley], v. 7, n. 1, p. 77–91, 1952. ISSN 00221082, 15406261. Disponível em: <http://www.jstor.org/stable/2975974>. Citado nas páginas 57, 70, 99, 100 e 106.

MARTI, G.; NIELSEN, F.; BIŃKOWSKI, M.; DONNAT, P. A review of two decades of correlations, hierarchies, networks and clustering in financial markets. **Progress in Information Geometry**, Springer, p. 245–274, 2021. Citado nas páginas 37, 47, 59, 60, 62 e 100.

MARTI, G.; VERY, P.; DONNAT, P.; NIELSEN, F. A proposal of a methodological framework with experimental guidelines to investigate clustering stability on financial time series. In: IEEE. **2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)**. [S.l.], 2015. p. 32–37. Citado na página 74.

MARTÍNEZ, V.; BERZAL, F.; CUBERO, J.-C. A survey of link prediction in complex networks. **ACM Computing Surveys (CSUR)**, ACM, v. 49, n. 4, p. 69, 2017. Citado na página 52.

MESQUITA, C. M.; VALLE, C. A.; PEREIRA, A. C. Dynamic portfolio optimization using a hybrid mlp-har approach. In: IEEE. **2020 IEEE Symposium Series on Computational Intelligence (SSCI)**. [S.l.], 2020. p. 1075–1082. Citado na página 102.

METZ, J.; CALVO, R.; SENO, E. R.; ROMERO, R. A.; LIANG, Z. Redes complexas: conceitos e aplicações. **Relatórios Técnicos do ICMC-USP São Carlos**, 2007. Citado na página 46.

MISHRA, S. K.; PANDA, G.; MAJHI, B. Prediction based mean-variance model for constrained portfolio assets selection using multiobjective evolutionary algorithms. **Swarm and Evolutionary Computation**, Elsevier, v. 28, p. 117–130, 2016. Citado nas páginas 100 e 102.

MITRA, G.; BARTOLOMEO, D. di; BANERJEE, A.; YU, X. Automated analysis of news to compute market sentiment: Its impact on liquidity and trading. 2015. Citado na página 52.

MITRA, G.; MITRA, L. **The Handbook of News Analytics in Finance**. [S.l.]: Wiley, 2011. (The Wiley Finance Series). Citado na página 52.

MITRA, L. R.; MITRA, G.; BARTOLOMEO, D. di. Equity portfolio risk (volatility) estimation using market information and sentiment. 2008. Citado na página 52.

MORALES, R.; MATTEO, T. D.; ASTE, T. Non-stationary multifractality in stock returns. **Physica A: Statistical Mechanics and its Applications**, v. 392, n. 24, p. 6470 – 6483, 2013. ISSN 0378-4371. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0378437113007668>. Citado na página 59.

MORALES, R.; MATTEO, T. D.; GRAMATICA, R.; ASTE, T. Dynamical generalized hurst exponent as a tool to monitor unstable periods in financial time series. **Physica A: Statistical Mechanics and its Applications**, North-Holland, v. 391, n. 11, p. 3180–3189, 2012. Citado na página 60.

MUSMECI, N.; ASTE, T.; MATTEO, T. D. Interplay between past market correlation structure changes and future volatility outbursts. **Scientific Reports**, Nature Publishing Group, v. 6, p. 36320, 2016. Citado nas páginas 37, 60 e 71.

MUSMECI, N.; ASTE, T.; MATTEO, T. di. Clustering and hierarchy of financial markets data: advantages of the dbht. **CoRR**, 2014. Citado nas páginas 48 e 62.

MUSMECI, N.; NICOSIA, V.; ASTE, T.; MATTEO, T. D.; LATORA, V. The multiplex dependency structure of financial markets. **Complexity**, Hindawi, v. 2017, 2017. Citado na página 60.

MUTLU, E. C.; OGHAZ, T. A. Review on graph feature learning and feature extraction techniques for link prediction. **Proceedings of ACM**, 2019. Citado nas páginas 53, 54 e 67.

NARASIMHAN, J. S. **Link prediction in dynamic networks**. [S.l.]: Washington State University, 2015. Citado na página 105.

NEWMAN, M. E. The structure and function of complex networks. **SIAM review**, SIAM, v. 45, n. 2, p. 167–256, 2003. Citado na página 46.

OLIVEIRA, M.; GAMA, J. An overview of social network analysis. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, Wiley Online Library, v. 2, n. 2, p. 99–115, 2012. Citado nas páginas 66, 67 e 105.

ONNELA, J.-P.; CHAKRABORTI, A.; KASKI, K.; KERTESZ, J.; KANTO, A. Asset trees and asset graphs in financial markets. **Physica Scripta**, IOP Publishing, v. 2003, n. T106, p. 48, 2003. Citado nas páginas 48, 62 e 63.

____. Dynamics of market correlations: Taxonomy and portfolio analysis. **Physical Review E**, APS, v. 68, n. 5, p. 056110, 2003. Citado nas páginas 47 e 101.

ONNELA, J.-P.; KASKI, K.; KERTÉSZ, J. Clustering and information in correlation based financial networks. **The European Physical Journal B-Condensed Matter and Complex Systems**, Springer, v. 38, n. 2, p. 353–362, 2004. Citado nas páginas 47, 62 e 63.

OU, M.; CUI, P.; PEI, J.; ZHANG, Z.; ZHU, W. Asymmetric transitivity preserving graph embedding. In: **Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining**. [S.l.: s.n.], 2016. p. 1105–1114. Citado na página 57.

PAFKA, S.; KONDOR, I. Estimated correlation matrices and portfolio optimization. **Physica A: Statistical Mechanics and Its Applications**, Elsevier, v. 343, p. 623–634, 2004. Citado na página 101.

PAGOLU, V. S.; REDDY, K. N.; PANDA, G.; MAJHI, B. Sentiment analysis of twitter data for predicting stock market movements. In: IEEE. **2016 international conference on signal processing, communication, power and embedded system (SCOPES)**. [S.l.], 2016. p. 1345–1350. Citado na página 35.

PARK, J. H.; CHANG, W.; SONG, J. W. Link prediction in the granger causality network of the global currency market. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 553, p. 124668, 2020. Citado nas páginas 37 e 60.

PEARSON, K. Note on regression and inheritance in the case of two parents. **Proceedings of the Royal Society of London**, JSTOR, v. 58, p. 240–242, 1895. Citado na página 103.

PERALTA, G.; ZAREEI, A. A network approach to portfolio selection. **Journal of Empirical Finance**, Elsevier, v. 38, p. 157–180, 2016. Citado nas páginas 100 e 101.

PEROZZI, B.; AL-RFOU, R.; SKIENA, S. Deepwalk: Online learning of social representations. In: **Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York, NY, USA: ACM, 2014. (KDD '14), p. 701–710. ISBN 978-1-4503-2956-9. Disponível em: <http://doi.acm.org/10.1145/2623330.2623732>. Citado na página 56.

POTVIN, J.-Y.; SORIANO, P.; VALLÉE, M. Generating trading rules on the stock markets with genetic programming. **Computers & Operations Research**, Elsevier, v. 31, n. 7, p. 1033–1047, 2004. Citado nas páginas 35 e 51.

POZZI, F.; MATTEO, T. D.; ASTE, T. Centrality and peripherality in filtered graphs from dynamical financial correlations. **Advances in Complex Systems**, World Scientific, v. 11, n. 06, p. 927–950, 2008. Citado nas páginas 37 e 101.

____. Spread of risk across financial markets: better to invest in the peripheries. **Scientific reports**, Nature Publishing Group, v. 3, 2013. Citado nas páginas 60, 101 e 128.

RIBEIRO, L. F.; SAVERESE, P. H.; FIGUEIREDO, D. R. struc2vec: Learning node representations from structural identity. In: **Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining**. [S.l.: s.n.], 2017. p. 385–394. Citado na página 56.

RISH, I. An empirical study of the naive bayes classifier. In: **IJCAI 2001 workshop on empirical methods in artificial intelligence**. [S.l.: s.n.], 2001. v. 3, n. 22, p. 41–46. Citado na página 91.

ROWEIS, S. T.; SAUL, L. K. Nonlinear dimensionality reduction by locally linear embedding. **science**, American Association for the Advancement of Science, v. 290, n. 5500, p. 2323–2326, 2000. Citado na página 56.

RUSSELL, S. J.; NORVIG, P. **Inteligência artificial**. [S.l.]: Elsevier, 2004. Citado na página 50.

SÁ, H. R. D.; PRUDÊNCIO, R. B. Supervised link prediction in weighted networks. In: IEEE. **The 2011 international joint conference on neural networks**. [S.l.], 2011. p. 2281–2288. Citado na página 55.

SHIHAVUDDIN, A.; AMBIA, M. N.; AREFIN, M. M. N.; HOSSAIN, M.; ANWAR, A. Prediction of stock price analyzing the online financial news using naive bayes classifier and local economic trends. In: IEEE. **Advanced Computer Theory and Engineering (ICACTE), 2010 3rd International Conference on**. [S.l.], 2010. v. 4, p. V4–22. Citado na página 51.

SHOJAIE, A. Link prediction in biological networks using multi-mode exponential random graph models. In: CITESEER. **11th Workshop on Mining and Learning with Graphs**. [S.l.], 2013. p. 987–991. Citado na página 52.

SILVA, E.; BRANDAO, H.; CASTILHO, D.; PEREIRA, A. C. A binary ensemble classifier for high-frequency trading. In: IEEE. **Neural Networks (IJCNN), 2015 International Joint Conference on**. [S.l.], 2015. p. 1–8. Citado nas páginas 15 e 45.

SILVA, E.; CASTILHO, D.; PEREIRA, A.; BRANDAO, H. A neural network based approach to support the market making strategies in high-frequency trading. In: IEEE. **Neural Networks (IJCNN), 2014 International Joint Conference on**. [S.l.], 2014. p. 845–852. Citado na página 43.

SOARES, P. R. da S.; PRUDÊNCIO, R. B. C. Time series based link prediction. In: IEEE. **Neural Networks (IJCNN), The 2012 International Joint Conference on**. [S.l.], 2012. p. 1–7. Citado na página 57.

SONG, W.-M.; ASTE, T.; MATTEO, T. D. Analysis on filtered correlation graph for information extraction. **Statistical Mechanics of Molecular Biophysics**, p. 88, 2008. Citado na página 60.

SONG, W.-M.; MATTEO, T. D.; ASTE, T. Hierarchical information clustering by means of topologically embedded graphs. **PLoS ONE**, Public Library of Science, v. 7, n. 3, p. 1–14, 03 2012. Disponível em: <http://dx.doi.org/10.1371%2Fjournal.pone.0031929>. Citado nas páginas 60 e 62.

SOUZA, T. T. P.; ASTE, T. Predicting future stock market structure by combining social and financial network information. **Physica A: Statistical Mechanics and its Applications**, v. 535, p. 122343, 2019. ISSN 0378-4371. Citado nas páginas 37, 60, 63, 70 e 71.

SPELTA, A. Financial market predictability with tensor decomposition and links forecast. **Applied network science**, Nature Publishing Group, v. 2, n. 1, p. 7, 2017. Citado na página 60.

SRILATHA, P.; MANJULA, R. Similarity index based link prediction algorithms in social networks: A survey. **Journal of Telecommunications and Information Technology**, 2016. Citado na página 70.

TABAK, B. M.; SERRA, T. R.; CAJUEIRO, D. O. Topological properties of stock market networks: The case of brazil. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 389, n. 16, p. 3240–3249, 2010. Citado nas páginas 48 e 49.

TRIPPI, R. R.; DESIENO, D. Trading equity index futures with a neural network. **The Journal of Portfolio Management**, Institutional Investor Journals, v. 19, n. 1, p. 27–33, 1992. Citado nas páginas 35 e 51.

TRIPPI, R. R.; TURBAN, E. **Neural networks in finance and investing: Using artificial intelligence to improve real world performance**. [S.l.]: McGraw-Hill, Inc., 1992. Citado na página 51.

TSAI, C.-F.; LIN, Y.-C.; YEN, D. C.; CHEN, Y.-M. Predicting stock returns by classifier ensembles. **Applied Soft Computing**, Elsevier, v. 11, n. 2, p. 2452–2459, 2011. Citado na página 51.

TUMMINELLO, M.; ASTE, T.; MATTEO, T. D.; MANTEGNA, R. N. A tool for filtering information in complex systems. **Proceedings of the National Academy of Sciences of the United States of America**, National Acad Sciences, v. 102, n. 30, p. 10421–10426, 2005. Citado na página 62.

TUMMINELLO, M.; LILLO, F.; MANTEGNA, R. N. Correlation, hierarchies, and networks in financial markets. **Journal of Economic Behavior & Organization**, v. 75, n. 1, p. 40 – 58, 2010. ISSN 0167-2681. Transdisciplinary Perspectives on Economic Complexity. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0167268110000077>. Citado nas páginas 59 e 60.

TUMMINELLO, M.; MATTEO, T. D.; ASTE, T.; MANTEGNA, R. N. Correlation based networks of equity returns sampled at different time horizons. **The European Physical Journal B**, Springer, v. 55, n. 2, p. 209–217, 2007. Citado na página 63.

WANG, D.; CHANG, P.-C.; WU, J.-L.; ZHOU, C. A partially connected neural evolutionary network for stock price index forecasting. In: SPRINGER. **International Conference on Intelligent Computing**. [S.l.], 2011. p. 14–19. Citado na página 51.

WANG, D.; CUI, P.; ZHU, W. Structural deep network embedding. In: **Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining**. [S.l.: s.n.], 2016. p. 1225–1234. Citado na página 55.

WANG, G.-J.; XIE, C.; HAN, F.; SUN, B. Similarity measure and topology evolution of foreign exchange markets using dynamic time warping method: Evidence from minimal spanning tree. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 391, n. 16, p. 4136–4146, 2012. Citado na página 48.

WANG, P.; XU, B.; WU, Y.; ZHOU, X. Link prediction in social networks: the state-of-the-art. **Science China Information Sciences**, Springer, v. 58, n. 1, p. 1–38, 2015. Citado na página 53.

WANG, Y.-H. Nonlinear neural network forecasting model for stock index option price: Hybrid gjr–garch approach. **Expert Systems with Applications**, Elsevier, v. 36, n. 1, p. 564–570, 2009. Citado na página 51.

WASSERMAN, S.; FAUST, K.; PRESS, C. U.; GRANOVETTER, M.; CAMBRIDGE, U. of; IACOBUCCI, D. **Social Network Analysis: Methods and Applications**. Cambridge University Press, 1994. (Structural Analysis in the Social Sciences). ISBN 9780521387071. Disponível em: <https://books.google.com.br/books?id=CAm2DpIqRUIC>. Citado na página 85.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of'small-world'networks. **nature**, Nature Publishing Group, v. 393, n. 6684, p. 440, 1998. Citado na página 47.

YANG, C.; CHEN, Y.; NIU, L.; LI, Q. Cointegration analysis and influence rank—a network approach to global stock markets. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 400, p. 168–185, 2014. Citado nas páginas 35, 48 e 101.

YANG, Y.; YANG, H. Complex network-based time series analysis. **Physica A: Statistical Mechanics and its Applications**, v. 387, n. 5-6, p. 1381–1386, 2008. ISSN 03784371. Citado na página 63.

YOON, Y.; SWALES, G. Predicting stock price performance: A neural network approach. In: IEEE. **System Sciences, 1991. Proceedings of the Twenty-Fourth Annual Hawaii International Conference on**. [S.l.], 1991. v. 4, p. 156–162. Citado na página 51.

ZHANG, G.; PATUWO, B. E.; HU, M. Y. Forecasting with artificial neural networks:: The state of the art. **International journal of forecasting**, Elsevier, v. 14, n. 1, p. 35–62, 1998. Citado na página 51.

ZHAO, L.; WANG, G.-j.; WANG, M.; BAO, W.; LI, W.; STANLEY, H. E. Stock market as temporal network. **Physica A**, Elsevier B.V., v. 506, p. 1104–1112, 2018. ISSN 0378-4371. Citado nas páginas 100 e 102.

ZHOU, T.; LÜ, L.; ZHANG, Y.-C. Predicting missing links via local information. **The European Physical Journal B**, Springer, v. 71, n. 4, p. 623–630, 2009. Citado na página 54.

ZITZLER, E. **Evolutionary algorithms for multiobjective optimization: Methods and applications**. [S.l.]: Citeseer, 1999. v. 63. Citado na página 109.

APÊNDICE

A

RESULTADOS COMPLEMENTARES DE PREVISÃO DE FORMAÇÃO DE LINKS

A.1 Pseudo-Algoritmo do Método de Previsão de Links

Algoritmo 1 – Abordagem de Previsão de Links

Requer: $G_T(V,E)$
$k \leftarrow 30$
$t \leftarrow k$
enquanto $t \leq T$ **faça**

Criando o conjunto de treinamento
para $i \leftarrow 1,k$ **faça**
$node\_features \leftarrow extractNodeFeatures(G_{t-i}(V,E))$
$link\_features \leftarrow extractLinkFeatures(G_{t-i}(V,E))$
$train\_set \leftarrow combine(train\_set, concatenate(node\_features, link\_features))$
fim para

Criando o conjunto de teste
$node\_features \leftarrow extractNodeFeatures(G_t(V,E))$
$link\_features \leftarrow extractLinkFeatures(G_t(V,E))$
$test\_set \leftarrow concatenate(node\_features, link\_features)$

Treino e teste do modelo de aprendizagem de máquina
$model \leftarrow train(train\_set)$
$prediction \leftarrow predict(model, test\_set)$
$evaluate(prediction)$

$t \leftarrow t + \delta T$
fim enquanto

A.2 Parâmetros dos Algoritmos de Aprendizagem de Máquina

Esta seção apresenta o conjunto de parâmetros usados no algoritmo XGboost para executar os experimentos. Os atributos de configuração são:

- booster = “gbtree”;
- objective = “reg:linear”;
- eta = 0.05;
- max_depth = 2;
- min_child_weight = 100;

A.3 Análises Descritivas Complementares

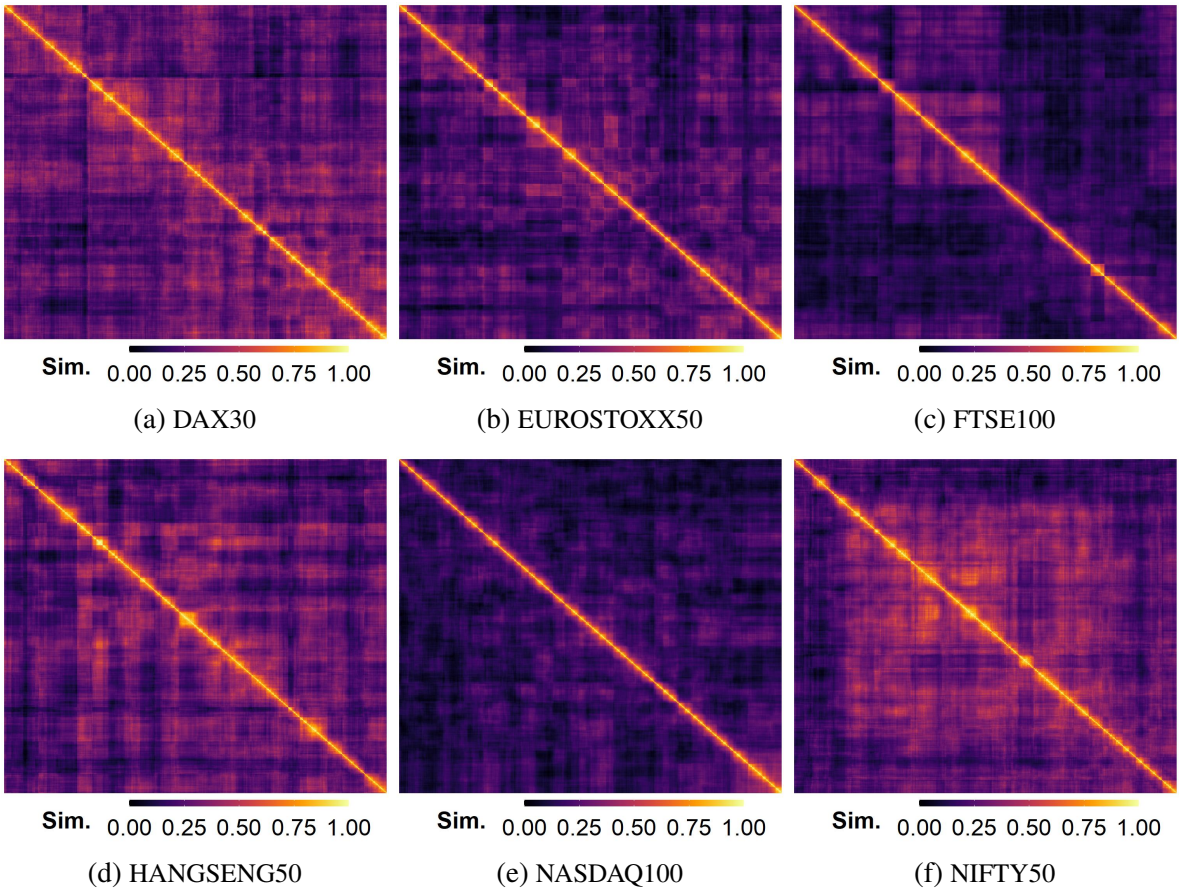
Essa seção apresenta um conjunto de resultados complementares relacionados a análises descritivas de redes financeiras criadas utilizando diferentes índices de mercado. Os resultados apresentados abaixo ilustram a persistência da rede financeira durante o período de testes, medida através da similaridade de Jaccard entre redes $G(t)$ e $G(t')$, considerando $L \in 126, 504$ dias de negociação para criar cada rede. Tais análises permitem medir como as redes financeiras mudam sua estrutura ao longo do tempo. As Figuras 38, 39 e 40 apresentam resultados para $L = 126$ utilizando os métodos DAG, DTN e DMST. As Figuras 41, 42 e 43 apresentam resultados para $L = 504$ utilizando os métodos DAG, DTN e DMST.

