Modelos Preditivos para Detecção de Violações de Dados: Uma Abordagem Comparativa entre Técnicas Clássicas e de Aprendizado Profundo

Evanei Gomes Dos Santos[1], **Gabriel Arquelau Pimenta Rodrigues**[1], **André Luiz Marques Serrano**[1], **Geraldo Pereira Rocha Filho**[2], **Felipe Barreto De Oliveira**[1], **Vinicius Pereira Goncalves**[1]

[1]Programa de Pós-Graduação Profissional em Engenharia Elétrica (PPEE)
Departamento de Engenharia Elétrica (ENE)
Faculdade de Tecnologia, Universidade de Brasília (UnB)
Brasília 70910-900, Brasil

[2]Departamento de Ciências Exatas e Tecnológicas (DCET)
Universidade Estadual do Sudoeste da Bahia (UESB)
Vitória da Conquista 45083-900, Brasil

evanei.santos@aluno.unb.br, gabriel.arquelau@redes.unb.br,
andrelms@unb.br, geraldof@unb.br,
felipe.barreto@aluno.unb.br, vpgvinicius@unb.br

Abstract. *Given the increase in data breaches and the high costs involved, this study performs a comparative analysis between prediction algorithms applied to information security in different organizational sectors. The LSTM, TCN, Prophet, SARIMA and XGBoost models were evaluated, based on incident data made available by the Privacy Rights Clearinghouse. The comparison considered the MAE, RMSE and MAPE metrics. The best MAPE results were achieved by TCN in the General Total (10.21%), Healthcare (23.52%) and Other Businesses (19.39%), and by LSTM in the Unknown sectors (11.95%), Financial Services (21.14%) and also in the General Total (12.13%). The results show good performance of models based on neural networks.*

Resumo. *Diante do aumento das violações de dados e dos altos custos envolvidos, este estudo realiza uma análise comparativa entre algoritmos de previsão aplicados à segurança da informação em diferentes setores organizacionais. Foram avaliados os modelos LSTM, TCN, Prophet, SARIMA e XGBoost, com base em dados de incidentes disponibilizados pela Privacy Rights Clearinghouse. A comparação considerou as métricas MAE, RMSE e MAPE. Os melhores resultados de MAPE foram alcançados pelo TCN no Total Geral (10,21%), Saúde (23,52%) e Outros Negócios (19,39%), e pelo LSTM nos setores Desconhecidos (11,95%), Serviços Financeiros (21,14%) e também no Total Geral (12,13%). Os resultados mostram um bom desempenho de modelos baseados em redes neurais.*

1. Introdução

A crescente sofisticação dos ataques cibernéticos e o volume elevado de informações sensíveis armazenadas digitalmente têm intensificado as preocupações com violações de

dados. Organizações públicas e privadas enfrentam desafios constantes na proteção de dados pessoais, financeiros e corporativos. Segundo o relatório da IBM Security e Ponemon Institute, o custo médio global de uma violação de dados em 2024 foi de US$ 4,88 milhões [IBM 2024].

Diversos incidentes relevantes ilustram essa realidade, como o caso dos hotéis Marriott, que expôs informações de 500 milhões de clientes [Yu et al. 2022]. Empresas como T-Mobile, Quora, Google, Orbitz e Facebook também sofreram ataques que comprometeram mais de 100 milhões de usuários [Privacy Rights Clearinghouse 2025]. Esses vazamentos comprometem a confidencialidade das organizações e impactam diretamente as legislações de privacidade vigentes [Gong et al. 2022].

Segundo [Pimenta Rodrigues et al. 2024], violações de dados podem levar à perda de informações cruciais, incluindo dados pessoais, de saúde e financeiros, que são sensíveis e privados. Essas violações configuram incidentes de segurança nos quais dados confidenciais são expostos a indivíduos não autorizados, gerando problemas de privacidade.

Para empresas e entidades públicas, a divulgação não intencional de dados pode acarretar danos à reputação [Perera et al. 2022], processos judiciais [Duggineni 2023] e perdas financeiras diretas [Foerderer and Schuetz 2022].

Dados sensíveis frequentemente envolvem informações bancárias, uso de redes sociais e identidade pessoal, conforme apontado por [Mangku et al. 2021]. Além disso, a literatura evidencia um aumento tanto na frequência dos incidentes quanto no custo médio das violações [Almulihi et al. 2022].

Considerando esse cenário, preocupações com a privacidade das informações tem aumentado [Varshney et al. 2020]. Para uma estratégia de prevenção mais eficaz, é fundamental gerenciar os riscos cibernéticos considerando tanto sua probabilidade quanto seu potencial de impacto [Alahmari and Duncan 2020].

Nesse contexto, a análise de séries temporais se destaca como uma abordagem promissora para compreender padrões, identificar tendências e antecipar o comportamento de eventos cibernéticos ao longo do tempo. Essa técnica analítica tem ganhado relevância à medida que cresce a necessidade de métodos capazes de capturar a evolução temporal dos incidentes e suas recorrências.

Dados de séries temporais consistem em observações sequenciais registradas em intervalos regulares e estão presentes em diversas aplicações do mundo real, como mercados financeiros, reconhecimento de fala, medições de tráfego, monitoramento climático, dados biomédicos e registros populacionais [Ahmed et al. 2023].

Para muitas dessas aplicações, os modelos de aprendizado profundo têm mostrado resultados surpreendentes, superando modelos estatísticos e modelos tradicionais de aprendizado de máquina. [Janiesch et al. 2021].

Em síntese, este trabalho realiza uma análise comparativa entre técnicas de Machine Learning para estimar a quantidade de violações de dados de segurança, visando identificar os modelos que apresentam melhor desempenho preditivo. As principais contribuições deste estudo são a aplicação de modelos preditivos avançados (LSTM, TCN, XGBoost, SARIMA e Prophet) para a previsão de violações em diferentes seto-

res organizacionais e a análise comparativa detalhada com base em métricas estatísticas (MAPE, MAE, RMSE), destacando a acurácia preditiva dos modelos.

Para esse estudo foi ultilizado o conjunto de dados Data Breach Chronology da [Privacy Rights Clearinghouse 2025] (PRC), que embora seja amplamente utilizado em estudos sobre segurança da informação, não foram identificadas pesquisas que realizem uma comparação sistemática entre diferentes modelos preditivos aplicados a esse contexto.

Este estudo está organizado em 5 seções. A Seção 2 aborda os trabalhos relacionados. A Seção 3 descreve a metodologia utilizada. A Seção 4 apresenta a análise exploratória, preparação dos dados e aplicação dos modelos preditivos. Por fim, a Seção 5 discute os resultados obtidos e expõe as conclusões e propostas para trabalhos futuros.

2. Trabalhos Relacionados

Esta seção apresenta os trabalhos relacionados, com foco na literatura sobre segurança da informação e violações de dados. A revisão bibliográfica foi conduzida por meio de uma busca sistemática nas bases de dados Scopus e Google Scholar, utilizando os termos *("data breach"OR "data leak") AND (forecasting OR predict)* como palavras-chave, com o objetivo de identificar estudos relevantes sobre previsão de violações de dados.

