UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

SÉLIO GUILHERME DA SILVA FILHO

PREDIÇÃO ZERO-SHOT DO H-INDEX DE PESQUISADORES VIA FOUNDATION MODELS

Ouro Preto
2026

SÉLIO GUILHERME DA SILVA FILHO

PREDIÇÃO ZERO-SHOT DO H-INDEX DE PESQUISADORES VIA FOUNDATION MODELS

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Dr. Vander Luis de Souza Freitas

Ouro Preto
2026

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E BIOLOGICAS
DEPARTAMENTO DE COMPUTACAO

FOLHA DE APROVAÇÃO

Sélio Guilherme da Silva Filho

Predição Zero-shot do H-index de Pesquisadores via Foundation Models

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Ciência da Computação

Aprovada em 25 de Fevereiro de 2026

Membros da banca

Vander Luis de Souza Freitas (Orientador) - Doutor - Universidade Federal de Ouro Preto
Eduardo José da Silva Luz (Examinador) - Doutor - Universidade Federal de Ouro Preto
Jadson Castro Gertrudes (Examinador) - Doutor - Universidade Federal de Ouro Preto

Vander Luis de Souza Freitas, Orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 04/03/2026

Documento assinado eletronicamente por **Vander Luis de Souza Freitas**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 04/03/2026, às 14:02, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **1062463** e o código CRC **9BE79B19**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: 3135591692 - www.ufop.br

Aos meus pais, amigos e mestres, que tornaram possível a conclusão deste ciclo. E a todos que acreditaram no meu potencial, mesmo quando os resultados ainda eram incertos.

Agradecimentos

A conclusão deste trabalho de conclusão de curso representa não apenas o fim de um ciclo acadêmico, mas o resultado de uma jornada de aprendizado, resiliência e apoio mútuo.

Agradeço aos meus familiares, pelo suporte e por acreditarem no meu potencial. A presença e o incentivo de vocês foram os pilares necessários para a realização deste trabalho.

Ao meu orientador, Professor Dr. Vander Luis, pela confiança depositada e pela orientação fundamental no desenvolvimento desta pesquisa. Agradeço pelas valiosas sugestões, pela paciência durante o processo de aprendizado e por compartilhar seu conhecimento técnico, que foi essencial para a construção deste trabalho.

Agradeço também aos professores do DECOM por todos os ensinamentos.

Aos meus amigos e companheiros de curso, pelas discussões acadêmicas, pelo companheirismo e por terem tornando a caminhada mais leve e proporcionando um ambiente de aprendizado que vai muito além dos livros.

Por fim, agradeço a todos que, de forma direta ou indireta, contribuíram para a realização deste trabalho e para o meu crescimento pessoal durante estes anos.

Resumo

A previsão do h-index, métrica fundamental para avaliar o impacto científico, apresenta desafios complexos devido à sua natureza multifatorial e não-estacionária. Esta monografia investiga a aplicação de modelos de fundação para séries temporais (do inglês *Time Series Foundation Models* - TSFMs) operando em regime *Zero-Shot* para esta tarefa, contrastando sua capacidade de generalização com modelos especialistas supervisionados. Utilizando dados históricos de pesquisadores brasileiros da Ciência da Computação, avaliou-se o desempenho dos modelos de fundação MOIRAI-MoE e TimesFM em comparação com uma rede neural LSTM (treinada especificamente no domínio) e um modelo de persistência (Toy Model). Para lidar com a não estacionariedade do h-index, é proposta uma transformação dos dados, substituindo o valor do h-index de um ano pela sua derivada correspondente, resultando em uma série temporal que não é monotonicamente crescente. A metodologia incluiu a transformação dos dados para garantir estacionariedade e uma validação rigorosa sobre um conjunto de teste de 295 pesquisadores feito a partir de um dataset completo com 1472 pesquisadores. O modelo baseado em LSTM atingiu um RMSE global de $2,3156$, enquanto o MOIRAI-MoE e o TimesFM tiveram RMSE de $2,6582$ e $3,1323$, respectivamente. Os resultados globais demonstraram a superioridade do modelo especialista LSTM, confirmando a importância do ajuste fino para capturar tendências de longo prazo. Entretanto, o MOIRAI-MoE apresentou desempenho notável, superando o TimesFM e alcançando paridade estatística com o LSTM em cenários de curto prazo e contextos históricos limitados. O estudo conclui que, embora modelos especialistas ofereçam maior precisão absoluta, os modelos de fundação baseados em mistura de especialistas (do inglês *Mixture of Experts* - MoE), como o MOIRAI-MoE, representam uma alternativa robusta e eficiente, capazes de prever trajetórias acadêmicas com alta competência sem a necessidade de treinamento específico.

Palavras-chave: H-index. Time Series Foundation Models. MOIRAI-MoE. Previsão de Séries Temporais.

Abstract

The prediction of the h-index, a fundamental metric for assessing scientific impact, presents complex challenges due to its multifactorial and non-stationary nature. This monograph investigates the application of Time Series Foundation Models (TSFMs) operating in a Zero-Shot regime for this task, contrasting their generalization capability with supervised specialist models. Using historical data from Brazilian Computer Science researchers, the performance of the MOIRAI-MoE and TimesFM foundation models was evaluated in comparison with a domain-specific trained LSTM neural network and a persistence model (Toy Model). To address the non-stationarity of the h-index, a data transformation is proposed, replacing the annual h-index value with its corresponding derivative, resulting in a time series that is not monotonically increasing. The methodology included this stationarity transformation and a rigorous validation on a test set of 295 researchers, derived from a complete dataset of 1,472 researchers. The LSTM-based model achieved a global RMSE of 2.3156, while MOIRAI-MoE and TimesFM reached RMSE values of 2.6582 and 3.1323, respectively. The overall results demonstrated the superiority of the specialist LSTM model, confirming the importance of fine-tuning to capture long-term trends. However, MOIRAI-MoE showed remarkable performance, outperforming TimesFM and achieving statistical parity with the LSTM in short-term scenarios and limited historical contexts. This study concludes that while specialist models offer higher absolute precision, Mixture of Experts (MoE) based foundation models, such as MOIRAI-MoE, represent a robust and efficient alternative, capable of predicting academic trajectories with high competence without the need for domain-specific training.

Keywords: H-index. Time Series Foundation Models. MOIRAI-MoE. Time Series Forecasting.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

TSFM	Time Series Foundation Models
LLMs	Large Language Models
MoE	Mixture of Experts
LLaMA	Large Language Model Meta AI
GPT	Generative Pre-trained Transformer
ARIMA	Autoregressive Integrated Moving Average
ARMA	AutoRegressive Moving Average
RNN	Recurrent Neural Network
TCN	Temporal Convolutional Networks
NLP	Natural Language Processing
CV	Computational Vision
FM	Foundation Model
LSTM	Long Short-Term Memory
SciSci	Science of Science
ETS	Error Trend and Seasonality (Suavização Exponencial)
MAE	Mean Absolute Error
EMD	Empirical Mode Decomposition
RMSE	Root Mean Squared Error
MSE	Mean Squared Error
SMAPE	Symmetric Mean Absolute Percentage Error
SVR	Support Vector Regression
RF	Random Forest
GB	Gradient Boosting
MLP	Multilayer Perceptron

FFN	Feed-Forward
LRC	Logistic Regression Classifier
BAG	Bootstrap Aggregating / Bagging
SVM	Support Vector Machines
AUC	Area Under the Curve
ID	Identity
PPG	Programa de Pós-Graduação
MEC	Ministério da Educação
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
ML	Machine learning
DL	Deep learning
CD	Critical Difference
GPU	Graphics Processing Unit
IQR	Inter Quartile Range

Sumário

1 Introdução

A análise de séries temporais tem aplicações em diversos domínios, como finanças, saúde, e ciência da computação (SIAMI-NAMINI; TAVAKOLI; NAMIN, 2018). A evolução de métodos de análise, desde abordagens estatísticas tradicionais até técnicas avançadas de deep learning, reflete a crescente complexidade dos dados e a necessidade de modelos mais eficazes (ZENG et al., 2022). Recentemente, os modelos de fundação (do inglês *Foudation Models* - FMs) surgiram como uma nova classe de modelos, capazes de generalizar o conhecimento adquirido a partir de grandes volumes de dados, e têm transformado a forma como abordamos diversas tarefas de análise, incluindo a previsão de séries temporais (LIANG et al., 2024), destacando-se propostas como o modelo MOIRAI-MoE (LIU et al., 2024) e o TimesFM (DAS et al., 2024).

O *h-index*, uma métrica bibliométrica que busca combinar a produtividade e o impacto científico de um pesquisador, tem se estabelecido como um indicador fundamental para avaliar o sucesso e a influência na pesquisa acadêmica (JENSEN; ROUQUIER; CROISSANT, 2009). A capacidade de prever a evolução do *h-index* de um pesquisador pode oferecer *insights* valiosos para pesquisadores, instituições acadêmicas e agências de fomento, auxiliando no planejamento de carreira, na avaliação de desempenho e na alocação de recursos (MOMENI; MAYR; DIETZE, 2023). No entanto, a previsão do *h-index* é um desafio complexo, influenciado por múltiplos fatores como a qualidade e o impacto das publicações, as colaborações, o reconhecimento na área, entre outros (DONG; JOHNSON; CHAWLA, 2016).

A literatura científica tem atacado o problema da predição do *h-index* através de diversas evoluções metodológicas. O trabalho de Acuna, Allesina e Kording (2012) estabeleceu as bases da predição algorítmica ao utilizar modelos de regressão linear com regularização *elastic net*. Eles identificaram cinco fatores importantes na predição: o *h-index* atual (h), a produtividade acumulada ($\sqrt{n}$), o tempo de carreira (y), a diversidade de periódicos (j) e a qualidade das publicações (q). Embora eficaz para previsões de curtíssimo prazo (alcançando $R^2 = 0,92$ para um ano em Neurociências), o desempenho desses modelos lineares decai significativamente em horizontes maiores, como 10 anos ($R^2 = 0,48$), evidenciando a dificuldade de capturar a dinâmica não linear das trajetórias científicas.

Com o avanço da computação, abordagens baseadas em aprendizado de máquina e redes neurais foram introduzidas para lidar com volumes massivos de dados. Algoritmos de *ensemble*, como *Random Forest* e *XGBoost*, têm sido aplicados em bases de dados de Ciência da Computação, conseguindo explicar até 87,5% da variação do *h-index* em horizontes de 5 anos (WEIHS; ETZIONI, 2017; KUPPLER, 2022). Paralelamente, redes neurais artificiais têm buscado identificar conjuntos ótimos de parâmetros sem restrições de inércia, enquanto técnicas de *network embedding* mapeiam autores, artigos e instituições em espaços vetoriais heterogêneos

para prever o sucesso acadêmico (OCHI et al., 2022).

Modelos recentes de *Time Series Foundation Models* (TSFMs), como o MOIRAI-MoE, propõem uma abordagem inovadora para modelar a diversidade de padrões em dados de séries temporais, usando uma arquitetura de *mixture of experts*. O MOIRAI-MoE, ao contrário de modelos anteriores que utilizam projeções de entrada/saída específicas para diferentes frequências, emprega uma única camada de projeção, delegando a modelagem da diversidade a *experts* especializados dentro da arquitetura *Transformer*. Esta abordagem permite uma especialização automática em nível de *token*, potencialmente melhorando a capacidade de adaptação do modelo a diferentes padrões temporais (LIU et al., 2024).

Apesar dos avanços na predição do impacto acadêmico via modelos supervisionados, os estudos tradicionais focam majoritariamente em métricas de acurácia global e, em sua maioria, negligenciam a natureza não estacionária do h-index ao tratá-lo como uma série absoluta e monotonicamente (quando sua evolução ocorre em um único sentido, apenas cresce ou apenas decresce.) crescente, prática que pode levar a resultados distorcidos (WEIHS; ETZIONI, 2017; ACUNA; ALLESINA; KORDING, 2012). Este trabalho propõe uma mudança de paradigma metodológico através da transformação dos dados em derivadas (Δh) para garantir a estacionariedade. A pesquisa investiga a hipótese de uma "inversão de liderança"entre modelos generalistas e especialistas conforme o tempo avança. Ataca-se, assim, a lacuna de conhecimento sobre a eficácia dos TSFMs em horizontes de curto prazo, avaliando se a vasta exposição prévia desses modelos a dinâmicas universais oferece predições mais robustas e com menor risco de sobreajuste (*overfitting*) local do que as abordagens especialistas convencionais.

O problema central desta pesquisa reside na previsão do *h-index* de pesquisadores ao longo do tempo, uma tarefa complexa dada a natureza multifatorial e não estacionária da progressão de carreira científica. Diante da necessidade de modelos especialistas (como a LSTM) se adaptarem estritamente aos padrões de um conjunto de dados local, surge a questão sobre a eficácia da generalização de modelos pré-treinados em larga escala. Dessa forma, este trabalho investiga a hipótese de que, em horizontes de previsão de curto prazo, os FMs podem superar modelos especialistas treinados no domínio. A fundamentação para tal proposição reside no fato de que os FMs, por serem expostos a uma diversidade massiva de séries temporais, podem capturar dinâmicas universais de curto prazo com maior robustez, enquanto modelos especialistas podem apresentar um viés excessivo aos padrões médios do conjunto de treino (sobreajuste local). A validação desta hipótese é conduzida através de uma análise comparativa utilizando os modelos MOIRAI-MoE e TimesFM frente a uma arquitetura LSTM dedicada.

1.1 Justificativa

A relevância deste trabalho fundamenta-se, primeiramente, na necessidade crescente de mecanismos objetivos para a gestão da ciência e da carreira acadêmica. A capacidade de

prever o *h-index* com precisão transcende o exercício computacional, configurando-se como uma ferramenta estratégica para a avaliação de produtividade em programas de pós-graduação e agências de fomento. Tais previsões fornecem subsídios para processos de promoção na carreira docente, auxiliam na alocação eficiente de recursos institucionais e permitem que o próprio pesquisador realize um planejamento de carreira fundamentado em dados, identificando tendências de crescimento ou possíveis estagnações em sua trajetória de impacto (DONG; JOHNSON; CHAWLA, 2016; MOMENI; MAYR; DIETZE, 2023).

Nesse contexto, a importância do *h-index* como métrica central de avaliação exige o desenvolvimento de ferramentas de previsão mais eficazes e adaptáveis (MOMENI; MAYR; DIETZE, 2023). A aplicação de FMs voltados para séries temporais (*Time Series Foundation Models*), com suas arquiteturas avançadas e capacidade de especialização em nível de *token*, pode contribuir significativamente para aprimorar a precisão e a robustez dessas estimativas. A investigação desses modelos em um domínio específico revela *insights* cruciais sobre a capacidade das arquiteturas de larga escala em lidar com dados complexos, heterogêneos e não lineares, típicos da produção científica.

Ademais, o uso de FMs introduz uma vantagem pragmática decisiva, a eliminação da necessidade de treinamento em dados específicos para cada nova base analisada. O processo de treinamento de modelos especialistas é frequentemente oneroso, demandando *hardware* de alto desempenho e tempo de processamento considerável. Ao aproveitar o enorme esforço computacional já despendido na criação desses grandes modelos, esta pesquisa investiga se a estratégia *zero-shot*, na qual o modelo realiza predições em dados nunca vistos anteriormente, é suficiente para entregar resultados competitivos de forma imediata e escalável, reduzindo barreiras de entrada para instituições com menor poder computacional.

1.2 Objetivos

O objetivo geral desta pesquisa é investigar a eficácia de dois *Foundation Models*, o MOIRAI-MOE e o TimesFM, ambos na estratégia *zero-shot*, para a previsão do *h-index* de pesquisadores, contrastando seu desempenho com o de baselines treinados com a base de dados em estudo e identificando seus pontos fortes e limitações.

- Comparar o desempenho do MOIRAI-MOE e do TimesFM com o modelo LSTM, dado um conjuntos de dados e cenários de previsão, utilizando métricas apropriadas, como o erro médio absoluto (MAE), Erro Quadrático Médio (RMSE) e Erro Percentual Absoluto Médio Simétrico (SMAPE).
- Avaliar os modelos de fundação na tarefa de predição de h-index, na estratégia *zero-shot*
- Investigar a capacidade dos modelos de fundação de capturar as dinâmicas temporais e os padrões complexos que caracterizam a progressão do *h-index* ao longo do tempo para

diferentes horizontes de predição.

Este estudo busca, portanto, contribuir para o avanço da área de previsão de séries temporais, além de oferecer *insights* valiosos para pesquisadores e instituições na avaliação do impacto científico.

1.2.1 Estrutura da Monografia

A presente monografia está subdividida como segue. O Capítulo 1 trouxe a introdução. A revisão bibliográfica, para compreensão da teoria necessária para entendimento do texto, e o posicionamento das contribuições com a literatura recente são apresentados no Capítulo 2. O Capítulo 3 detalha a metodologia proposta, descrevendo o processo de coleta e pré-processamento do conjunto de dados de pesquisadores brasileiros, a aplicação da transformação delta para mitigar a não estacionariedade e a implementação técnica da rede especialista LSTM frente aos modelos de fundação TimesFM e MOIRAI-MOE. A análise experimental e a discussão detalhada dos resultados são conduzidas no Capítulo 4, onde o desempenho é mensurado por meio das métricas de erro MAE, RMSE e SMAPE, além de ser submetido a uma rigorosa validação estatística sobre um conjnto de teste de 295 autores. Por fim, o Capítulo 5 apresenta as considerações finais do estudo, sintetizando as conclusões sobre a eficácia dos modelos *zero-shot* na predição de trajetórias acadêmicas e sugerindo direções para trabalhos futuros na área.

2 Revisão Bibliográfica

Este Capítulo introduz os conceitos necessários para entendimento da pesquisa conduzida, como *h-index*, predição de séries temporais, e foundation models.

2.1 Fundamentação Teórica

A Seção está organizada em cinco Subseções, cada uma abordando um dos pilares fundamentais investigados nesta pesquisa. Na Subseção 2.1.1, discute-se o conceito de *h-index*, sua definição e relevância como métrica acadêmica. Na Subseção 2.1.3, exploram-se os *Time Series Foundation Models* (TSFMs), destacando suas características e aplicações. Em seguida, na Subseção 2.1.4 apresenta-se o TimesFM e seus principais aspectos. Na Subseção 2.1.6, define-se o que são as redes *Long Short-Term Memory* (LSTM) e seu funcionamento. Por fim, na Subseção 2.1.5, detalha-se o funcionamento do MOIRAI-MoE, modelo selecionado como objeto de estudo desta pesquisa.

2.1.1 H-index

A Ciência da Ciência (SciSci) surge como um campo multidisciplinar que busca compreender e aprimorar a prática científica por meio da análise de grandes volumes de dados. Integrando disciplinas como cientometria, sociologia, ciência da computação e estatística, a SciSci emprega métodos quantitativos avançados, incluindo aprendizado de máquina, análise de redes e simulações computacionais. Seus tópicos de investigação abrangem a dinâmica das redes científicas, a seleção de problemas de pesquisa, a inovação, as trajetórias de carreira, o impacto das equipes e a dinâmica de citação (FORTUNATO et al., 2018). Dentre as diversas métricas utilizadas para avaliar o impacto acadêmico e a produtividade dos pesquisadores, destaca-se o *h-index*.

Proposto por J.E. Hirsch (HIRSCH, 2005), o *h-index* é uma métrica amplamente utilizada para medir o impacto e a produtividade de um pesquisador. A definição formal estabelece que um cientista possui *h-index* se de seus artigos tiverem recebido pelo menos h citações cada, enquanto os demais artigos não excedem esse número de citações. Por exemplo, um pesquisador com *h-index* igual a 20 possui 20 artigos citados pelo menos 20 vezes cada, e nenhum outro com mais de 20 citações (DONG; JOHNSON; CHAWLA, 2016).

Em Momeni, Mayr e Dietze (2023) é demonstrado como a predição do *h-index* é uma tarefa complexa que requer a integração de fatores históricos, demográficos e de colaboração, pois características não baseadas no impacto anterior oferecem maior robustez em previsões de longo prazo, especialmente para jovens pesquisadores. O uso de aprendizado de máquina tem

desempenhado um papel crucial ao permitir análises mais precisas e escaláveis, destacando-se como uma abordagem promissora para lidar com a evolução dinâmica do impacto científico permitindo um estudo que vise melhorar o alcance e relevância de novos estudos.