A.4 Análises Preditivas Complementares

Esta seção fornece resultados experimentais relacionados a previsão de redes financeiras usando o tamanho da janela deslizante $L = 126$ e $L = 504$ dias de negociação para construir as redes. Para cada intervalo de tempo $\{_{1,2,\ldots,20\}}^{h=}\}$, realizamos a previsão da rede $G(t+h)$ e calculamos a AUC média de cada método e seu respectivo erro padrão durante o período de teste compreendido entre 5 maio 2007 e 18 de dezembro de 2019.

Considerando $L = 126$, Figuras 44, 45 e 46 mostram a medida AUC do método de aprendizagem de máquina proposto em comparação com algoritmos de *benchmark* para métodos de filtragem de rede DAG, DTN e DMST, respectivamente.

Considerando $L = 504$, Figuras 48, 49 e 50 mostram a medida AUC do método de aprendizagem de máquina proposto em comparação com algoritmos de *benchmark* para métodos de filtragem de rede DAG, DTN e DMST, respectivamente.



(a) DAX30 (b) EUROSTOXX50 (c) FTSE100

(d) HANGSENG50 (e) NASDAQ100 (f) NIFTY50

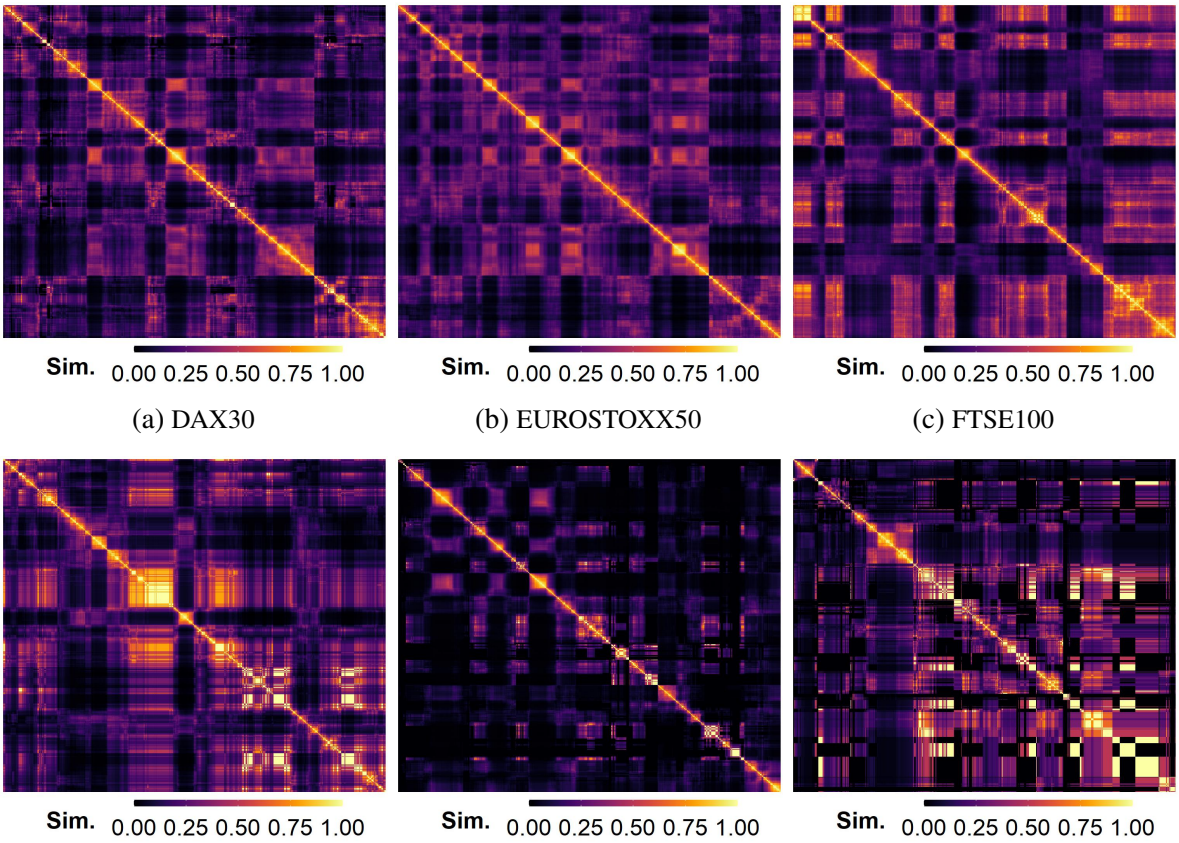
Figura 38 – **DAG - Matriz de similaridade cruzada para cada índice de mercado.** Calculamos a persistência da rede usando a matriz de similaridade entre as redes $G(t)$ e $G(t')$ usando o coeficiente de Jaccard, considerando $L = 126$. Para criar as matrizes de similaridade, utilizamos dados de cada índice de mercado durante o período entre 12 de maio de 2006 e 18 de dezembro de 2019. Para cada figura de índice de mercado, a primeira rede em 12 de maio de 2006 é representada na parte superior esquerda e a última rede em 18 de dezembro de 2019 na parte inferior direita de cada figura individual.

As Figuras 47 e 51 apresentam o desempenho AUC e a melhoria AUC * do método proposto usando $L = 126$ e $L = 504$, respectivamente, para h semanas de negociação à frente ($1 \leq h \leq 20$). Os resultados são fornecidos para os métodos de filtragem de rede DAG, DTN e DMST. A melhoria de AUC * é calculada sobre o método de *benchmark* invariante no tempo (TI).

A.4.1 Comparação entre Algoritmos de Aprendizado de Máquina

Essa seção apresenta um conjunto de resultados experimentais visando comparar o desempenho preditivo de diferente algoritmos de aprendizagem de máquinas.

Para $L = 126$, as Figuras 58, 59 e 60 mostram a AUC e o erro padrão dos algoritmos de aprendizagem de máquina para previsão de h semanas à frente ($1 \leq h \leq 20$), considerando os métodos de filtragem de rede DAG, DTN e DMST, respectivamente.



(a) DAX30 (b) EUROSTOXX50 (c) FTSE100

(d) HANGSENG50 (e) NASDAQ100 (f) NIFTY50

Figura 39 – **DTN - Matriz de similaridade cruzada para cada índice de mercado.** Calculamos a persistência da rede usando a matriz de similaridade entre as redes $G(t)$ e $G(t')$ usando o coeficiente de Jaccard, considerando $L = 126$. Para criar as matrizes de similaridade, utilizamos dados de cada índice de mercado durante o período entre 12 de maio de 2006 e 18 de dezembro de 2019. Para cada figura de índice de mercado, a primeira rede em 12 de maio de 2006 é representada na parte superior esquerda e a última rede em 18 de dezembro de 2019 na parte inferior direita de cada figura individual.

Para $L = 504$, as Figuras 61, 62 e 63 apresentam a AUC e o erro padrão dos algoritmos de aprendizagem de máquina para previsão de h semanas à frente ($1 \leq h \leq 20$), considerando os métodos de filtragem de rede DAG, DTN e DMST, respectivamente.

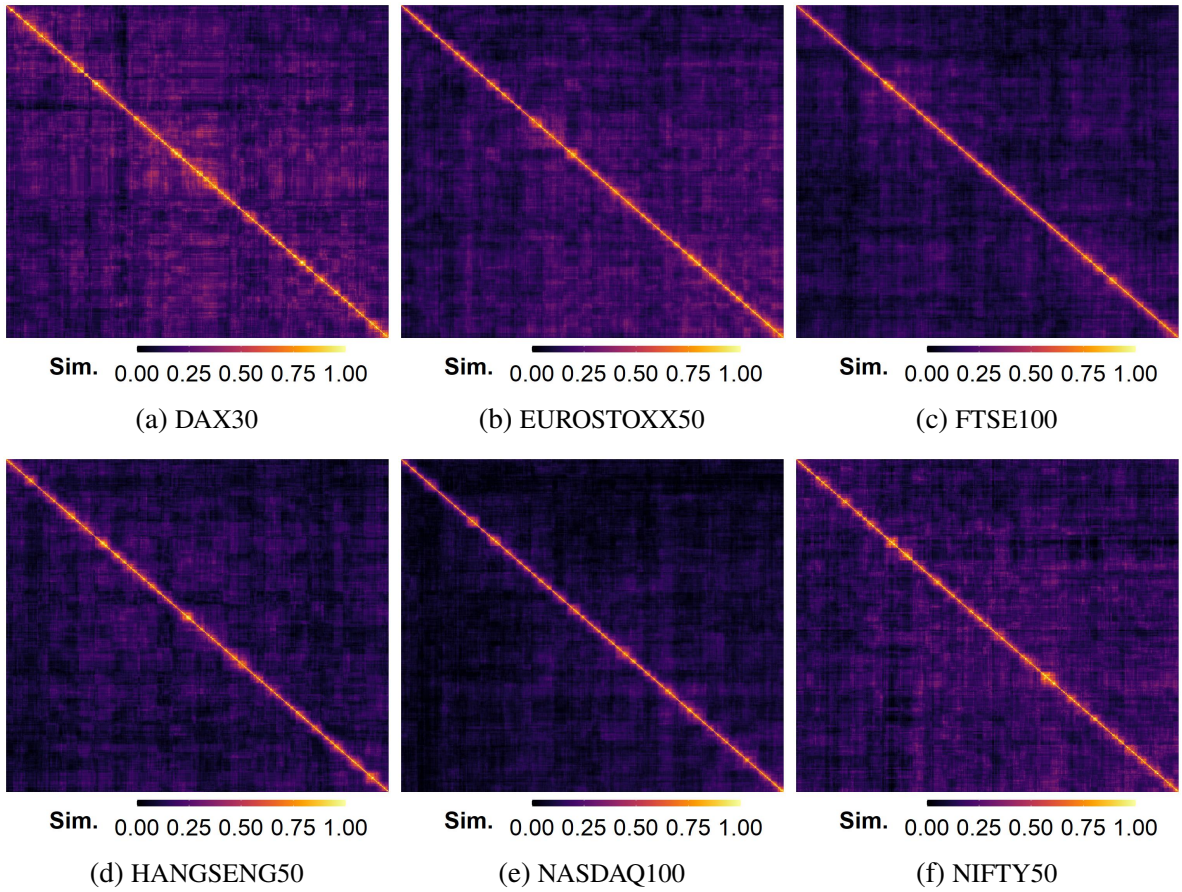
A.4.2 Comparação Algoritmos de Embedding

Essa seção apresenta uma análise comparativa utilizando algoritmos de *embedding* para previsão de links em redes de ações, considerando $L \in 126, 504$.

análise comparativa utilizando algoritmos de *embedding* para previsão de links

Para $L = 126$, as Figuras 64, 65 e 66 mostram a AUC e o erro padrão dos algoritmos de aprendizagem de máquina e de *embedding* para previsão de h semanas à frente ($1 \leq h \leq 20$), considerando os métodos de filtragem de rede DAG, DTN e DMST, respectivamente.

Para $L = 504$, as Figuras 67, 68 e 69 apresentam a AUC e o erro padrão dos algoritmos

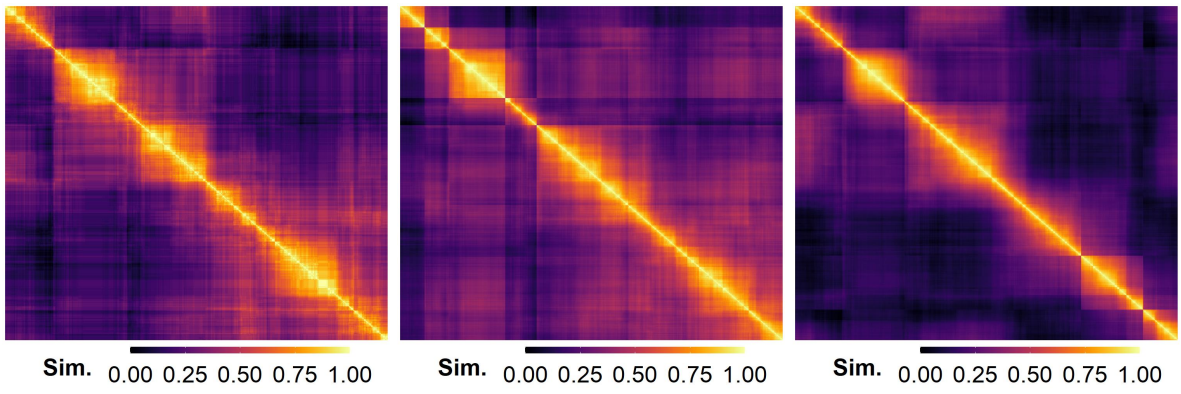


(a) DAX30 (b) EUROSTOXX50 (c) FTSE100

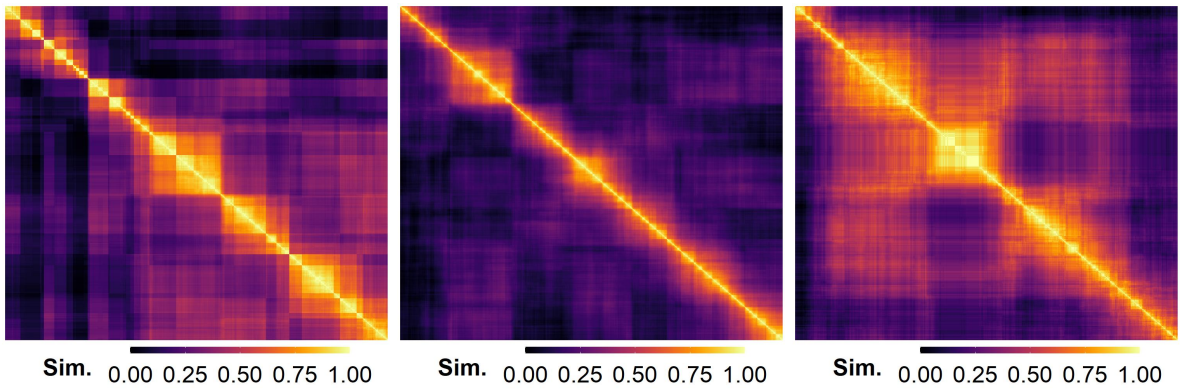
(d) HANGSENG50 (e) NASDAQ100 (f) NIFTY50

Figura 40 – **DMST - Matriz de similaridade cruzada para cada índice de mercado.** Calculamos a persistência da rede usando a matriz de similaridade entre as redes $G(t)$ e $G(t')$ usando o coeficiente de Jaccard, considerando $L = 126$. Para criar as matrizes de similaridade, utilizamos dados de cada índice de mercado durante o período entre 12 de maio de 2006 e 18 de dezembro de 2019. Para cada figura de índice de mercado, a primeira rede em 12 de maio de 2006 é representada na parte superior esquerda e a última rede em 18 de dezembro de 2019 na parte inferior direita de cada figura individual.

de aprendizagem de máquina e de *embedding* para previsão de h semanas à frente ($1 \leq h \leq 20$), considerando os métodos de filtragem de rede DAG, DTN e DMST, respectivamente.

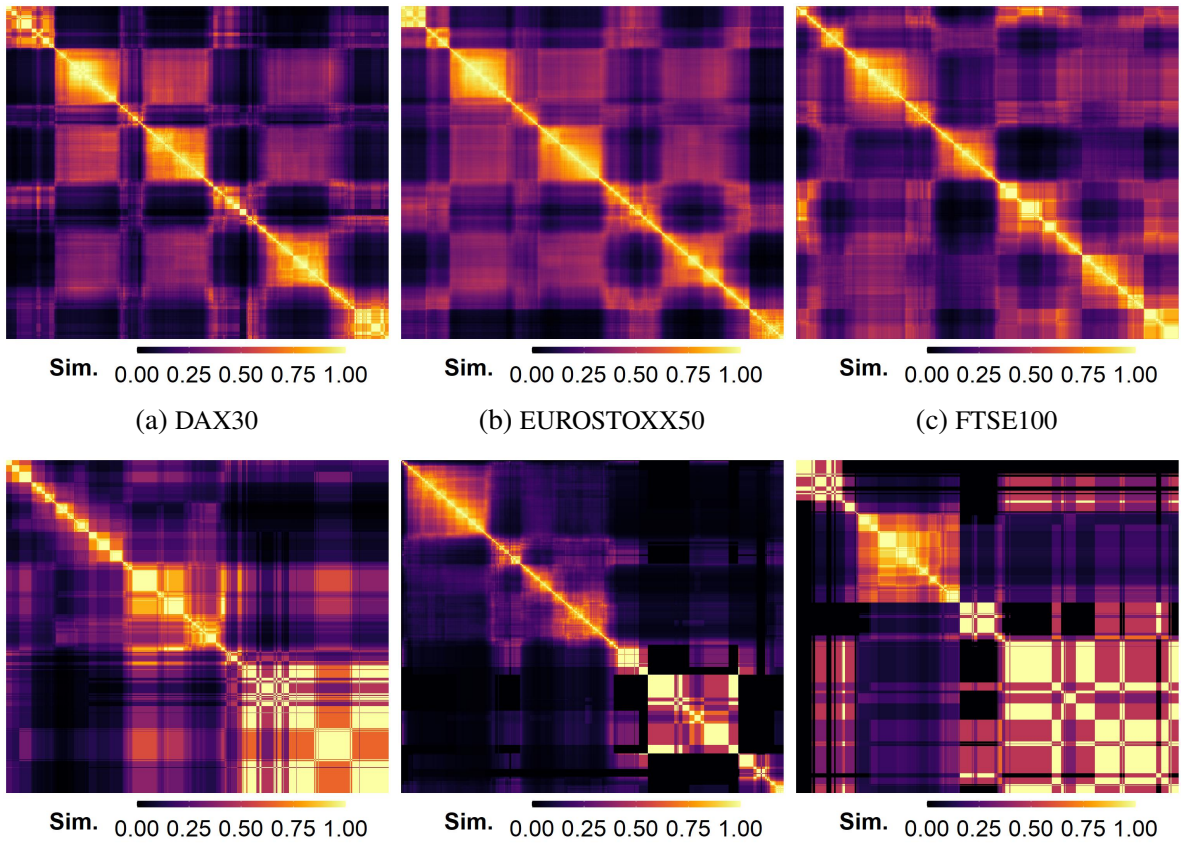


(a) DAX30 (b) EUROSTOXX50 (c) FTSE100



(d) HANGSENG50 (e) NASDAQ100 (f) NIFTY50

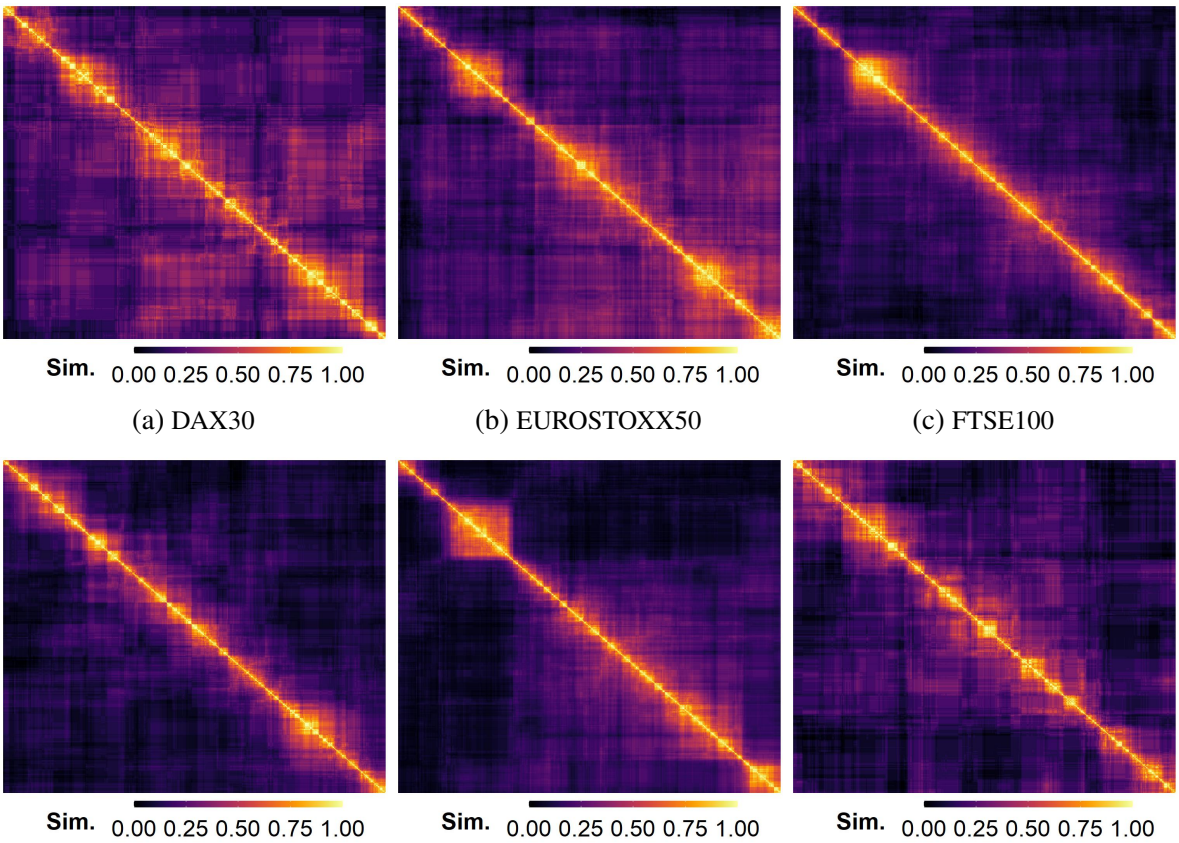
Figura 41 – **DAG - Matriz de similaridade cruzada para cada índice de mercado.** Calculamos a persistência da rede usando a matriz de similaridade entre as redes $G(t)$ e $G(t')$ usando o coeficiente de Jaccard, considerando $L = 504$. Para criar as matrizes de similaridade, utilizamos dados de cada índice de mercado durante o período entre 12 de maio de 2006 e 18 de dezembro de 2019. Para cada figura de índice de mercado, a primeira rede em 12 de maio de 2006 é representada na parte superior esquerda e a última rede em 18 de dezembro de 2019 na parte inferior direita de cada figura individual.