[Avanzi et al. 2025] explora a evolução dos padrões de notificação e a frequência de violações de dados nos Estados Unidos, com base em relatórios de Procuradores-Gerais de oito estados. O estudo destaca variações nos padrões de notificação, níveis de severidade e períodos de tempo, discutindo suas implicações para o seguro cibernético. Os autores identificam um aumento nos atrasos das notificações e na frequência das violações após 2020, utilizando o conjunto de dados da PRC como base empírica para modelar essas tendências.

[Barati and Yankson 2022] investigam a previsão de violações de dados com base em registros históricos reportados nos EUA entre 2005 e 2019, obtidos pela PRC. Os autores aplicaram modelos de Poisson e Binomial Negativo para estimar a frequência e o tamanho das violações. Os resultados mostraram que, apesar do aumento da atenção e advertências, não houve crescimento significativo nas ocorrências. Ambos os modelos apresentaram baixa divergência em relação aos dados reais, demonstrando desempenho promissor.

[Africk and Levy 2021] analisaram 9.015 violações de dados registradas pela PRC entre 2005 e 2019, utilizando Big Data e visualização de séries temporais para identificar tendências em segurança cibernética. Destacam a importância da categorização por tipo de violação e setor afetado, além do papel das regulamentações e medidas de segurança na mitigação de incidentes. As visualizações demonstram o potencial da análise temporal para apoiar decisões estratégicas.

[Carfora and Orlando 2022] propuseram a quantificação do risco cibernético, com foco em violações de dados maliciosas e negligentes. Utilizando dados da PRC, os autores modelam a frequência das violações com uma distribuição binomial negativa e a gravidade com uma distribuição skew-normal. O estudo destaca a importância de incorporar o risco cibernético na gestão empresarial e propõe o uso do Valor em Risco (VaR) para estimar o impacto financeiro potencial, reforçando a necessidade de distinguir entre

tipos de violações para maior precisão na análise.

[Sun et al. 2020] desenvolveram um modelo atuarial de frequência-severidade para agregar dados de violações de nível empresarial, demonstrando como a formação de taxas pode ser fundamentada durante a subscrição de seguros de segurança cibernética. O modelo proposto possibilita a integração de informações sobre a frequência dos incidentes e a gravidade dos danos, fornecendo uma base para a avaliação de riscos e a precificação dos seguros nesse segmento.

Os estudos analisados avançam na compreensão das violações de dados, com foco em modelos estatísticos, atuariais e descritivos baseados nos dados da PRC. Contudo, há uma lacuna quanto ao uso de modelos de aprendizado de máquina na base Data Breach Chronology, limitando a capacidade preditiva. Nesse contexto, este estudo se destaca ao comparar sistematicamente técnicas clássicas e de aprendizado de máquina aplicadas a séries temporais, visando identificar os modelos mais eficazes para apoiar a gestão da segurança da informação.

3. Metodologia

Esta seção propõe uma arquitetura metodológica para análise comparativa entre modelos de aprendizado de máquina.

O conjunto de dados utilizado, Data Breach Chronology da PRC, reúne informações detalhadas sobre incidentes de violação de dados, incluindo tipo de ataque, número de pessoas afetadas, setores organizacionais, data da ocorrência, estado, cidade, entre outros atributos. Para esta pesquisa, foi selecionada a variável tipo de organização que reflete o setor de atuação daquela organização. A Figura 1 apresenta uma visão geral da arquitetura proposta.

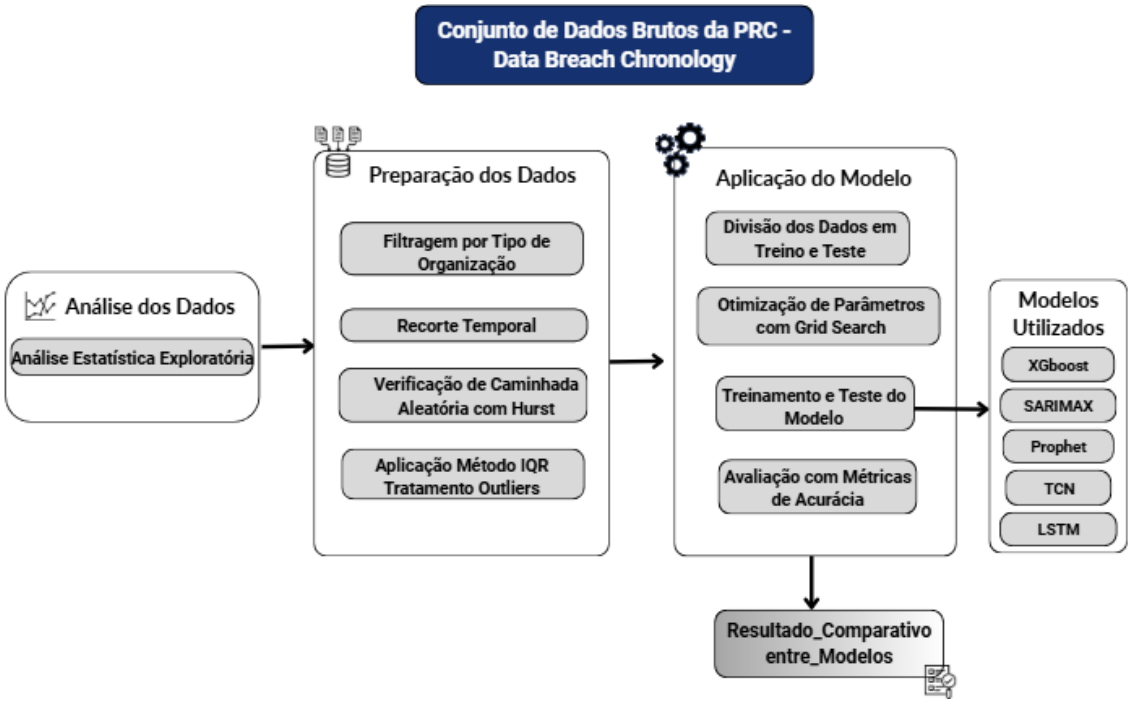

Figura 1. Arquitetura Metodológica para Aplicação dos Modelos Preditivos

Inicialmente, foi realizada a Análise Estatística Exploratória (AEE), essencial para a compreensão preliminar dos dados.

A preparação dos dados envolveu limpeza, ajuste, transformação e organização, seguidos do filtro por tipo de organização e período para construção das séries temporais mensais. Aplicou-se a métrica de Hurst para verificar a presença de caminhada aleatória, caracterizada por independência entre valores passados e futuros. Por fim, outliers foram removidos pelo método do intervalo interquartil (IQR), aumentando a assertividade das análises.

Com os dados preparados, adotou-se uma abordagem preditiva com séries temporais mensais, aplicando modelos estatísticos, algoritmos de árvores de decisão e redes neurais. A calibração dos modelos foi feita com grid search, e os dados foram divididos em treino e teste. O desempenho foi avaliado com as métricas MAPE, MAE e RMSE, permitindo comparar e identificar o modelo mais preciso.

Todo o processo de análise, modelagem e previsão foi realizado em Python, utilizando bibliotecas como *statsmodels*, *fbprophet*, *xgboost*, *tensorflow*, *keras*, *pandas*, *numpy*, *matplotlib* e *scikit-learn*. Os artefatos completos do projeto, incluindo scripts, pseudocódigo e documentação, estão disponíveis em: `https://github.com/evaneigomes/predicao_violacao_dados`. Devido à natureza proprietária e restrições de licença, os dados não serão disponibilizados.