A avaliação do desempenho de pesquisadores é crucial para comitês de contratação e órgãos de financiamento. Geralmente, essa avaliação baseia-se na produtividade científica, citações ou em métricas combinadas, como o *h-index*. Nesse contexto, prever o *h-index* pode auxiliar na compreensão do impacto científico de pesquisadores e em tomadas de decisão relacionadas a sua carreira acadêmica (MOMENI; MAYR; DIETZE, 2023).

O *h-index* combina dois elementos fundamentais: a quantidade de publicações e a sua qualidade. Isso significa que não basta publicar muitos artigos, é necessário que essas publicações sejam amplamente citadas por outros pesquisadores. Esse balanço entre volume e impacto transformou o *h-index* em um padrão para medir desempenho acadêmico e impacto científico (DONG; JOHNSON; CHAWLA, 2016), permitindo comparar o desempenho de diferentes pesquisadores, especialmente dentro de uma mesma área de pesquisa (DONG; JOHNSON; CHAWLA, 2015).

Diversas ferramentas online estão disponíveis para calcular o *h-index*, utilizando bases de dados como o *Web of Science* (KELLY; JENNIONS, 2006). A previsão do impacto científico é uma área de interesse crescente, pois é essencial para antecipar trajetórias de carreira e compreender os mecanismos que influenciam o impacto futuro no processo científico (MOMENI; MAYR; DIETZE, 2023).

O *h-index* busca mitigar as limitações de outras métricas, como o número total de artigos ou o número total de citações, que podem ser influenciadas por fatores que não refletem diretamente a qualidade ou o impacto de uma pesquisa. Ao combinar produtividade e impacto em uma única métrica, o *h-index* fornece uma avaliação mais equilibrada do desempenho acadêmico (JENSEN; ROUQUIER; CROISSANT, 2009).

Diversos estudos têm explorado a previsão do *h-index* de pesquisadores, utilizando abordagens como modelos de regressão linear, classificação e técnicas de aprendizado de máquina (DONG; JOHNSON; CHAWLA, 2016). Esses estudos indicam que é possível prever o *h-index* futuro com boa precisão, especialmente no caso de pesquisadores em estágios iniciais de suas carreiras (MOMENI; MAYR; DIETZE, 2023).

Entretanto, a precisão das previsões do *h-index* tende a diminuir conforme aumenta o intervalo de tempo considerado. Isso se deve às complexidades e incertezas associadas ao processo de citação em horizontes temporais mais longos (PENNER et al., 2013).

A previsão do impacto científico é um dos desafios abordados pela SciSci, sendo fundamental para antecipar trajetórias de carreira e revelar mecanismos subjacentes ao processo de citação. No entanto, prever o *h-index* é uma tarefa complexa, pois ele é uma medida não estacionária, ou seja, não possui média e variância constantes ao longo do tempo. O uso de

medidas cumulativas como variáveis dependentes em modelos de regressão pode superestimar a capacidade preditiva dos modelos, levando a resultados distorcidos. (PENNER et al., 2013).

2.1.2 Time Series Data

Uma série temporal é definida, em sua essência, como uma sequência de pontos de dados ordenados cronologicamente (THAKUR, 2024; LIANG et al., 2024). Diferentemente de conjuntos de dados convencionais, nos quais a ordem das observações pode ser arbitrária, as séries temporais são caracterizadas por uma ordenação sequencial estrita, a qual encapsula dependências temporais que refletem a dinâmica e a evolução de sistemas complexos no mundo real (LIANG et al., 2024). O objetivo primordial da análise desses dados reside em modelar sua estrutura temporal subjacente, que frequentemente exibe padrões estruturais distintos, tais como as tendências, que indicam a direção geral ou o movimento de longo prazo dos dados, a sazonalidade e os ciclos, que compreendem flutuações repetitivas em intervalos regulares, e o ruído, que engloba variações aleatórias desprovidas de padrões específicos (THAKUR, 2024).

No que concerne à sua classificação, as séries temporais podem assumir diferentes formatos dependendo da complexidade das variáveis registradas. São denominadas univariadas quando monitoram apenas uma única variável ao longo do tempo, e multivariadas quando capturam simultaneamente múltiplos indicadores a cada passo temporal (LIANG et al., 2024). Adicionalmente, o conceito pode ser expandido para séries temporais espaciais, que associam dados evolutivos a localizações geográficas específicas, frequentemente modeladas por grafos espaço-temporais ou matrizes de grade, e para trajetórias ou sequências de eventos, que incluem dados contínuos de movimentação espacial ou conjuntos de eventos ordenados por suas respectivas datas de ocorrência (LIANG et al., 2024).

Uma série temporal estacionária é caracterizada por possuir propriedades estatísticas fundamentais que permanecem invariantes com o passar do tempo (PENNER et al., 2013). Sob a definição de estacionariedade fraca, a série deve apresentar média e variância independentes do tempo, enquanto a autocovariância entre um ponto t e um ponto futuro $t+\Delta t$ deve ser uma função que dependa exclusivamente da defasagem temporal Δt (PENNER et al., 2013). Esse conceito está intrinsecamente ligado à premissa de homocedasticidade, que determina a constância da variância dos dados ao longo da sequência (THAKUR, 2024). No contexto desta pesquisa, o *h-index* é identificado como uma medida inerentemente não estacionária, uma vez que não possui média e variância constantes ao longo do tempo e apresenta uma tendência natural de crescimento acumulado. Esta característica define o *h-index* como uma série temporal monotonicamente crescente, na qual os valores sucessivos são sempre iguais ou superiores aos anteriores devido à sua natureza cumulativa.

A importância da estacionariedade reside no fato de ela servir como base para modelos clássicos de previsão, como o ARIMA e a Suavização Exponencial (ETS), que são construídos sob suposições de linearidade e relações temporais estáveis (THAKUR, 2024). Contudo, séries

temporais do mundo real frequentemente se opõem a essa premissa, exibindo mudanças de distribuição e comportamento que variam continuamente, mesmo em janelas de contexto reduzidas (LIU et al., 2024).

Para mitigar tais efeitos, séries não estacionárias exigem transformações matemáticas, como a diferenciação. No modelo ARIMA, a etapa de "Integração"(I) aplica esse método para converter a série original, onde o parâmetro d indica a ordem de diferenciação necessária para atingir a estacionariedade (SIAMI-NAMINI; TAVAKOLI; NAMIN, 2018). Cabe ressaltar que, devido à dependência dessas premissas estritas, abordagens tradicionais tendem a ser menos eficazes na captura de dependências não-lineares complexas quando comparadas a modelos modernos de aprendizado de máquina (THAKUR, 2024).

2.1.3 Time Series Foundation Models

Os *Time Series Foundation Models* (TSFMs) são modelos de aprendizado de máquina pré-treinados em grandes quantidades de dados, projetados para capturar padrões e características amplas em séries temporais. Esses modelos são altamente generalizáveis e adaptáveis a diversas tarefas de análise de séries temporais, possibilitando aplicações em previsão, detecção de anomalias e classificação de padrões (THAKUR, 2024; LIANG et al., 2024).

A principal ideia por trás dos TSFMs é que, ao aprender com um grande volume de dados, eles possam ser ajustados para tarefas específicas com menos necessidade de dados rotulados. Isso os torna uma solução robusta para lidar com os desafios impostos pela diversidade e complexidade dos dados temporais (THAKUR, 2024; LIANG et al., 2024).

2.1.3.1 Arquiteturas de TSFMs

Muitos TSFMs utilizam arquiteturas baseadas em *Transformers* devido à capacidade desses modelos de capturar dependências de longo alcance nos dados por meio de mecanismos de autoatenção (*self-attention*) (THAKUR, 2024). É importante notar, entretanto, que a autoatenção processa as informações de forma paralela e é inerentemente invariante à permutação, o que, isoladamente, resultaria na perda da noção de ordem cronológica essencial às séries temporais. Para mitigar esse problema e tornar a arquitetura apta ao tratamento de dados sequenciais, técnicas como a codificação posicional (*positional encoding*) e a organização de dados em blocos significativos (*patching* ou *tokens* de sub-séries) são amplamente empregadas, na tentativa de garantir que a estrutura temporal seja preservada e compreendida pelo modelo (ZENG et al., 2022).

A codificação posicional é a responsável por injetar informações sobre a posição relativa ou absoluta de cada ponto na série temporal diretamente nos vetores de entrada, permitindo que o modelo diferencie a ordem dos eventos e aprenda relações temporais de forma eficaz. Assim, a combinação da autoatenção global com a codificação posicional permite que o modelo mantenha

a sensibilidade à sequência enquanto beneficia-se de uma maior eficiência computacional (ZENG et al., 2022).

Os TSFMs podem ser classificados em três tipos principais de arquitetura:

- **Somente Encoder:** Focam no processamento da entrada para criar representações contextualizadas, geralmente aprendendo por meio de reconstrução de máscara (estilo BERT). Um exemplo proeminente é o *MOMENT*, que utiliza *patches* de comprimento fixo e arquitetura *Transformer* para representação de séries temporais, empregando técnicas de normalização para centralizar os dados antes do processamento.
- **Somente Decoder:** Projetados para a geração sequencial de saídas, operando *token* por *token* em um regime autorregressivo. Exemplos notáveis incluem o *TimesFM*, que utiliza blocos residuais para previsões determinísticas, o *LagLlama*, focado em predição probabilística, e o *MOIRAI-MoE*, que utiliza uma arquitetura de Mistura de Especialistas para lidar com a heterogeneidade das séries temporais (LIU et al., 2024; MULAYIM et al., 2024).
- **Encoder-Decoder:** Combinam ambas as funcionalidades para traduzir ou transformar sequências complexas, sendo úteis em tarefas que exigem um mapeamento direto entre janelas de contexto e horizontes de previsão distintos (LIANG et al., 2024).

Outras arquiteturas inovadoras também reprogramam *embeddings* de modelos de linguagem para tarefas de séries temporais, como o TimeLLM e o TimeGPT, que lidam com dados faltantes e *timestamps* irregulares (MULAYIM et al., 2024).

2.1.3.2 Técnicas de Pré-Treinamento e Adaptação

Os TSFMs são treinados em grandes conjuntos de dados que incluem séries temporais reais e sintéticas de diversas fontes, como *Google Trends*, estatísticas de páginas da Wikipédia e *datasets* como o M4. Esses conjuntos de dados permitem capturar padrões variados e granularidades (DAS et al., 2024).

O pré-treinamento geralmente utiliza técnicas de aprendizado autônomo (*self-supervised learning*), como *contrastive learning* e abordagem generativa, em combinação com *mini-batches* e máscaras aleatórias para lidar com diferentes comprimentos de contexto (DAS et al., 2024). Após o pré-treinamento, os modelos podem ser ajustados para tarefas específicas através de *fine-tuning* (processo de realizar o ajuste fino dos pesos de um modelo, utilizando um conjunto de dados específico e rotulado do domínio alvo) ou aprendizado *few-shot* (quando o modelo realiza previsões após receber uma pequena quantidade de exemplos ou amostras de referência da tarefa específica), permitindo aplicações eficazes mesmo com poucos dados rotulados (LIANG et al., 2024).

2.1.3.3 Características e Aplicações

Uma característica marcante dos TSFMs é sua capacidade de realizar predição sem *fine-tuning*, dependendo do pré-treinamento e da arquitetura do modelo. Essa abordagem, chamada de predição *zero-shot*, permite que os modelos desempenhem bem mesmo em cenários novos e desconhecidos (DAS et al., 2024).

Esses modelos são amplamente utilizados para prever valores futuros com base em dados históricos, capturando dependências temporais e tendências. Entre os TSFMs, destacam-se o TimesFM e o MOIRAI, ambos com capacidades notáveis de previsão *zero-shot* e alta generalização (DAS et al., 2024). Além disso, suas aplicações incluem previsão de demanda, detecção de anomalias, previsão de preços e análise de tráfego, tornando-os ferramentas versáteis (LIANG et al., 2024).

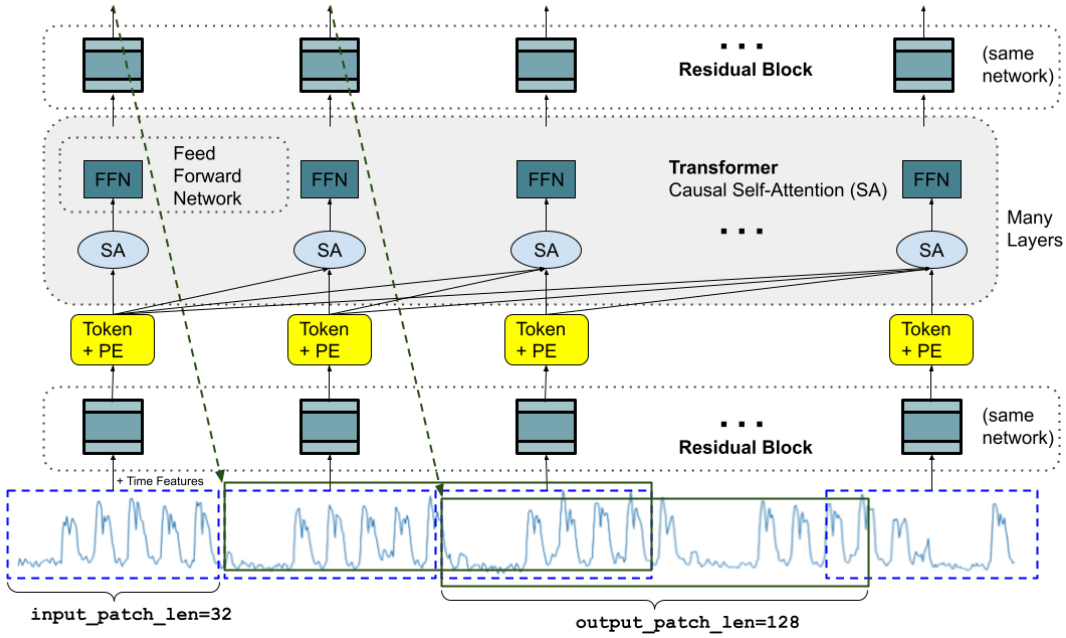
2.1.4 TimesFM

O TimesFM é um modelo de fundação para previsão de séries temporais, desenvolvido com o objetivo de atingir um desempenho de *zero-shot* próximo ao de modelos supervisionados, sem a necessidade de treinamento individual para cada conjunto de dados. Ele se destaca por ser um modelo único, pré-treinado, capaz de generalizar bem em diferentes domínios, granularidades temporais e horizontes de previsão (DAS et al., 2024).

O TimesFM é treinado com um grande volume de dados de séries temporais, incluindo dados do mundo real e sintéticos. Os dados reais incluem buscas do Google, visualizações de páginas da Wikipedia e conjuntos de dados públicos como M4, Electricity e Traffic. Os dados sintéticos foram gerados por combinações de processos ARMA, padrões sazonais, tendências e funções degrau (DAS et al., 2024).

A arquitetura (Figura 2.1) e funcionamento do TimesFM são descritos da seguinte forma:

Figura 2.1 – Arquitetura do modelo TimesFm.



Fonte: (DAS et al., 2024)

O modelo TimesFM é fundamentado em uma estrutura *decoder-only* baseada em atenção, assemelhando-se aos grandes modelos de linguagem (LLMs), porém otimizada especificamente para o processamento eficiente de séries temporais. Inspirado por avanços em visão computacional e sistemas de previsão de longo prazo, o modelo segmenta as séries temporais de entrada em blocos menores denominados *patches*, que operam como os *tokens* fundamentais da rede. Cada *patch* é processado por um bloco residual que o converte em um vetor de tamanho fixo, o qual é posteriormente alimentado nas camadas do *Transformer*; essa abordagem não apenas possibilita o tratamento de sequências extensas, como também acelera a inferência ao reduzir significativamente o volume de *tokens* processados. Para assegurar a integridade temporal das previsões, o modelo emprega um mecanismo de atenção causal, garantindo que cada *token* de saída dependa exclusivamente de informações passadas na sequência. Adicionalmente, durante a fase de treinamento, utiliza-se uma estratégia de *masking* que consiste em ocultar aleatoriamente porções iniciais dos *patches* de entrada, capacitando o modelo a realizar predições robustas sob variados comprimentos de histórico contextual (DAS et al., 2024).

O TimesFM é treinado para prever o próximo *patch* com base nos *patches* anteriores, permitindo previsões flexíveis com diferentes comprimentos de histórico e horizontes de previsão. Os *patches* de saída podem ser mais longos que os *patches* de entrada, reduzindo a necessidade de passos auto-regressivos durante a inferência (DAS et al., 2024).

Estudos de ablação mostram que o desempenho do modelo melhora com o aumento do número de parâmetros e com o comprimento dos *patches* de saída. Os resultados também evidenciam a importância dos dados sintéticos para a generalização do modelo (DAS et al., 2024).

O modelo demonstrou desempenho significativamente superior ao de LLMs como GPT-3 e LLaMA-2 em tarefas *zero-shot*, aliado a um custo computacional drasticamente reduzido. Adicionalmente, seu desempenho pode ser ainda mais aprimorado através do refinamento com porções mínimas de um conjunto de dados específicos (DAS et al., 2024).

Em suma, o TimesFM é um modelo de fundação promissor para previsão de séries temporais, destacando-se pela sua alta precisão, flexibilidade e capacidade de generalização. Ele alcança um desempenho notável em diversas tarefas sem a necessidade de treinamento específico para cada conjunto de dados, além de possuir um tamanho e custo computacional significativamente menores em relação a grandes modelos de linguagem (LLMs) (DAS et al., 2024).

2.1.5 MOIRAI-MoE

O modelo MOIRAI-MoE é uma versão aprimorada do MOIRAI (WOO et al., 2024), utilizando uma arquitetura de *sparse mixture of experts* (MoE) para lidar com a diversidade em dados de séries temporais. A especialização ocorre no nível do *token*, permitindo que diferentes partes do modelo se concentrem em padrões específicos. Enquanto o MOIRAI utiliza múltiplas camadas de projeção de entrada/saída especializadas para diferentes frequências de dados, o MOIRAI-MoE adota apenas uma camada de projeção combinada com uma MoE, aumentando sua eficiência e flexibilidade. Este modelo demonstrou desempenho superior em comparação a outros, como TimesFM e Chronos (LIU et al., 2024).

Essa mudança de paradigma é fundamentada em duas limitações críticas da especialização por frequência: (1) a frequência não é um indicador absoluto dos padrões subjacentes, visto que séries de frequências distintas podem apresentar comportamentos similares, e (2) a não-estacionariedade inerente aos dados reais exige uma modelagem mais granular do que a permitida por categorias fixas (LIU et al., 2024).

Na arquitetura *sparse mixture of experts* (MoE), cada *token* é processado por um subconjunto de especialistas dentro do *Transformer*. A camada MoE é composta por múltiplas redes de especialistas e uma função de *gating*, que gera um vetor indicando a atribuição dos especialistas. Essa atribuição é esparsa, garantindo eficiência computacional ao ativar apenas um subconjunto de especialistas para cada *token* processado. Essa abordagem é altamente eficaz em domínios diversos, como finanças, saúde, gerenciamento de energia, análise de tráfego, previsão de preços de ações e detecção de anomalias em dados de sensores. (LIU et al., 2024).

Além do mecanismo de esparsidade, a arquitetura incorpora outras estratégias para eficiência e representação temporal. Entre elas, destaca-se a técnica de *patching*, que segmenta a série temporal de comprimento S em *patches* não sobrepostos de tamanho P para capturar informações semânticas locais e reduzir o custo computacional. Após a normalização para mitigar desvios de distribuição, esses *patches* passam por um *Multilayer Perceptron* (MLP) residual

que gera os *tokens* temporais. O treinamento adota um objetivo *decoder-only* (autoregressivo), treinada sob o objetivo de previsão do próximo *token* (*Next-Token Prediction*), o que otimiza a eficiência ao permitir o aprendizado paralelo de variados comprimentos de contexto em uma única atualização do modelo (LIU et al., 2024).

O MOIRAI-MoE apresenta um desempenho superior tanto em cenários in-distribution (quando os dados de teste possuem distribuição semelhante à dos dados de treinamento) quanto *zero-shot*. Em avaliações experimentais, ele superou o MOIRAI em até 17% e demonstrou resultados superiores a modelos como TimesFM e Chronos, mesmo utilizando um número significativamente menor de parâmetros ativados (LIU et al., 2024).