(a) DAX30 (b) EUROSTOXX50 (c) FTSE100

(d) HANGSENG50 (e) NASDAQ100 (f) NIFTY50

Figura 42 – **DTN - Matriz de similaridade cruzada para cada índice de mercado.** Calculamos a persistência da rede usando a matriz de similaridade entre as redes $G(t)$ e $G(t')$ usando o coeficiente de Jaccard, considerando $L = 504$. Para criar as matrizes de similaridade, utilizamos dados de cada índice de mercado durante o período entre 12 de maio de 2006 e 18 de dezembro de 2019. Para cada figura de índice de mercado, a primeira rede em 12 de maio de 2006 é representada na parte superior esquerda e a última rede em 18 de dezembro de 2019 na parte inferior direita de cada figura individual.



(a) DAX30 (b) EUROSTOXX50 (c) FTSE100

(d) HANGSENG50 (e) NASDAQ100 (f) NIFTY50

Figura 43 – **DMST - Matriz de similaridade cruzada para cada índice de mercado.** Calculamos a persistência da rede usando a matriz de similaridade entre as redes $G(t)$ e $G(t')$ usando o coeficiente de Jaccard, considerando $L = 504$. Para criar as matrizes de similaridade, utilizamos dados de cada índice de mercado durante o período entre 12 de maio de 2006 e 18 de dezembro de 2019. Para cada figura de índice de mercado, a primeira rede em 12 de maio de 2006 é representada na parte superior esquerda e a última rede em 18 de dezembro de 2019 na parte inferior direita de cada figura individual.

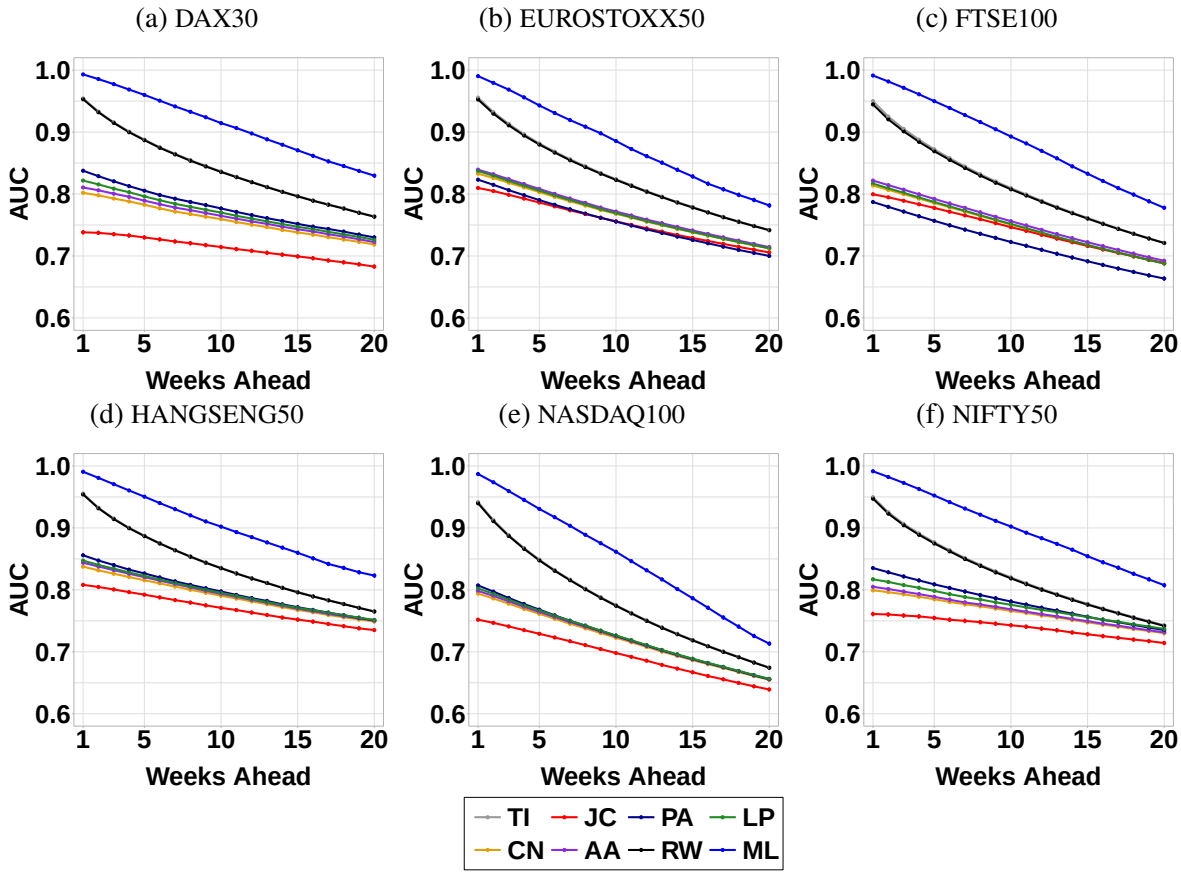


Figura 44 – **DAG - Comparação de desempenho preditiva de todos os métodos.** A figura mostra a medida de AUC e o desvio do método de aprendizado de máquina em comparação com os métodos base para previsão de links. Para cada intervalo de tempo, calculamos a média AUC de cada método e seu respectivo erro padrão ao longo de todo o período de testes. O método de aprendizado de máquina supera os métodos base em todos os índices de mercado.

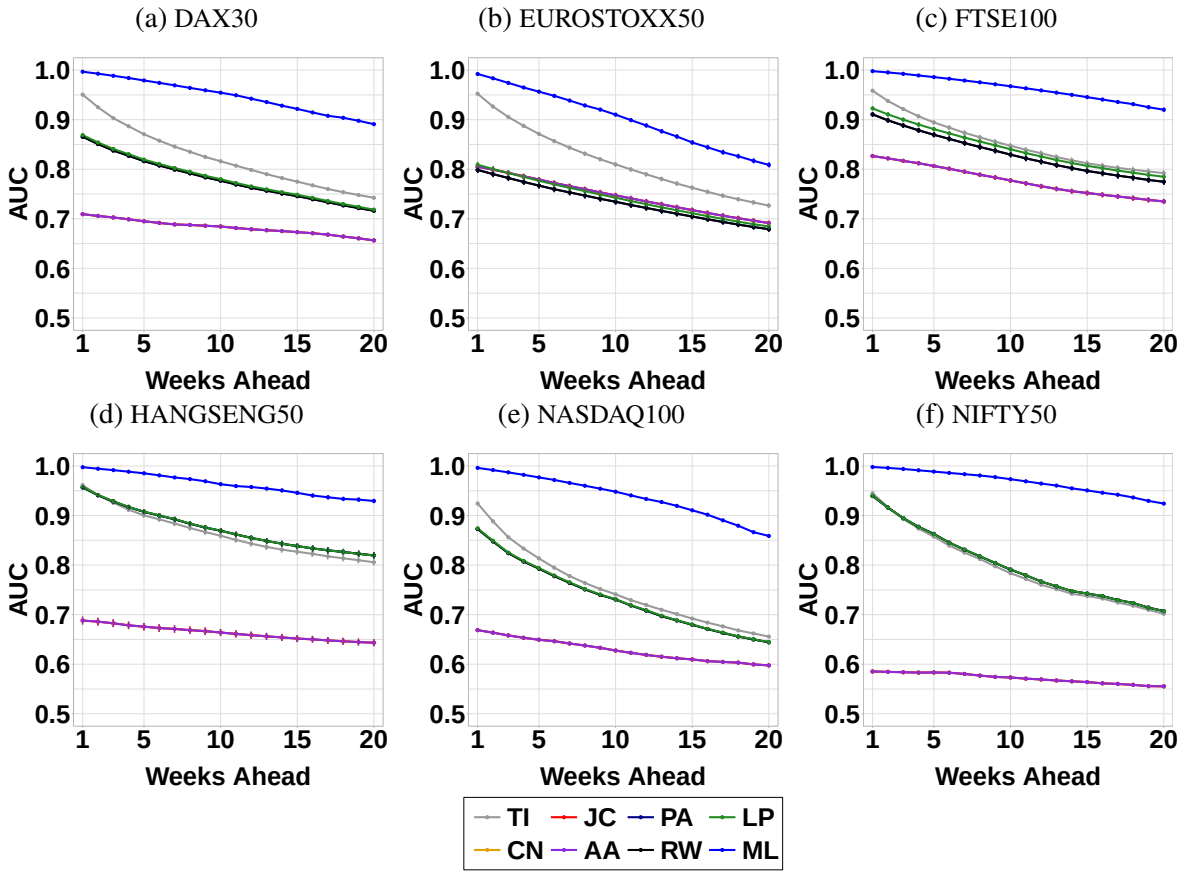


Figura 45 – **DTN - Comparação de desempenho preditiva de todos os métodos.** A figura mostra a medida de AUC e o desvio do método de aprendizado de máquina em comparação com os métodos base para previsão de links. Para cada intervalo de tempo, calculamos a média AUC de cada método e seu respectivo erro padrão ao longo de todo o período de testes. O método de aprendizado de máquina supera os métodos base em todos os índices de mercado.

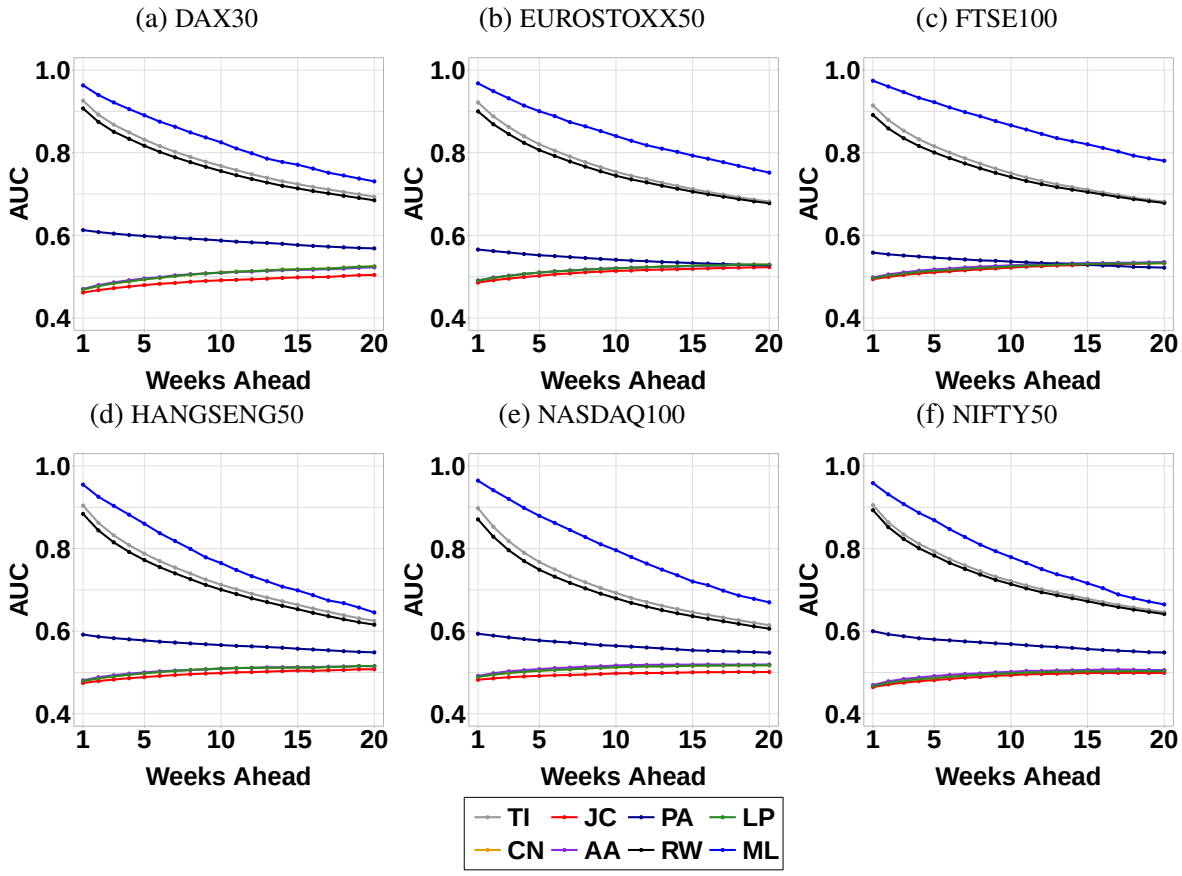
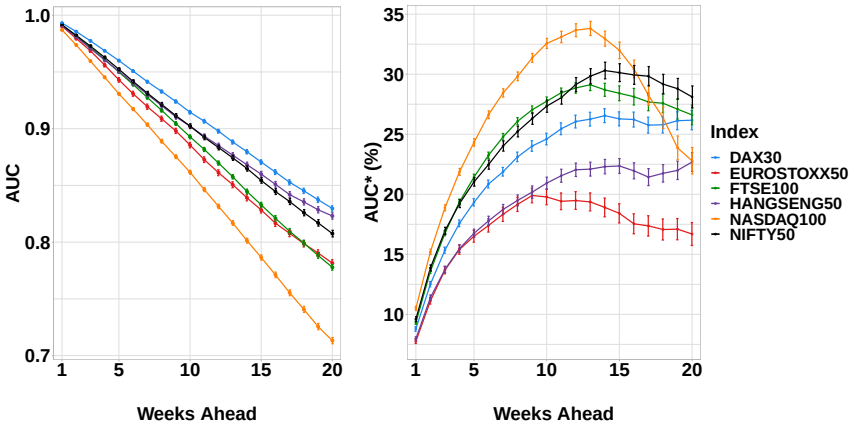
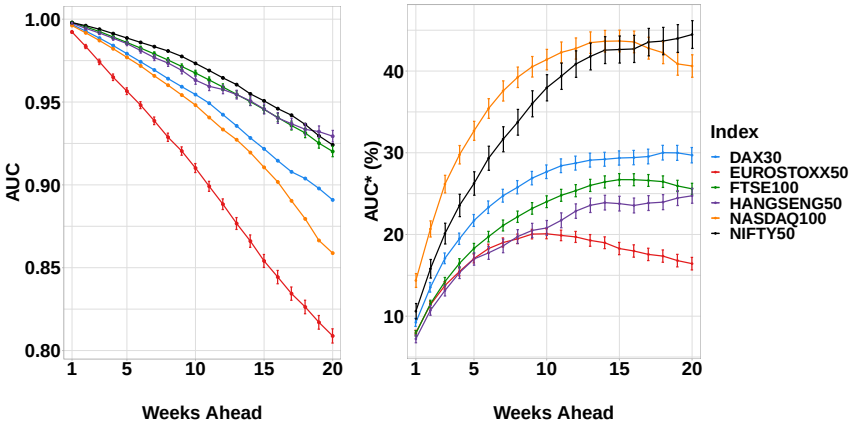


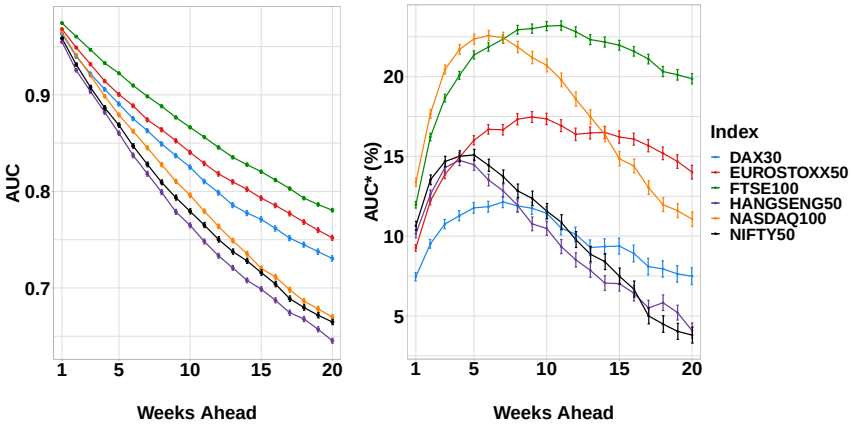
Figura 46 – **DMST - Comparação de desempenho preditiva de todos os métodos.** A figura mostra a medida de AUC e o desvio do método de aprendizado de máquina em comparação com os métodos base para previsão de links. Para cada intervalo de tempo, calculamos a média AUC de cada método e seu respectivo erro padrão ao longo de todo o período de testes. O método de aprendizado de máquina supera os métodos base em todos os índices de mercado.



(a) DAG (AUC) (b) DAG (AUC*)



(c) DTN (AUC) (d) DTN (AUC*)



(e) DMST (AUC) (f) DMST (AUC*)

Figura 47 – **AUC e AUC* do algoritmo ML para métodos de filtragem de rede DAG, DTN e DMST.** Os painéis (a), (c) e (e) apresentam a métrica AUC do aprendizado de máquina e seu erro padrão para previsão de h semanas de negociação à frente ($1 \leq h \leq 20$). Os painéis (b), (d) e (f) apresentam a melhoria da AUC em relação ao método invariante no tempo e seu erro padrão. Os resultados apresentados são relacionados a $L = 126$.