Como forma de complementar o entendimento da abordagem proposta, foi desenvolvido um pseudocódigo, apresentado a seguir:

Algorithm 1: Fluxo Geral da Arquitetura Metodológica

```
1  1. Análise Exploratória
2      Analise_Estatistica_Exploratoria(dados_brutos)
3  2. Preparação dos Dados
4      dados_brutos ← LER_ARQUIVO("Data_Breach_Chronology.xlsx")
5      AJUSTAR_FORMATO_DATAS(dados_brutos["Date_Breach"])
6      REMOVER_REGISTROS_VAZIOS(dados_brutos, "Date_Breach")
7      DEFINIR_INDICE_TEMPORAL(dados_brutos, "Date_Breach")
8      recorte_temporal ← FILTRAR_PERIODO_ANALISE(dados_brutos, "2010-01-01",
         "2023-12-31")
9      EXPONENTE_HURST ← CALCULAR_HURST(recorte_temporal)
10     colunas_numericas ← IDENTIFICAR_COLUNAS_NUMERICAS(recorte_temporal)
11     dados_preparados ← APLICAR_TRATAMENTO_OUTLIERS_IQR(recorte_temporal,
         colunas_numericas)
12 3. Aplicação dos Modelos
13     LISTA_SETORES ← [BSF, BSO, BSR, EDU, GOV, MED, NGO, Total_Geral, UNKN]
14     GRADE_PARAMETROS ← COMBINAR(P1, P2, P3)
15     foreach setor ∈ LISTA_SETORES do
16         serie ← dados_preparados[setor]
17         foreach (P1, P2, P3) ∈ GRADE_PARAMETROS do
18             modelo ← INICIALIZAR_MODELO(P1, P2, P3)
19             dados_treino, dados_teste ← DIVIDIR_SERIE(serie, 0.85)
20             TREINAR_MODELO(modelo, dados_treino)
21             previsoes ← PREVER(modelo, dados_teste.X)
22             metricas ← AVALIAR_PREVISOES(dados_teste.Y, previsoes)
23             resultados ← RESULTADOS(setor, P1, P2, P3, metricas)
24 avaliação ← Avaliação_comparativa_de_modelos(resultados)
25 Exibir_Resultados(avaliação)
```

O script realiza previsões de séries temporais para os setores `BSF`, `BSO`, `BSR`, `EDU`, `GOV`, `MED`, `NGO`, `UNKN` e `Total_Geral`. Inicialmente, os dados são carregados a partir de um arquivo excel para um ambiente de programação python. Em seguida, é realizada

análise estatística exploratória, logo depois é feito ajustes no formato das datas, remoção de registros incompletos, definição do índice temporal e aplicação de um filtro para o período de 2010 a 2023. Também são calculados os expoentes de Hurst e aplicado o tratamento de outliers com base no método do IQR.

Os parâmetros P1, P2 e P3 utilizados no grid search foram definidos conforme o modelo. Para TCN e LSTM: look_back, epochs e batch_size. Para o Prophet: changepoint_prior_scale, seasonality_prior_scale e holidays_prior_scale. No SARIMA: ordens (p, d, q), (P, D, Q) e o período sazonal s. Já no XGBoost: n_estimators, max_depth e learning_rate. Esses parâmetros foram ajustados para otimizar o desempenho preditivo de cada abordagem.

Por fim, são definidos os setores a serem analisados, a grade de combinações de parâmetros e as proporções dos conjuntos de treino (85%) e teste (15%). O script executa um laço para cada setor e combinação de parâmetros, no qual a série temporal é dividida, o modelo é treinado, as previsões são geradas e avaliadas, e as métricas de desempenho são registradas para posterior análise comparativa entre os modelos. Dessa forma, a metodologia adotada — que integra análise exploratória e aplicação de modelos preditivos sobre séries temporais — permite tanto compreender o comportamento histórico das violações de dados quanto antecipar possíveis evoluções futuras.

4. Análise, Preparação dos Dados e Aplicação dos Modelos

Esta seção detalha as etapas de análise exploratória, preparação dos dados e aplicação dos modelos preditivos, conforme descrito na metodologia e ilustrado na Figura 1.

4.1. Análise dos Dados

Com o intuito de compreender o comportamento das violações de dados ao longo do tempo, foi realizada uma Análise Estatística Exploratória (AEE). Essa etapa permitiu identificar padrões, sazonalidades e variações relevantes, subsidiando a construção de modelos preditivos mais consistentes e sensíveis às dinâmicas temporais.

A análise da série histórica indicou um crescimento acentuado nas violações de dados a partir de 2010, sugerindo aumento no uso de serviços digitais [Silveira et al. 2023], maior exposição organizacional a riscos de segurança. A Figura 2 ilustra essa evolução.

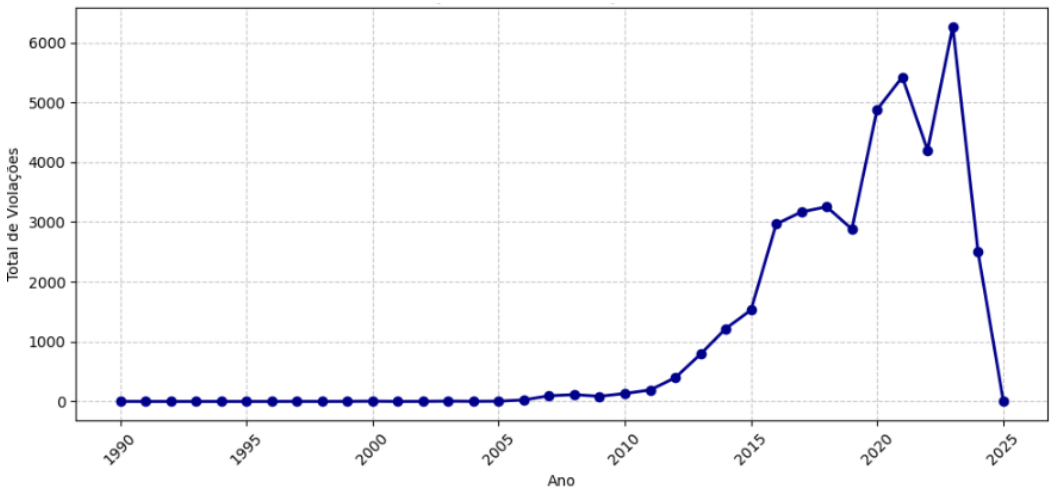

Figura 2. Evolução Anual das Violações de Dados

A Figura 3 apresenta a evolução anual das violações de dados classificadas por tipo de organização, permitindo a identificação de tendências históricas específicas para cada setor. A análise evidencia picos e quedas na ocorrência de incidentes, sugerindo possíveis relações com fatores externos, como mudanças regulatórias, avanços tecnológicos e alterações no perfil das ameaças cibernéticas. A descrição detalhada dos tipos de organização considerados encontra-se na Tabela 1.