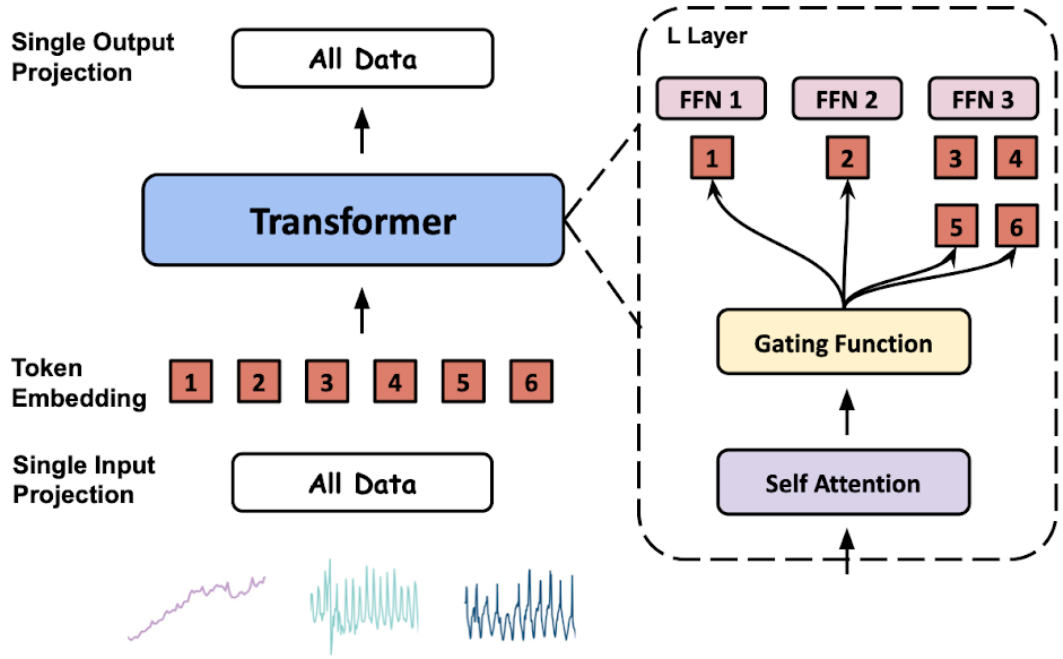
As especificações do modelo MOIRAI-MoE referente aos parâmetros nativos da arquitetura Base se econtra na Tabela 2.1.

Tabela 2.1 – Parâmetros de arquitetura do modelo MOIRAI-MoE Base.

Parâmetro de Arquitetura	**Valor/Configuração**
Número de Camadas (*Layers*)	12
Dimensão do Modelo (d_{model})	768
Dimensão *Feed-Forward* (d_{ff})	1024
Total de Especialistas (M)	32
Especialistas Ativados por *Token* (K)	2
Parâmetros Totais	935 Milhões
Parâmetros Ativos (Inferência)	86 Milhões
Objetivo de Treinamento	*Decoder-Only*

Na Figura 2.2 se encontra a arquitetura visual do MOIRAI-MoE:

Figura 2.2 – Arquitetura do modelo MOIRAI-MoE.



Fonte: (LIU et al., 2024)

O processamento inicial dos dados no MOIRAI-MoE é estruturado em quatro etapas fundamentais voltadas a assegurar a robustez preditiva. Primeiramente, a série temporal é submetida ao *patching*, no qual é segmentada em blocos não sobrepostos de tamanho fixo P (tipicamente $P = 16$), tratados como unidades semânticas locais equivalentes a *tokens* em modelos de linguagem. Em seguida, ocorre o processo de achatamento (*flattening*), que unifica todas as variáveis em uma sequência única para permitir que o mecanismo de autoatenção capture correlações temporais e entre variáveis de forma simultânea. A terceira etapa consiste na normalização causal via *Instance Normalization*, visando mitigar desvios de distribuição (*distribution shifts*); por tratar-se de um modelo *decoder-only*, utiliza-se uma taxa de mascaramento ($r = 0.3$) para evitar o vazamento de informações futuras durante o cálculo do normalizador. Por fim, os *patches* passam por uma projeção residual via Perceptron Multicamadas (MLP), que os converte em vetores de dimensão D, focando estritamente em padrões estruturais ao ocultar a frequência original dos dados.

O núcleo da inovação do MOIRAI-MoE reside na substituição das camadas convencionais de redes *feed-forward* (FFN) por camadas MoE. Uma camada MoE é composta por um conjunto de M redes de especialistas $\{E_1, \ldots, E_M\}$, onde cada especialista possui a estrutura de uma FFN padrão, e uma função de roteamento ou *gating* (G). A principal vantagem desta abordagem é a escalabilidade: o modelo pode aumentar drasticamente sua capacidade total de parâmetros sem elevar proporcionalmente o custo computacional por *token*, uma vez que apenas um subconjunto K de especialistas é ativado para cada processamento (LIU et al., 2024).

Para um dado *token* com representação $\tilde{x}^{l-1}$, a saída da camada MoE é a soma ponderada

das saídas dos especialistas selecionados:

$$\text{Saída} = \sum_{i=1}^{M} G(\tilde{x}^{l-1})_i \cdot E_i(\tilde{x}^{l-1}), \tag{2.1}$$

sendo $E_i(\tilde{x}^{l-1})$ a representação da saída da i-ésima rede de especialistas e $G(\tilde{x}^{l-1})_i$ consiste na pontuação de afinidade *token-to-expert* gerada pela função de *gating*.

Inclusive, utiliza-se $M = 32$ especialistas totais, mas apenas $K = 2$ são ativados por *token*. No modelo Base, isso resulta em um volume de 935 milhões de parâmetros totais, mas apenas 86 milhões de parâmetros ativos durante a inferência, garantindo uma eficiência computacional superior a modelos densos.

A função de *gating* é responsável por gerar um vetor de afinidade que determina quais especialistas processarão determinado *token*. No MOIRAI-MoE, utiliza-se geralmente $K = 2$, o que significa que cada *token* é enviado para os dois especialistas com maior pontuação de afinidade. Uma inovação específica deste modelo é a introdução do *Gating* por *Clusters* de *Tokens*. Em vez de uma projeção linear aleatória, o modelo utiliza centroides de *clusters* derivados de representações de um modelo denso pré-treinado (MOIRAI) submetidas ao algoritmo *k-means* para identificar centroides que representam padrões temporais reais. O roteador calcula a distância Euclidiana entre o *token* atual e esses centroides de *clusters* ($C \in R^{M \times D}$), direcionando o dado para o especialista cujo *cluster* é semanticamente mais próximo. Isso garante que *tokens* com comportamentos similares, como um crescimento explosivo inicial no *h-index*, sejam processados pelos mesmos especialistas, independentemente da escala ou domínio dos dados (LIU et al., 2024).

Essas distâncias atuam como pontuações de afinidade *token-to-expert* para a atribuição dos especialistas, conforme a função de *gating* definida por:

$$G(\tilde{x}^{l-1}) = \text{Softmax}(\text{TopK}(\text{Euclidean}(\tilde{x}^{l-1}, C))), \tag{2.2}$$

sendo $\tilde{x}^{l-1}$ o vetor de ativação do *token* de entrada ou a representação latente proveniente da camada anterior $l-1$, a variável C denota a matriz de centroides de *clusters* $C \in R^{M \times D}$, na qual M representa o total de especialistas disponíveis (32 na versão *Base*) e D a dimensão do modelo. A função Euclidean atua no cálculo da distância geométrica entre a representação do *token* e cada um dos padrões prototípicos contidos em C, enquanto o operador TopK seleciona os K especialistas (com $K = 2$) que apresentam as menores distâncias em relação ao dado atual. Por fim, a função Softmax realiza a normalização desses valores, convertendo-os em pesos de probabilidade que definem a contribuição relativa de cada especialista selecionado para a composição da saída final da camada.

Para evitar que poucos especialistas fiquem sobrecarregados enquanto outros permanecem subutilizados (problema de colapso de especialistas), introduz-se uma perda auxiliar de

balanceamento de carga (*load balancing loss*). Formamente, para um lote $\mathcal{B}$ contendo T *tokens*, a perda de balanceamento de carga é definida como:

$$\mathcal{L}_{load} = M \sum_{i=1}^{M} D_i P_i, \tag{2.3}$$

em que:

$$D_i = \frac{1}{T} \sum_{x \in \mathcal{B}} 1\{\text{argmax } G(\tilde{x}^{l-1}) = i\}, \quad P_i = \frac{1}{T} \sum_{x \in \mathcal{B}} G(\tilde{x}^{l-1})_i. \tag{2.4}$$

Nesta formulação, D_i denota a fração de *tokens* roteados para o especialista i, enquanto P_i indica a proporção da probabilidade de *gating* alocada a esse mesmo especialista.

Análises do modelo revelam que a especialização ocorre de forma hierárquica: com camadas rasas, que apresentam uma seleção de especialistas altamente diversa para lidar com a variabilidade de curto prazo, ruídos e mudanças sazonais abruptas. E camadas profundas onde a seleção de especialistas converge, sugerindo um processo de *denoising* progressivo onde o modelo abstrai as séries temporais em representações de alto nível e invariantes à frequência, focando em tendências globais e padrões de longo prazo.

Dessa forma, o MOIRAI-MoE não apenas supera modelos densos muito maiores em cenários *zero-shot*, utilizando uma fração dos parâmetros ativos (até 65 vezes menos que competidores como o Chronos) e mantendo eficiência de inferência com precisão superior, mas também representa, em síntese, um avanço significativo no campo dos *Time Series Foundation Models*. Sua arquitetura inovadora permite especialização automática em nível de *token*, alcançando resultados superiores em previsões de séries temporais tanto em cenários *in-distribution* quanto *zero-shot* (LIU et al., 2024).

2.1.6 Arquitetura Long Short-Term Memory (LSTM)

As Redes Neurais Recorrentes (RNNs) são uma extensão das redes neurais feedforward, projetadas para lidar com sequências de comprimento variável. Diferente das redes feedforward, nas quais todas as entradas e saídas são tratadas de forma independente, as RNNs utilizam estados ocultos recorrentes para armazenar informações das entradas anteriores, permitindo capturar dependências temporais nos dados (SIAMI-NAMINI; TAVAKOLI; NAMIN, 2019).

Embora, em teoria, as RNNs possam reter informações sequenciais arbitrariamente longas, na prática, sua memória é limitada devido a problemas como o desaparecimento e a explosão dos gradientes. O desaparecimento dos gradientes ocorre quando os valores dos gradientes tornam-se extremamente pequenos à medida que a informação passa por várias camadas, tornando difícil a aprendizagem de dependências de longo prazo. Por outro lado, a explosão dos gradientes

ocorre quando esses valores crescem exponencialmente, resultando em um aprendizado instável (SIAMI-NAMINI; TAVAKOLI; NAMIN, 2019).

As redes *Long Short-Term Memory* (LSTM) foram desenvolvidas para mitigar essas limitações das RNNs, permitindo a retenção de informações por longos intervalos de tempo. A diferença fundamental entre uma RNN convencional e uma LSTM está na sua capacidade de armazenar informações de dependência temporal de longo prazo e de mapear corretamente relações entre dados de entrada e saída (SIAMI-NAMINI; TAVAKOLI; NAMIN, 2018; ABBASIMEHR; SHABANI; YOUSEFI, 2020).

A arquitetura das LSTMs difere da arquitetura perceptron convencional por incluir uma célula de memória e três portões que regulam o fluxo de informações:

- **Portão de Esquecimento (Forget Gate):** Decide quais informações devem ser descartadas da memória da célula. A operação desse portão é descrita por:

$$f_t = \sigma(W_{fh}[h_{t-1}], W_{fx}[x_t], b_f). \tag{2.5}$$

  onde b_f é um viés constante (SIAMI-NAMINI; TAVAKOLI; NAMIN, 2019).

- **Portão de Entrada (Input Gate):** Determina quais novos dados serão armazenados na memória. Esse portão consiste em:

  Uma camada decide quais valores serão modificados:

$$i_t = \sigma(W_{i_h}[h_{t-1}], W_{i_x}[x_t], b_i). \tag{2.6}$$

  Com função de ativação tanh cria um vetor de novos valores candidatos que podem ser adicionados ao estado:

$$\tilde{c}_t = \tanh(W_{c_h}[h_{t-1}], W_{c_x}[x_t], b_c). \tag{2.7}$$

  A atualização da memória da LSTM, onde o valor atual é esquecido usando a camada do portão de esquecimento, ocorre conforme a equação:

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t. \tag{2.8}$$

- **Portão de Saída (Output Gate):** Decide qual será a saída da célula. A operação desse portão é descrita por:

$$o_t = \sigma(W_{o_h}[h_{t-1}], W_{o_x}[x_t], b_o), \tag{2.9}$$

  e a saída saída final da célula é calculada como:

$$h_t = o_t * \tanh(c_t). \tag{2.10}$$

O processo de treinamento das redes LSTM é estruturado em três etapas fundamentais. Inicialmente, realiza-se o cálculo da saída da rede por meio do aprendizado para frente (*forward pass*), fundamentado em suas equações constitutivas. Subsequentemente, efetua-se o cálculo do erro ao confrontar a saída predita com os dados reais observados. Por fim, ocorre a etapa de propagação reversa do erro (*backpropagation through time*), na qual os pesos dos portões e da célula de memória são ajustados através de algoritmos de otimização, como o SGD ou o Adam, visando a minimização da função de perda e o refinamento da capacidade preditiva do modelo.

Esse processo é repetido por um determinado número de iterações até que os valores ótimos dos pesos e vieses sejam alcançados (ABBASIMEHR; SHABANI; YOUSEFI, 2020).

As arquiteturas LSTM apresentam vantagens significativas em relação às RNNs tradicionais, destacando-se, primordialmente, pela sua robusta capacidade de modelar padrões complexos intrínsecos a séries temporais. Esta estrutura demonstra uma eficácia superior na captura de dependências temporais tanto de curto quanto de longo prazo. Além disso, a LSTM possui a competência de modelar relações não lineares entre os parâmetros de entrada, permitindo uma representação mais fiel das nuances da evolução das series. Por fim, ressalta-se a versatilidade destes modelos, que permitem a integração com técnicas complementares, como a decomposição de modo empírico (EMD), visando o aprimoramento contínuo da precisão das previsões geradas.

Ao construir um modelo baseado em LSTM, a otimização criteriosa de seus hiperparâmetros torna-se fundamental para garantir o desempenho e a capacidade de generalização da rede. O número de camadas ocultas desempenha um papel central nesse processo, pois afeta diretamente a complexidade estrutural e a capacidade de aprendizado do modelo frente aos padrões intrínsecos das séries temporais de produtividade acadêmica.

Para mitigar o risco de *overfitting*, emprega-se a taxa de *dropout*, técnica que consiste na eliminação aleatória de conexões neurais durante o treinamento, forçando a rede a aprender representações mais robustas. Adicionalmente, a taxa de aprendizado (*learning rate*) atua como um controlador da velocidade de atualização dos pesos, sendo um fator determinante para a convergência estável do algoritmo de otimização. Por fim, o tamanho do lote (*batch size*) define a quantidade de instâncias processadas simultaneamente a cada iteração, influenciando tanto a estabilidade da estimativa do gradiente quanto a eficiência do processamento paralelo em GPU.

Além disso, a escolha do algoritmo de otimização e da função de perda é crucial para o desempenho do modelo. Métricas como RMSE e SMAPE são comumente utilizadas para avaliar a precisão das previsões (ABBASIMEHR; SHABANI; YOUSEFI, 2020).

2.1.7 Métricas de Avaliação dos Modelos

2.1.7.1 Mean Absolute Error (MAE)

O Erro Médio Absoluto (MAE), é uma métrica que quantifica a magnitude média dos erros em um conjunto de previsões sem considerar a direção desses desvios (JADON; PATIL; JADON, 2022). Matematicamente, o MAE é definido pela média aritmética das diferenças absolutas entre os valores reais observados e os valores previstos pelo modelo, conforme expresso na Equação 2.11:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|, \tag{2.11}$$

onde y_i é o valor real (esperado) e $\hat{y}_i$ o valor previsto para a amostra i. Valores baixos significam que, em média, a distância entre o que o modelo previu e o valor real é pequena. Se o MAE é $1,5$, o modelo erra a produtividade futura em aproximadamente 1 ou 2 pontos para mais ou para menos.

2.1.7.2 Root Mean Square Error (RMSE)

A Raiz do Erro Quadrático Médio (RMSE), é uma métrica calculada como a raiz quadrada da média dos erros elevados ao quadrado. Por elevar as diferenças ao quadrado antes de extrair a média, o RMSE penaliza de forma mais severa os erros de maior magnitude (outliers), tornando-se um indicador valioso para avaliar a dispersão dos erros e a confiabilidade das previsões. Formalmente, o RMSE é definido pela Equação 2.12:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}. \tag{2.12}$$

Nesta formulação, y_i representa o valor real observado, $\hat{y}_i$ corresponde ao valor previsto pelo modelo e N denota o número total de observações no conjunto de teste.

Valores altos indicam que modelo não consegue lidar com crescimento explosivo ou series que mudaram bruscamente de comportamento, gerando erros pontuais muito altos que "puxam" a média para cima.

Uma das principais vantagens práticas desta métrica é que a aplicação da raiz quadrada reverte a penalização quadrática dos erros de volta à escala original dos dados. Como consequência, o RMSE expressa a magnitude do erro nas mesmas unidades da variável observada, o que o torna uma métrica altamente interpretável (CHEN, 2026; HODSON, 2022).

2.1.7.3 Symmetric Mean Absolute Percentage Error (SMAPE)

O Erro Percentual Absoluto Médio Simétrico (SMAPE) mede o erro percentual médio, sendo ajustado para ser simétrico em relação à escala dos dados. No nosso estudo, essa característica o torna ideal para comparar o desempenho dos modelos entre diferentes perfis de pesquisadores, permitindo uma avaliação justa tanto para autores com baixo índice h quanto para pesquisadores seniores com métricas elevadas. O cálculo do SMAPE é expresso pela Equação 2.13:

$$SMAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|y_t - \hat{y}_t|}{(|y_t| + |\hat{y}_t|)/2} \times 100\%. \tag{2.13}$$

Nesta formulação, y_t representa o valor real, $\hat{y}_t$ o valor previsto e n o número total de observações.

Vale destacar que a qualidade da previsão é mantida independentemente do tamanho total da serie. Valores altos sugerem que o modelo está falhando em lidar com a proporcionalidade. Geralmente, SMAPE alto indica que o modelo está indo muito mal com os series de valor baixo (onde qualquer erro pequeno de 1 ou 2 pontos representa uma porcentagem enorme do total).

2.2 Validação Estatística

2.2.1 Testes de Friedman

O teste de Friedman é um procedimento estatístico não paramétrico amplamente empregado em estudos de aprendizado de máquina para comparar o desempenho de múltiplos algoritmos, geralmente três ou mais, avaliados em diversos conjuntos de dados. Considerado o equivalente não paramétrico da ANOVA de medidas repetidas (*repeated measures ANOVA*), este teste opera fundamentalmente através de um mecanismo de rankeamento. Ao contrário de abordagens que utilizam valores brutos de desempenho, o Friedman utiliza a posição relativa de cada modelo, o algoritmo com melhor desempenho em um conjunto de dados específico recebe o *rank* 1, o segundo melhor recebe o *rank* 2, e assim sucessivamente . Em situações de empate, atribui-se um posto médio aos competidores envolvidos (CALVO; SANTAFé, 2015; DEMŠAR, 2006).

A hipótese nula (H_0) do teste de Friedman postula que todos os algoritmos avaliados são estatisticamente equivalentes. Sob essa premissa, seus *ranks* médios (R_j), calculados pela média dos postos obtidos em todos os N conjuntos de dados, deveriam ser similares. A estatística original do teste (χ^2_F) quantifica a variância desses *ranks* e segue uma distribuição qui-quadrado com $k-1$ graus de liberdade, onde k representa o número de modelos testados.

Por ser um teste global, a rejeição da hipótese nula indica apenas a existência de uma diferença significativa em pelo menos um dos modelos, sem especificar onde ela reside. Con-

sequentemente, é obrigatória a realização de testes *post hoc* para identificar as disparidades específicas. O Teste de Nemenyi é a escolha padrão quando o objetivo é a comparação par a par entre todos os algoritmos da amostra. Em contrapartida, procedimentos como Bonferroni Dunn, Holm ou Hochberg são indicados para comparar múltiplos modelos contra um único algoritmo de controle.