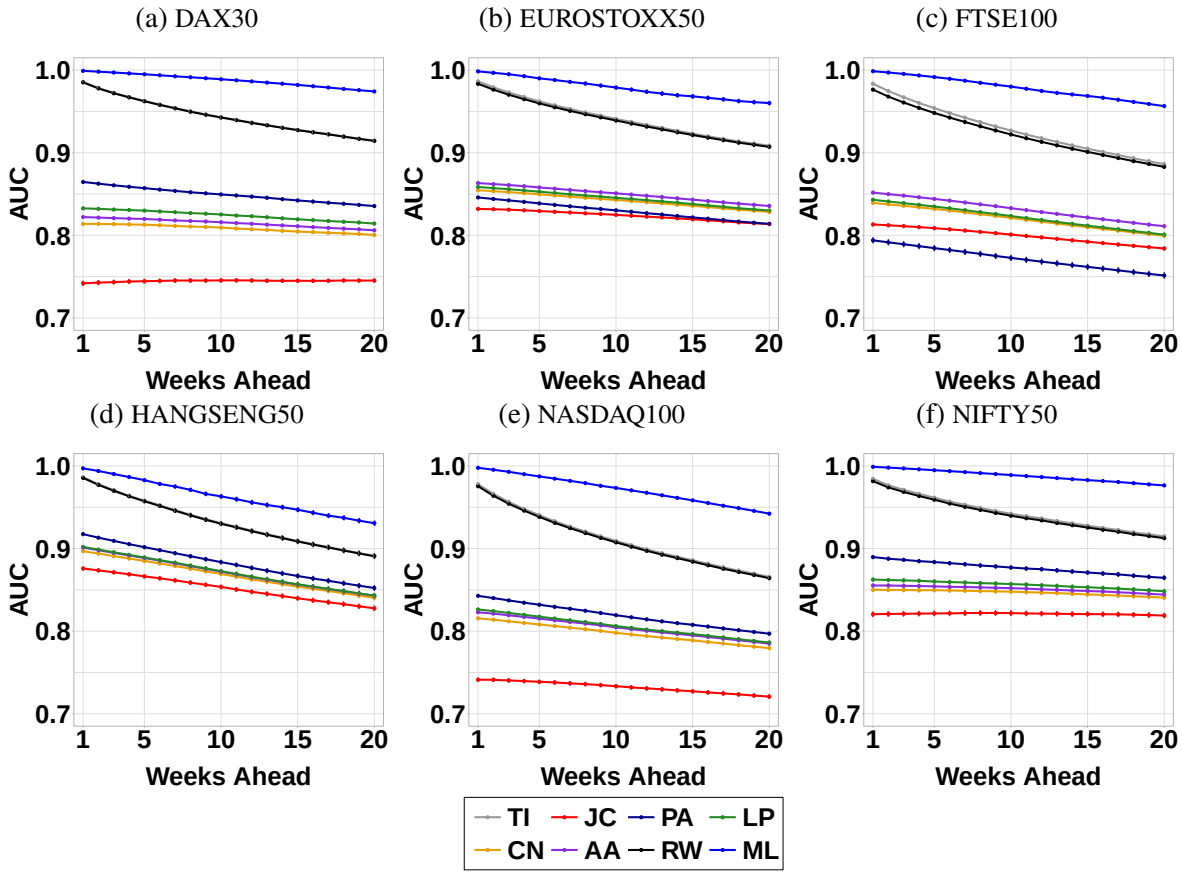


Figura 48 – **DAG - Comparação de desempenho preditiva de todos os métodos.** A figura mostra a medida de AUC e o desvio do método de aprendizado de máquina em comparação com os métodos base para previsão de links. Para cada intervalo de tempo, calculamos a média AUC de cada método e seu respectivo erro padrão ao longo de todo o período de testes. O método de aprendizado de máquina supera os métodos base em todos os índices de mercado.

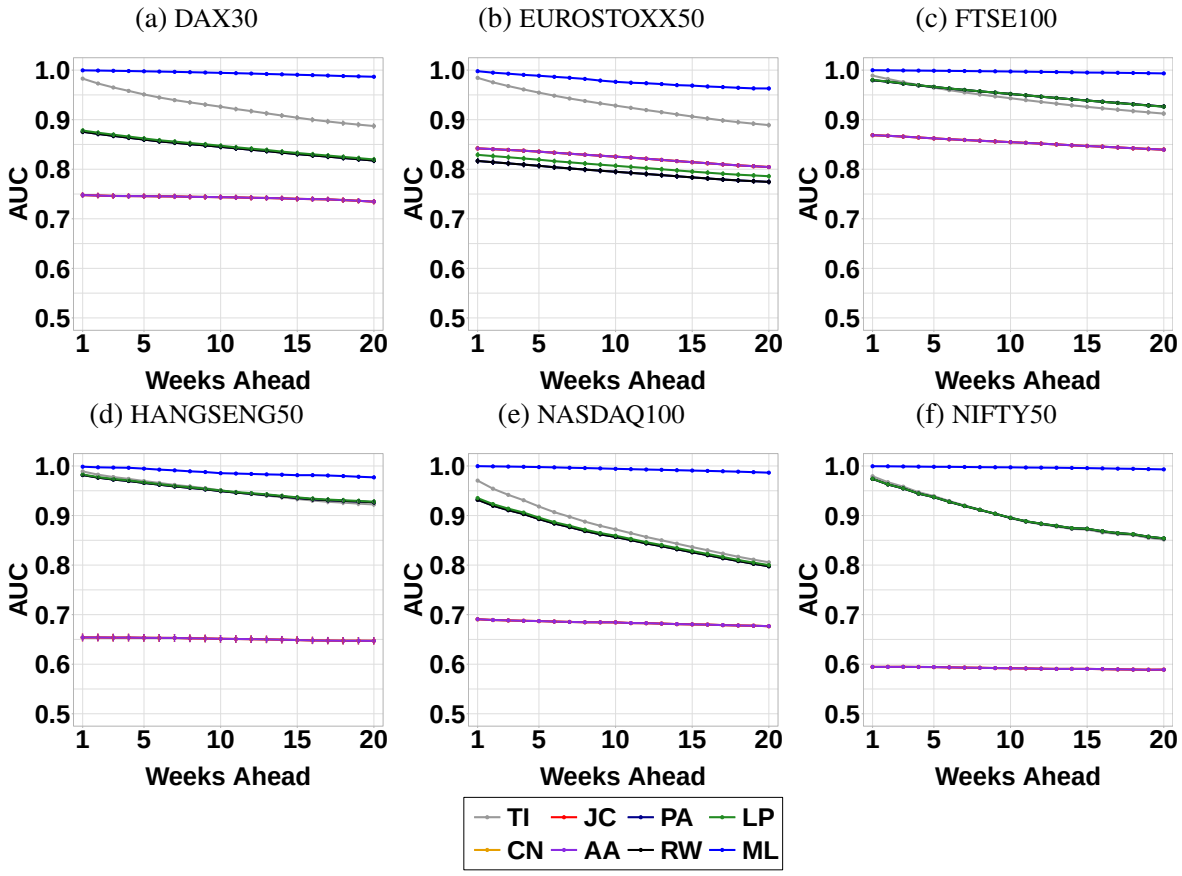


Figura 49 – **DTN - Comparação de desempenho preditiva de todos os métodos.** A figura mostra a medida de AUC e o desvio do método de aprendizado de máquina em comparação com os métodos base para previsão de links. Para cada intervalo de tempo, calculamos a média AUC de cada método e seu respectivo erro padrão ao longo de todo o período de testes. O método de aprendizado de máquina supera os métodos base em todos os índices de mercado.

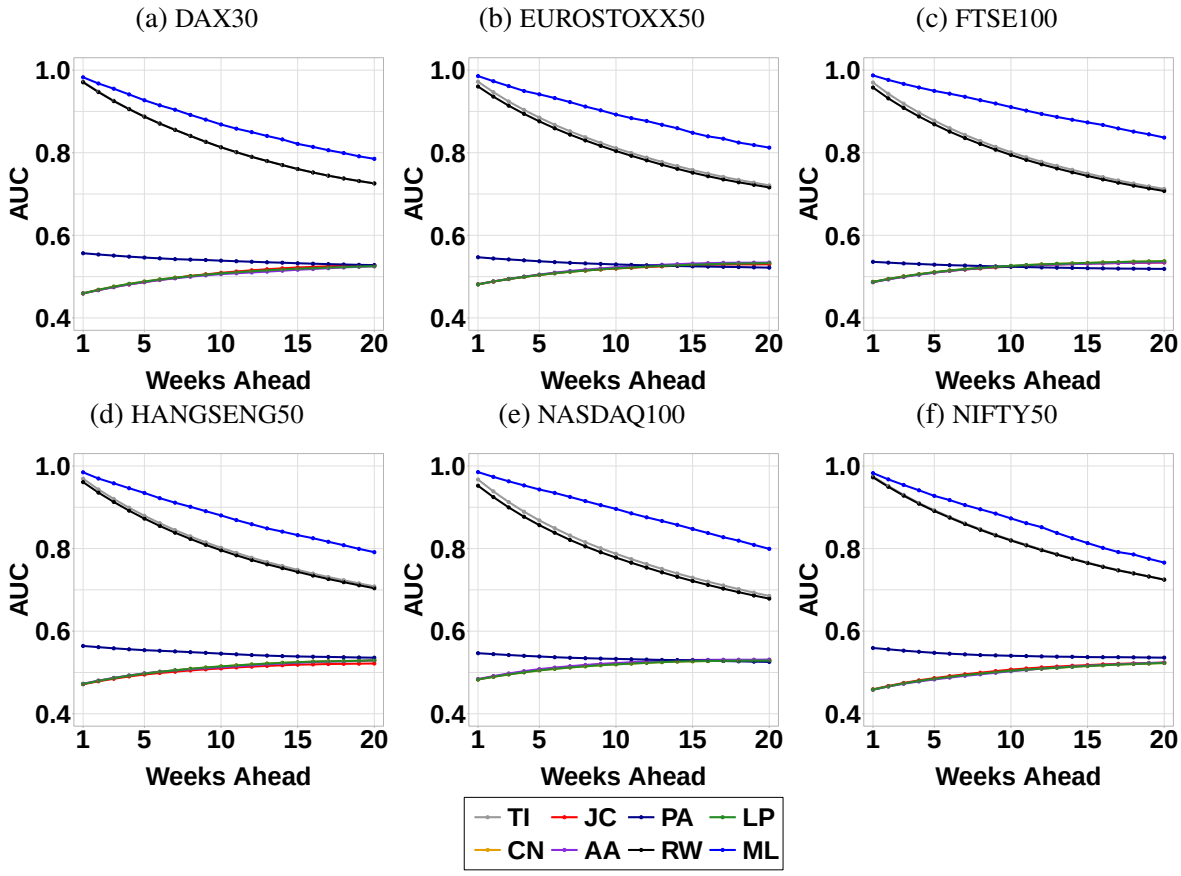
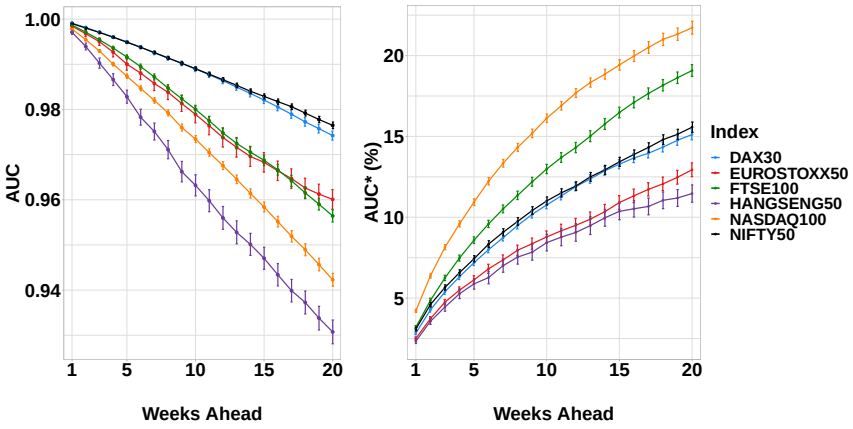
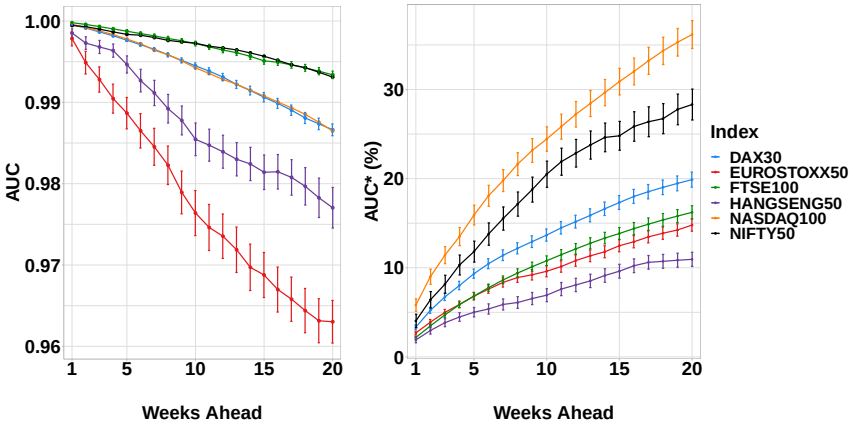


Figura 50 – **DMST - Comparação de desempenho preditiva de todos os métodos.** A figura mostra a medida de AUC e o desvio do método de aprendizado de máquina em comparação com os métodos base para previsão de links. Para cada intervalo de tempo, calculamos a média AUC de cada método e seu respectivo erro padrão ao longo de todo o período de testes. O método de aprendizado de máquina supera os métodos base em todos os índices de mercado.



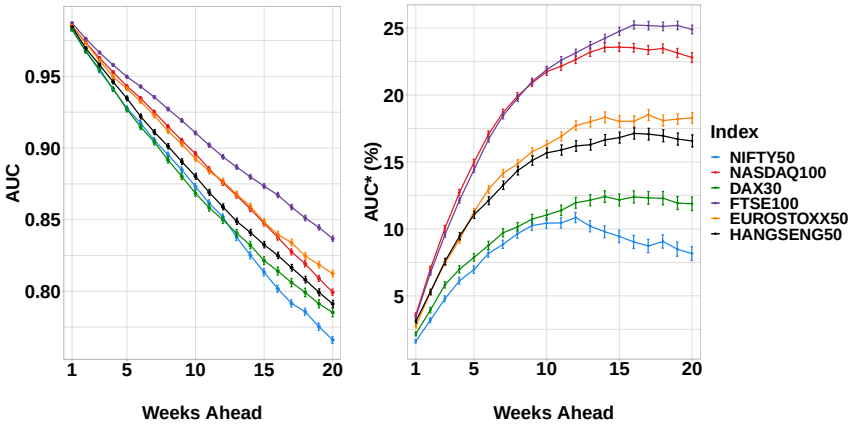
(a) DAG (AUC)

(b) DAG (AUC*)



(c) DTN (AUC)

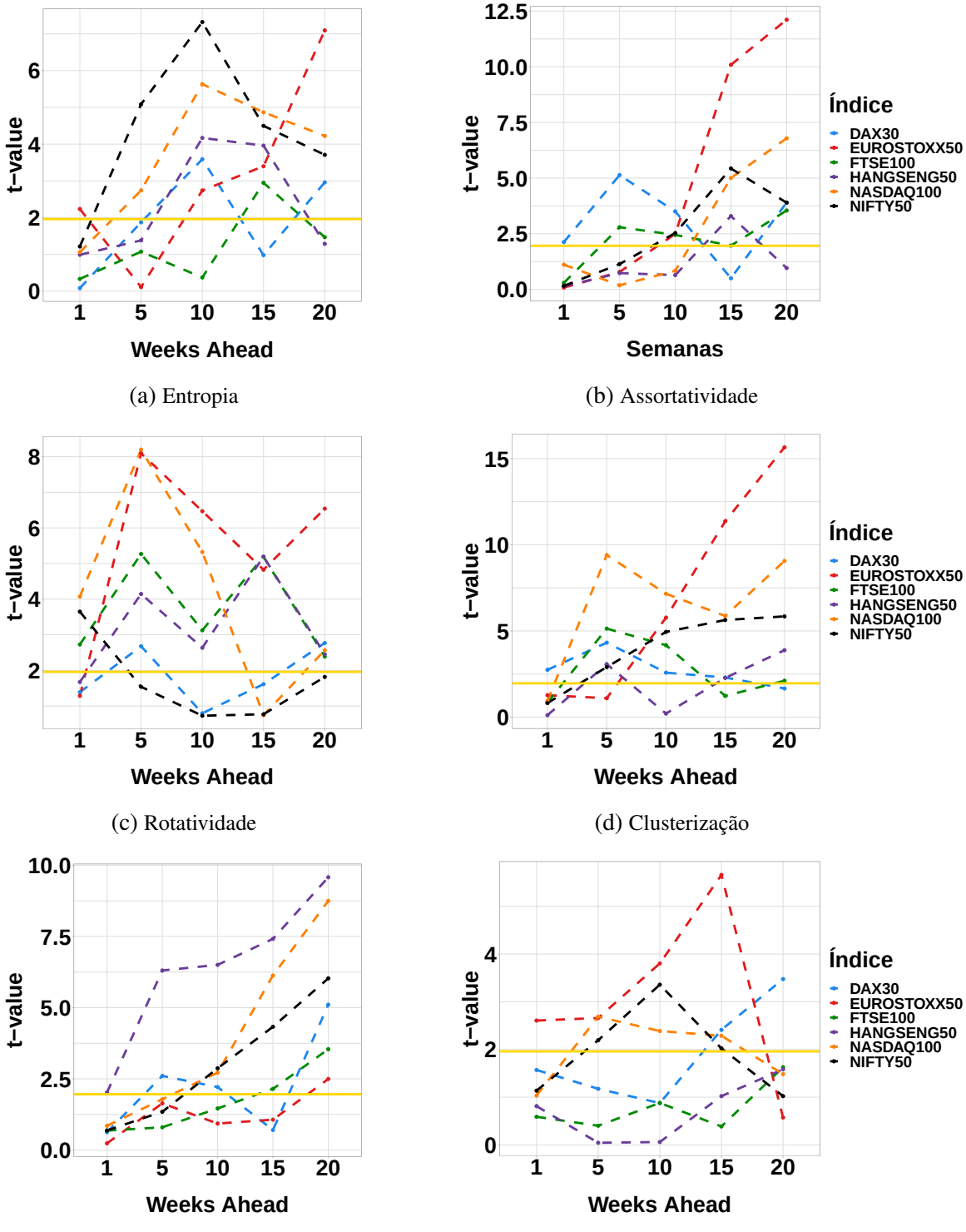
(d) DTN (AUC*)



(e) DMST (AUC)

(f) DMST (AUC*)

Figura 51 – **AUC e AUC* do algoritmo ML para métodos de filtragem de rede DAG, DTN e DMST.** Os painéis (a), (c) e (e) apresentam a métrica AUC do aprendizado de máquina e seu erro padrão para previsão de h semanas de negociação à frente ($1 \leq h \leq 20$). Os painéis (b), (d) e (f) apresentam a melhoria da AUC em relação ao método invariante no tempo e seu erro padrão. Os resultados apresentados são relacionados a $L = 504$.



(a) Entropia

(b) Assortatividade

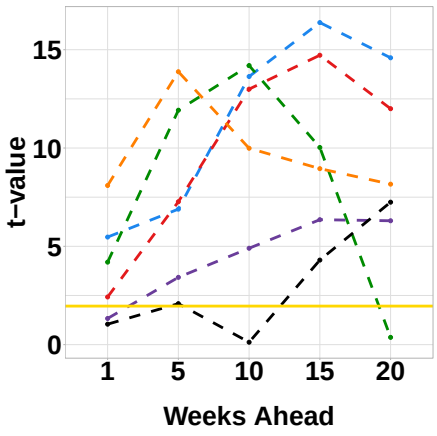
(c) Rotatividade

(d) Clusterização

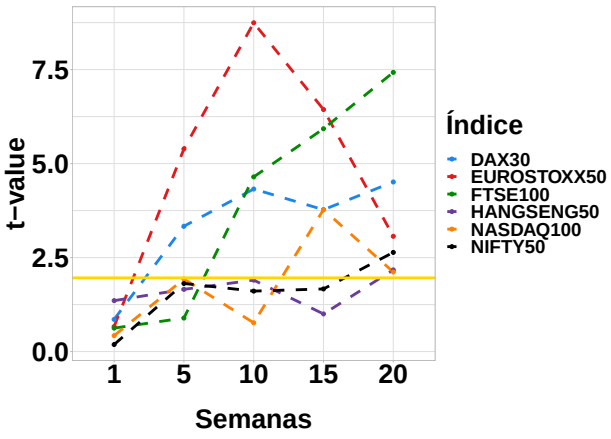
(e) Modularidade

(f) Diâmetro

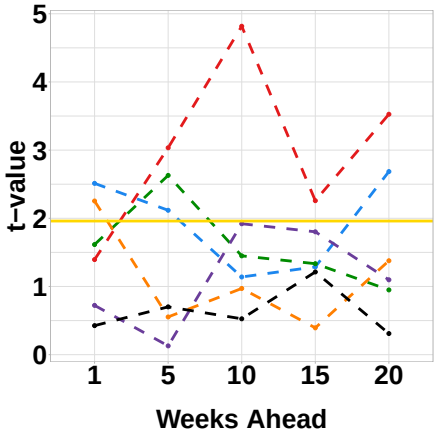
Figura 52 – **DAG - Variáveis de Rede vs. Previsibilidade da Estrutura de Mercado.** A figura apresenta o valor t de uma regressão linear múltipla usando variáveis de rede como variáveis explicativas e a AUC do modelo de aprendizado de máquina como resultado. Os resultados de cada índice de mercado são apresentados individualmente, considerando $L = 126$ e $h = \{1, 5, 10, 15, 20\}$. De acordo com t-student, os valores acima da linha amarela em $1,96$ apresentam relevância estatística.