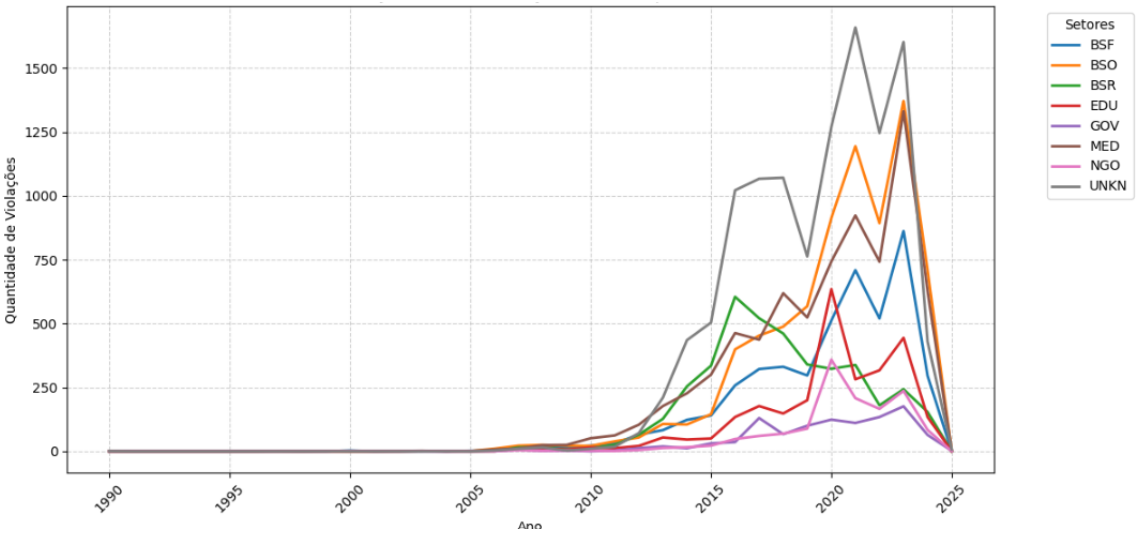

Figura 3. Evolução Anual das Violações por Tipo de Organização

Com isso, esta análise assume papel importante na definição da modelagem, ao direcionar a seleção de métodos preditivos ajustados às especificidades temporais e setoriais dos dados.

4.2. Preparação dos Dados

Dos 72.553 registros iniciais de violações, 40.142 foram validados após limpeza e exclusão de registros inconsistentes. A preparação envolveu ajustes, filtragem por tipo de organização e recorte temporal, organizando a série temporal de forma estruturada. Utilizou-se o expoente de Hurst para analisar padrões temporais, e outliers foram tratados via IQR para melhorar a qualidade preditiva.

4.2.1. Filtragem do Conjunto de Dados

Nesta etapa, foi realizada a filtragem da base de dados, considerando apenas as variáveis de tipo de organização e data da violação. Essa seleção teve como objetivo focar em informações relevantes para a construção da série temporal.

A filtragem permitiu organizar os dados de forma consistente por tipo de organização, garantindo maior homogeneidade ao longo do período analisado. Os setores organizacionais utilizadas estão apresentadas na Tabela 1.

Tabela 1. Tipos de Organização

Tipo de Organização	Descrição
BSF	Serviços financeiros (bancos, corretoras, seguradoras não-sanitárias)
BSO	Outros negócios (tecnologia, manufatura, utilidades, serviços profissionais)
BSR	Varejo (lojas físicas e online)
EDU	Instituições educacionais (escolas, universidades, serviços educacionais)
GOV	Governo e militares (agências públicas e forças armadas)
MED	Saúde (hospitais e clínicas)
NGO	Organizações sem fins lucrativos (ONGs, igrejas, grupos de advocacia)
UNKN	Setor desconhecido devido a informações insuficientes para classificar

4.2.2. Delimitação Temporal da Análise

Para assegurar maior consistência à análise, o período de estudo foi delimitado entre 2010 e 2023, conforme evidenciado na Figura 4. Registros anteriores a 2010 apresentavam volume reduzido, provavelmente em função da menor maturidade dos mecanismos de detecção e da inexistência de regulamentações específicas. Embora o conjunto de dados contenha registros até 2025, optou-se por considerar apenas as informações até 2023, a fim de evitar distorções decorrentes de possíveis atrasos nas notificações que poderiam comprometer a confiabilidade da análise.

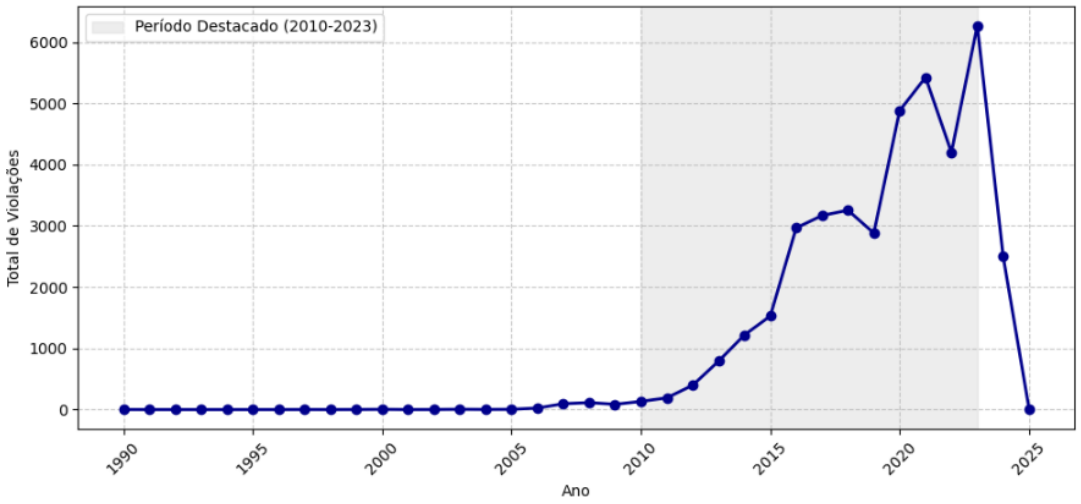

Figura 4. Delimitação Temporal

4.2.3. Avaliação de Comportamento Aleatório

Para analisar o comportamento dinâmico das séries temporais, utilizou-se o expoente de Hurst, uma medida estatística que quantifica a dependência de longo prazo e a previsibilidade da série, indicando se os retornos apresentam persistência ou antipersistência. Esse indicador permite classificar a série como tendencial ($H > 0{,}5$), aleatória ($H \approx 0{,}5$) ou revertente à média ($H < 0{,}5$), conforme [Partners 2022]. Essa abordagem contribui para uma compreensão mais profunda da estrutura temporal dos dados e auxilia na escolha dos modelos preditivos mais adequados. Os resultados estão apresentados na Tabela 2.

Tabela 2. Expoente de Hurst por Setor Organizacional

Setor Organizacional	Expoente de Hurst (H)
BSF	0.5237
BSO	0.5042
BSR	0.5773
EDU	0.5939
GOV	0.6508
MED	0.5703
NGO	0.6027
UNKN	0.4629
Total Geral	0.5269

4.2.4. Tratamento de Valores Extremos

Durante a análise, identificaram-se valores discrepantes (*outliers*) ao longo do tempo, os quais podem prejudicar a precisão e generalização dos modelos. Para minimizar esse impacto, utilizou-se o método do Intervalo Interquartil (IQR), eficaz na detecção de outliers, especialmente em distribuições não paramétricas.

Para aplicar o método do Intervalo Interquartil (IQR), deve-se inicialmente calcular o primeiro quartil (Q_1), que corresponde ao valor que separa os 25% menores dados, e o terceiro quartil (Q_3), que separa os 75% menores dados.