A principal vantagem do teste de Friedman reside em sua robustez e segurança, visto que, por ser não paramétrico, ele não exige a normalidade dos dados nem a homogeneidade da variância condições raramente encontradas em experimentos de aprendizado de máquina. Além disso, ao basear-se em *ranks*, o teste resolve o problema da incomensurabilidade e evita que conclusões sejam distorcidas por valores atípicos (*outliers*) que poderiam inflar as médias de erro (CALVO; SANTAFé, 2015; DEMŠAR, 2006).

2.2.2 Testes de Nemenyi

O teste de Nemenyi caracteriza-se como um procedimento estatístico não paramétrico de comparações múltiplas, atuando obrigatoriamente como um teste *post hoc* após a detecção de significância global pelo teste de Friedman. Considerado o equivalente não paramétrico do teste de Tukey (comumente associado à ANOVA), ele é especificamente projetado para situações que exigem comparações par a par entre todos os algoritmos avaliados em múltiplos conjuntos de dados. O funcionamento do teste baseia-se na estipulação de um limite matemático denominado Diferença Crítica (CD — *Critical Difference*), que serve como critério de distinção estatística (CALVO; SANTAFé, 2015; DEMŠAR, 2006).

O desempenho de dois classificadores ou modelos é considerado significativamente diferente se a disparidade entre seus ranks médios for igual ou superior ao valor de CD. Este limiar é calculado por meio da seguinte formulação matemática:

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}}. \tag{2.14}$$

Nesta equação, k representa o número total de algoritmos em comparação, N denota o número de instâncias ou conjuntos de dados avaliados e q_α consiste no valor crítico derivado da estatística de amplitude estudentizada para um determinado nível de confiança α. Se a diferença absoluta entre os postos médios for menor que o CD, não há evidência estatística suficiente, com 95% de confiança, para afirmar que um modelo é superior ao outro.

Uma das principais vantagens do teste de Nemenyi reside na síntese visual de seus resultados através do Diagrama de Diferença Crítica (CD *plot*). Neste diagrama, um eixo horizontal representa a escala dos ranks médios, no qual os modelos são posicionados de acordo com seu desempenho relativo, valores menores à esquerda indicam melhor desempenho. Algoritmos cujas diferenças não atingem o limiar de CD são agrupados e conectados por linhas horizontais, formando “cliques” de equivalência estatística.

Apesar de sua ampla utilização e clareza visual, a literatura ressalta que o teste de Nemenyi é consideravelmente conservador e possui baixo poder estatístico. Essa fragilidade decorre do ajuste dos valores críticos para comportar um elevado número de comparações simultâneas ($k(k-1)/2$), o que torna a rejeição da hipótese nula mais difícil. Devido a esse conservadorismo, em cenários onde o objetivo é comparar todos os modelos contra um único algoritmo de controle (baseline), procedimentos como Bonferroni-Dunn ou Holm são preferíveis, pois realizam menos comparações e oferecem maior sensibilidade estatística (CALVO; SANTAFé, 2015; DEMŠAR, 2006).

2.2.3 Análise de Performance Ordinal (*Ranking*)

Para avaliar a robustez dos modelos, calculou-se o *Ranking* Médio em função do horizonte de previsão (h) e ao tamanho de contexto (c). Esta métrica estima a posição esperada de um modelo se ele fosse escolhido aleatoriamente para uma tarefa de previsão futura.

A Análise de Performance Ordinal atua como uma técnica de normalização não-paramétrica, atribuindo pesos idênticos a cada instância de teste, independentemente da escala do valor alvo. O posto $r_{m,i}$ do modelo m na instância i é definido pela sua posição na sequência ordenada dos erros absolutos:

$$r_{m,i} = \text{Posição de } m \text{ na sequência } \{|e_{m_1,i}|, \ldots, |e_{m_k,i}|\}_{\text{asc}} \tag{2.15}$$

A partir dessa distribuição, deriva-se a Taxa de Vitória (*Win Rate*), correspondente à probabilidade empírica $P(r_m = 1)$, que identifica a frequência com que um modelo é estritamente superior aos demais.

2.2.4 Visualização de Distribuição via Boxplot

O *boxplot*, originalmente batizado por John W. Tukey como *schematic plot* ou *box-and-whisker plot*, constitui uma ferramenta gráfica fundamental da Análise Exploratória de Dados. Sua finalidade primordial é resumir distribuições de dados e, simultaneamente, sinalizar possíveis valores atípicos (*outliers*) para investigação adicional. O núcleo do gráfico é composto por uma caixa retangular cujas extremidades, denominadas *hinges* (dobradiças), funcionam como aproximações para os quartis dos dados: a base situa-se no limite inferior (H_L ou $Q1$, representando 25% dos dados) e o topo no limite superior (H_U ou $Q3$, representando 75% dos dados). No interior dessa estrutura, uma linha horizontal demarca o valor exato da mediana ($Q2$) do conjunto de dados.A amplitude vertical da caixa, calculada pela diferença entre o limite superior e o inferior ($H_U - H_L$), é definida como *H-spread* ou Intervalo Interquartil (IQR). Essa medida é matematicamente essencial para a detecção de anomalias, servindo de base para a construção de "cercas"(*fences*) invisíveis que delimitam a normalidade dos dados. O método projeta um múltiplo (k) do IQR além das extremidades da caixa para identificar desvios, as

cercas internas utilizam $k = 1,5$, classificando valores excedentes como *outside* (externos), enquanto as cercas externas adotam $k = 3,0$ para rotular valores *far out* (extremos). As hastes, ou *whiskers*, estendem-se verticalmente a partir da caixa até o último ponto contido dentro do limite das cercas internas. Qualquer observação que rompa essa barreira não é conectada pelas hastes, sendo plotada individualmente, geralmente por pequenos círculos, para evidenciar visualmente quais pontos destoam do grupo principal. No âmbito desta pesquisa, o *boxplot* permite visualizar a distribuição completa dos resíduos, servindo como métrica para avaliar a estabilidade e a confiabilidade das arquiteturas preditivas. Uma caixa mais achatada indica um modelo mais consistente, ao passo que a presença de uma média aritmética superior à mediana revela uma assimetria positiva (*skewness*), indicando que o modelo sofre de erros esporádicos de alta magnitude.

2.3 Trabalhos Relacionados

A previsão do *h-index* é um campo de pesquisa ativo, com diversas abordagens propostas para modelar a evolução dessa métrica e identificar os fatores que influenciam o seu crescimento. A seguir, apresenta-se uma análise detalhada das principais abordagens, destacando estudos que contribuem para o desenvolvimento de modelos preditivos mais robustos.

Acuna, Allesina e Kording (2012) propuseram um modelo para prever o futuro *h-index* de cientistas na área de ciências da vida, com foco em neurocientistas, pesquisadores de Drosophila e cientistas evolutivos. O objetivo do estudo era estimar o impacto futuro de pesquisadores com base em informações extraídas de seus currículos, utilizando técnicas de aprendizado de máquina. A previsão do *h-index* é uma tarefa desafiadora, pois essa métrica é retrospectiva, refletindo apenas o desempenho passado, mas seu valor preditivo pode ser crucial para decisões de contratação, promoção e financiamento.

Os autores coletaram dados de cientistas a partir do site David e Hayden (2005) e do banco de dados Scopus, considerando pesquisadores com *h-index* maior que 4 e publicações após 1995. As informações analisadas incluíram o número de publicações, o *h-index* atual, a quantidade de artigos publicados em revistas de alto impacto, a diversidade de periódicos em que o cientista publicou e o tempo desde a primeira publicação. Com esses dados, os autores desenvolveram um modelo de regressão linear com regularização elastic net para prever o *h-index* futuro em diferentes horizontes temporais (1 ano, 5 anos e 10 anos), possuindo uma equação para cada horizonte de evento.

De acordo Acuna, Allesina e Kording (2012), as equações aproximadas abaixo são para prever o *h-index* futuro de neurocientistas. Embora sejam razoavelmente precisas para cientistas da área de biologia e saúde, tendem a ser menos significativas para outras áreas da ciência:

Predição para o próximo ano ($R^2 = 0,92$):

$$h_{+1} = 0,76 + 0,37\sqrt{n} + 0,97h - 0,07y + 0,02j + 0,03q. \tag{2.16}$$

Predição para os próximos 5 anos ($R^2 = 0,67$):

$$h_{+5} = 4 + 1,58\sqrt{n} + 0,86h - 0,35y + 0,06j + 0,2q. \quad (2.17)$$

Predição para os próximos 10 anos ($R^2 = 0,48$):

$$h_{+10} = 8,73 + 1,33\sqrt{n} + 0,48h - 0,41y + 0,52j + 0,82q. \quad (2.18)$$

Para as Equações supramencionadas, n é o número de artigos escritos, h representa o *h-index* atual, y corresponde aos anos decorridos desde a publicação do primeiro artigo e j indica o número de periódicos distintos nos quais o autor publicou. Adicionalmente, a variável q refere-se ao número de artigos publicados em veículos de alto impacto, abrangendo periódicos de prestígio como *Nature*, *Science*, *Nature Neuroscience*, *Proceedings of the National Academy of Sciences* e *Neuron*.

Os resultados demonstraram um desempenho significativo do modelo, especialmente para neurocientistas, alcançando um coeficiente de determinação (R^2) de 0.67 para previsões de 5 anos e 0.52 para previsões de 10 anos. Um modelo simplificado, baseado apenas em um subconjunto das variáveis, apresentou um desempenho similar ao modelo completo (R^2 = 0.66). Para pesquisadores de Drosophila e cientistas evolucionários, as previsões foram menos precisas (R^2 = 0.54 e R^2 = 0.61, respectivamente), mas ainda assim superiores a previsões baseadas apenas no *h-index* atual.

Entretanto, o estudo apresenta algumas limitações. A generalização dos resultados para outras áreas do conhecimento não foi explorada, e a relação causal entre as variáveis analisadas e o sucesso acadêmico não foi estabelecida. Além disso, questões como viés de gênero e desigualdade racial não foram aprofundadas, apesar de sua relevância no contexto acadêmico.

Os autores sugerem que estudos futuros possam expandir a abordagem para outras disciplinas, incluir novas variáveis preditivas, como colaborações internacionais e impacto social das pesquisas, e investigar a influência de viés estruturais no desenvolvimento da carreira científica (ACUNA; ALLESINA; KORDING, 2012).

Em Momeni, Mayr e Dietze (2023), os autores previram o *h-index* de pesquisadores classificados em três fases distintas da carreira: júnior (primeira publicação entre 2005 e 2008), intermediária (primeira publicação entre 2000 e 2004) e sênior (primeira publicação entre 1995 e 1999), considerando tanto previsões de curto quanto de longo prazo. Para isso, os autores categorizaram as características utilizadas na previsão em dois grupos: (i) características baseadas em impacto prévio, como o número de publicações, citações recebidas e o *h-index* atual, e (ii) características não baseadas em impacto prévio, incluindo informações sobre coautoria, papel do autor e o tipo de veículo de publicação.

Para realizar a previsão, o estudo utilizou uma abordagem de aprendizado de máquina, especificamente o modelo XGBoost, devido ao seu desempenho superior em comparação com

outros métodos, como SVR (*Support Vector Regression*), RF (*Random Forest*) e GB (*Gradient Boosting*). O conjunto de dados foi obtido a partir da base Scopus-KB e incluiu informações de mais de 1,8 milhão de autores que iniciaram suas publicações após 1994 e permaneceram ativos até 2008. O *dataset* foi filtrado para incluir apenas pesquisadores com pelo menos cinco anos de carreira, um *h-index* maior que zero e uma taxa de publicação mínima de um artigo a cada três anos.

Os resultados do estudo indicaram que as características baseadas em impacto prévio são mais eficazes para previsões de curto prazo em todas as fases da carreira. No entanto, em previsões de longo prazo, as características não baseadas em impacto prévio demonstraram maior robustez, especialmente para pesquisadores júnior. O estudo também analisou o impacto de variáveis como gênero e mobilidade acadêmica, concluindo que o gênero teve impacto mínimo na previsão do *h-index*, enquanto a mobilidade internacional apresentou uma correlação positiva moderada, mas limitada devido à baixa proporção de pesquisadores móveis no conjunto de dados.

Apesar dos resultados promissores, algumas limitações foram identificadas. O estudo não considerou artigos de conferência, o que pode introduzir viés em disciplinas onde tais publicações são predominantes. Além disso, a análise de redes de coautoria não foi explorada de forma intensiva, o que poderia melhorar a previsão, especialmente para pesquisadores em estágios iniciais da carreira.

Pesquisas futuras poderiam expandir a análise para incluir redes de coautoria, explorar o conteúdo textual dos artigos e ampliar o *dataset* para incluir diferentes formas de publicação (MOMENI; MAYR; DIETZE, 2023).

Diferentemente de estudos que tentam prever o número absoluto de citações, que segue uma distribuição de lei de potência e é de difícil previsão, em Dong, Johnson e Chawla (2015) os autores reformulam o problema como uma tarefa de classificação binária. O objetivo é determinar se um artigo contribuirá para o aumento do *h-index* do autor principal dentro de um período de cinco anos após sua publicação. Para isso, foram investigados diversos fatores, incluindo a autoridade do autor no tópico da publicação, o local de publicação, a popularidade do tópico e os h-indices dos coautores.

O estudo foi conduzido utilizando dados da plataforma ArnetMiner, abrangendo informações de 1,7 milhão de autores, 2 milhões de artigos e 8 milhões de relações de citação, cobrindo o período de 1960 a 2012. O *dataset* inclui títulos, resumos, autoria, referências, locais de publicação e anos de publicação, permitindo uma análise abrangente dos padrões de impacto acadêmico.

Para a previsão, foram utilizados três modelos principais de classificação: Regressão Logística (LRC), Florestas Aleatórias (RF) e Árvores de Decisão Ensacadas (BAG). Outros modelos, como *Support Vector Machines* (SVM), *Naive Bayes* e redes neurais, também foram testados, mas apresentaram desempenho inferior e foram descartados. O processo de previsão

envolveu a extração de características dos artigos e autores, o treinamento dos modelos de classificação e a avaliação do desempenho utilizando métricas como precisão, recall, F1-score, AUC e acurácia.

Os resultados indicaram que os modelos de classificação tiveram um desempenho significativamente melhor do que uma previsão aleatória, alcançando um F1-score de 0,776 e uma acurácia de 87,5%. Os fatores mais determinantes para a previsão foram a autoridade do autor no tópico e o local de publicação. Curiosamente, a popularidade do tópico e os h-indices dos coautores tiveram pouca influência na previsão do impacto do artigo.

Apesar dos resultados positivos, o estudo apresenta algumas limitações. A análise foi restrita à área da ciência da computação, levantando a necessidade de expansão para outras disciplinas, como física, biologia e matemática, para verificar se os padrões observados são consistentes em diferentes áreas do conhecimento. Além disso, o estudo considerou o *h-index* dos autores como uma métrica estática, não levando em conta sua evolução ao longo do tempo, o que poderia melhorar a precisão das previsões e explorar fatores adicionais, como colaboração internacional e impacto de artigos de revisão.

Os estudos revisados demonstram que a previsão do *h-index* é uma tarefa complexa, mas viável por meio de técnicas de aprendizado de máquina. Abordagens como regressão linear, modelos de classificação e algoritmos de *boosting* têm sido exploradas para prever o impacto acadêmico futuro. Os resultados indicam que variáveis baseadas em impacto prévio são altamente eficazes em previsões de curto prazo, enquanto fatores mais amplos, como colaboração acadêmica e a evolução temporal da produção científica, podem melhorar a previsão no longo prazo.

No entanto, as limitações dos modelos existentes indicam a necessidade de abordagens mais sofisticadas, como modelos baseados em séries temporais e arquiteturas mais robustas, incluindo *Transformers*, TSFMs e LSTMs. Nesta monografia, essas abordagens serão exploradas e comparadas, visando avaliar seu desempenho na previsão do *h-index* (DONG; JOHNSON; CHAWLA, 2015).

3 Materiais e Métodos

Neste Capítulo, serão detalhados os procedimentos metodológicos adotados para a realização da pesquisa, incluindo a descrição dos modelos utilizados, as métricas de avaliação, a base de dados, e os parâmetros dos modelos. O objetivo principal é comparar o desempenho de diferentes abordagens para a previsão do *h-index*, utilizando diversos modelos, como o MOIRAI-MOE, TimesFM, LSTM, e um *Toy model*, que servirá como limite inferior de desempenho.

3.1 Dados e Pré-Processamento

O conjunto de dados utilizado nesta pesquisa foi desenvolvido por Fernandes (2024) e Vasconcelos (2024), e consiste em informações sobre 1472 pesquisadores brasileiros da área de computação, extraídas da plataforma Lattes e do banco de dados OpenAlex[1] (PRIEM; PIWOWAR; ORR, 2022). Trata-se do conjunto de professores credenciados em Programas de Pós-Graduação strictu senso em Computação no Brasil no ano de 2022, considerando apenas os programas acadêmicos. Embora o *dataset* original contenha diversas informações detalhadas sobre cada pesquisador (como nome, afiliação institucional, dados de doutorado, regime de trabalho e etc...), para os propósitos desta pesquisa será utilizado somente o *author_id* (identificador único do pesquisador na OpenAlex).

Utilizando esse identificador, coletamos as obras e suas respectivas citações anuais por meio da API da OpenAlex. Com base nesses dados, calculamos o *h-index* anual de cada pesquisador desde o início de sua carreira até 2024. Esse processo possibilitou a construção de séries temporais completas, que serviram de base para a realização das previsões.

Para garantir uma análise comparativa robusta, foram realizadas previsões com horizontes de 1 a 10 anos e janelas de contexto variando de 4 a 12 anos. Utilizou-se a técnica de janela deslizante (*sliding window*) para todos os modelos, considerando todas as combinações possíveis de horizonte, contexto e ano de referência, respeitando o tempo de carreira de cada pesquisador. Por exemplo, em uma configuração com contexto de 10 anos e horizonte de 4 anos para o ano de referência 2020, o modelo recebeu o *h-index* anual de 2010 a 2020 com o objetivo de prever o valor de 2024.

Dado que o *h-index* é uma métrica não estacionária, ou seja, tende a crescer ao longo do tempo, foi aplicada uma transformação nos dados para reduzir a autocorrelação. Substitui-se o valor do *h-index* de um ano *h(t)* por *h(t)-h(t-1)*, ou seja, a diferença entre o *h-index* do ano atual e o ano anterior, resultando em uma série temporal que não é monotonicamente crescente, visando melhorar a capacidade dos modelos de capturar padrões temporais mais sutis. Após a predição,

[1] <https://docs.openalex.org/how-to-use-the-api/api-overview>

os valores são reintegrados para a obtenção do *h-index* acumulado final para fins de validação.

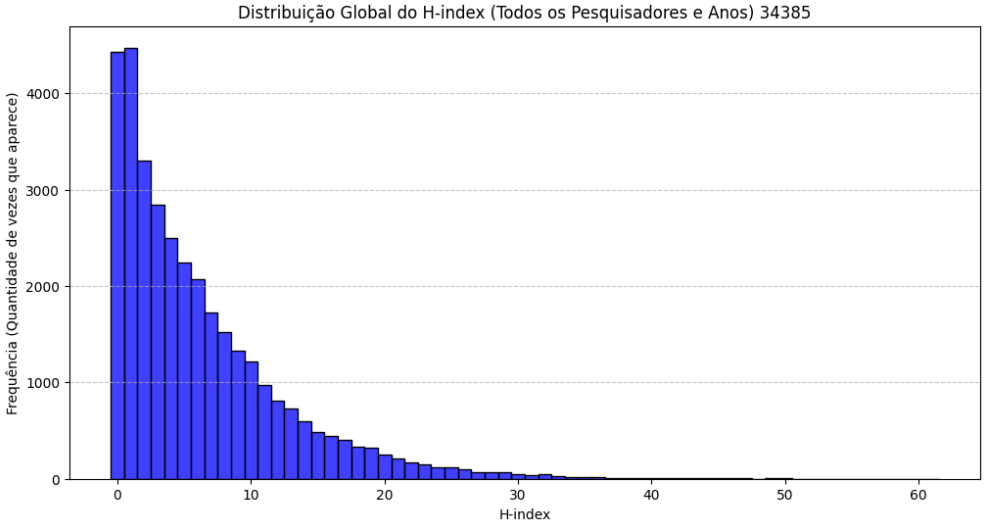
A fim de assegurar a isonomia na comparação entre os modelos, aplicou-se um critério de interseção estrita (conjunto de teste com 20% do conjunto total), no qual foram consideradas apenas as instâncias de teste definidas pela tripla (Autor, Ano de Corte, Horizonte de Previsão) para as quais todos os modelos avaliados geraram previsões com sucesso. Esse procedimento eliminou potenciais vieses de seleção que poderiam favorecer modelos com falhas em casos complexos, resultando em um conjunto final de $N = 270.156$ pontos de previsão individuais e 5.143 casos independentes (pares Autor-Ano de Corte). Consequentemente, este recorte delimitou a análise a um subconjunto de 295 pesquisadores que únicamente são utilizados como conjunto de teste, extraídos do universo inicial de 1.472 autores, garantindo que as comparações de desempenho sejam fundamentadas exatamente sobre o mesmo conjunto de dados. Ressalta-se que o objetivo principal desta filtragem foi garantir a equidade experimental, uma vez que o modelo baseado na arquitetura LSTM, por ser supervisionado, exige a separação de um conjunto de teste independente. Embora os modelos MOIRAI-MoE e TimesFM pudessem ser aplicados a toda a base de dados em regime *zero-shot*, a adoção deste subconjunto comum é indispensável para uma validação estatística rigorosa e comparável entre todas as abordagens.