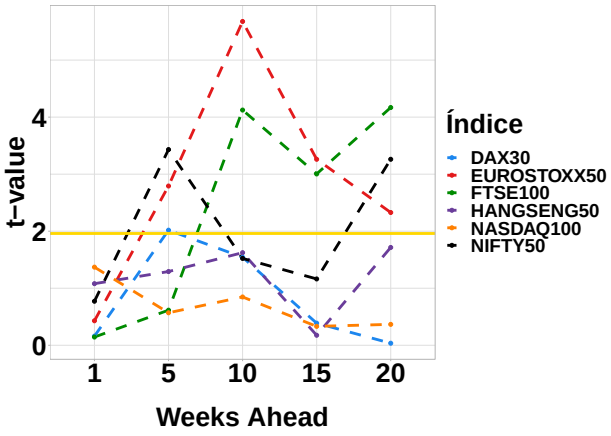
(a) Entropia



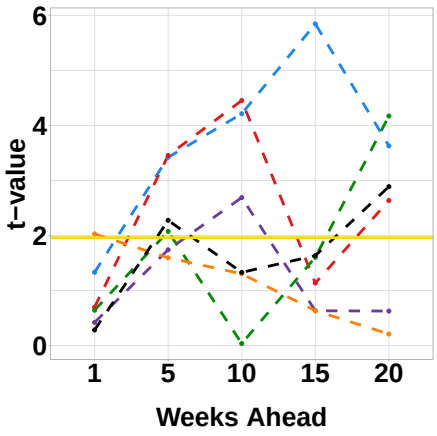
(b) Assortatividade



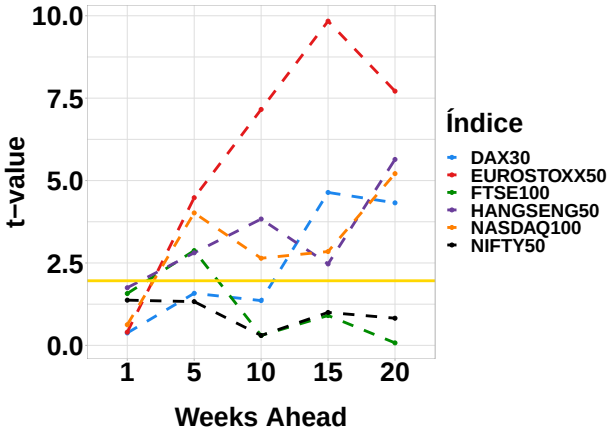
(c) Rotatividade



(d) Clusterização

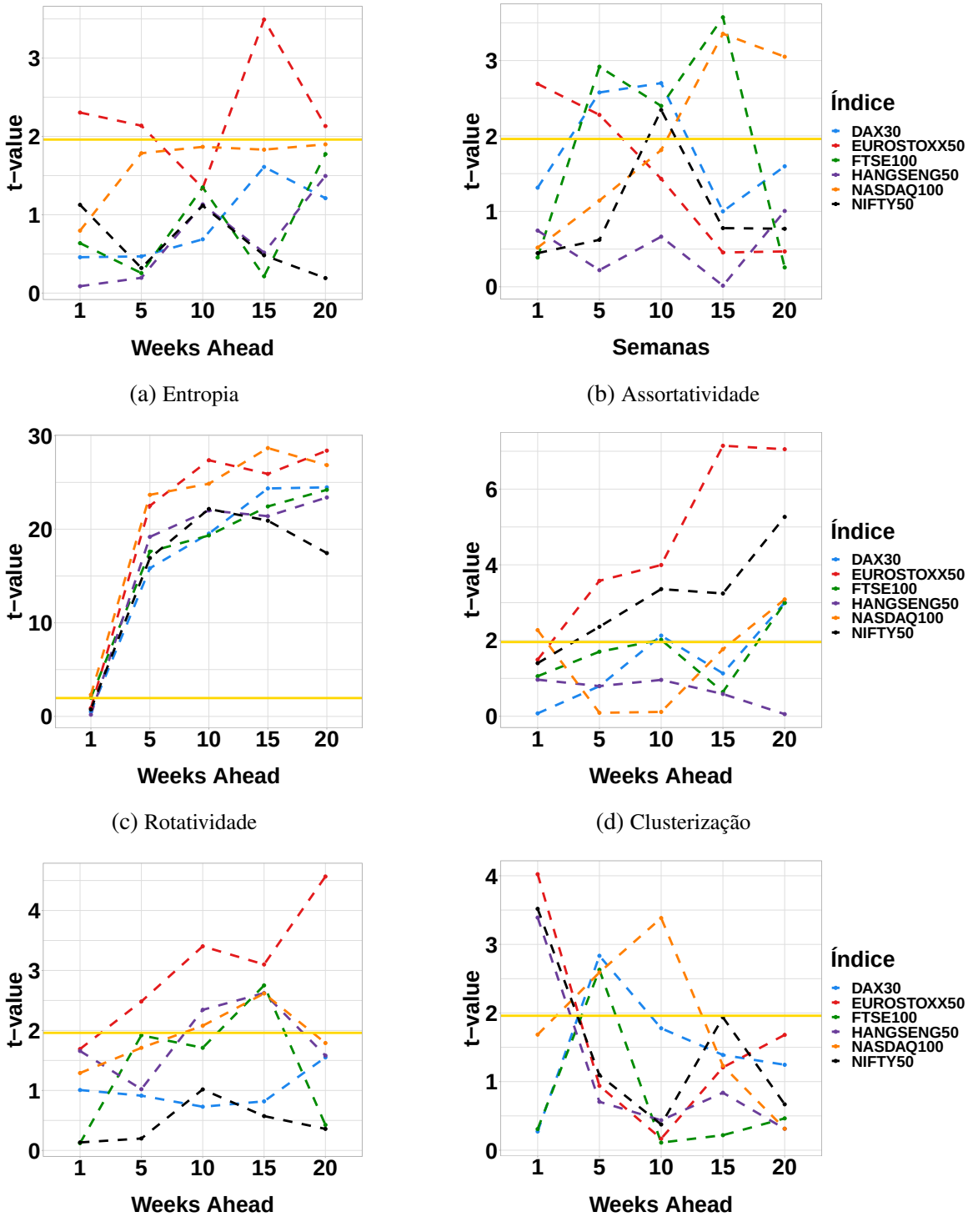


(e) Modularidade



(f) Diâmetro

Figura 53 – **DTN - Variáveis de Rede vs. Previsibilidade da Estrutura de Mercado.** A figura apresenta o valor t de uma regressão linear múltipla usando variáveis de rede como variáveis explicativas e a AUC do modelo de aprendizado de máquina como resultado. Os resultados de cada índice de mercado são apresentados individualmente, considerando $L = 126$ e $h = \{1, 5, 10, 15, 20\}$. De acordo com t-student, os valores acima da linha amarela em $1,96$ têm relevância estatística.



(a) Entropia

(b) Assortatividade

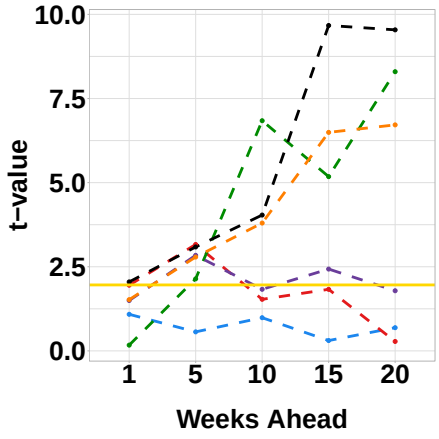
(c) Rotatividade

(d) Clusterização

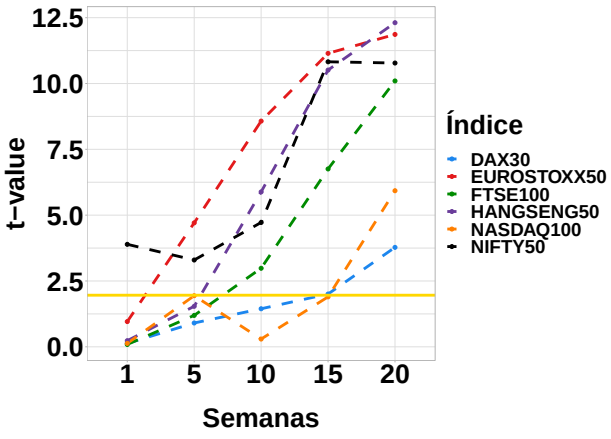
(e) Modularidade

(f) Diâmetro

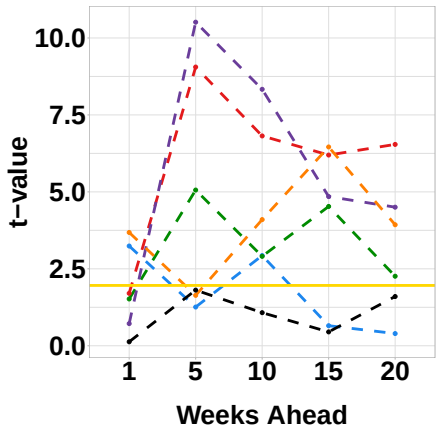
Figura 54 – **DMST - Variáveis de Rede vs. Previsibilidade da Estrutura de Mercado.** A figura apresenta o valor t de uma regressão linear múltipla usando variáveis de rede como variáveis explicativas e a AUC do modelo de aprendizado de máquina como resultado. Os resultados de cada índice de mercado são apresentados individualmente, considerando $L = 126$ e $h = \{1, 5, 10, 15, 20\}$. De acordo com t-student, os valores acima da linha amarela em $1,96$ têm relevância estatística.



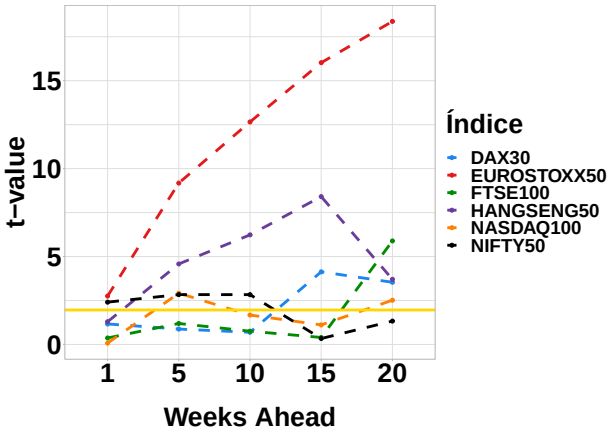
(a) Entropia



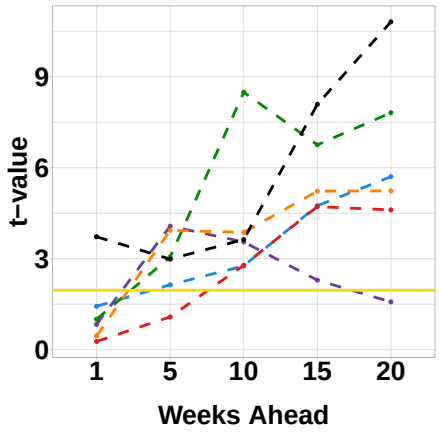
(b) Assortatividade



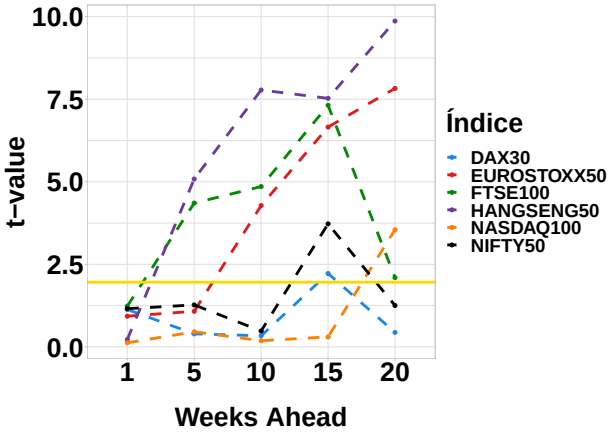
(c) Rotatividade



(d) Clusterização

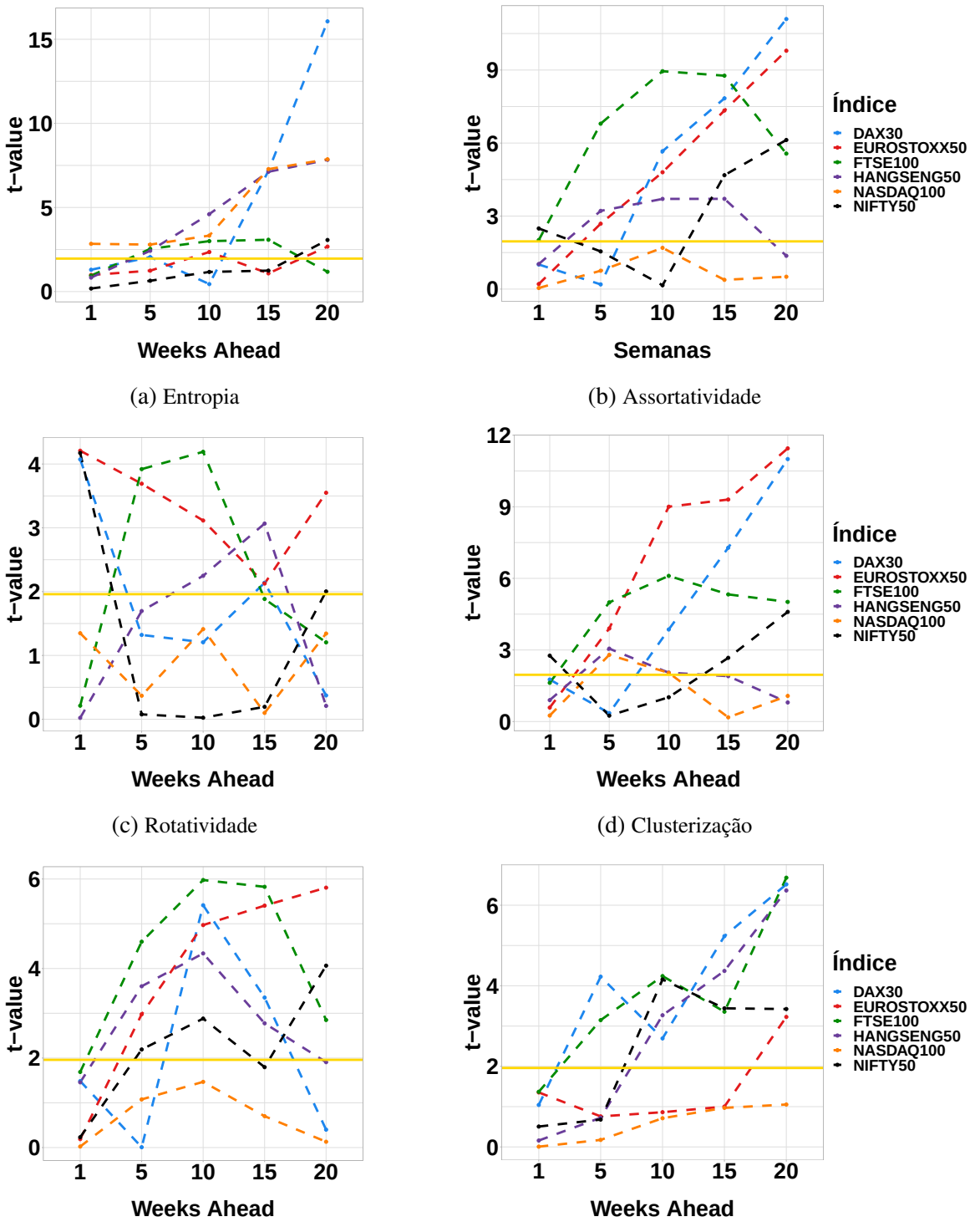


(e) Modularidade



(f) Diâmetro

Figura 55 – **DAG - Variáveis de Rede vs. Previsibilidade da Estrutura de Mercado.** A figura apresenta o valor t de uma regressão linear múltipla usando variáveis de rede como variáveis explicativas e a AUC do modelo de aprendizado de máquina como resultado. Os resultados de cada índice de mercado são apresentados individualmente, considerando $L = 504$ e $h = \{1, 5, 10, 15, 20\}$. De acordo com t-student, os valores acima da linha amarela em $1,96$ apresentam relevância estatística.



(a) Entropia

(b) Assortatividade

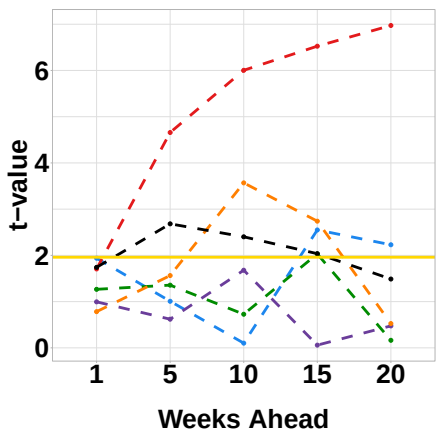
(c) Rotatividade

(d) Clusterização

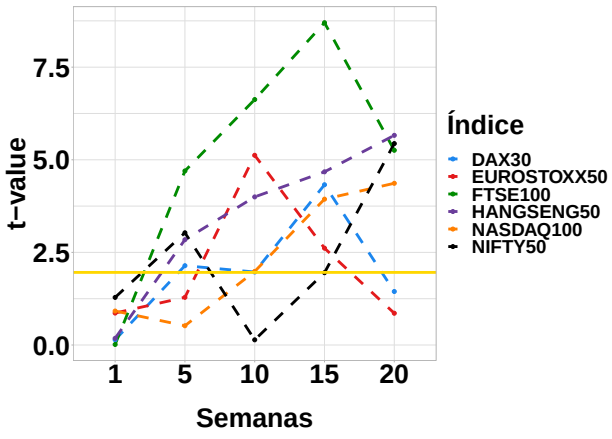
(e) Modularidade

(f) Diâmetro

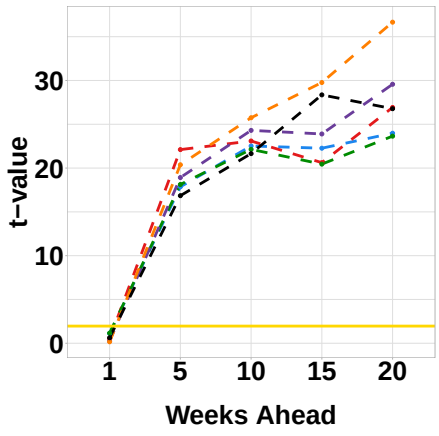
Figura 56 – **DTN - Variáveis de Rede vs. Previsibilidade da Estrutura de Mercado.** A figura apresenta o valor t de uma regressão linear múltipla usando variáveis de rede como variáveis explicativas e a AUC do modelo de aprendizado de máquina como resultado. Os resultados de cada índice de mercado são apresentados individualmente, considerando $L = 504$ e $h = \{1, 5, 10, 15, 20\}$. De acordo com t-student, os valores acima da linha amarela em $1,96$ têm relevância estatística.