A seguir, apresenta-se o procedimento para calcular o intervalo interquartil (IQR) e determinar os limites inferior e superior utilizados na identificação de outliers:

$$\text{IQR} = \text{Q}_3 - Q_1$$
$$\text{Limite Inferior} = \text{Q}_1 - 1{,}5 \times IQR$$
$$\text{Limite Superior} = \text{Q}_3 + 1{,}5 \times IQR$$

Qualquer valor abaixo do limite inferior ou acima do limite superior é considerado um *outlier*.

Essa abordagem permite lidar com valores extremos sem removê-los, preservando a estrutura dos dados e aprimorando a qualidade das análises. De acordo com [Kumar et al. 2023], a combinação do método do intervalo interquartil (IQR) com a *winsorização* possibilita a identificação e o ajuste eficiente de outliers, promovendo resultados mais consistentes, especialmente em contextos que demandam alta qualidade dos dados. A *winsorização* substitui valores extremos — situados além de determinados percentis — pelos próprios valores dos percentis-limite, reduzindo a influência de observações atípicas sem necessidade de exclusão. Essa técnica é especialmente útil em análises sensíveis à presença de outliers, como séries temporais e dados financeiros.

4.3. Aplicação do Modelo

A aplicação dos modelos será realizada individualmente para cada setor, com o objetivo de avaliar comparativamente seu desempenho na previsão da quantidade de violações de dados ao longo do tempo.

4.3.1. Divisão do Dados em Treino e Teste

Para avaliar o desempenho preditivo dos modelos, os dados foram particionados sequencialmente em conjuntos de treino e teste, de forma a preservar a ordem temporal da série

— um requisito essencial em análises de séries temporais. Cada modelo considerado (LSTM, TCN, Prophet, SARIMA e XGBoost) foi treinado com aproximadamente 85% dos dados, correspondentes ao período de 2010 a 2021, permitindo o aprendizado dos padrões históricos e tendências relacionados às violações de dados. Os 15% restantes, referentes ao intervalo de 2022 a 2023, compuseram o conjunto de teste, possibilitando uma avaliação imparcial da capacidade preditiva dos modelos em relação a eventos futuros não observados durante o treinamento. Esse procedimento foi realizado individualmente para cada setor analisado, assegurando uma comparação justa e contextualizada entre os diferentes modelos aplicados.

4.3.2. Otimização de Parâmetros

A etapa de ajuste de hiperparâmetros é fundamental para garantir o melhor desempenho possível dos modelos preditivos. Neste estudo, utilizou-se a técnica de Grid Search, um método exaustivo que realiza uma busca sistemática por combinações ideais de parâmetros dentro de um espaço definido previamente. Essa abordagem permite testar todas as possíveis configurações especificadas em uma grade, avaliando o desempenho de cada modelo com base em métricas de validação, como o erro médio absoluto (MAE), o erro quadrático médio (RMSE) e o erro percentual absoluto médio (MAPE).

O Grid Search foi aplicado em conjunto com validação cruzada, a fim de mitigar o risco de sobreajuste (overfitting) e garantir maior generalização dos resultados. Esse processo foi executado individualmente para cada modelo e para cada setor organizacional analisado, respeitando as particularidades das séries temporais envolvidas. Como resultado, foi possível identificar configurações mais adequadas aos dados específicos de cada segmento, contribuindo diretamente para a robustez e acurácia das previsões geradas.

4.3.3. Treinamento e Teste do Modelo

Dando continuidade à etapa de preparação e com os dados devidamente estruturados para análise de séries temporais, procedeu-se à aplicação dos modelos preditivos. Neste estudo, foram considerados cinco modelos de previsão com diferentes abordagens e níveis de complexidade: SARIMA (Seasonal Autoregressive Integrated Moving Average), um modelo estatístico clássico que incorpora componentes sazonais; Prophet, um modelo estatístico aditivo desenvolvido pelo Facebook, que decompõe a série em tendência, sazonalidade e feriados, facilitando o ajuste por especialistas; XGBoost, um algoritmo de aprendizado de máquina baseado em gradient boosting de árvores de decisão, aplicado com engenharia de atributos para capturar padrões temporais; LSTM (Long Short-Term Memory), uma rede neural recorrente de aprendizado profundo, eficaz para modelar dependências de longo prazo; e TCN (Temporal Convolutional Network), uma arquitetura convolucional que utiliza convoluções causais e dilatadas para capturar dependências temporais de longo alcance. A Tabela 3 apresenta a comparação entre os modelos considerados.

Tabela 3. Descrição dos Modelos Preditivos Utilizados

Modelo	Tipo	Descrição e Características Relevantes
SARIMA	Modelo estatístico clássico	Lida com componentes de tendência, sazonalidade e resíduos. Requer séries estacionárias e pode apresentar alta complexidade para ajuste de parâmetros.
Prophet	Modelo estatístico aditivo com heurísticas	Desenvolvido para dados de séries temporais com fortes efeitos sazonais e de tendência. Automatiza a detecção de tendências e sazonalidades e é tolerante a falhas ou lacunas nos dados. Pode, contudo, superestimar tendências e ter menor desempenho com dados com ruídos ou outliers extremos.
XGBoost	Ensemble de árvores de decisão (Boosting)	Algoritmo baseado em árvores de decisão que constrói um modelo preditivo de forma aditiva. Oferece alta performance com dados estruturados e boa explicabilidade de variáveis. Contudo, não modela diretamente a sequência temporal e requer engenharia de features para dados de séries temporais.
LSTM	Rede neural recorrente (Deep Learning)	Capaz de aprender dependências de longo prazo em dados sequenciais. É eficaz com dados não lineares e sazonais. Suas desvantagens incluem a necessidade de muitos dados e tempo de treinamento, além de ser sensível ao ajuste de hiperparâmetros.
TCN	Rede neural convolucional temporal (Deep Learning)	Utiliza convoluções para modelar dependências em séries temporais. Apresenta melhor paralelismo que LSTM e é capaz de captar padrões de longo prazo de forma mais estável. No entanto, é mais complexo para configurar e ainda menos difundido na literatura de séries temporais.

4.3.4. Métricas de Acurácia

Para mensurar o desempenho dos modelos aplicados à previsão de violações de dados organizacionais, foram utilizadas três métricas estatísticas: o Erro Médio Absoluto (MAE - Mean Absolute Error), o Erro Quadrático Médio da Raiz (RMSE - Root Mean Squared Error) e o Erro Percentual Absoluto Médio (MAPE - Mean Absolute Percentage Error).

O MAE representa a média dos erros absolutos entre os valores reais e previstos. Por estar na mesma unidade da variável prevista, é de fácil interpretação e fornece uma estimativa direta do desvio médio. O RMSE calcula a raiz quadrada da média dos quadrados dos erros. Esta métrica penaliza desvios maiores de forma mais intensa que o MAE, sendo sensível a outliers. Por fim, o MAPE mede o erro percentual médio em relação aos valores reais. Por ser expresso em porcentagem, facilita a comparação entre diferentes séries ou modelos, sendo muito utilizado por sua interpretabilidade.

Neste estudo, o MAPE é adotado como a principal métrica de avaliação da acurácia preditiva dos modelos, por sua capacidade de comparação entre diferentes séries temporais e fácil interpretação. Seguindo os critérios propostos por [Lewis 1982], os valores de MAPE são classificados de acordo com a tabela 4.