3.1.1 Caracterização do Dataset

3.1.1.1 Distribuição Global do H-Index

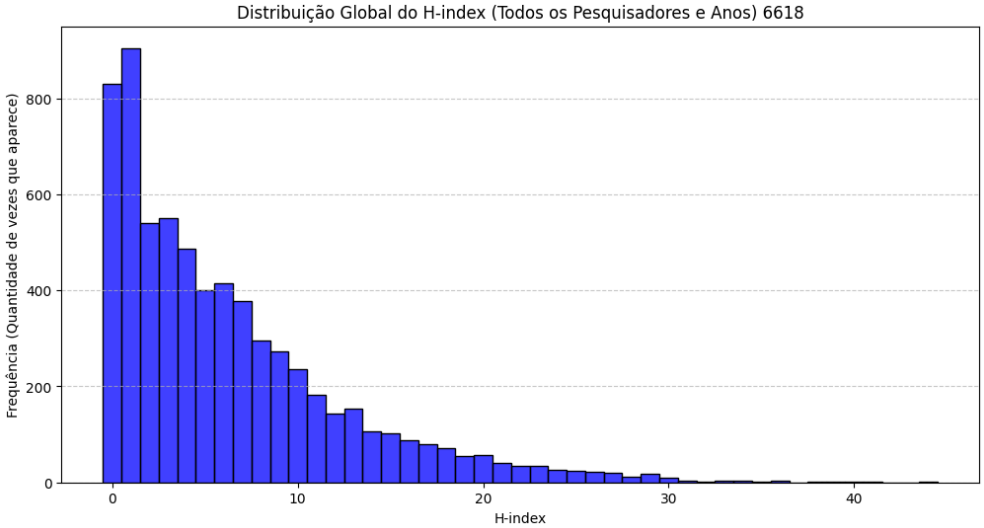
A Figura 3.1 apresenta o histograma da distribuição global dos valores de *h-index* presentes no conjunto de dados geral. Esta visualização agrega todos os registros anuais de 1.472 pesquisadores, totalizando 34.385 pontos de dados, enquanto a Figura 3.2 apresenta o histograma da distribuição global dos valores de *h-index* presentes no conjunto de teste. Esta visualização agrega todos os registros anuais dos 295 pesquisadores do *dataset*, totalizando 6.618 pontos de dados. O objetivo desta análise é compreender a dispersão e a tendência central da variável alvo (*h-index*) que o modelo proposto buscará predizer.

Figura 3.1 – Distribuição global do *h-index*: *dataset* geral.



Fonte: Autoria própria.

Figura 3.2 – Distribuição global do *h-index* no conjunto de teste.



Fonte: Autoria própria.

No conjunto de dados geral, o fato de o desvio padrão ($\sigma \approx 6,50$) ser ligeiramente superior à média ($\approx 6,26$) evidencia uma alta variabilidade. Essa característica indica que o *dataset* é heterogêneo e capaz de representar desde pesquisadores iniciantes até seniores, ainda que este último grupo apresente uma frequência menor. Em contrapartida, no conjunto de teste, o desvio padrão ($\sigma \approx 6,297$) é praticamente idêntico à média ($\approx 6,298$), o que reforça a elevada variabilidade relativa dos dados. Em ambos os casos, a mediana manteve-se estável em 4.

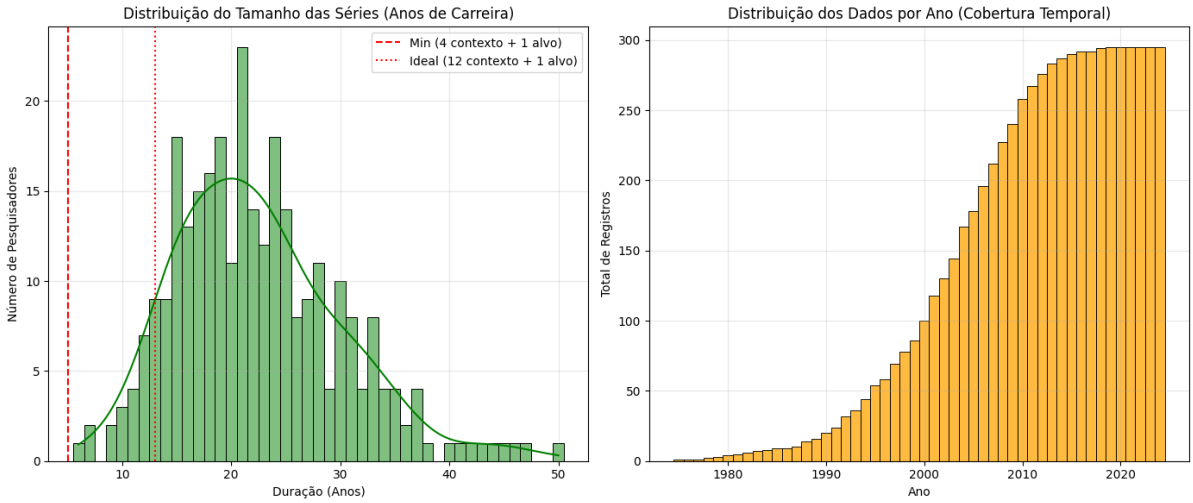
Essa consistência nos indicadores de tendência central (média e mediana) e de dispersão (desvio padrão) sugere que a seleção dos 295 pesquisadores não introduziu um viés de distribuição significativo. Para os modelos de aprendizado de máquina, o desafio computacional persiste:

a necessidade de generalizar padrões a partir de uma base predominantemente composta por valores baixos (intervalo de 0 a 4) para projetar com precisão a evolução de carreiras mais longas e produtivas.

3.1.1.2 Análise da Extensão de Histórico Temporal

A viabilidade de um modelo de predição temporal depende diretamente da extensão do histórico disponível para treinamento. O lado Esquerdo da Figura 3.3 apresenta a distribuição do tamanho das séries temporais (anos de carreira) para os 295 pesquisadores selecionados. A análise estatística desta distribuição revela um *dataset* composto majoritariamente por perfis de carreira consolidados. A média de duração das séries é de $22,43$ anos, com uma mediana muito próxima de 21 anos. A proximidade entre média e mediana sugere uma distribuição equilibrada, sem o domínio excessivo de carreiras extremamente curtas ou longas que poderiam enviesar o treinamento.

Figura 3.3 – Conjunto de teste. a) Frequências dos tamanhos das carreiras dos pesquisadores, em anos. b) Frequência de dados disponíveis por ano



Fonte: Autoria própria.

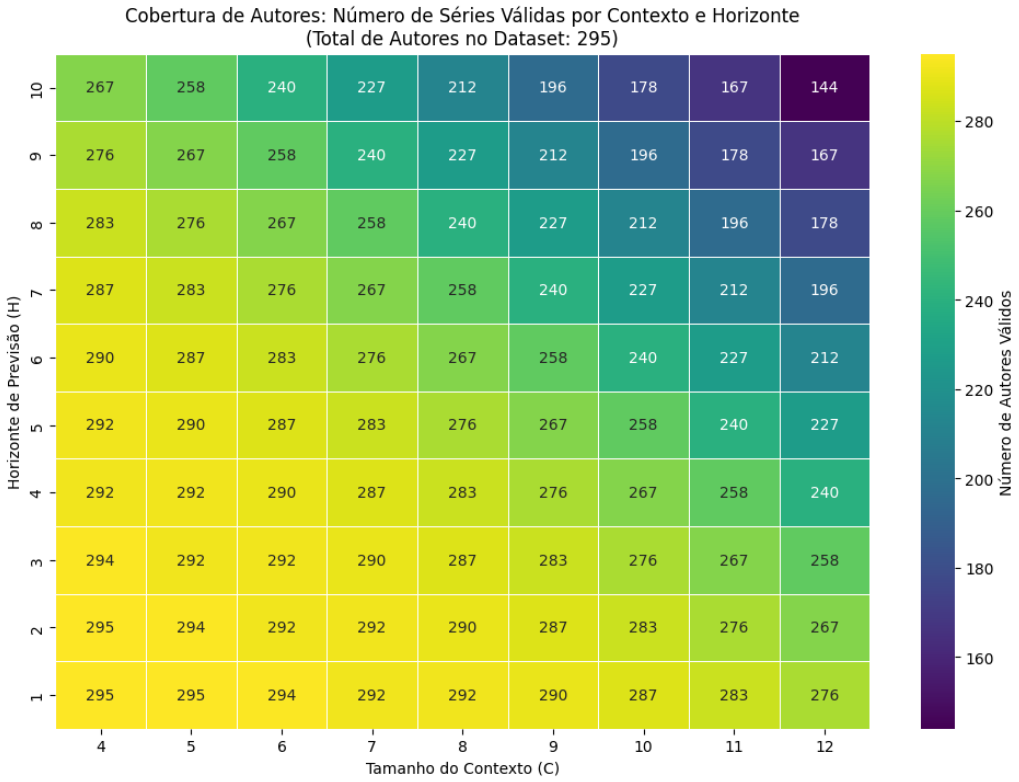
Complementarmente, o lado direito da Figura 3.3 ilustra a densidade de registros ao longo do tempo. Observa-se que o volume de dados acompanha a tendência natural de expansão da produção científica e a digitalização de registros nas últimas décadas. A maior concentração de dados nos anos mais recentes é benéfica para a tarefa de predição, pois garante que o modelo será treinado com padrões de publicação e citação vigentes, capturando a dinâmica atual da comunidade acadêmica até o ano de 2024.

3.1.1.3 Número de Séries Válidas

A definição dos parâmetros de entrada e saída dos modelos preditivos, especificamente o tamanho da janela de contexto (C) e o horizonte de predição (H), impõe restrições diretas sobre a

quantidade de amostras disponíveis para treinamento e teste. A Figura 3.4 ilustra a relação entre a exigência temporal do modelo e a disponibilidade de dados no subconjunto de 295 pesquisadores.

Figura 3.4 – Número de séries válidas no conjunto de teste para cada combinação de tamanho de contexto e horizonte de predição.



Fonte: Autoria própria.

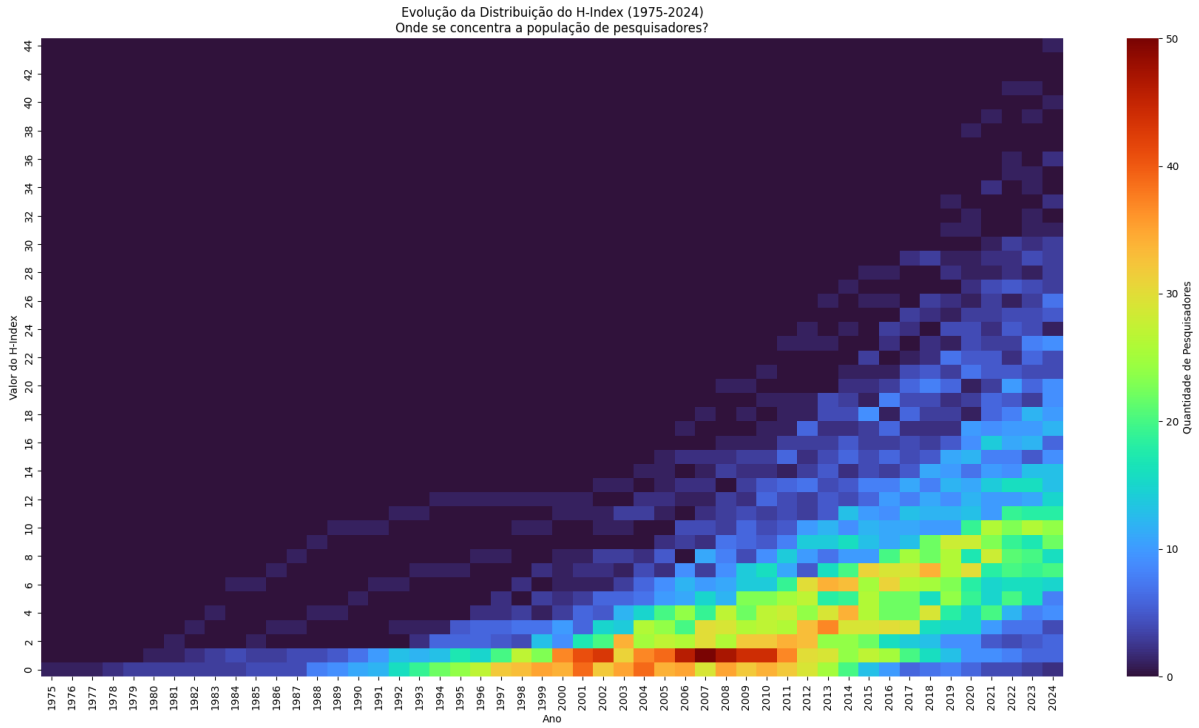
A condição para que um pesquisador seja incluído em um experimento definido pelo par (C, H) é que sua série temporal possua comprimento $L \geq C + H$. A análise do mapa de calor revela um *trade-off* claro entre a complexidade temporal do experimento e a representatividade do *dataset*.

A análise demonstra que experimentos que exigem uma janela total $(C + H)$ superior a 20 anos resultam na perda de aproximadamente 40% a 50% da base de dados. Por exemplo, a configuração extrema $(C = 12, H = 10)$ utiliza apenas 48,8% dos pesquisadores originais.

3.1.1.4 Dinâmica Evolutiva e Concentração da População de Pesquisadores

A Figura 3.5 apresenta a distribuição evolutiva do *h-index* ao longo das cinco décadas cobertas pelo *dataset* (1975–2024). Este mapa de calor permite visualizar a trajetória coletiva dos 295 pesquisadores selecionados, onde o eixo horizontal representa o tempo, o eixo vertical o valor do *h-index*, e a intensidade da cor (ou valor numérico) indica a quantidade de autores naquele estado específico (h, t).

Figura 3.5 – Número de regsitros no conjunto de teste considerando cada ano e *h-index* presente.



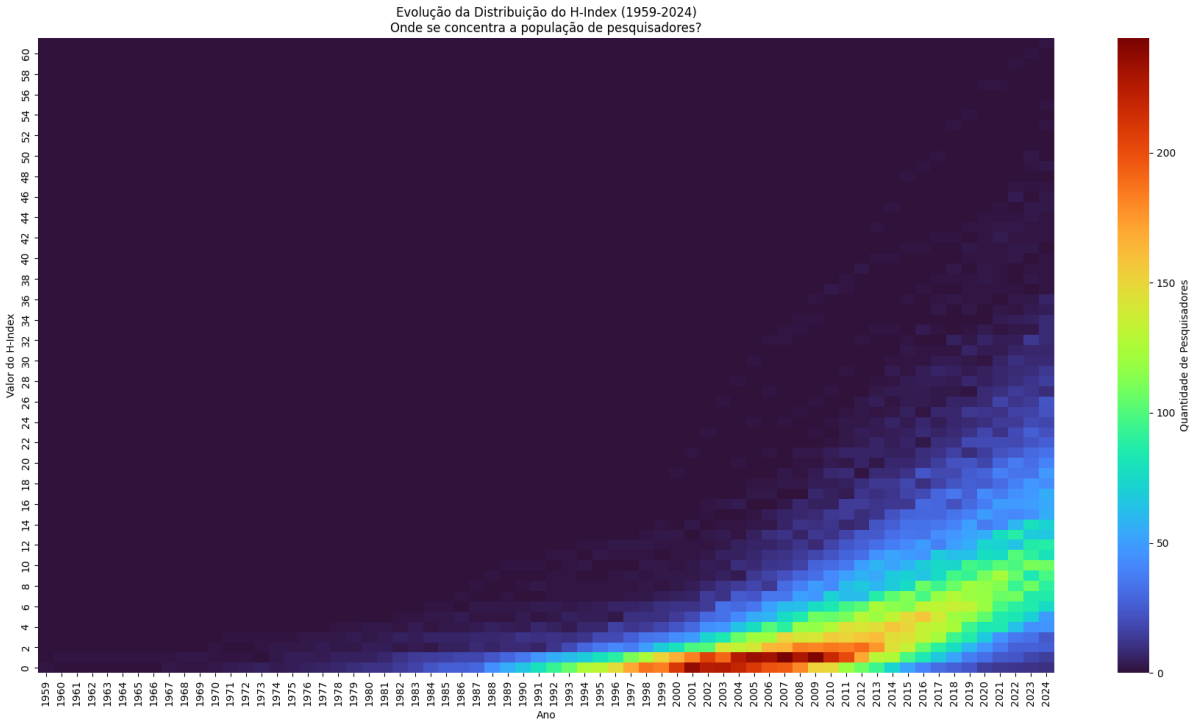
Fonte: Autoria própria.

Na fase inicial (1975-1995), a concentração é densa nos valores $h = 0$ e $h = 1$. Isso caracteriza o início da carreira da maioria dos pesquisadores da amostra. Por exemplo, em 1990, embora existissem pesquisadores com h até 10, a moda da distribuição ainda era $h = 0$ (9 ocorrências) e $h = 1$ (5 ocorrências).

Em se tratando dos últimos anos (2010-2024), nota-se uma inversão na densidade. Em 2024, a frequência de pesquisadores com $h = 0$ é residual (apenas 2 casos), enquanto a distribuição se espalha verticalmente, com concentrações significativas em faixas intermediárias (ex: 24 pesquisadores com $h = 10$ e 22 com $h = 9$).

Com o objetivo de demonstrar que o conjunto de teste é representativo, a Figura 3.6 apresenta a concentração de pesquisadores a partir do *dataset* geral, permitindo uma comparação visual da evolução das carreiras entre os conjuntos de dados. Observa-se a manutenção dos mesmos padrões de distribuição e dispersão do *h-index*, evidenciando que a integridade estrutural dos dados foi preservada, mesmo com o subconjunto representando apenas 20% do volume total original.

Figura 3.6 – Concentração de pesquisadores: *dataset* geral



Fonte: Autoria própria.

Conclui-se, portanto, que o conjunto de teste é uma amostra estratificada natural do universo de pesquisa. A redução do número de séries temporais impactou apenas a densidade absoluta (escala quantitativa), mas não a distribuição de frequências relativas (comportamento qualitativo), legitimando o uso deste subconjunto para o treinamento e validação dos modelos preditivos sem risco de viés de seleção.

3.2 Implementação dos Modelos

Foram avaliadas cinco abordagens distintas, divididas em três categorias: *Baseline*, Especialista e Modelos de Fundação.

3.2.1 Toy Model

Como *baseline* para a avaliação, utilizou-se um modelo de persistência ingênua (*Naive Persistence*). Este modelo assume que o valor futuro da série será igual ao último valor observado no histórico. Embora simples, o *Toy model* é fundamental para quantificar o ganho real de inteligência das abordagens de aprendizado de máquina, a partir da seguinte formulação:

$$f(t + H) = f(t - 1), \tag{3.1}$$

ou seja, o valor para predição no tempo H é simplesmente o valor do último ponto do contexto.

3.2.2 LSTM

Representando a abordagem especialista, treina-se um modelo a partir da arquitetura doravante chamado de modelo LSTM. Diferente dos modelos de fundação, o LSTM foi treinado especificamente com os dados deste domínio. A arquitetura foi escolhida devido à sua capacidade de capturar dependências de longo prazo e mitigar o problema do desvanecimento do gradiente (*vanishing gradient*) comum em RNNs tradicionais.