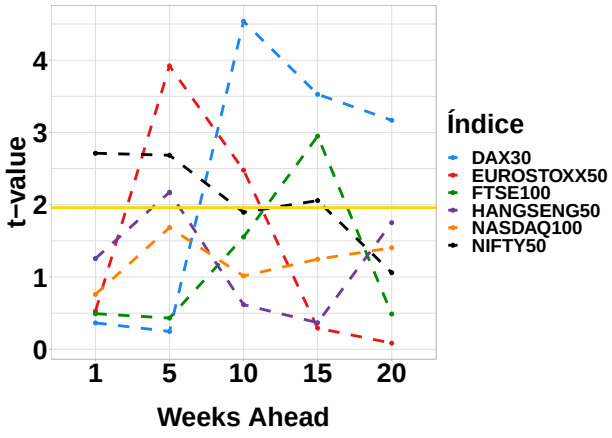
(a) Entropia



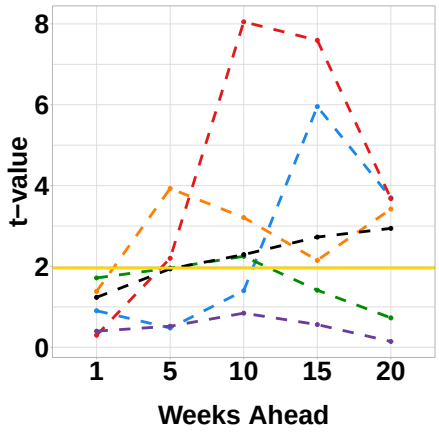
(b) Assortatividade



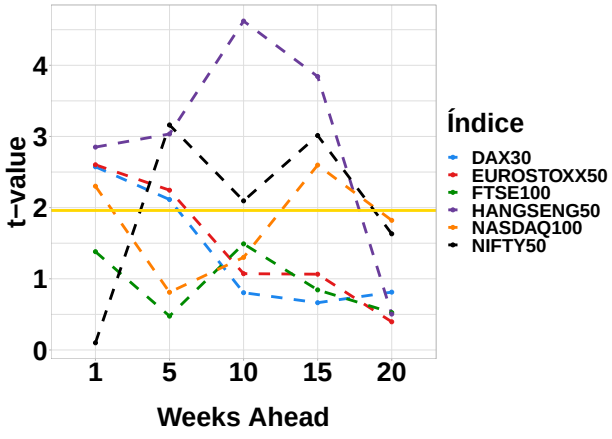
(c) Rotatividade



(d) Clusterização



(e) Modularidade



(f) Diâmetro

Figura 57 – **DMST - Variáveis de Rede vs. Previsibilidade da Estrutura de Mercado.** A figura apresenta o valor t de uma regressão linear múltipla usando variáveis de rede como variáveis explicativas e a AUC do modelo de aprendizado de máquina como resultado. Os resultados de cada índice de mercado são apresentados individualmente, considerando $L = 504$ e $h = \{1, 5, 10, 15, 20\}$. De acordo com t-student, os valores acima da linha amarela em $1,96$ têm relevância estatística.

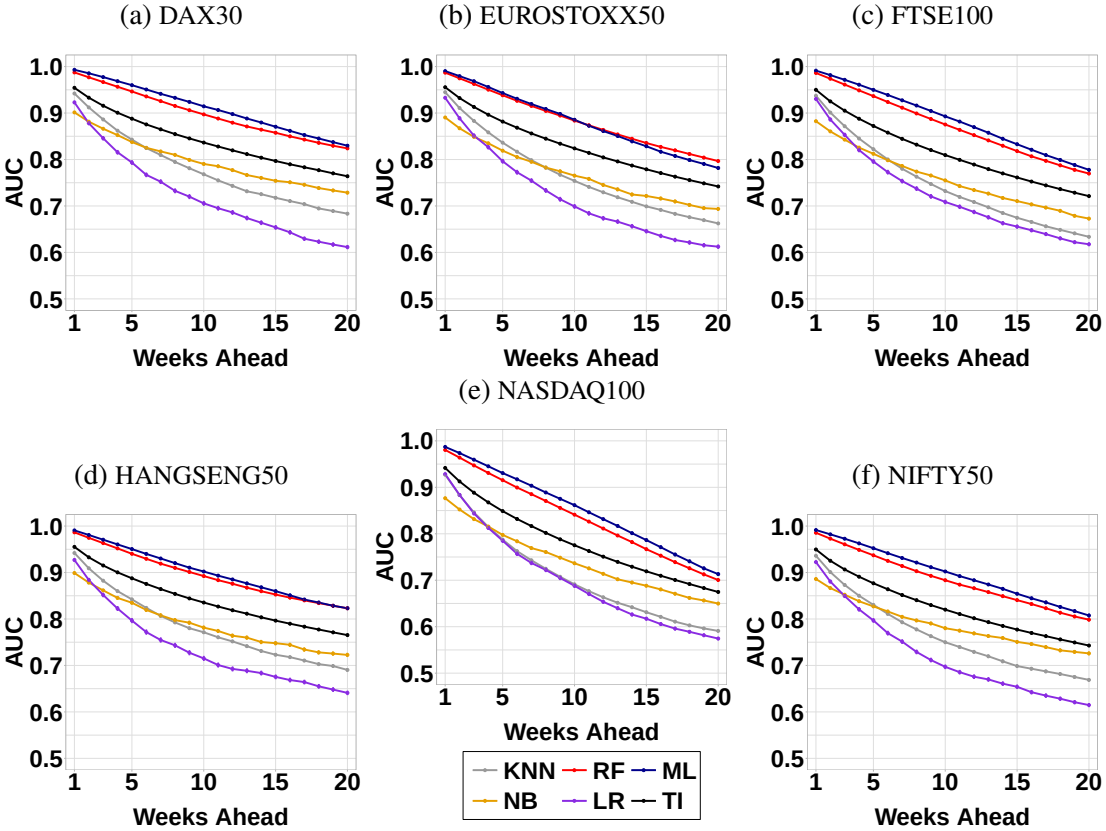


Figura 58 – **DAG - Comparação entre algoritmos de aprendizagem de máquina para $L = 126$.** A figura mostra a média da AUC e o erro padrão do algoritmo ML em comparação com outros algoritmos de aprendizagem de máquina. ML e RF superam o algoritmo de *benchmark* TI.

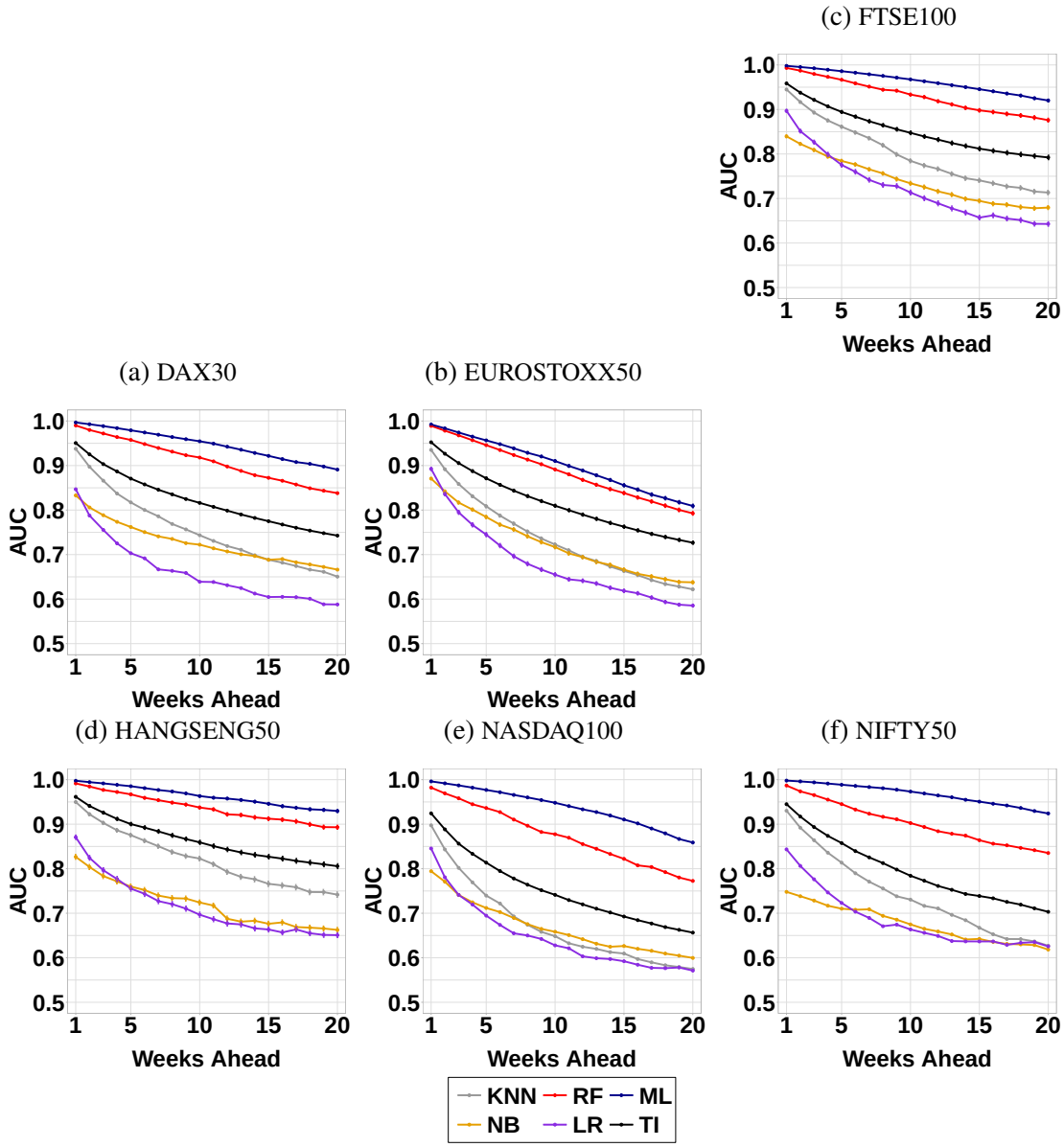


Figura 59 – **DTN - Comparação entre algoritmos de aprendizagem de máquina para $L = 126$.** A figura mostra a média da AUC e o erro padrão do algoritmo ML em comparação com outros algoritmos de aprendizagem de máquina. ML e RF superam o algoritmo de *benchmark* TI.

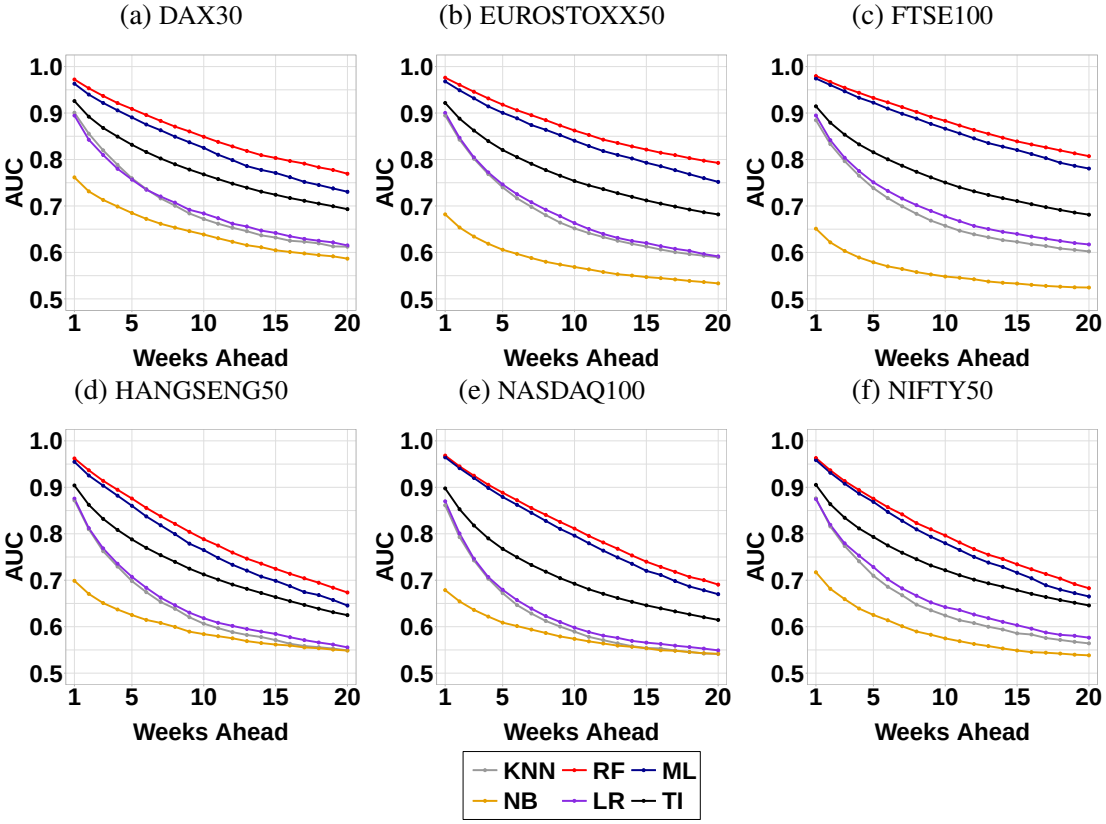


Figura 60 – **DMST - Comparação entre algoritmos de aprendizagem de máquina para** $L = 126$**.** A figura mostra a média da AUC e o erro padrão do algoritmo ML em comparação com outros algoritmos de aprendizagem de máquina. ML e RF superam o algoritmo de *benchmark* TI.

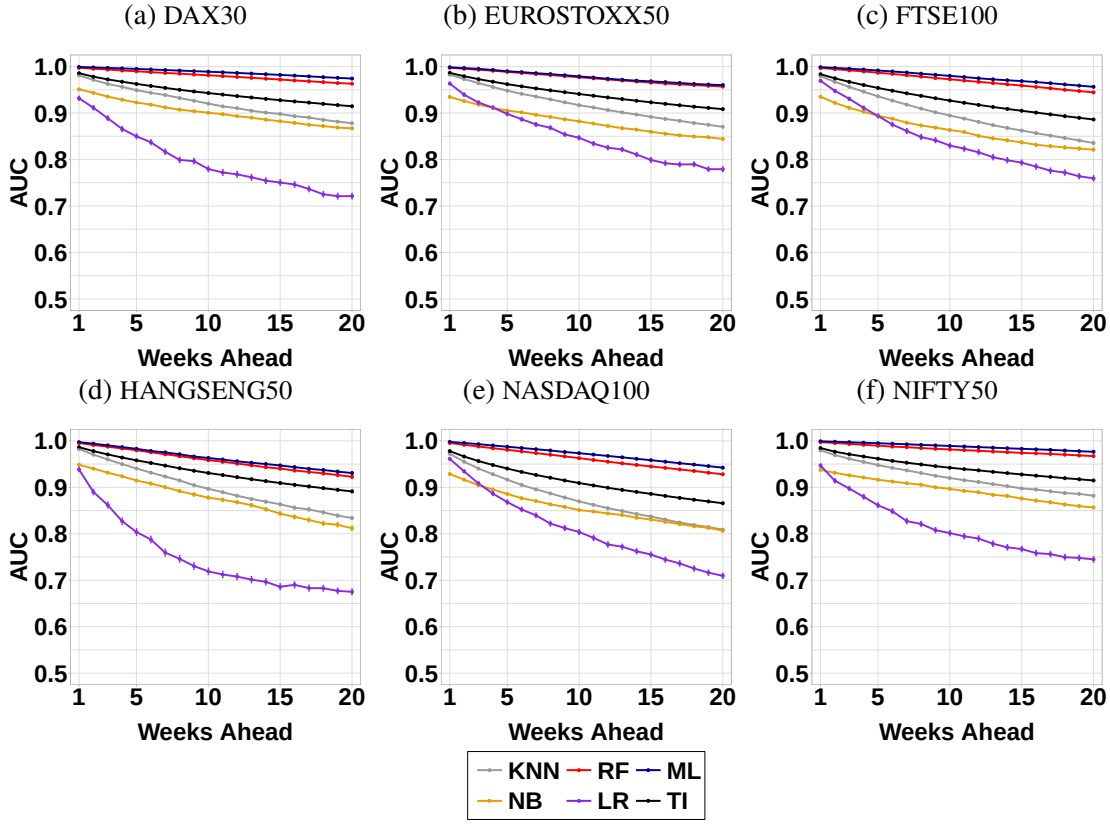


Figura 61 – **DAG - Comparação entre algoritmos de aprendizagem de máquina para** $L = 504$**.** A figura mostra a média da AUC e o erro padrão do algoritmo ML em comparação com outros algoritmos de aprendizagem de máquina. ML e RF superam o algoritmo de *benchmark* TI.

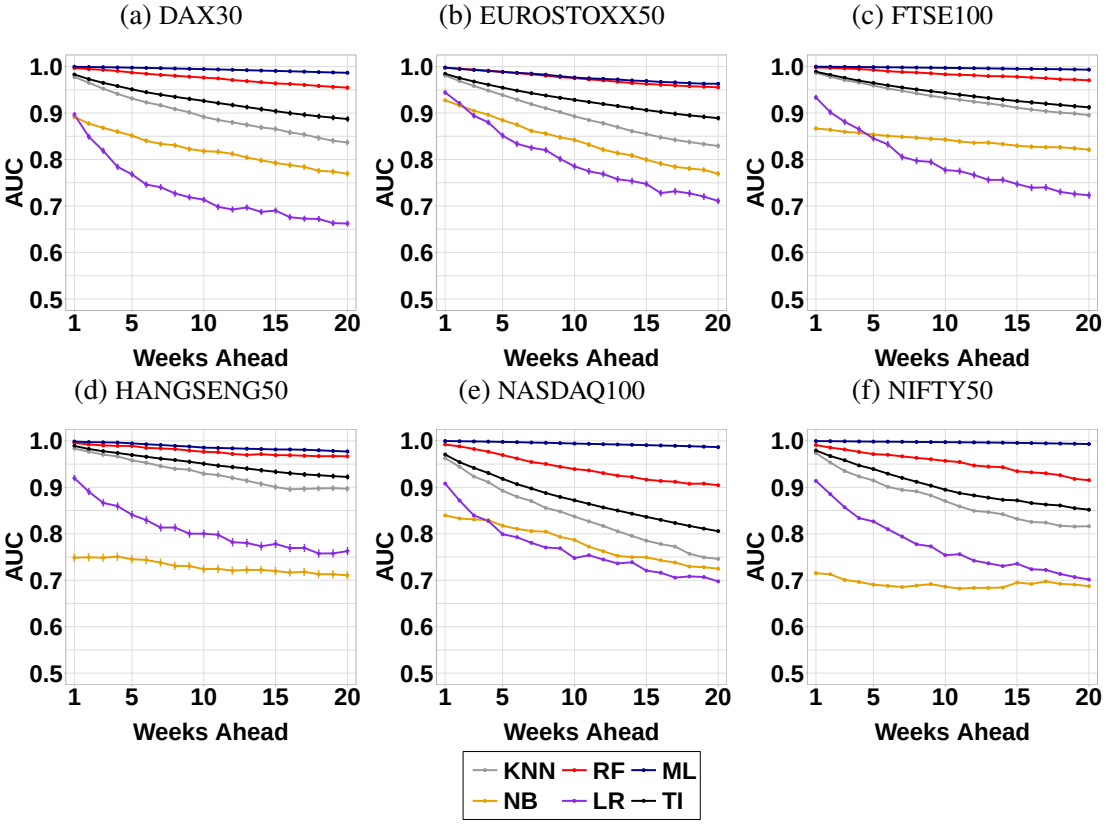


Figura 62 – **DTN - Comparação entre algoritmos de aprendizagem de máquina para $L = 504$.** A figura mostra a média da AUC e o erro padrão do algoritmo ML em comparação com outros algoritmos de aprendizagem de máquina. ML e RF superam o algoritmo de *benchmark* TI.

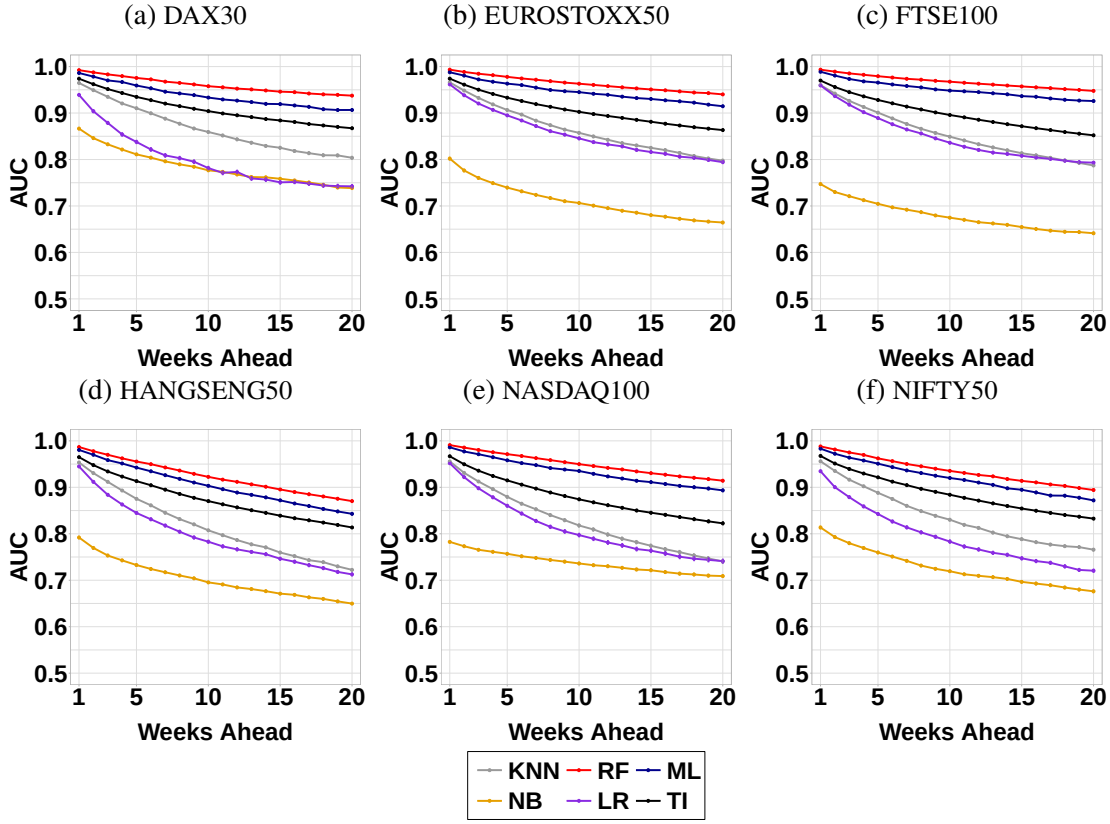


Figura 63 – **DMST - Comparação entre algoritmos de aprendizagem de máquina para $L = 504$.** A figura mostra a média da AUC e o erro padrão do algoritmo ML em comparação com outros algoritmos de aprendizagem de máquina. ML e RF superam o algoritmo de *benchmark* TI.