Tabela 4. Classificação da Precisão das Previsões com base no MAPE

Intervalo do MAPE (%)	Classificação	Descrição
< 10	Previsão Altamente Precisa	Indica previsões com erro percentual muito baixo; trata-se de um desempenho excelente para fins analíticos e operacionais.
10 – 19,99	Boa Previsão	Indica modelos com boa acurácia, confiáveis para aplicações práticas, embora com margem de erro perceptível.
20 – 49,99	Previsão Razoável	As previsões possuem erro moderado; podem ser utilizadas em contextos exploratórios, mas com cautela nas decisões.
50 ou mais	Previsão Imprecisa	Reflete alto grau de erro percentual, tornando o modelo inadequado para aplicações que exigem confiabilidade nas estimativas.

5. Resultados e Conclusão

Esta seção apresenta os principais resultados da análise dos dados da PRC, com ênfase na comparação do desempenho preditivo dos modelos, utilizando o MAPE como métrica

principal. Ao final, são apresentadas a conclusão do estudo, suas limitações e propostas para trabalhos futuros.

5.1. Resultados da Avaliação Comparativa da Precisão entre os Modelos

A análise dos resultados obtidos com os modelos LSTM, Prophet, SARIMA, TCN, e XGBoost foi realizada com base nas métricas MAE, RMSE e MAPE, considerando a previsão de violações de dados em diferentes tipos de organização, tendo o MAPE como métrica principal de avaliação. Foi incluída a Figura **??** com resultados do MAPE por setor e por modelo aplicado:

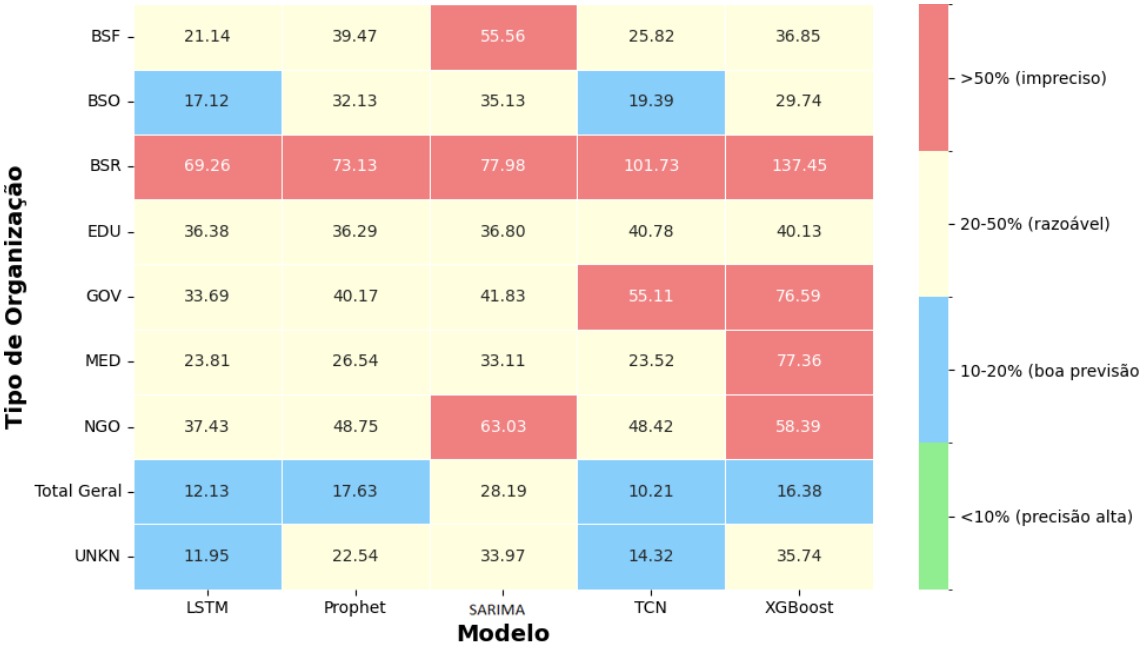

Figura 5. MAPE por Modelo com Comparação entre Setores Organizacionais

A seguir, são apresentados os resultados obtidos para cada setor, com base na aplicação dos diferentes modelos preditivos, considerando a classificação de [Lewis 1982], tendo o MAPE como referência para avaliação da acurácia.

No setor BSF (Business – Financial), o modelo LSTM apresentou o melhor desempenho relativo, com um MAPE de 21,14% (classificado como previsão razoável), além de baixos valores absolutos de erro: MAE de 7,36 e RMSE de 11,37. Embora o XGBoost tenha alcançado o menor MAE (6,30) e o menor RMSE (7,22) entre todos os modelos, seu MAPE elevado (36,85%) indica menor precisão relativa, sugerindo maior sensibilidade a variações percentuais, especialmente em valores baixos da série. Os modelos TCN, Prophet e SARIMA tiveram desempenho inferior, com MAPE de 25,82%, 39,47% e 55,56%, respectivamente — sendo os dois primeiros classificados como previsões razoáveis e o último como impreciso. Essa variação de desempenho pode ser explicada pelas particularidades do setor financeiro, cujos dados geralmente são mais estruturados, consistentes e submetidos a padrões regulatórios. Tais características favorecem modelos como o LSTM, capazes de capturar padrões temporais complexos e de longo prazo. Já o XGBoost, apesar de sua eficiência em erros absolutos, pode ter sido

impactado por outliers ou variações abruptas, que influenciam negativamente o MAPE devido à natureza percentual dessa métrica.

No setor BSO (Business – Other), o modelo LSTM apresentou o melhor desempenho entre os avaliados, com um MAPE de 17,12% (classificado como boa previsão), seguido de perto pelo TCN (19,39% – boa previsão) e pelo XGBoost (29,74% – previsão razoável). Os modelos Prophet e SARIMA, por sua vez, registraram os piores desempenhos preditivos, com MAPE de 32,13% e 35,13%, respectivamente (ambos classificados como previsões razoáveis), o que sugere uma menor capacidade de aderência aos padrões da série temporal para este setor. A superioridade de modelos de aprendizado profundo, como LSTM e TCN, neste setor, pode ser atribuída à sua capacidade de lidar com a diversidade e os padrões menos homogêneos de ”Outros Negócios”, que englobam tecnologia, manufatura e serviços profissionais. Estes modelos são mais aptos a identificar e aprender com as complexas relações não lineares e as variações atípicas que podem ocorrer em um segmento tão abrangente, superando abordagens estatísticas mais tradicionais na captura dessas dinâmicas.

No caso do setor BSR (Business - Retail), o desempenho de todos os modelos foi relativamente fraco, com MAPE elevado em todas as abordagens. O modelo LSTM apresentou o menor MAPE (69,26% – *previsão imprecisa*), seguido por Prophet (73,13% – *previsão imprecisa*) e SARIMA (77,98% – *previsão imprecisa*). O TCN e o XGBoost tiveram desempenhos ainda piores, com MAPE de 101,73% e 137,45%, respectivamente (*previsões imprecisas*). Essa baixa precisão pode ser atribuída à alta volatilidade e heterogeneidade das violações nesse setor, além da possível influência de eventos atípicos e do ruído inerente aos dados.