A implementação adotada baseia-se em uma arquitetura de codificador-decodificador (Sequence-to-Sequence ou Seq2Seq), desenvolvida de forma personalizada utilizando o *framework PyTorch*, com foco em reprodutibilidade e robustez estatística. Para garantir que os experimentos pudessem ser replicados fielmente, fixou-se uma semente aleatória (*seed*) de valor 42 para todas as operações do sistema, incluindo a inicialização de pesos e a divisão dos dados. Diferente dos modelos de fundação, esta rede foi treinada e validada especificamente utilizando 80% do conjunto de dados de pesquisadores brasileiros (conjunto de treinamento), sendo assim, restando 20% do conjunto para teste (com autores que estão unicamente nesse conjunto e nunca antes visto pelo modelo na fase de treino), e uso na comparação entre os modelos.

O modelo processa séries temporais da variação do *h-index* ($\Delta h(t)$), utilizando janelas de contexto que variam de 4 a 12 anos. Devido a essa variação no comprimento das sequências de entrada, aplicou-se a técnica de *Padding* por meio da função `pad_sequence` no `DataLoader` em cada ponto de previsão gerado pela janela deslizante com contexto menor que 12. Essa técnica garante que todas as sequências dentro de um mesmo lote (*batch*) possuam o mesmo comprimento através do preenchimento com zeros no espaço normalizado, permitindo o processamento paralelo eficiente via GPU. O funcionamento do modelo está estruturado em dois componentes principais:

1. **Codificador (Encoder):** Processa a série temporal de entrada de comprimentos variados (janelas de 4 a 12 anos de $\Delta h(t)$). Sua função é comprimir a sequência histórica em vetores de estado oculto (*hidden states*) e de célula (*cell states*), que servem como o contexto resumido da trajetória do pesquisador.

2. **Decodificador (Decoder):** Utiliza os estados finais do codificador como condição inicial para gerar, de forma autoregressiva, a sequência de valores previstos para os horizontes de 1 a 10 anos. Durante o treinamento, emprega-se a técnica de Teacher Forcing com uma taxa de 0,5 para estabilizar a convergência do modelo.

Para a otimização dos pesos, foi utilizado o algoritmo AdamW com decaimento de aprendizado seguindo um cronograma de *Cosine Annealing*. A função de perda minimizada foi o Erro Quadrático Médio (MSE), calculada sobre os valores normalizados através do *StandardScaler* (*z-score*). E para evitar o sobreajuste (*overfitting*) e otimizar o tempo de computação, implementou-se um mecanismo de *Early Stopping*. Este mecanismo monitora a perda no conjunto de validação e interrompe o treinamento caso não haja melhoria superior a um limiar mínimo por

10 épocas consecutivas (patience = 10). As configurações técnicas do modelo estão na Tabela 3.1.

Tabela 3.1 – Configurações técnicas e de treinamento.

Configuração Técnica	**Valor/Algoritmo**
Semente Aleatória (*Seed*)	42
Épocas Máximas	300
Paciência (*Early Stopping*)	10
Horizonte de Previsão	10 anos
Otimizador	AdamW
Função de Perda	*MSE Loss*

Após a realização de um *grid search* para a sintonização fina do modelo, as configurações de hiperparâmetros que apresentaram o melhor desempenho no conjunto de validação e que foram utilizadas para a geração dos resultados finais desta monografia estão detalhadas na Tabela 3.2.

Tabela 3.2 – Hiperparâmetros da rede LSTM Seq2Seq.

Hiperparâmetro	**Valor**
Tamanho do Estado Oculto (*Hidden Size*)	16
Número de Camadas	1
Taxa de Abandono (*Dropout*)	0
Taxa de Aprendizado (*Learning Rate*)	0,001
Weight Decay	1×10^{-4}
Teacher Forcing Ratio	0,5
Tamanho do Lote (*Batch Size*)	64

Abaixo, detalha-se a função e a justificativa para a escolha de cada hiperparâmetro no contexto da predição do *h-index*:

Tamanho do Estado Oculto (*Hidden Size*): Define a dimensionalidade dos vetores de estado oculto e de célula da LSTM. Este valor representa a capacidade de memória interna da rede para codificar os padrões da série temporal no espaço latente.

Número de Camadas: Refere-se ao empilhamento vertical de células LSTM. O uso de 2 camadas permite que o modelo capture uma hierarquia de características, onde as camadas superiores podem aprender representações mais abstratas da evolução da carreira do pesquisador.

Taxa de Abandono (*Dropout*): Técnica de regularização que desativa neurônios aleatoriamente para evitar a dependência excessiva em conexões específicas. O valor 0 indica que a rede manteve a generalização sem a necessidade dessa técnica, dada a dimensionalidade controlada do modelo.

Taxa de Aprendizado (*Learning Rate*): Determina o tamanho do passo dado pelo otimizador na direção do mínimo da função de perda. O valor de 0,001 foi selecionado para garantir um equilíbrio entre a velocidade de convergência e a estabilidade do aprendizado.

Decaimento de Pesos (*Weight Decay*): Aplica uma penalidade L_2 à magnitude dos pesos, incentivando o modelo a manter parâmetros menores e evitando que ele se ajuste excessivamente a ruídos nos dados históricos.

Razão de *Teacher Forcing*: Parâmetro crítico em arquiteturas Seq2Seq, define a probabilidade de utilizar o valor real do passo anterior como entrada para o próximo passo do decodificador. A taxa de 0,5 auxilia a rede a aprender a estrutura da sequência de forma mais eficiente durante as épocas iniciais.

Tamanho do Lote (*Batch Size*): Define a quantidade de janelas temporais processadas antes de uma atualização dos pesos. O valor 64 foi escolhido para otimizar o uso da memória computacional e estabilizar a estimativa do gradiente.

3.2.3 TimesFM

Para a realização dos experimentos de predição do *h-index*, utilizou-se a versão TimesFM 2.5 com 200 milhões de parâmetros, disponibilizada publicamente pelo Google através do repositório *Hugging Face*[2] (Google Research, 2024). A implementação foi conduzida em ambiente *Python*, utilizando a biblioteca *PyTorch* para a execução das inferências em regime *zero-shot*. Neste trabalho, o TimesFM foi aplicado em configuração *zero-shot*, sem *fine-tuning* nos dados proprietários.

A execução foi configurada através da classe `ForecastConfig`, cujos parâmetros técnicos estão resumidos na Tabela 3.3.

Tabela 3.3 – Especificações técnicas e de inferência do TimesFM.

Parâmetro	**Configuração/Valor**
Versão do Modelo	TimesFM 2.5 (200M)
Semente Aleatória (*Seed*)	42
Janela de Contexto Máxima	12 anos
Horizonte de Previsão Máximo	10 anos
Normalização de Entrada	Ativada (`True`)
Restrição de Positividade	Ativada (`True`)

Algumas definições da configuração do modelo:

1. **Tokenização por Patches:** O modelo segmenta a série temporal de entrada em blocos (*patches*), que funcionam como unidades semânticas locais processadas pelo Transformer Decoder-only.

[2] <https://huggingface.co/google/timesfm-2.5-200m-pytorch>

2. **Garantia de Monotonicidade:** Através do parâmetro `infer_is_positive=True`, garantiu-se que as predições de crescimento fossem sempre não-negativas, respeitando a natureza matemática cumulativa do *h-index*.

3. **Normalização de Entrada:** Através do parâmetro `normalize_inputs=True`, o modelo realiza a padronização dos dados de entrada antes do processamento. Esta etapa é fundamental para garantir que as séries temporais de diferentes pesquisadores, que possuem ordens de magnitude distintas no *h-index*, sejam trazidas para uma escala comum, o que estabiliza o aprendizado e melhora a precisão das previsões *zero-shot*.

3.2.4 MOIRAI-MoE

O MOIRAI-MoE é um modelo de fundação baseado em uma arquitetura de *Masked Encoder* com *Mixture of Experts* (MoE). Desenvolvido pela *Salesforce AI Research*. A versão adotada foi a moirai-moe-1.0-R-base, composta por 935 milhões de parâmetros totais e 86 milhões de parâmetros ativos por *token*. O modelo foi acessado através do repositório oficial no *Hugging Face*[3] (Salesforce Research, 2024)

A implementação foi conduzida em linguagem *Python*, utilizando a biblioteca uni2ts[4] (versão 2.0.0) para a manipulação do modelo e a GluonTS[5] (versão 0.14.4) para a estruturação dos conjuntos de dados em formato de lista (*ListDataset*). O MOIRAI-MOE opera sob o paradigma de previsão probabilística, o que exige a geração de múltiplas amostras para cada passo temporal previsto para a posterior agregação dos resultados. Para garantir que os experimentos pudessem ser replicados fielmente, fixou-se uma semente aleatória (*seed*) de valor 42 para a divisão dos dados (treino 80% e teste 20%), e foi estabelecido um batch de 64.

Com o intuito de avaliar os melhores hiperparâmetros para o modelo utilizando o conjunto de treino, realizou-se uma busca em grade (*grid search*) para identificar a combinação ótima de tamanho de *patch* (*patch size*), método de agregação e número de amostras (*num samples*). As configurações que apresentaram o melhor desempenho, fundamentadas no menor Erro Quadrático Médio (MSE), foram aplicadas para a predição no conjunto de teste. Os hiperparâmetros finais adotados para a comparação com os demais modelos encontram-se detalhados na Tabela 3.4.

Tabela 3.4 – Melhores configurações de hiperparâmetros para o MOIRAI-MOE.

Contexto Máximo	Tamanho do Patch	Agregação	Nº de Amostras
12 anos	8	Mediana	100

O fluxo de dados para a inferência seguiu as etapas de pré-processamento padronizadas para garantir a consistência experimental:

[3] <https://huggingface.co/Salesforce/moirai-moe-1.0-R-base>
[4] <https://github.com/SalesforceAIResearch/uni2ts>
[5] <https://github.com/awslabs/gluonts/>

1. **Janela Deslizante:** Foram geradas janelas de contexto variando de 4 até o limite máximo 12 para prever um horizonte de até 10 anos.

2. **Inferência Probabilística:** Para cada janela, o modelo gerou o número configurado de trajetórias possíveis (samples). A previsão final foi extraída através da mediana (percentil 0,5) dessas amostras, garantindo maior robustez contra valores atípicos (*outliers*) gerados pelo modelo.

3. **Restrição de Domínio:** Aplicou-se uma restrição de não-negatividade nos deltas previstos (`np.maximum(pred_diff, 0)`), assegurando que o *h-index* reconstruído fosse monotonicamente não-decrescente.

A classe `MoiraiMoEForecast`, utilizada para a geração das previsões, foi instanciada com parâmetros específicos que definem a estrutura da tarefa de inferência. A Tabela 3.5 resume essas configurações conforme implementado.

Tabela 3.5 – Parâmetros de configuração da inferência MOIRAI-MOE.

Parâmetro	**Configuração/Valor**
`prediction_length`	10 (anos)
`context_length`	12
`patch_size`	8
`num_samples`	100
`target_dim`	1 (Univariado)
`feat_dynamic_real_dim`	0
`past_feat_dynamic_real_dim`	0
`batch_size`	64

Abaixo, detalha-se o significado técnico de cada um desses parâmetros no contexto da predição do *h-index*:

`prediction_length` **(Horizonte de Previsão):** Define o número de passos de tempo futuros que o modelo deve prever a partir do ponto de corte. Neste trabalho, fixou-se o valor em 10 para abranger previsões de curto e longo prazo.

`context_length` **(Janela de Contexto):** Especifica o comprimento máximo da série histórica que o modelo utiliza como entrada para gerar a previsão. Foram testados contextos de 4 a 12 anos.

`patch_size` **(Tamanho do Patch):** Refere-se à técnica de agregação de pontos temporais á um único *token*, para reduzir o custo computacional e melhorar a captura de padrões locais. O valor 8 foi selecionado como ótimo após os testes de *grid search*.

`num_samples` **(Número de Amostras):** Por ser um modelo probabilístico, o MOIRAI-MOE não gera um único valor, mas sim trajetórias possíveis extraídas da distribuição de mistura

aprendida. Este parâmetro define quantas dessas trajetórias são geradas antes da agregação final pela mediana .

`target_dim` **(Dimensão do Alvo):** Define a dimensionalidade da série temporal de saída. O valor 1 indica que o modelo está realizando uma previsão univariada, focada exclusivamente no valor do *h-index* .

`feat_dynamic_real_dim` **(Covariáveis Dinâmicas Futuras):** Indica a quantidade de variáveis externas que variam no tempo e são conhecidas no futuro (como feriados ou eventos planejados). O valor 0 indica que a previsão baseia-se puramente no histórico do *h-index*, sem variáveis exógenas futuras .

`past_feat_dynamic_real_dim` **(Covariáveis Dinâmicas Passadas):** Indica a quantidade de variáveis externas conhecidas apenas para o passado. O valor 0 confirma que o modelo opera de forma univariada clássica, utilizando apenas a série temporal alvo (Δh) para informar a predição .

`batch_size` **(Tamanho do Lote):** Determina a quantidade de janelas de pesquisadores processadas simultaneamente pela GPU durante a inferência. O valor 64 foi selecionado para otimizar o paralelismo computacional e garantir a eficiência no tempo de execução dos testes.

3.3 Métodos de Avaliação

3.3.1 Métricas

Para avaliação dos resultados, utiliza-se três métricas, o erro absoluto médio (MAE), o erro quadrático médio (RMSE) e o erro percentual absoluto médio simétrico (SMAPE).

3.3.2 Validação Estatística

Para garantir a robustez das conclusões e a significância das comparações entre os modelos, foram adotadas diversas técnicas de análise estatística e visualização, detalhadas a seguir.

3.3.2.1 Teste de Friedman

Para este teste, os dados foram agregados por instância independente (média dos erros de todos os horizontes para um mesmo autor), resultando em 5.143 amostras independentes. Esse agrupamento foi necessário para garantir a premissa de independência das amostras e evitar a autocorrelação serial das previsões (pseudoréplica).

Aplicou-se o teste de Friedman sobre os erros absolutos médios (MAE) agregados por instância de teste (par Autor-Ano). O agrupamento foi realizado para garantir a independência

das amostras, evitando a pseudoréplica que ocorreria se utilizássemos cada passo do horizonte de previsão individualmente. Se o p-valor resultante for menor que o nível de significância ($\alpha = 0.05$), rejeita-se a hipótese de que os modelos são iguais.

3.3.2.2 Teste de Nemenyi (Post-hoc)

Após a detecção de significância pelo Friedman, aplicou-se o Nemenyi para comparar cada modelo contra todos os outros (*Toy* vs LSTM, LSTM vs MOIRAI-MoE, etc.). O resultado é apresentado na forma de uma matriz de p-valores, visualizada através de um mapa de calor (heatmap).

3.3.2.3 Avaliação de Dominância Estatística Estratificada

Diferente da avaliação global, que determina o melhor modelo com base no desempenho médio geral, esta análise visa identificar qual arquitetura oferece superioridade estatisticamente comprovada para cada cenário específico de previsão. Esses cenários são definidos pela combinação do Horizonte de Previsão ($h \in \{1, \ldots, 10\}$) e do Tamanho de Contexto ($c \in \{4, \ldots, 12\}$).

Para garantir a robustez das conclusões e atender às premissas de independência das amostras, os dados foram agregados por instância independente (par Autor-Ano), totalizando $N = 5.143$ observações. Este procedimento de agrupamento foi essencial para mitigar o risco de autocorrelação serial das previsões e evitar o fenômeno da pseudoréplica, garantindo que cada ponto de dado no teste de hipótese represente uma trajetória de carreira distinta.

Para cada par (h, c) na grade de resultados, aplicou-se um procedimento de teste de hipótese independente, seguindo as etapas abaixo:

Filtragem e Pareamento: Selecionaram-se as instâncias de teste correspondentes ao horizonte h e contexto c específicos. Os erros absolutos (MAE) de todos os modelos foram organizados em uma estrutura de blocos pareados por instância.

Teste de Significância: Aplicou-se o teste de Friedman seguido pelo teste post-hoc de Nemenyi (quando aplicável) para verificar a existência de diferenças estatisticamente significantes entre os rankings das arquiteturas.

A determinação da dominância estatística fundamenta-se no critério de distinção do p-valor ($\alpha = 0,05$). Caso o modelo com o menor erro médio apresente uma diferença em relação aos demais competidores que exceda o limiar da Diferença Crítica ($p < 0,05$), ele é declarado Vencedor Isolado. Por outro lado, se o desempenho deste modelo não for estatisticamente distinguível de um ou mais rivais ($p > 0,05$), ele é classificado como Vencedor Nominal. Nessa situação de equivalência, o resultado é sinalizado com um asterisco (*), indicando a ocorrência de um empate técnico e demonstrando que a superioridade numérica observada não possui significância estatística suficiente para separá-los.

Os resultados dessa análise foram compilados em um Mapa de Calor de Dominância, no qual cada célula (h, c) é colorida conforme o modelo dominante. Esta visualização permite identificar padrões de comportamento e "zonas de competência", revelando como a variação do histórico disponível e do tempo de projeção influencia a eficácia de cada arquitetura.

3.3.2.4 Análise de Dominância por Cenário (Mapa de Calor de Menor Erro)

Complementando a avaliação global, que fornece uma perspectiva macroscópica do desempenho, esta análise foi conduzida para identificar especializações locais e nichos de superioridade de cada arquitetura. O objectivo principal consistiu em delimitar a "zona de competência"de cada modelo, diferenciando o desempenho em cenários de curto prazo com histórico limitado versus predições de longo prazo com contexto abundante.

A metodologia baseou-se no cálculo estratificado do Erro Absoluto Médio (MAE) para cada modelo m dentro de uma grelha experimental exaustiva, estruturada através da variação sistemática de duas variáveis de controlo: o Horizonte de Previsão (h), que representa o tempo de projecção futura num intervalo de 1 a 10 anos, e o Tamanho de Contexto (c), que compreende a janela de histórico disponível, variando entre 4 e 12 anos.

Para cada combinação única (h, c) presente no conjunto de teste, os erros de predição foram isolados e a média aritmética dos erros absolutos foi computada. O resultado deste processamento foi consolidado numa matriz de desempenho, na qual a cor de cada célula identifica o modelo que obteve o menor MAE nominal para aquele cenário específico. Esta abordagem permite observar visualmente a transição de eficácia entre os *Foundation Models* e o modelo especialista à medida que a complexidade da tarefa de previsão aumenta.

3.3.2.5 Análise de Ranking:

Essa análise permite identificar a degradação relativa de desempenho. Um modelo cuja curva de *ranking* médio apresenta inclinação descendente ao longo do eixo do horizonte indica que ele perde competitividade para seus pares em previsões de longo prazo, enquanto uma curva plana indica estabilidade temporal

Enquanto o *ranking* médio mede a tendência central, a Distribuição de Frequência dos Postos avalia a consistência. Calculou-se a probabilidade empírica de um modelo m ocupar a posição k:

$$P(r_m = k) = \frac{\text{Contagem de vezes que } r_m = k}{\text{Total de Instâncias}}$$

Desta distribuição, deriva-se a métrica de Taxa de Vitória (*Win Rate*), correspondente a $P(r_m = 1)$. Esta métrica responde à pergunta pragmática: “Em qual porcentagem dos casos este modelo foi estritamente superior a todos os concorrentes?”. Esta análise é crucial para distinguir entre modelos que são consistentemente “segundos colocados” (baixo risco, mas raramente os

melhores) e modelos que são "especialistas"(vencem frequentemente em nichos específicos, mas falham gravemente em outros).

3.3.2.6 Diagrama de Diferença Crítica

Para validar formalmente a superioridade de um modelo sobre os demais, não basta comparar as médias de erro, pois estas não consideram a variância e a correlação entre as amostras. Adotou-se, portanto, uma metodologia para comparação de múltiplos preditores em múltiplos conjuntos de dados.

O primeiro passo consistiu na aplicação do teste de Friedman. Rejeitada a hipótese nula pelo Friedman ($p < 0.05$), procedeu-se com o teste de Nemenyi para comparações par-a-par.

Os resultados foram sintetizados visualmente através do Diagrama de Diferença Crítica. O eixo horizontal representa o *Ranking* Médio de cada modelo (valores menores à esquerda indicam melhor desempenho). Uma linha horizontal grossa conectando dois ou mais modelos indica que a diferença entre seus *rankings* é menor que o CD. Em termos estatísticos, esses modelos formam um “clique” de equivalência, ou seja, não há evidência suficiente para afirmar que um é melhor que o outro dentro desse grupo. Modelos isolados (não conectados) são estatisticamente distintos dos demais.