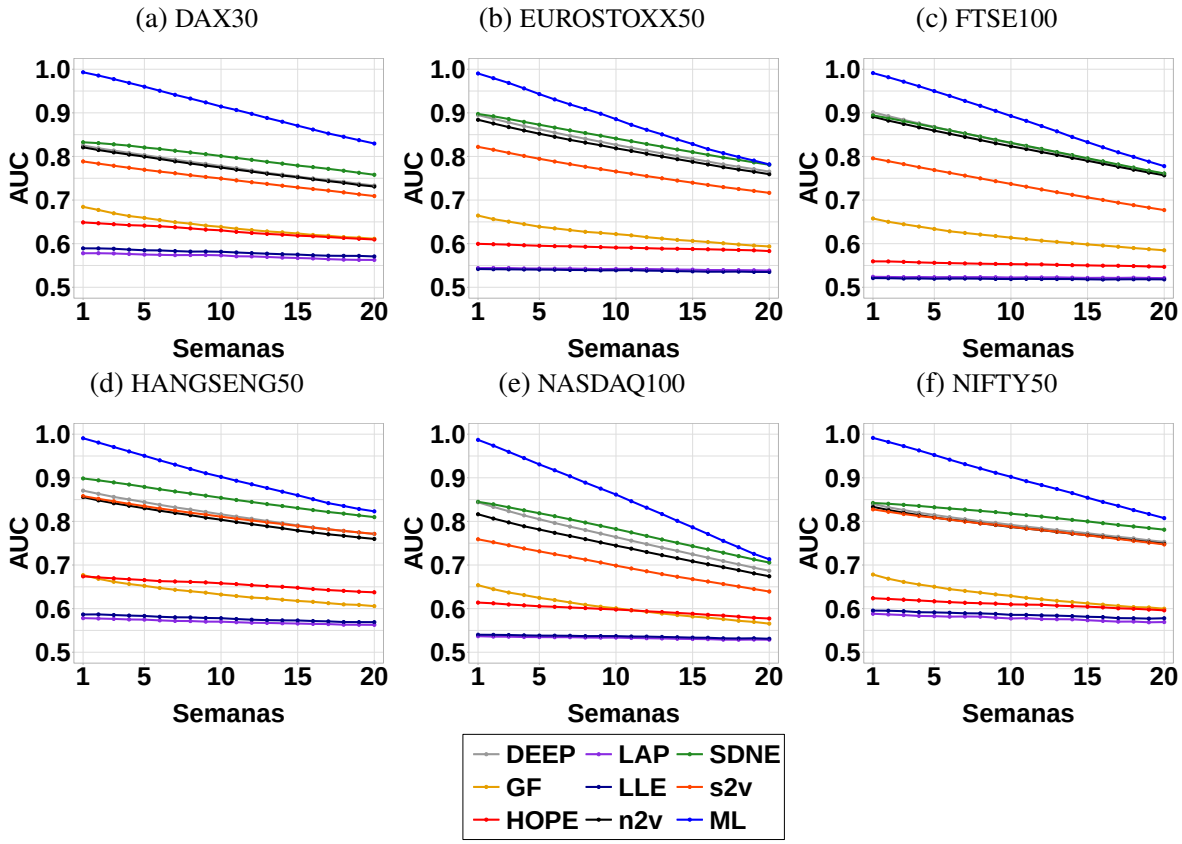


Figura 64 – **DAG - Comparação entre algoritmos de *embedding* usando $L = 126$.** A figura mostra a média da AUC e o erro padrão do algoritmo ML em comparação com outros algoritmos de *embedding*. ML supera os algoritmos de *embedding* em todos os índices de mercado.

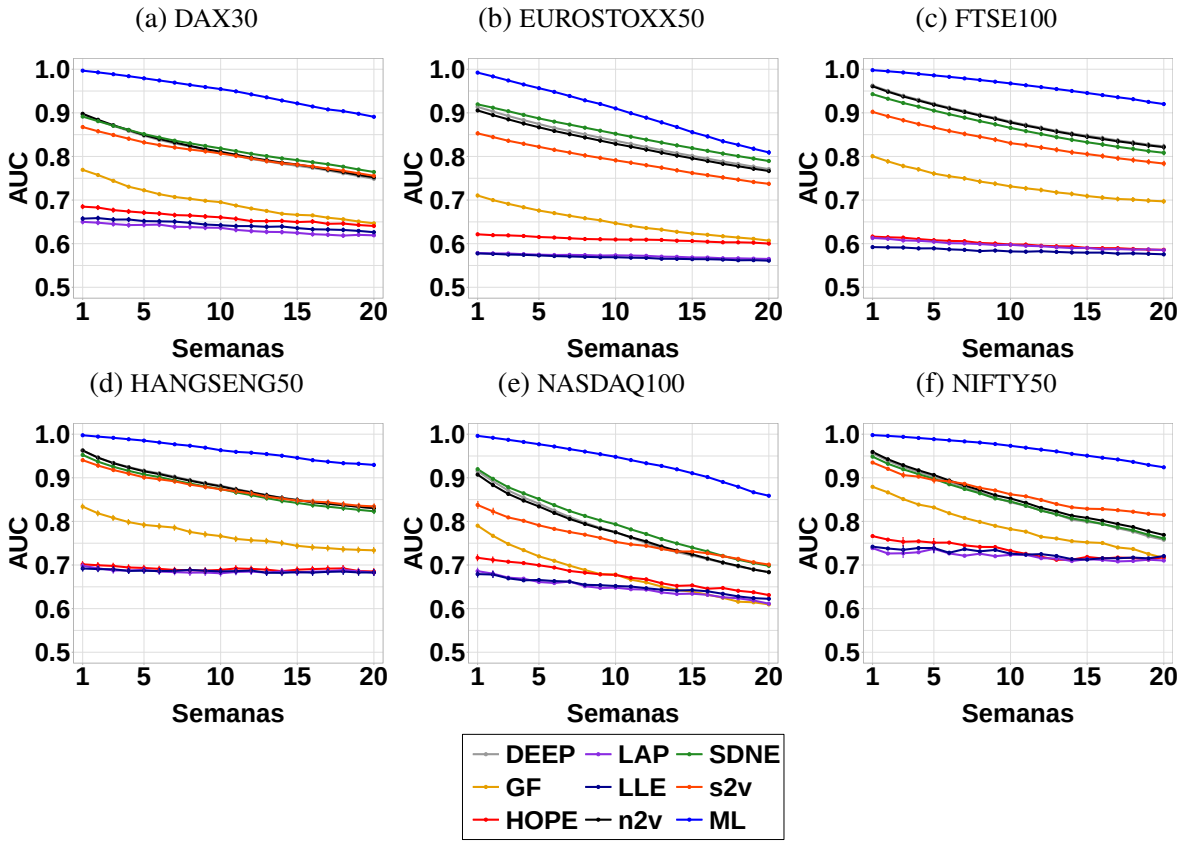


Figura 65 – **DTN - Comparação entre algoritmos de *embedding* usando $L = 126$.** A figura mostra a média da AUC e o erro padrão do algoritmo ML em comparação com outros algoritmos de *embedding*. ML supera os algoritmos de *embedding* em todos os índices de mercado.

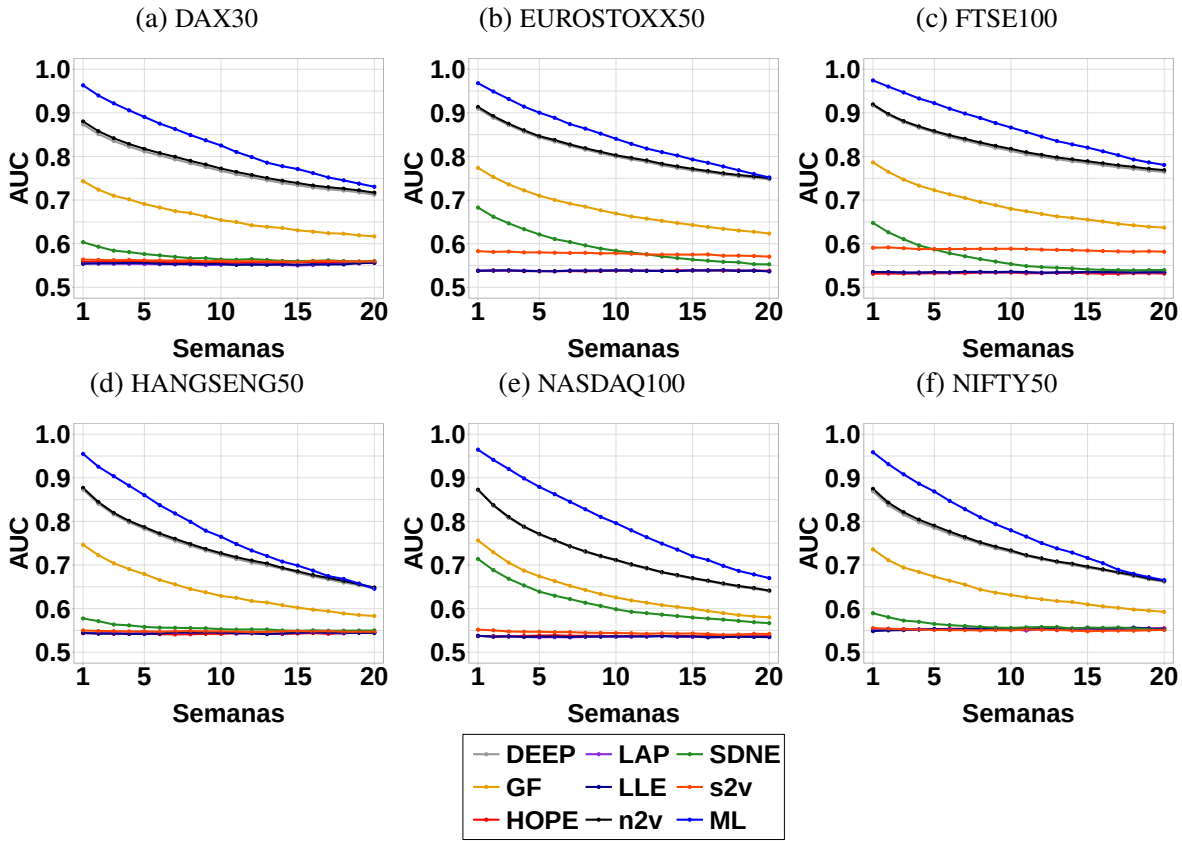


Figura 66 – **DMST - Comparação entre algoritmos de *embedding* usando $L = 126$.** A figura mostra a média da AUC e o erro padrão do algoritmo ML em comparação com outros algoritmos de *embedding*. ML supera os algoritmos de *embedding* em todos os índices de mercado.

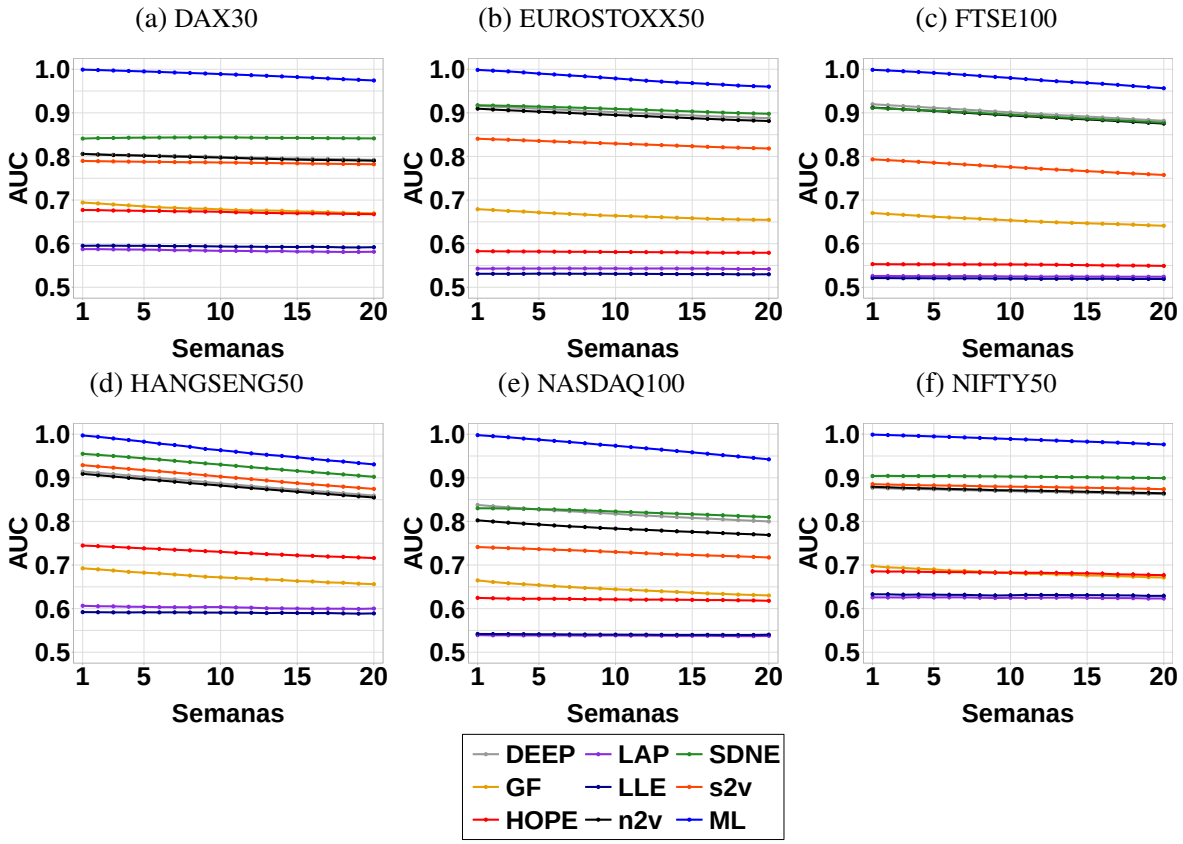


Figura 67 – **DAG - Comparação entre algoritmos de *embedding* usando $L = 504$.** A figura mostra a média da AUC e o erro padrão do algoritmo ML em comparação com outros algoritmos de *embedding*. ML supera os algoritmos de *embedding* em todos os índices de mercado.

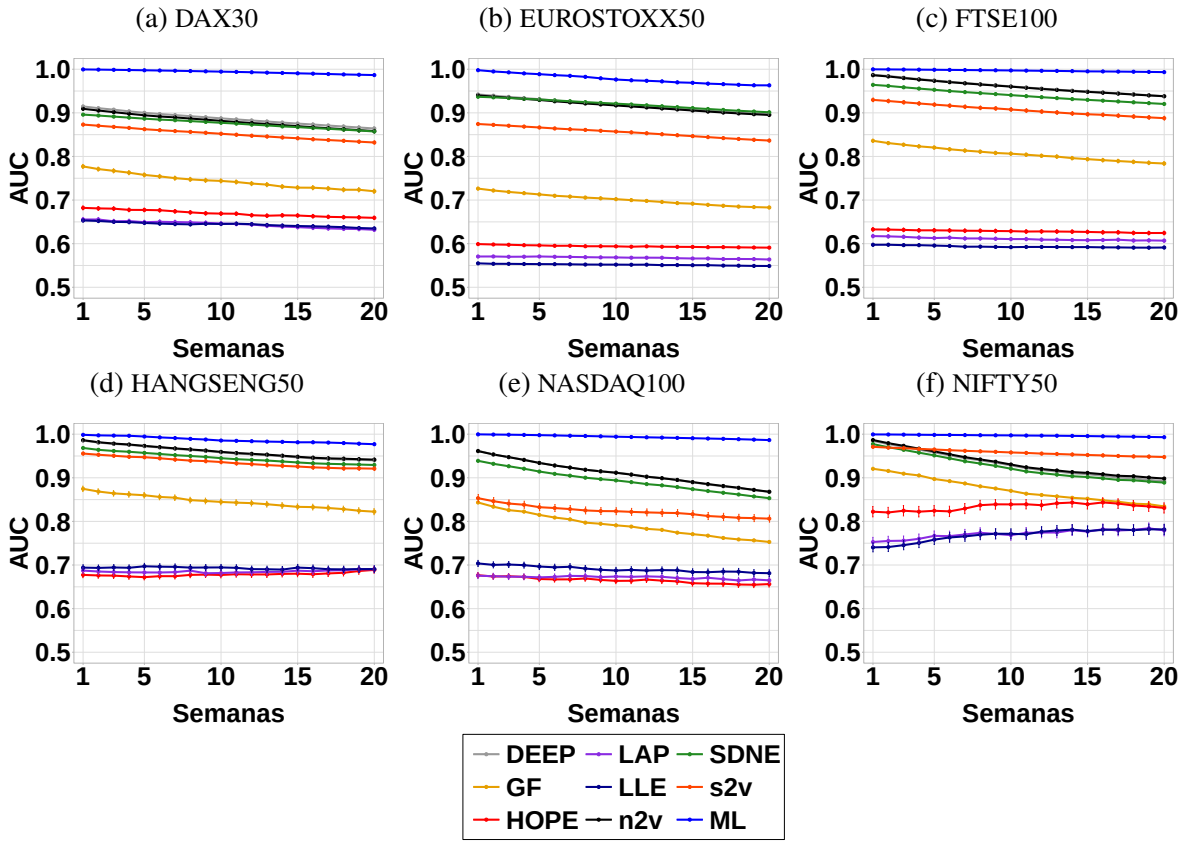


Figura 68 – **DTN - Comparação entre algoritmos de *embedding* usando $L = 126$.** A figura mostra a média da AUC e o erro padrão do algoritmo ML em comparação com outros algoritmos de *embedding*. ML supera os algoritmos de *embedding* em todos os índices de mercado.

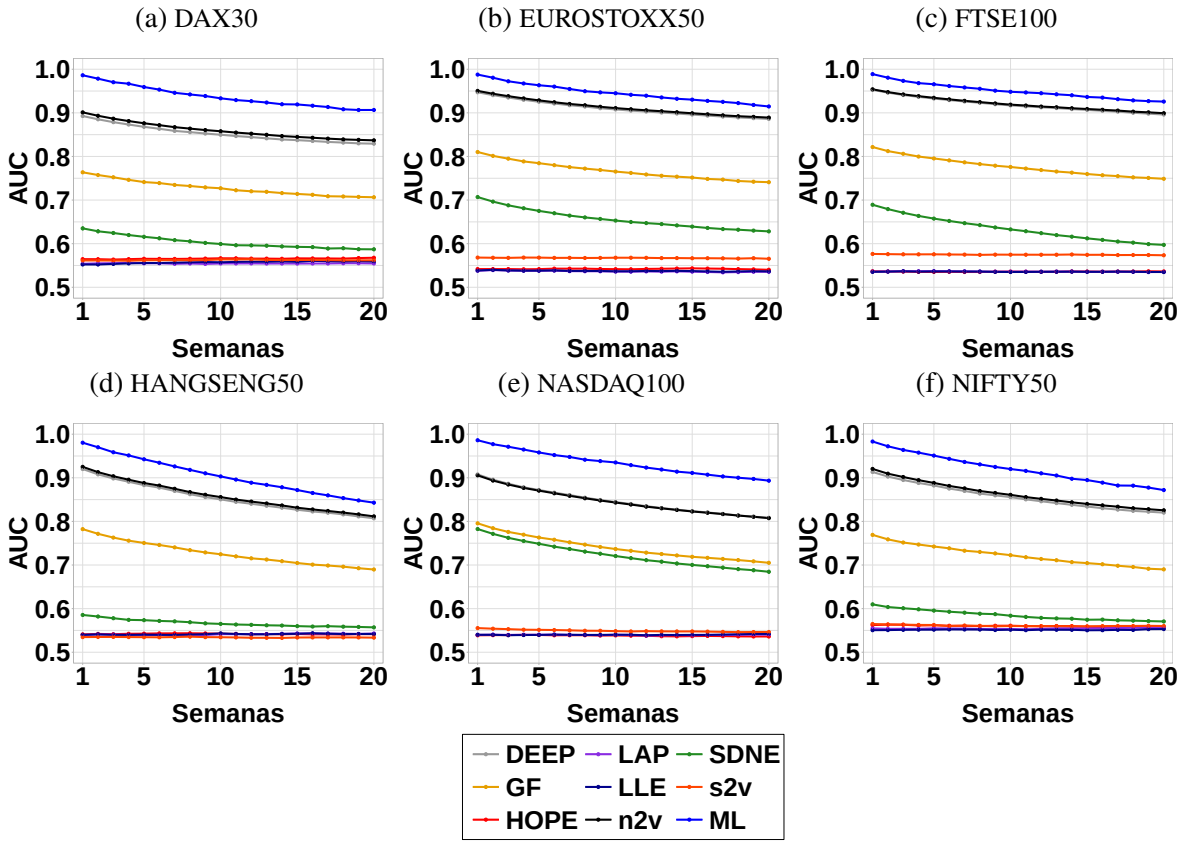


Figura 69 – **DMST - Comparação entre algoritmos de *embedding* usando $L = 504$.** A figura mostra a média da AUC e o erro padrão do algoritmo ML em comparação com outros algoritmos de *embedding*. ML supera os algoritmos de *embedding* em todos os índices de mercado.