No setor EDU (Education), o modelo Prophet apresentou o melhor desempenho relativo, com um MAPE de 36,29% (classificado como previsão razoável), seguido de perto por LSTM (36,38% – previsão razoável) e SARIMA (36,80% – previsão razoável). Apesar de o XGBoost ter alcançado os menores valores absolutos de erro (MAE de 9,29 e RMSE de 14,89), seu MAPE de 40,13% indica menor precisão proporcional (previsão razoável). Já o modelo TCN registrou o pior resultado preditivo, com um MAPE de 40,78% (previsão razoável). A relativa similaridade de desempenho entre Prophet, LSTM e SARIMA neste setor sugere que as séries temporais de violações em instituições educacionais podem apresentar padrões sazonais e de tendência mais pronunciados e menos voláteis, características que são bem capturadas por modelos estatísticos como Prophet e SARIMA. O XGBoost, embora precise em termos de erros absolutos, teve seu MAPE elevado, possivelmente devido à sensibilidade a poucos valores discrepantes que, percentualmente, representam grandes erros. A performance ligeiramente inferior do TCN, em comparação com o LSTM, pode indicar que as dependências de longo prazo ou a estrutura temporal específica desse setor foram mais bem modeladas pela arquitetura recorrente do LSTM.

No setor GOV (Government), o modelo LSTM destacou-se com um MAPE de 33,69% (classificado como *previsão razoável*), apresentando desempenho superior ao TCN (55,11% – *previsão imprecisa*), Prophet (40,17% – *previsão razoável*), SARIMA (41,83% – *previsão razoável*) e XGBoost (76,59% – *previsão imprecisa*)[cite: 491]. A consistência entre os valores de erro absoluto (MAE de 4,09) e erro quadrático médio (RMSE de 6,03) reforça a robustez do LSTM nesse contexto[cite: 492]. A superioridade do LSTM no setor governamental pode estar relacionada à capacidade desse modelo de capturar padrões

temporais complexos em ambientes com dados que, embora sensíveis, podem exibir dependências de longo prazo influenciadas por regulamentações ou processos burocráticos. Em contraste, o baixo desempenho do TCN e, notavelmente, do XGBoost e SARIMA, sugere que as abordagens desses modelos podem ter dificuldades em adaptar-se às particularidades das séries de violações governamentais, que podem incluir variações menos previsíveis ou maior impacto de eventos pontuais que distorcem tendências lineares ou sazonais mais simples.

No setor MED (Medical), os modelos LSTM e TCN apresentaram desempenhos semelhantes, com MAPE de 23,81% e 23,52% (classificados como *previsões razoáveis*), além de MAE em torno de 15,5 e RMSE inferiores a 22. O modelo Prophet teve desempenho intermediário (MAPE de 26,54% – *previsão razoável*), enquanto o SARIMA apresentou um MAPE de 33,11% (classificado como *previsão razoável*), o segundo pior entre os avaliados. Embora o XGBoost tenha registrado os menores valores absolutos de erro (MAE de 10,49 e RMSE de 13,09), seu MAPE extremamente elevado (77,36%) indica uma baixa precisão relativa (*previsão imprecisa*). [cite: 494] A performance robusta das redes neurais (LSTM e TCN) neste setor sugere sua capacidade de modelar a complexidade e a sensibilidade dos dados de saúde, que podem incluir padrões não-lineares e eventos de violação de dados com características temporais específicas. A dicotomia no desempenho do XGBoost, com baixos erros absolutos mas alto erro percentual, reitera sua suscetibilidade a picos ou descontinuidades na série que, em termos proporcionais, geram grande impacto na acurácia preditiva.

No setor NGO (Non-Governmental Organizations), o modelo LSTM apresentou o melhor desempenho relativo, com um MAPE de 37,43% (classificado como *previsão razoável*), além de erros absolutos moderados (MAE de 5,35 e RMSE de 6,61). Embora o XGBoost tenha obtido os menores valores de MAE (4,27) e RMSE (5,32), seu MAPE elevado (58,39%) compromete a precisão proporcional (*previsão imprecisa*). Os modelos TCN, Prophet e SARIMA também registraram desempenhos inferiores, com MAPE de 48,42%, 48,75% e 63,03%, respectivamente (classificados como *previsões razoáveis* e *imprecisa*, no caso do SARIMA). A performance do LSTM neste setor, embora ”razoável”, destaca sua adaptabilidade a séries que podem ser mais irregulares ou com menor volume de dados padronizados, características comuns a organizações não-governamentais. O XGBoost, novamente, demonstra sua capacidade de minimizar erros absolutos, mas sua ineficácia em termos percentuais (MAPE impreciso) pode ser reflexo da sensibilidade a flutuações pontuais ou da dificuldade em capturar a variabilidade inerente a um setor que pode ter padrões de violação menos consistentes devido à diversidade de operações e recursos.

No Total Geral, os modelos TCN e LSTM apresentaram os melhores desempenhos relativos, com MAPE de 10,21% e 12,13% (ambos classificados como *boa previsão*). Embora o XGBoost tenha alcançado os menores erros absolutos (MAE de 33,89 e RMSE de 42,01), seu MAPE de 16,38% foi superior ao dos modelos neurais, indicando menor precisão proporcional. Prophet e SARIMA obtiveram os piores resultados agregados, com MAPE de 17,63% (*boa previsão*) e 28,19% (*previsão razoável*). O destaque de TCN e LSTM reforça a robustez dessas redes neurais na modelagem de séries temporais complexas e heterogêneas, típicas de um conjunto que abrange múltiplos setores. Sua capacidade de capturar padrões temporais não lineares, sem depender de características específicas de um único setor, contribui para a alta precisão observada. Já o XGBoost, apesar da

eficácia em minimizar erros absolutos, demonstra maior sensibilidade no MAPE, refletindo limitações na captura da dinâmica temporal e variações percentuais em contextos amplos.

O setor UNKN (Setores Desconhecidos) apresentou o melhor desempenho geral, com o modelo LSTM destacando-se (MAPE de 11,95% – *boa previsão*), seguido pelo TCN (14,32% – *boa previsão*). Esse resultado sugere maior previsibilidade nesse agrupamento, possivelmente devido à heterogeneidade e ao volume agregado de registros não classificados em setores específicos, o que pode ter gerado uma série temporal mais suavizada e com padrões menos voláteis. A capacidade dos modelos LSTM e TCN de lidar com grandes volumes de dados e aprender com a diversidade do setor "desconhecido"parece ter sido determinante para seu desempenho superior.

De modo geral, os modelos LSTM e TCN demonstraram maior consistência preditiva entre os setores analisados. O LSTM destacou-se nos setores BSF, BSO, BSR, GOV, NGO e UNKN, com melhor desempenho neste último (MAPE de 11,95%, *boa previsão*). O TCN obteve resultados competitivos, liderando no desempenho agregado (MAPE de 10,21%) e com bons resultados nos setores MED e UNKN. O XGBoost apresentou os menores erros absolutos em vários setores, mas foi penalizado por altos valores de MAPE, indicando *baixa precisão proporcional*. Já os modelos Prophet e SARIMA apresentaram desempenho inferior recorrente, com destaque negativo para o setor BSR, onde todos os modelos mostraram *previsões imprecisas*. Esses resultados sugerem que, embora o XGBoost seja eficaz em termos absolutos, LSTM e TCN são mais robustos para previsões proporcionais, especialmente em contextos complexos e com alta variabilidade setorial.