3.3.2.7 Boxplot (Média e Desvio Padrão)

Utilizou-se o *boxplot* para representar graficamente a distribuição dos erros absolutos $|y - \hat{y}|$ gerados por cada modelo no conjunto de dados pareado. Os componentes do gráfico foram definidos conforme o padrão estatístico de Tukey.

Adicionalmente, plotou-se a média aritmética (triângulo) para permitir a comparação direta com a mediana e identificar a assimetria (*skewness*) da distribuição. Se a média é muito superior à mediana, indica que o modelo sofre de erros esporádicos de alta magnitude (cauda longa à direita).

4 Resultados

Este Capítulo apresenta a avaliação experimental das abordagens propostas para a predição do *h-index*. A análise foi conduzida sobre o *conjunto de teste* (detalhado na Seção 3.1), garantindo que todas as comparações fossem realizadas exatamente sobre as mesmas instâncias. Os resultados estão organizados em cinco dimensões: desempenho global, validação estatística, análise de dominância por cenário, sensibilidade ao horizonte de previsão e distribuição de erros.

4.1 Desempenho Global dos Modelos

A Tabela 4.1 sumariza as métricas de erro (MAE, RMSE e SMAPE) agregadas para todos os cenários de teste. O objetivo desta análise macroscópica é estabelecer a hierarquia de desempenho entre as abordagens de *baseline* (*Toy Model*), especialista (LSTM) e *Foundation Models* (TimesFM e MOIRAI-MOE).

Tabela 4.1 – Comparativo global de métricas de erro entre os modelos avaliados.

Modelo	**MAE**	**RMSE**	**SMAPE (%)**
Toy Model	3,2850	4,6058	45,46%
TimesFM	2,0948	3,1323	32,26%
MOIRAI-MOE	1,7211	2,6582	24,88%
LSTM	**1,5194**	**2,3156**	**20,86%**

Fonte: Autoria própria.

Os resultados experimentais revelam que o modelo de persistência ingênua (*Toy Model*), utilizado como *baseline*, estabeleceu o limite inferior de desempenho ao apresentar os maiores erros em todas as métricas analisadas, com um $MAE \approx 3,28$. O elevado valor de *SMAPE* $(45,46\%)$ confirma que a estratégia de assumir a estagnação do *h-index* é ineficaz, dada a natureza cumulativa e a dinâmica de crescimento da métrica. Em contrapartida, o modelo especialista LSTM alcançou a dominância no desempenho global, superando as demais abordagens com um MAE de $1,5194$ e um *SMAPE* de $20,86\%$. Este resultado evidencia a eficácia do treinamento supervisionado específico no domínio, uma vez que a rede LSTM teve acesso direto aos padrões históricos dos pesquisadores brasileiros, permitindo um *fine-tuning* orientado aos perfis de publicação locais.

No que concerne aos *Foundation Models* operando em regime *zero-shot*, o MOIRAI-MOE demonstrou uma superioridade clara sobre o TimesFM, reduzindo o erro absoluto médio em aproximadamente $17,8\%$ ($1,72$ contra $2,09$). É notável que o MOIRAI-MOE, mesmo sem treinamento prévio no conjunto de dados alvo, aproximou-se significativamente do desempenho do especialista LSTM, apresentando uma diferença de apenas $\approx 0,2$ no MAE. Esta competitividade

valida a alta capacidade de generalização da arquitetura *Mixture of Experts*, sugerindo que tais modelos representam alternativas robustas e eficientes para cenários onde o treinamento de redes dedicadas é computacionalmente inviável ou limitado pela escassez de dados.

Observa-se também que, para todos os modelos, o RMSE é consistentemente maior que o MAE, penalizando erros maiores. A diferença entre RMSE e MAE é mais acentuada no *Toy Model* (4.60 vs 3.28) e menor no LSTM (2.31 vs 1.51), indicando que o modelo especialista não apenas acerta mais na média, mas também comete menos erros grosseiros (outliers) do que as abordagens generalistas.

4.2 Validação Estatística

Para mitigar o risco de conclusões baseadas apenas em médias aritméticas, que podem ser influenciadas por *outliers*, aplicou-se o teste não-paramétrico de Friedman para avaliar a significância estatística das diferenças de desempenho entre os modelos. O teste foi conduzido considerando o *ranking* de erro de cada modelo para cada uma das $N = 5.143$ amostras independentes (agrupamento por autor).

O teste de Friedman resultou em uma estatística de $\chi^2_F = 4817.89$ e um p-valor de 0.00, rejeitando fortemente a hipótese nula com nível de confiança de 95% ($\alpha = 0.05$). Isso confirma que existe uma diferença estatisticamente significativa no desempenho entre pelo menos dois dos modelos avaliados.

A Tabela 4.2 apresenta o *Ranking* Médio (*Mean Rank*) obtido por cada modelo, onde valores menores indicam melhor desempenho (posicionamento consistente nas primeiras colocações).

Tabela 4.2 – *Ranking* médio dos modelos (Teste de Friedman)

Modelo	***Ranking* Médio (Menor é Melhor)**
LSTM (Especialista)	**1.416**
MOIRAI-MoE	1.610
TimesFM	1.949
Toy Model	2.985

Fonte: Autoria própria.

A análise dos *rankings* corrobora os resultados das métricas globais.

A análise dos resultados demonstra que o modelo LSTM se consolidou como a abordagem estatisticamente superior, atingindo um *ranking* médio de 1,41. Tal índice indica que, na maioria dos cenários de teste, o modelo obteve o menor erro ou apresentou um desempenho extremamente próximo do ideal.

Posicionando-se em um segundo patamar de performance, o modelo MOIRAI-MOE registrou um *ranking* médio de 1,61. Em contrapartida, as abordagens TimesFM e *Toy Model* apresentaram os índices menos favoráveis, com *rankings* de 1,94 e 2,98, respectivamente, distanciando-se de forma significativa dos líderes da comparação.

4.2.1 Diagrama de Diferença Crítica (CD Diagram)

Para identificar onde exatamente residem as diferenças significativas, aplicou-se o teste pós-hoc de Nemenyi. A Figura 4.1 apresenta o Diagrama de Diferença Crítica. A barra horizontal superior conecta modelos que não apresentam diferença estatística significativa entre si (equivalência estatística), enquanto modelos não conectados são considerados distintos.

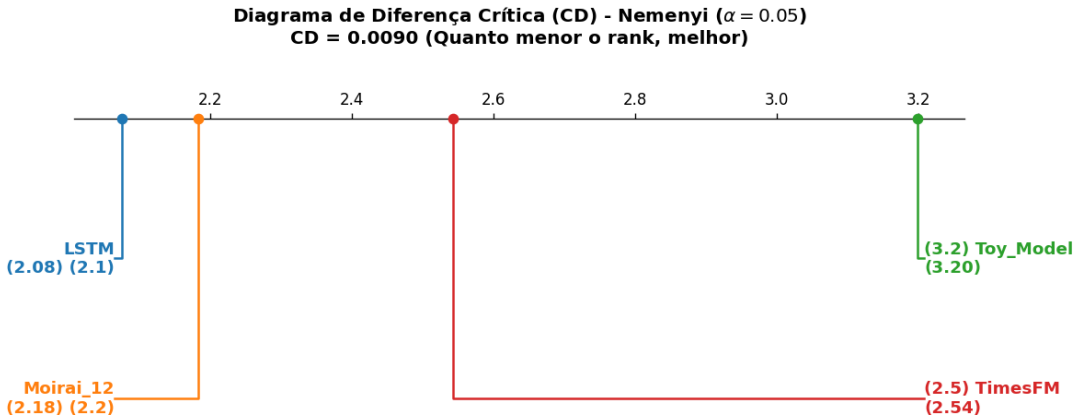


Figura 4.1 – Diagrama de diferença crítica (Teste de Nemenyi).

Fonte: Autoria própria.

Considerando o Limiar de Diferença Crítica ($CD \approx 0.009$, calculado sobre o total de casos pareados), a análise do diagrama e da matriz de p-valores permite as seguintes inferências:

A análise do diagrama de diferença crítica revela a dominância isolada da arquitetura LSTM, a qual não compartilha barra de ligação com nenhum outro modelo avaliado, o que evidencia uma distinção estatística clara de desempenho. Note-se que a distância entre o LSTM e o segundo colocado, o MOIRAI-MoE, é superior ao valor crítico ($|2,08 - 2,18| > CD$), confirmando que sua superioridade não decorre do acaso. Finalmente, o TimesFM posiciona-se em um patamar estatístico isolado e inferior aos modelos MOIRAI-MoE, reforçando que a arquitetura *Mixture of Experts* demonstrou maior competência na generalização *Zero-Shot* para este domínio específico do que a abordagem adotada pelo TimesFM.

4.3 Análise de Dominância por Cenário

Enquanto as métricas globais fornecem uma visão macroscópica, a Análise de Dominância permite identificar a "zona de competência" de cada modelo. Esta análise foi estratificada por

duas variáveis de controle: o Tamanho do Contexto Histórico ($C \in [4, 12]$) e o Horizonte de Previsão ($H \in [1, 10]$).

4.3.1 Dominância Estatística

A Figura 4.2 apresenta o Mapa de Calor de Dominância Estatística. Cada célula (C, H) indica o modelo vencedor após a aplicação do teste de Friedman seguido de Nemenyi ($\alpha = 0.05$) para aquele cenário específico. Células marcadas com asterisco (*) ou indicando múltiplos modelos representam Empate Técnico, onde a diferença de desempenho entre o primeiro colocado e os demais listados não é estatisticamente significativa.

Figura 4.2 – Mapa de calor de dominância estatística por contexto e horizonte.

Fonte: Autoria própria.

A análise da matriz de dominância revela dois regimes distintos de predição, categorizados pela extensão do horizonte temporal. Primeiramente, observa-se a hegemonia do modelo especialista LSTM em horizontes de médio e longo prazo ($H \geq 2$), nos quais a abordagem obteve vitória isolada na vasta maioria dos cenários, independentemente do comprimento do contexto. Esse padrão sugere que, para projetar a trajetória de carreira em prazos mais extensos,

o *fine-tuning* aos dados específicos realizado pelo LSTM é crucial para capturar a inércia e as tendências não-lineares do *h-index* que os modelos generalistas tendem a suavizar.

Em contrapartida, nota-se uma competitividade significativa dos *Foundation Models* em regime *zero-shot* no curto prazo ($H = 1$), regime este caracterizado pela predominância de empates técnicos. Em contextos intermediários, como $C = 6, 7$ e 8, o modelo MOIRAI-MOE chega a figurar nominalmente como vencedor estatístico, o que valida a hipótese de que, para predições imediatas, a capacidade de transferência de aprendizado desses modelos é tão eficaz quanto o treinamento dedicado.

4.3.2 Minimização do Erro Absoluto (MAE)

Para complementar a visão estatística, a Figura 4.3 apresenta o modelo que obteve o menor Erro Absoluto Médio (MAE) nominal em cada cenário, juntamente com a magnitude do erro.

Figura 4.3 – Matriz de menor erro absoluto (MAE).

Fonte: Autoria própria.

A análise quantitativa dos erros confirma e refina as observações estatísticas, além de corroborar com a hipótese levantada na Introdução. O MOIRAI-MoE registrou os menores erros absolutos de todo o experimento nos cenários de curto prazo e contexto médio, especificamente para $H = 1$ e contextos de 6, 7 e 8 anos, o MOIRAI-MoE atingiu valores de MAE situados entre $0,534$ e $0,539$, superando marginalmente o modelo LSTM. Tal desempenho é particularmente notável para uma abordagem *zero-shot*, pois indica uma alta precisão na detecção de tendências imediatas. Paralelamente, observa-se uma clara progressão na dificuldade da tarefa conforme o horizonte de previsão se estende, com o erro do melhor modelo, geralmente a rede LSTM, saltando de aproximadamente $0,53$ no primeiro horizonte para cerca de $3,01$ no horizonte $H = 10$. Surge ainda uma observação contra-intuitiva ao analisar o comportamento do LSTM no longo prazo: o modelo obteve um erro menor utilizando um contexto curto ($C = 4$, $MAE \approx 2,79$) do que

com o contexto máximo ($C = 12$, $MAE \approx 3,01$). Esse fenômeno sugere que, para a arquitetura recorrente treinada neste conjunto de dados, históricos excessivamente longos podem introduzir ruído ou dificultar a generalização para horizontes distantes, caracterizando o que se define como *context overfitting*.

4.4 Análise de Ranking e Sensibilidade

Além das métricas de erro absoluto, conduziu-se uma Análise de Performance Ordinal (*Rank-Based Performance Analysis*). Esta abordagem normaliza as diferenças de escala entre pesquisadores, focando na posição relativa de cada modelo frente aos seus competidores em cada cenário de teste. O *Ranking* Médio (RM) é calculado de forma que $RM = 1$ indicaria um modelo que venceu todos os testes.

4.4.1 Sensibilidade ao Horizonte de Previsão

A Figura 4.4 ilustra a evolução do *Ranking* Médio conforme o horizonte de previsão se estende de 1 a 10 anos.

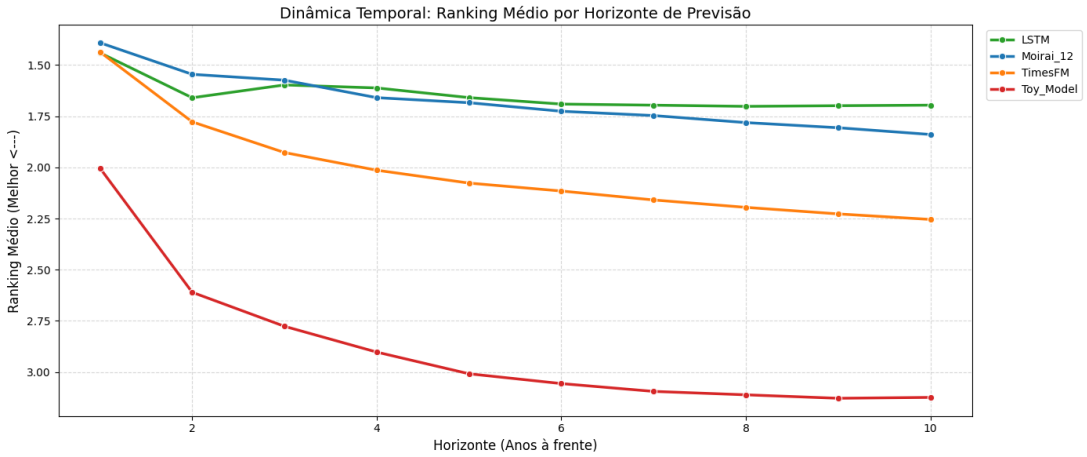


Figura 4.4 – Evolução do *ranking* médio por horizonte de previsão.

Fonte: Autoria própria.

A análise temporal do *ranking* revela uma inversão de liderança crítica conforme o horizonte de previsão se estende. No horizonte imediato ($H = 1$), o modelo MOIRAI-MoE apresenta o melhor *ranking* médio ($1,39$), superando tanto o especialista LSTM ($1,44$) quanto o TimesFM ($1,44$). Este comportamento sugere que, para prever o próximo passo imediato na trajetória do pesquisador, a capacidade de generalização do *Foundation Model* demonstra-se mais eficaz que o aprendizado específico, possivelmente por capturar nuances sutis e universais em séries temporais que transcendem o domínio local. Entretanto, à medida que o horizonte se expande ($H \geq 2$), o modelo LSTM retoma e consolida a liderança experimental. No horizonte mais distante ($H = 10$), o LSTM mantém um *ranking* de $1,70$, enquanto o MOIRAI-MoE

apresenta uma degradação para $1,84$. O modelo treinado especificamente demonstra, portanto, maior robustez na preservação de tendências de longo prazo, ao passo que os modelos *zero-shot* tendem a perder precisão acumulada. Adicionalmente, nota-se que o TimesFM apresenta consistentemente *rankings* inferiores às variantes do MOIRAI-MoE, registrando $2,25$ contra $1,84$ no horizonte $H = 10$, o que reforça a superioridade da arquitetura *Mixture of Experts* para este domínio específico.

4.4.2 Sensibilidade ao Tamanho do Contexto

A Figura 4.5 avalia como a disponibilidade de dados históricos ($C \in [4, 12]$) afeta o desempenho relativo.

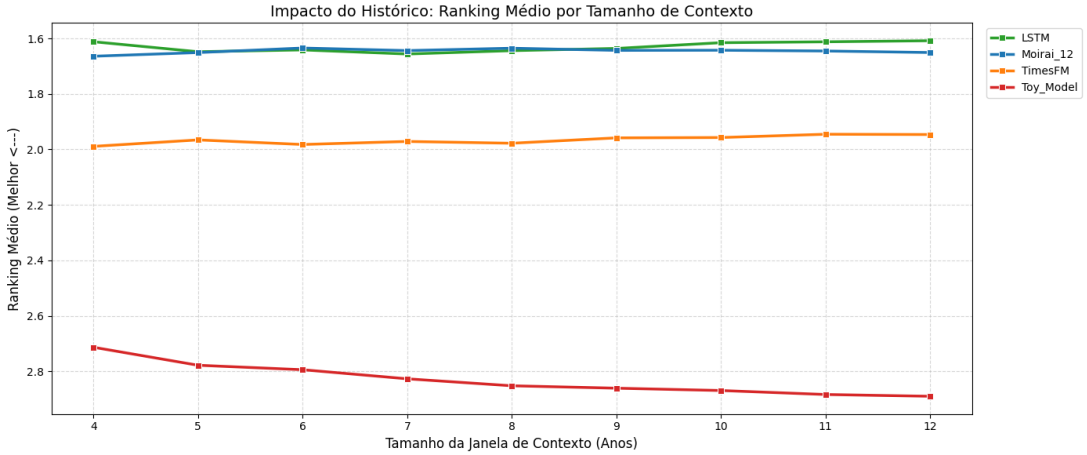


Figura 4.5 – *Ranking* médio em função do tamanho do contexto histórico disponível.

Fonte: Autoria própria.

Observa-se uma acentuada competitividade entre os modelos LSTM e MOIRAI-MOE ao longo dos diferentes comprimentos de contexto analisados. Em janelas de contexto curto e médio ($C \in [4, 9]$), o MOIRAI-MOE frequentemente supera o especialista LSTM em termos de *ranking* médio; como ilustrado no cenário $C = 6$, em que o MOIRAI-MOE registrou um índice de $1,63$ frente aos $1,64$ do LSTM. Esse comportamento reforça a premissa de que os *Foundation Models* são particularmente eficazes para cenários de *warm-start*, nos quais o histórico disponível para um item específico é limitado.Por outro lado, em contextos mais extensos ($C \geq 10$), a disparidade de desempenho entre as duas abordagens virtualmente se extingue. No cenário de contexto máximo ($C = 12$), identifica-se um empate técnico notável, com o LSTM atingindo um *ranking* de $1,60$ e o MOIRAI-MOE situando-se em $1,65$. Tal convergência sugere que, à medida que o volume de dados históricos aumenta, a vantagem da generalização inerente aos modelos de fundação é equalizada pela capacidade do modelo especialista em processar sequências mais longas e específicas do domínio acadêmico.

4.4.3 Distribuição de Posicionamento (Win Rate)

Por fim, a Figura 4.6 e a Tabela 4.3 detalha a frequência com que cada modelo ocupou cada posição no *ranking* (1º ao 4º lugar).

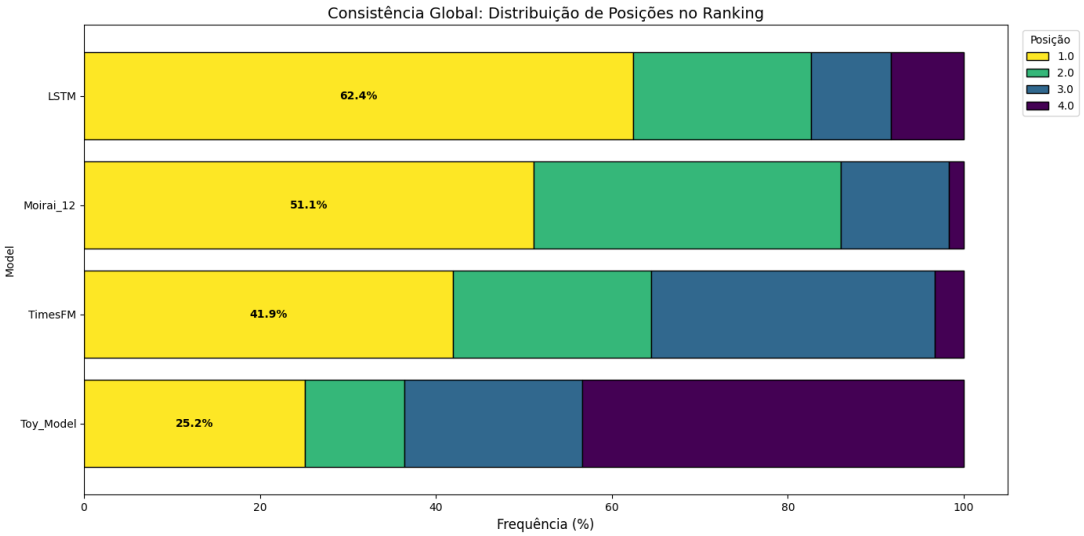


Figura 4.6 – Distribuição de posições no *ranking*.