APÊNDICE

B

RESULTADOS COMPLEMENTARES DE OTIMIZAÇÃO DE PORTFÓLIO

B.1 Parâmetros do Modelo

Esta seção apresenta o conjunto de parâmetros usados no algoritmo XGboost para executar os experimentos. Os atributos de configuração são:

- booster = “gbtree”;
- objective = “reg:linear”;
- eta = 0.05;
- max_depth = 2;
- min_child_weight = 100;

B.2 Resultados Complementares

Esse seção apresenta resultados complementares relacionados à previsão de links em redes ponderadas usando $L \in \{63, 126, 504\}$ dias de negociação para construir as redes. Para cada intervalo de tempo $h = \{1, 2, ..., 20\}$, realizamos a previsão da rede $G(t+h)$ e calculamos a MAE e RMSE média de cada método e seu respectivo erro padrão. Os experimentos compreendem um período de tempoentre 5 maio 2007 e 18 de dezembro de 2019.

As Figuras 70, 72 e 74 apresentam os resultados relacionados à métrica MAE, considerando $L = \{63, 126, 504\}$, respectivamente.

As Figuras 71, 73 e 75 apresentam os resultados relacionados à métrica RMSE, considerando $L = \{63, 126, 504\}$, respectivamente.

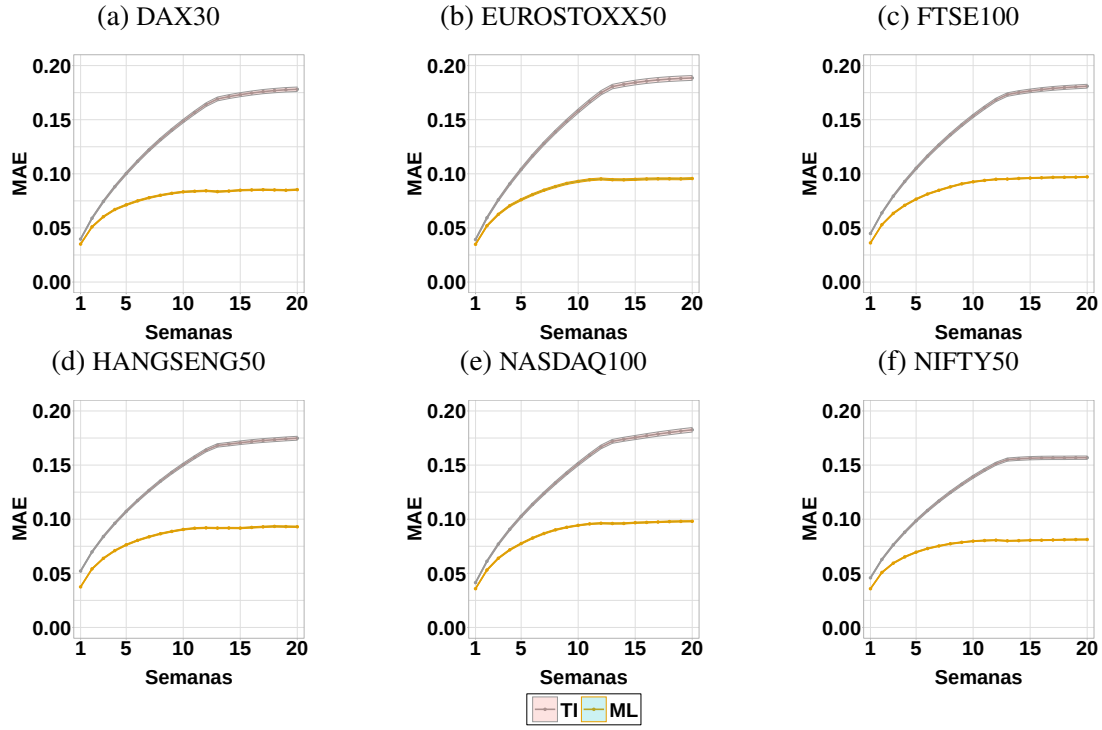


Figura 70 – **MAE - Comparação do desempenho preditivo dos métodos TI e ML usando** $L = 63$**.** A figura mostra a métrica MAE dos métodos ML e TI relacionada à previsão de links ponderados. Para cada intervalo de tempo, calculamos a média da métrica MAE de cada método e seu respectivo erro padrão ao longo de todo o período de testes.

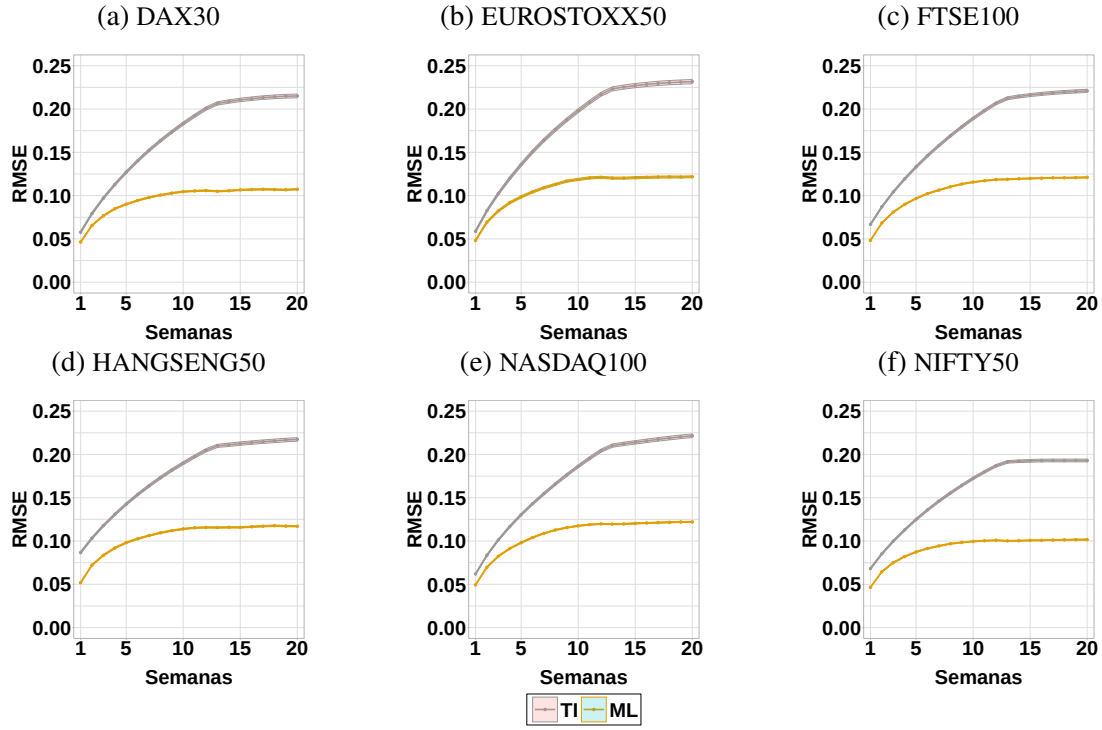


Figura 71 – **RMSE - Comparação do desempenho preditivo dos métodos TI e ML usando** $L = 63$**.** A figura mostra a métrica RMSE dos métodos ML e TI relacionada à previsão de links ponderados. Para cada intervalo de tempo, calculamos a média da RMSE de cada método e seu respectivo erro padrão ao longo de todo o período de testes.

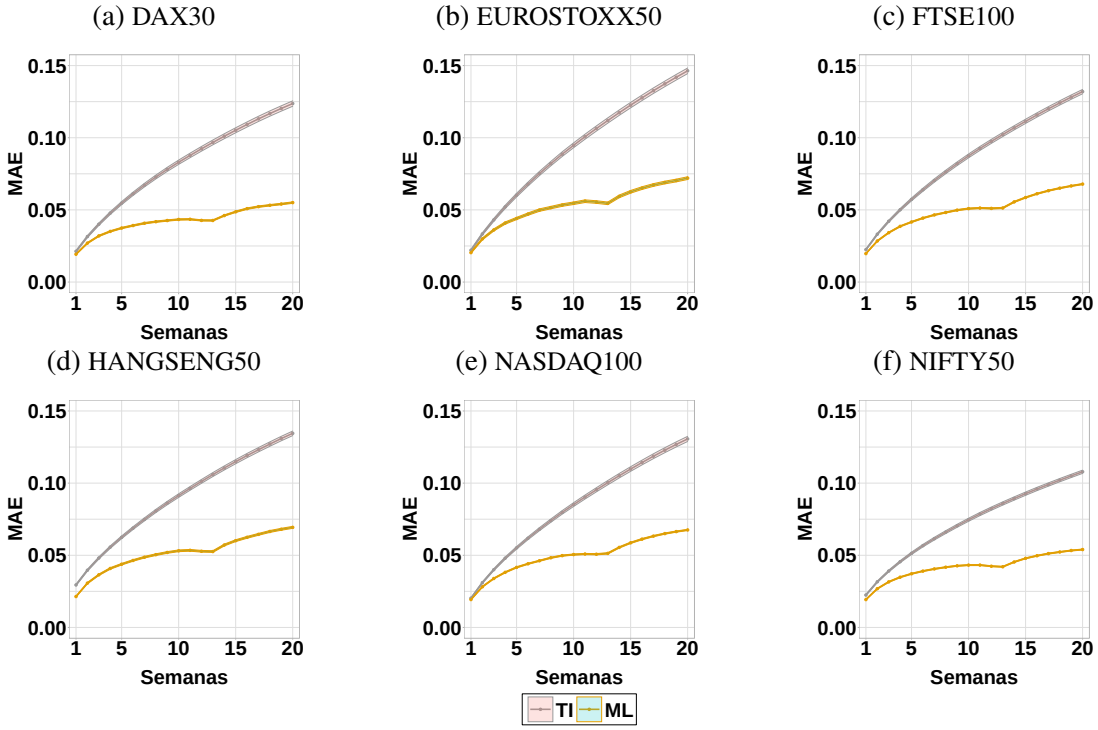


Figura 72 – **MAE - Comparação do desempenho preditivo dos métodos TI e ML usando** $L = 126$**.** A figura mostra a métrica MAE dos métodos ML e TI relacionada à previsão de links ponderados. Para cada intervalo de tempo, calculamos a média da métrica MAE de cada método e seu respectivo erro padrão ao longo de todo o período de testes.

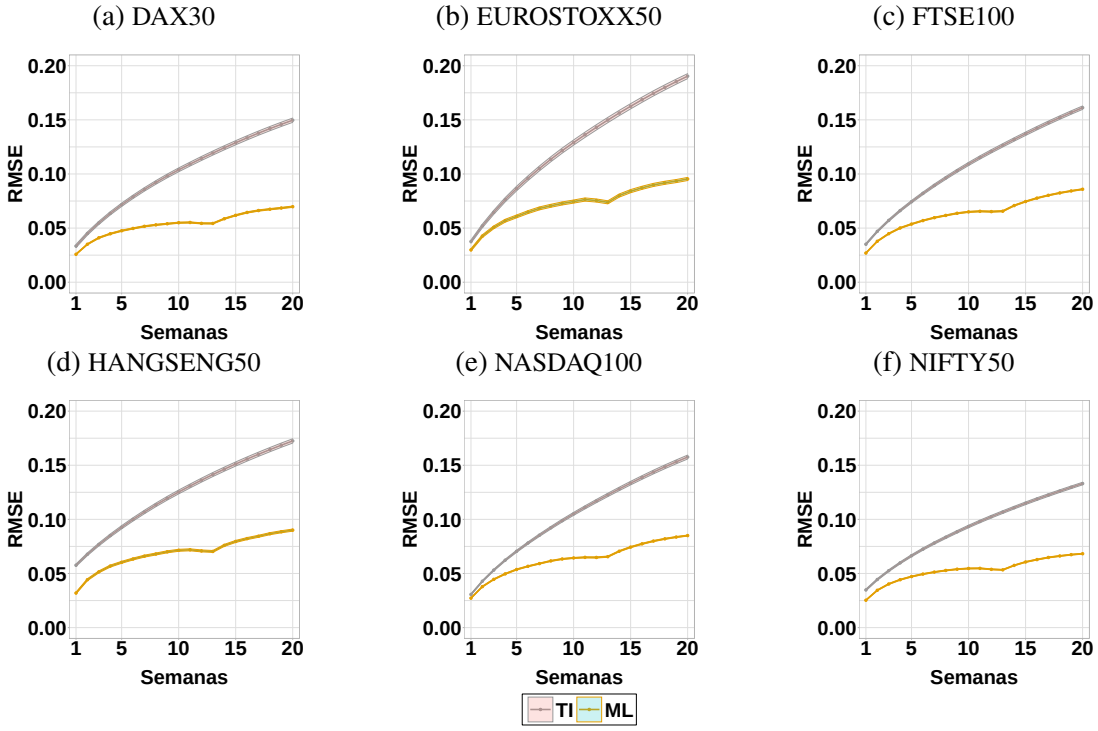


Figura 73 – **RMSE - Comparação do desempenho preditivo dos métodos TI e ML usando** $L = 126$**.** A figura mostra a métrica RMSE dos métodos ML e TI relacionada à previsão de links ponderados. Para cada intervalo de tempo, calculamos a média da RMSE de cada método e seu respectivo erro padrão ao longo de todo o período de testes.

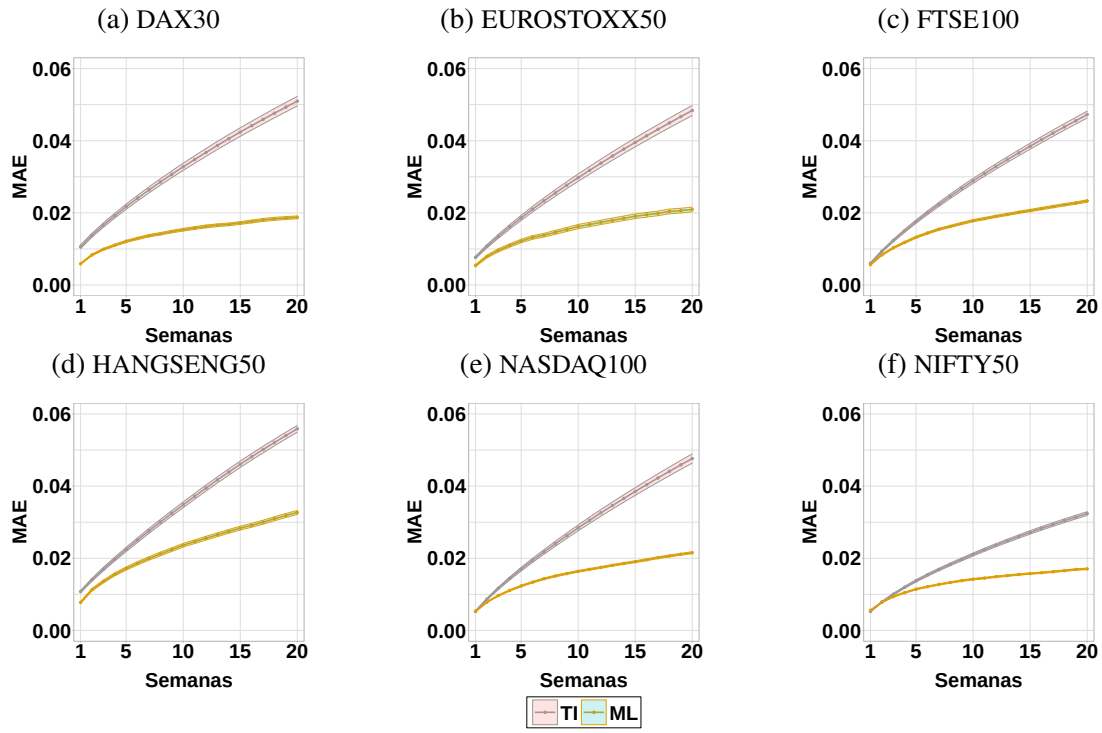


Figura 74 – **MAE - Comparação do desempenho preditivo dos métodos TI e ML usando** $L = 504$**.** A figura mostra a métrica MAE dos métodos ML e TI relacionada à previsão de links ponderados. Para cada intervalo de tempo, calculamos a média da métrica MAE de cada método e seu respectivo erro padrão ao longo de todo o período de testes.

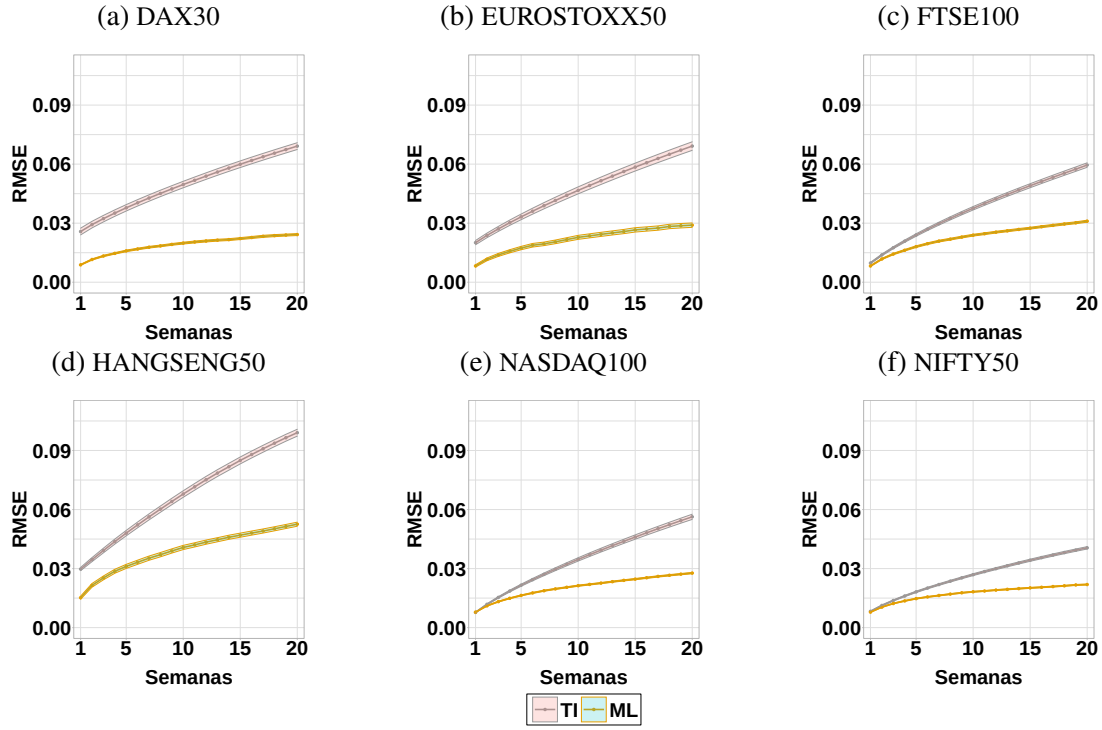


Figura 75 – **RMSE - Comparação do desempenho preditivo dos métodos TI e ML usando** $L = 504$**.** A figura mostra a métrica RMSE dos métodos ML e TI relacionada à previsão de links ponderados. Para cada intervalo de tempo, calculamos a média da RMSE de cada método e seu respectivo erro padrão ao longo de todo o período de testes.


ICMC
USP
SÃO CARLOS