5.2. Conclusão

Este estudo comparou os modelos LSTM, TCN, Prophet, SARIMA e XGBoost na previsão de violações de segurança setoriais, com base no MAPE. O LSTM se destacou pela maior precisão na maioria dos setores, seguido pelo TCN no resultado agregado. SARIMA teve bom desempenho no setor BSR, enquanto o XGBoost apresentou baixos erros absolutos, porém com alta variabilidade. Prophet e SARIMA foram menos eficazes em cenários de alta variabilidade. Os resultados evidenciam a superioridade das redes neurais recorrentes em contextos temporais complexos, contribuindo para a antecipação de incidentes e a otimização de estratégias de mitigação.

Apesar das contribuições, este estudo apresenta algumas limitações. A base de dados utilizada restringe-se a registros históricos de um único país, o que pode limitar a generalização dos resultados para outros contextos geográficos. Além disso, a avaliação foi baseada em um conjunto restrito de métricas quantitativas, o que reforça a necessidade de, em pesquisas futuras, considerar aspectos complementares como a interpretabilidade dos modelos, o custo computacional e a capacidade de adaptação a fluxos de dados em tempo real.

Como perspectivas futuras, recomenda-se a ampliação do escopo geográfico da análise, incorporando incidentes registrados em diferentes países. Essa abordagem possibilitaria comparações internacionais mais robustas e aumentaria a aplicabilidade dos modelos desenvolvidos. Destaca-se, ainda, a importância de manter a base de dados constantemente atualizada, em função da rápida evolução das ameaças cibernéticas e das

mudanças nas regulamentações de segurança da informação.

Além disso, sugere-se expandir o modelo para contemplar a análise por tipo de vazamento, uma vez que a base Data Breach Chronology inclui essa variável. Tal expansão permitirá uma abordagem mais granular, possibilitando a identificação de padrões específicos e recorrentes de violações em diferentes setores econômicos. Nesse sentido, é igualmente relevante explorar o desenvolvimento de modelos preditivos específicos por setor, considerando características como o tipo de dado manipulado, o perfil dos atacantes e o nível de maturidade em segurança da informação.

A inclusão de variáveis qualitativas — como causas das violações, impactos econômicos e eficácia das respostas — pode enriquecer a modelagem, permitindo análises mais completas e acionáveis. Destaca-se, ainda, o potencial das abordagens híbridas, que combinam aprendizado profundo com métodos de explicabilidade, unindo precisão preditiva e transparência, especialmente em contextos que exigem confiabilidade e interpretabilidade dos resultados.

Dessa forma, este estudo contribui com evidências relevantes sobre o desempenho de modelos de previsão aplicados à segurança da informação, ao mesmo tempo em que abre caminhos promissores para avanços na modelagem de riscos e no apoio à tomada de decisão baseada em dados no campo da cibersegurança.

Referências

Africk, E. and Levy, Y. (2021). An examination of historic data breach incidents: What cybersecurity big data visualization and analytics can tell us? *Online Journal of Applied Knowledge Management (OJAKM)*, 9(1):31–45.

Ahmed, S., Nielsen, I. E., Tripathi, A., Siddiqui, S., Ramachandran, R. P., and Rasool, G. (2023). Transformers in time-series analysis: A tutorial. *Circuits, Systems, and Signal Processing*, 42(12):7433–7466.

Alahmari, A. and Duncan, B. (2020). Cybersecurity risk management in small and medium-sized enterprises: A systematic review of recent evidence. In *2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA)*, pages 1–5. IEEE.

Almulihi, A. H., Alassery, F., Khan, A. I., Shukla, S., Gupta, B. K., and Kumar, R. (2022). Analyzing the implications of healthcare data breaches through computational technique. *Intelligent Automation & Soft Computing*, 32(3).

Avanzi, B., Eling, M., Hurley, M., and Schanz, K.-U. (2025). On the evolution of data breach reporting patterns and frequency in the united states: A cross-state analysis. *North American Actuarial Journal*, pages 1–32.

Barati, M. and Yankson, B. (2022). Predicting the occurrence of a data breach. *International Journal of Information Management Data Insights*, 2(2):100128.

Carfora, M. F. and Orlando, A. (2022). Algumas observações sobre estimativas de distribuição de violações de dados maliciosas e negligentes. *Computação*, 10(208).

Duggineni, S. (2023). Impact of controls on data integrity and information systems. *Science and Technology*, 13(2):29–35.

Foerderer, J. and Schuetz, S. W. (2022). Data breach announcements and stock market reactions: a matter of timing? *Management Science*, 68(10):7298–7322.

Gong, X., Chen, Y., Wang, Q., Wang, M., and Li, S. (2022). Private data inference attacks against cloud: Model, technologies, and research directions. *IEEE Communications Magazine*, 60(9):46–52.

IBM (2024). Cost of a data breach report 2024. Accessed: 2025-04-21.

Janiesch, C., Zschech, P., and Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, 31(3):685–695.

Kumar, S., Kaur, A., and Kumar, R. (2023). A hybrid oversampling approach to deal with data imbalance and outliers for credit card fraud detection. *Applications in Computing and Mathematics for Engineering*, 2(1):100004.

Lewis, C. D. (1982). *Métodos de Previsão Industrial e Empresarial: Um Guia Prático para Suavização Exponencial e Ajuste de Curvas*. Butterworth Scientific, Oxford, Reino Unido. [Google Scholar].

Mangku, D. G. S., Yuliartini, N. P. R., Suastika, I. G. N., and Wirawan, I. G. M. A. S. (2021). The personal data protection of internet users in indonesia. *Journal of Southwest Jiaotong University*, 56(1):202–209.

Partners, M. (2022). Detecting trends and mean reversion with the hurst exponent. Acesso em: 7 abr. 2025.

Perera, S., Jin, X., Maurushat, A., and Opoku, D. G. J. (2022). Factors affecting reputational damage to organisations due to cyberattacks. *Informatics*, 9(1):28.

Pimenta Rodrigues, G. A., Marques Serrano, A. L., Lopes Espiñeira Lemos, A. N., Canedo, E. D., Mendonça, F. L. L. D., de Oliveira Albuquerque, R., and García Villalba, L. J. (2024). Understanding data breach from a global perspective: Incident visualization and data protection law review. *Data*, 9(2):27.

Privacy Rights Clearinghouse (2025). Privacy rights clearinghouse: Chronology of data breaches. Accessed: 2025-04-21.

Silveira, M., Portela, A., Souza, M., Silva, D., Mesquita, M., Silva, D., Menezes, R., and Gomes, R. (2023). Aplicação de técnicas de encriptação e anonimização em nuvem para proteção de dados. In *Anais do XXIII Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais (SBSEG 2023)*, pages 111–124, Porto Alegre. SBC.

Sun, H., Xu, M., and Zhao, P. (2020). Modeling malicious hacking data breach risks. *North American Actuarial Journal*, 25(4):484–502.

Varshney, S., Munjal, D., Bhattacharya, O., Saboo, S., and Aggarwal, N. (2020). Big data privacy breach prevention strategies. In *2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC)*, pages 1–6. IEEE.

Yu, J., Moon, H., Chua, B.-L., and Han, H. (2022). Hotel data privacy: strategies to reduce customers' emotional violations, privacy concerns, and switching intention. *Journal of Travel & Tourism Marketing*, 39(2):213–225.