Fonte: Autoria própria.

Tabela 4.3 – Distribuição de posições no *ranking* (% de vezes em cada colocação)

Modelo	**1º (Vencedor)**	**2º**	**3º**	**4º** (Pior)
LSTM	**62.39%**	20.24%	9.11%	8.25%
MOIRAI-MoE	51.11%	**34.87%**	12.31%	1.71%
TimesFM	41.90%	22.57%	**32.27%**	3.26%
Toy Model	25.16%	11.26%	20.21%	**43.36%**

Fonte: Autoria própria.

A análise da taxa de vitória (Tabela 4.3) consolida o modelo LSTM como a abordagem predominante, alcançando o primeiro posto em $62,39\%$ dos cenários avaliados. O MOIRAI-MoE, por sua vez, demonstra uma estabilidade posicional notável, figurando na última posição (4º lugar) em apenas $1,71\%$ das instâncias. É imperativo ressaltar, contudo, que a métrica de ranqueamento, ao focar na ordem relativa entre os modelos, pode mascarar a magnitude real dos resíduos, nesse sentido, modelos que apresentam erros numericamente baixos podem ocupar posições inferiores sem que isso reflita uma falha crítica de predição. Por outro lado, observa-se que o LSTM apresenta uma taxa de falha de $8,25\%$, casos nos quais obteve desempenho inferior até mesmo ao *Toy Model*. Tal evidência corrobora a hipótese de que o modelo especialista enfrenta suas maiores limitações em horizontes de curto prazo, onde o ruído local, e tendências gerais do conjunto de dados, podem comprometer a eficácia das predições do modelo.

4.5 Distribuição de Erros e Estabilidade

Métricas de tendência central, como o erro médio (MAE), embora úteis, são sensíveis a valores atípicos (*outliers*) e podem ocultar a verdadeira variabilidade do desempenho dos modelos. Para avaliar a consistência e a estabilidade das predições, analisou-se a distribuição dos erros absolutos ($|y - \hat{y}|$) através de um Boxplot e estatísticas descritivas detalhadas.

A Tabela 4.4 resume a média, o desvio padrão e a mediana dos erros para cada modelo, calculados sobre o total de $N = 270.156$ amostras pareadas. Complementarmente, a Figura 4.7 exibe a distribuição desses erros de forma visual, permitindo identificar a dispersão dos resíduos e a presença de *outliers* em cada abordagem avaliada.

Tabela 4.4 – Estatísticas descritivas da distribuição de erros absolutos

Modelo	**Média** (μ)	**Desvio Padrão** (σ)	**Mediana** (Md)
Toy Model	3.285	3.228	2.0
TimesFM	2.095	2.329	1.0
MOIRAI-MoE	1.721	2.026	1.0
LSTM	**1.519**	**1.747**	**1.0**

Fonte: Autoria própria.

A Figura 4.7 apresenta a visualização gráfica destas distribuições.

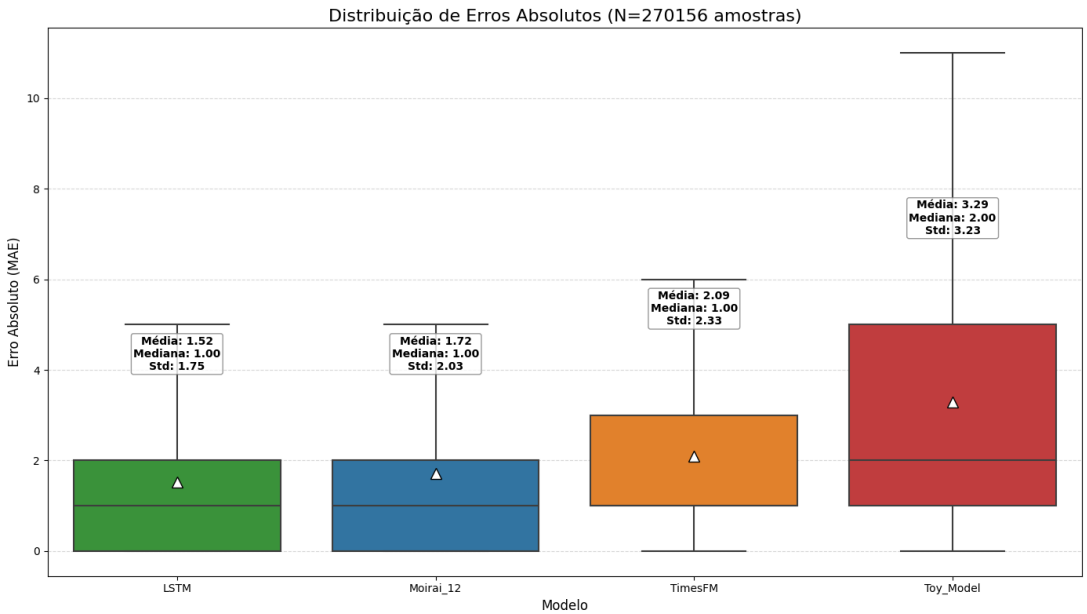


Figura 4.7 – Boxplot da distribuição de erros absolutos.

Fonte: Autoria própria.

A análise da distribuição dos resíduos revela uma alta precisão típica entre os modelos de aprendizado de máquina, visto que a mediana do erro absoluto para as arquiteturas LSTM, MOIRAI-MOE e TimesFM situou-se em $1,0$. Tal fato indica que, para pelo menos 50% das

instâncias de teste, os modelos apresentaram um desvio de apenas uma unidade no *h-index* futuro ou atingiram a predição exata, o que demonstra que a tarefa é tratável e que as dinâmicas básicas das carreiras foram capturadas.

Contudo, observa-se em todos os modelos que a média (μ) é sistematicamente superior à mediana (Md), padrão que se intensifica no *Toy Model*, onde a média de $3,28$ supera largamente a mediana de $2,0$. Essa configuração caracteriza uma distribuição com forte assimetria positiva (*right-skewed*), evidenciando uma cauda longa de erros de alta magnitude. Esses *outliers* provavelmente correspondem a pesquisadores com trajetórias atípicas, denominadas estrelas acadêmicas, cujo crescimento explosivo foge aos padrões estatísticos convencionais aprendidos pelas redes.

No que tange à consistência das predições, o modelo LSTM consolidou-se como a abordagem mais robusta, registrando não apenas a menor média, mas também o menor desvio padrão ($\sigma \approx 1,75$). Essa menor variabilidade em torno da média indica uma maior confiabilidade do modelo especialista, fato corroborado visualmente pela caixa mais compacta em seu *Boxplot*. Quanto à estabilidade em regime *zero-shot*, o MOIRAI-MOE apresentou um desvio padrão de $2,02$, valor significativamente inferior aos registrados pelo TimesFM ($2,33$) e pelo *Toy Model* ($3,23$). Tais resultados reforçam que, dentre as opções de *Foundation Models* avaliadas, o MOIRAI-MOE oferece o melhor compromisso entre precisão e risco, mantendo-se estatisticamente estável mesmo sem passar por um processo de treinamento específico para o domínio.

Em suma, enquanto o *Toy Model* falha catastroficamente em casos complexos (elevando sua média e desvio padrão), os modelos baseados em Redes Neurais (LSTM) e *Transformers* (MOIRAI-MoE) conseguem manter a maior parte das predições dentro de uma margem de erro mínima ($|e| \leq 1$), com o LSTM oferecendo a maior robustez contra casos extremos.

5 Considerações Finais

A previsão do impacto científico, quantificado pelo *h-index*, permanece um desafio aberto na cientometria devido à natureza não estacionária e multifatorial das trajetórias acadêmicas. Esta monografia investigou a viabilidade da aplicação de *Time Series Foundation Models* (TSFMs) operando em regime *zero-shot* para esta tarefa, contrastando-os com abordagens especialistas supervisionadas.

O objetivo central de investigar se modelos de fundação pré-treinados poderiam generalizar para a predição de *h-index* sem *fine-tuning* foi validado positivamente, e a hipótese levantada na Introdução acabou se revelando bem fundamentada. O modelo MOIRAI-MoE demonstrou um desempenho robusto, superando o *baseline* (*Toy Model*) em todas as métricas e alcançando uma redução de erro de aproximadamente $17,8\%$ em relação ao TimesFM. Mais notavelmente, em cenários de curto prazo ($H = 1$) e contextos limitados ($C \leq 8$), o MOIRAI-MoE atingiu a paridade estatística ou até superioridade nominal sobre o modelo especialista, com um *Ranking* Médio de $1,39$ contra $1,44$ do LSTM. Isso confirma que a arquitetura *Mixture of Experts* é capaz de transferir padrões universais de séries temporais para o domínio acadêmico com alta eficácia.

A comparação direta entre o MOIRAI-MoE e a rede LSTM (treinada especificamente com dados de pesquisadores brasileiros) revelou um *trade-off* claro. Enquanto o modelo especialista (LSTM) garantiu a liderança nas métricas globais ($MAE \approx 1.52$ e vitória em $62,39\%$ dos casos), sua vantagem reside majoritariamente na preservação da tendência de longo prazo ($H \geq 2$). Em contrapartida, os modelos de fundação provaram-se mais estáveis para predições imediatas (para o próximo ano), sugerindo que são a escolha ideal para sistemas de recomendação ou análise rápida onde o retreinamento constante de modelos especialistas é inviável.

A análise de distribuição de erros demonstrou que a tarefa de predição do *h-index* é tratável por modelos de aprendizado de máquina. Tanto o LSTM quanto o MOIRAI-MoE apresentaram uma mediana de erro absoluto igual a 1.0, indicando que, na maioria dos casos, a predição desvia-se minimamente do valor real. A consistência do MOIRAI-MoE (que raramente figura entre os piores modelos no *ranking*) destaca a segurança da abordagem *Foundation Model* como um preditor de propósito geral confiável.

Em suma, este estudo conclui que, embora modelos especialistas ainda detenham a supremacia na precisão absoluta para horizontes longos, os *Time Series Foundation Models* representam uma mudança de paradigma viável e promissora. Eles oferecem uma alternativa de baixo custo operacional (sem necessidade de treino) e alta competitividade para a análise imediata de impacto científico.

5.1 Trabalhos Futuros

A partir das limitações identificadas e dos resultados obtidos, sugerem-se algumas direções para a continuidade desta pesquisa. Uma proposta promissora reside em investigar se o ajuste fino (*fine-tuning*) do modelo MOIRAI-MoE, utilizando uma fração dos dados brasileiros (*few shot*) permitiria que ele superasse o LSTM também no longo prazo, combinando o conhecimento prévio universal com a especificidade do domínio. Outra possibilidade, seria enriquecer o *dataset* com variáveis além da série histórica pura, como a área de atuação do pesquisador, redes de coautoria ou métricas de impacto do veículo de publicação, permitindo que modelos multivariados capturem melhor os saltos de carreira atípicos.

Adicionalmente, pode-se expandir a validação para além da Ciência da Computação, testando a generalização dos modelos em áreas com dinâmicas de citação distintas, como Medicina (citação rápida) ou Humanidades (citação tardia).

Por fim, sugere-se explorar a natureza probabilística do MOIRAI-MoE para fornecer não apenas o valor pontual do *h-index*, mas intervalos de confiança que auxiliem gestores na avaliação de risco das projeções de carreira.

Referências

ABBASIMEHR, H.; SHABANI, M.; YOUSEFI, M. An optimized model using lstm network for demand forecasting. *Computers Industrial Engineering*, v. 143, p. 106435, 2020. ISSN 0360-8352. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0360835220301698>.

ACUNA, D. E.; ALLESINA, S.; KORDING, K. P. Predicting scientific success. *Nature*, v. 489, n. 7415, p. 201–202, set. 2012.

CALVO, B.; SANTAFé, G. scmamp: Statistical comparison of multiple algorithms in multiple problems. *The R Journal*, v. 8, p. 248–256, 2015. ISSN 2073-4859. Https://doi.org/10.32614/RJ-2016-017.

CHEN, W. On the mathematical relationship between rmse and nse across evaluation scenarios. *Preprints*, Preprints, February 2026. Disponível em: <https://doi.org/10.20944/preprints202509.2032.v2>.

DAS, A.; KONG, W.; SEN, R.; ZHOU, Y. *A decoder-only foundation model for time-series forecasting*. 2024. Disponível em: <https://arxiv.org/abs/2310.10688>.

DAVID, S. V.; HAYDEN, B. Y. *The Academic Family Tree*. 2005. Acesso em: 10 mar. 2025. Disponível em: <https://academictree.org/>.

DEMŠAR, J. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, v. 7, n. Jan, p. 1–30, 2006.

DONG, Y.; JOHNSON, R. A.; CHAWLA, N. V. Will this paper increase your h-index?: Scientific impact prediction. In: *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. ACM, 2015. (WSDM 2015), p. 149–158. Disponível em: <http://dx.doi.org/10.1145/2684822.2685314>.

DONG, Y.; JOHNSON, R. A.; CHAWLA, N. V. Can scientific impact be predicted? *IEEE Transactions on Big Data*, Institute of Electrical and Electronics Engineers (IEEE), v. 2, n. 1, p. 18–30, mar. 2016. ISSN 2332-7790. Disponível em: <http://dx.doi.org/10.1109/TBDATA.2016.2521657>.

FERNANDES, F. H. A. *Caracterização dos programas de pós-graduação em Ciência da Computação no Brasil a partir de dados bibliométricos da OpenAlex.* 2024. Acesso em: 15 mar. 2025. Disponível em: <http://www.monografias.ufop.br/handle/35400000/7098>.

FORTUNATO, S.; BERGSTROM, C. T.; BöRNER, K.; EVANS, J. A.; HELBING, D.; MILOJEVIć, S.; PETERSEN, A. M.; RADICCHI, F.; SINATRA, R.; UZZI, B.; VESPIGNANI, A.; WALTMAN, L.; WANG, D.; BARABáSI, A.-L. Science of science. *Science*, v. 359, n. 6379, p. eaao0185, 2018. Disponível em: <https://www.science.org/doi/abs/10.1126/science.aao0185>.

Google Research. *TimesFM 2.5 200M PyTorch Model*. [S.l.]: Hugging Face, 2024. <https://huggingface.co/google/timesfm-2.5-200m-pytorch>. Acessado em: 13 de fevereiro de 2026.

HIRSCH, J. E. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, v. 102, n. 46, p. 16569–16572, 2005. Disponível em: <https://www.pnas.org/doi/abs/10.1073/pnas.0507655102>.

HODSON, T. O. Root mean square error (rmse) or mean absolute error (mae): When to use them or not. *Geoscientific Model Development Discussions*, Göttingen, Germany, v. 2022, p. 1–10, 2022.

JADON, A.; PATIL, A.; JADON, S. *A Comprehensive Survey of Regression Based Loss Functions for Time Series Forecasting*. 2022. Disponível em: <https://arxiv.org/abs/2211.02989>.

JENSEN, P.; ROUQUIER, J.-B.; CROISSANT, Y. Testing bibliometric indicators by their prediction of scientists promotions. *Scientometrics*, Akadémiai Kiadó, co-published with Springer Science+Business Media B.V., Formerly Kluwer Academic Publishers B.V., Budapest, Hungary, v. 78, n. 3, p. 467 – 479, 2009. Disponível em: <https://akjournals.com/view/journals/11192/78/3/article-p467.xml>.

KELLY, C. D.; JENNIONS, M. D. The h index and career assessment by numbers. *Trends in Ecology & Evolution*, Elsevier, v. 21, n. 4, p. 167–170, abr. 2006.

KUPPLER, M. Predicting the future impact of computer science researchers: Is there a gender bias? *Scientometrics*, Springer, v. 127, n. 11, p. 6695–6732, 2022. Disponível em: <https://doi.org/10.1007/s11192-022-04337-2>.

LIANG, Y.; WEN, H.; NIE, Y.; JIANG, Y.; JIN, M.; SONG, D.; PAN, S.; WEN, Q. Foundation models for time series analysis: A tutorial and survey. In: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, 2024. (KDD '24), p. 6555–6565. Disponível em: <http://dx.doi.org/10.1145/3637528.3671451>.

LIU, X.; LIU, J.; WOO, G.; AKSU, T.; LIANG, Y.; ZIMMERMANN, R.; LIU, C.; SAVARESE, S.; XIONG, C.; SAHOO, D. *Moirai-MoE: Empowering Time Series Foundation Models with Sparse Mixture of Experts*. 2024. Disponível em: <https://arxiv.org/abs/2410.10469>.

MOMENI, F.; MAYR, P.; DIETZE, S. Investigating the contribution of author- and publication-specific features to scholars' h-index prediction. *EPJ Data Sci.*, Springer Science and Business Media LLC, v. 12, n. 1, out. 2023.

MULAYIM, O. B.; QUAN, P.; HAN, L.; OUYANG, X.; HONG, D.; BERGéS, M.; SRIVASTAVA, M. Are time series foundation models ready to revolutionize predictive building analytics? In: *Proceedings of the 11th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*. New York, NY, USA: Association for Computing Machinery, 2024. (BuildSys '24), p. 169–173. ISBN 9798400707063. Disponível em: <https://doi.org/10.1145/3671127.3698177>.

OCHI, M.; SHIRO, M.; MORI, J.; SAKATA, I. Predictive analysis of multiple future scientific impacts by embedding a heterogeneous network. *Plos one*, Public Library of Science San Francisco, CA USA, v. 17, n. 9, p. 1–22, 2022. Disponível em: <https://doi.org/10.1371/journal.pone.0274253>.

PENNER, O.; PAN, R. K.; PETERSEN, A. M.; KASKI, K.; FORTUNATO, S. On the predictability of future impact in science. *Scientific Reports*, v. 3, n. 1, p. 3052, out. 2013.

PRIEM, J.; PIWOWAR, H.; ORR, R. *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts*. 2022. Disponível em: <https://arxiv.org/abs/2205.01833>.

Salesforce Research. *Unified Training of Universal Time Series Forecasting Transformers (MOIRAI-MoE)*. [S.l.]: Hugging Face, 2024. <https://huggingface.co/Salesforce/moirai-moe-1.0-R-base>. Acessado em: 13 de fevereiro de 2026.

SIAMI-NAMINI, S.; TAVAKOLI, N.; NAMIN, A. S. A comparison of arima and lstm in forecasting time series. In: *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. [S.l.: s.n.], 2018. p. 1394–1401.

SIAMI-NAMINI, S.; TAVAKOLI, N.; NAMIN, A. S. The performance of lstm and bilstm in forecasting time series. In: *2019 IEEE International Conference on Big Data (Big Data)*. [S.l.: s.n.], 2019. p. 3285–3292.

THAKUR, S. C. Foundation models for time series forecasting. *International IT Journal of Research, ISSN: 3007-6706*, v. 2, n. 4, p. 144–156, Oct. 2024. Disponível em: <https://itjournal.org/index.php/itjournal/article/view/81>.

VASCONCELOS Ícaro L. L. *Mapeando as conexões acadêmicas: uma análise da rede de pesquisadores em ciência da computação via OpenAlex.* 2024. Acesso em: 15 mar. 2025. Disponível em: <http://www.monografias.ufop.br/handle/35400000/7078>.

WEIHS, L.; ETZIONI, O. Learning to predict citation-based impact measures. In: IEEE. *2017 ACM/IEEE joint conference on digital libraries (JCDL)*. [S.l.], 2017. p. 1–10.

WOO, G.; LIU, C.; KUMAR, A.; XIONG, C.; SAVARESE, S.; SAHOO, D. Unified training of universal time series forecasting transformers. In: *Forty-first International Conference on Machine Learning*. [s.n.], 2024. Disponível em: <https://openreview.net/forum?id=Yd8eHMY1wz>.

ZENG, A.; CHEN, M.; ZHANG, L.; XU, Q. *Are Transformers Effective for Time Series Forecasting?* 2022. Disponível em: <https://arxiv.org/abs/2205.13504>.