UNIVERSIDADE FEDERAL DO MARANHÃO

Programa de Pós-Graduação em Ciência da Computação

Alisson Emanuel Goes de Mendonça

Modelo de Aprendizagem de Máquina Segmentado para Previsão Analítica de Arrecadação Fiscal Baseada em Informações de Notas Fiscais Eletrônicas

São Luís
2024

Alisson Emanuel Goes de Mendonça

Modelo de Aprendizagem de Máquina Segmentado para Previsão Analítica de Arrecadação Fiscal Baseada em Informações de Notas Fiscais Eletrônicas

Dissertação apresentada como requisito parcial para obtenção do título de Mestre em Ciência da Computação, ao Programa de Pós-Graduação em Ciência da Computação, da Universidade Federal do Maranhão.

Programa de Pós-Graduação em Ciência da Computação
Universidade Federal do Maranhão

Orientador: Prof. Dr. Luciano Reis Coutinho

São Luís - MA
2024

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Diretoria Integrada de Bibliotecas/UFMA

Mendonça, Alisson Emanuel Goes de.
Modelo de Aprendizagem de Máquina Segmentado Para Previsão Analítica de Arrecadação Fiscal Baseada Em Informações de Notas Fiscais Eletrônicas / Alisson Emanuel Goes de Mendonça. - 2024.
69 p.

Orientador(a): Luciano Reis Coutinho.
Dissertação (Mestrado) - Programa de Pós-graduação em Ciência da Computação/ccet, Universidade Federal do Maranhão, Universidade Federal do Maranhão, 2024.

1. Regressão. 2. Cnn. 3. Mlp. 4. Aprendizagem Segmentada. 5. Previsão de Receita Fiscal. I. Coutinho, Luciano Reis. II. Título.

Alisson Emanuel Goes de Mendonça

Modelo de Aprendizagem de Máquina Segmentado para Previsão Analítica de Arrecadação Fiscal Baseada em Informações de Notas Fiscais Eletrônicas

Dissertação apresentada como requisito parcial para obtenção do título de Mestre em Ciência da Computação, ao Programa de Pós-Graduação em Ciência da Computação, da Universidade Federal do Maranhão.

Trabalho aprovado. São Luís - MA, 02 de agosto de 2024:

Prof. Dr. Luciano Reis Coutinho
Orientador
Universidade Federal do Maranhão

Prof. Dr. Francisco José da Silva e Silva
Examinador Interno
Universidade Federal do Maranhão

Prof. Dr. Evandro de Barros Costa
Examinador Externo
Universidade Federal de Alagoas

São Luís - MA
2024

Dedico este trabalho a todos aqueles que corajosamente se dedicam a enfrentar os desafios de uma pesquisa científica, buscando incessantemente a evolução do conhecimento em prol da sociedade.

Agradecimentos

Agradeço, em primeiro lugar, a Deus, pela sabedoria, orientação e força que me concedeu ao longo desta jornada de mestrado. Sem Sua graça e misericórdia, não teria sido possível alcançar este marco em minha vida profissional.

À minha amada esposa e meus preciosos filhos, agradeço o apoio incondicional, compreensão e paciência durante todos esses meses de dedicação intensa à pesquisa e aos estudos. Suas palavras de encorajamento e gestos de carinho foram fundamentais para que eu superasse os desafios e me mantivesse motivado em todos os momentos.

Quero expressar minha profunda gratidão aos meus pais e demais familiares, cujo amor, incentivo e apoio constantes têm sido uma fonte inesgotável de inspiração. Sou muito grato por cada palavra de estímulo, cada gesto de confiança e cada oração dedicada a mim.

Aos meus estimados professores, em especial ao meu orientador Luciano Reis Coutinho, meu mais sincero agradecimento. Vocês foram pilares fundamentais neste percurso de aprendizado e crescimento. Suas orientações especializadas, conhecimento compartilhado e desafios intelectuais contribuíram imensamente para o meu desenvolvimento acadêmico e profissional.

Agradeço também aos meus amigos e colegas, que estiveram presentes nesta caminhada. Suas conversas, debates e trocas de experiências enriqueceram meu conhecimento e me fizeram perceber a importância do trabalho em equipe e da colaboração mútua.

Por fim, expresso minha gratidão a todas as pessoas que, direta ou indiretamente, contribuíram para a realização deste trabalho. Seja por meio de apoio emocional, revisões, sugestões valiosas ou simplesmente por acreditarem em mim, cada um de vocês teve um papel importante nessa conquista.

"A ciência é um esforço colaborativo, onde cada pequena contribuição se soma para formar um imenso edifício do conhecimento."

(Neil deGrasse Tyson)

Resumo

No Brasil, o Imposto sobre Operações relativas à Circulação de Mercadorias e sobre Prestações de Serviços de Transporte Interestadual e Intermunicipal e de Comunicação, conhecido pela sigla ICMS, tem alta representatividade na receita dos entes subnacionais, cerca de 90%. Por ser um imposto sobre o consumo, seu valor está diretamente relacionado à atividade econômica, cujas informações são registradas em notas fiscais eletrônicas emitidas pelos contribuintes e recebidas pelos órgãos fiscalizadores de cada estado. Este trabalho propõe um modelo de aprendizagem para prever a arrecadação de ICMS por meio de um conjunto de dados derivados de informações fiscais em notas fiscais eletrônicas. O modelo de aprendizagem usa uma abordagem segmentada que começa com a divisão dos conjuntos de dados de treinamento e validação em diversos subconjuntos menores de acordo com um determinado parâmetro de segmentação. Depois disso, o modelo ajusta vários algoritmos para cada subconjunto dividido (segmento). Finalmente, o modelo seleciona o algoritmo de aprendizado de máquina adequado (instância de aprendizagem) que produz o melhor resultado preditivo em cada segmento. Essas instâncias de aprendizagem selecionadas compõem um conjunto de instâncias híbridas para prever os registros de um conjunto de dados de teste. Ao se comparar os resultados da abordagem tradicional sem segmentação das bases de treinamento e validação com os do modelo proposto, obteve-se uma melhora na métrica utilizada de 29,52% e 18,4% para os anos de 2021 e 2022, respectivamente. E ao se comparar com os resultados do modelo atual do órgão tributário que apoiou esta pesquisa, o modelo proposto melhorou a métrica utilizada em 60,34% e 51,9 % para os anos de 2021 e 2022, respetivamente. A melhora na métrica de assertividade utilizada sugere que o modelo proposto é promissor no fornecimento de informações para apoiar a tomada de decisões na gestão pública. Além disso, o modelo pode auxiliar em soluções para combater a evasão fiscal, analisar impactos resultantes de mudanças no sistema tributário e enfrentar outros desafios relacionados.

Palavras-chave: regressão, CNN, MLP, aprendizagem segmentada, previsão de receita fiscal

Abstract

In Brazil, the Tax on Operations Related to the Circulation of Goods and the Rendering of Interstate and Intermunicipal Transportation and Communication Services, known by the acronym ICMS, holds significant prominence in the revenue of the federative units, accounting for approximately 90%. As a consumption tax, its value depends on economic activity, whose tax information the taxpayers record in electronic invoices issued to the tax agencies of each federative unit. This work proposes a learning model to predict ICMS revenue through a dataset derived from tax information in electronic invoices. The learning model uses a segmented approach that starts with splitting the training and validation datasets according to a given parameter. After that, the architecture fits several machine learning models for each split subset (segment). Finally, the architecture chooses the fit machine learning model (learning instance) that produces the best prediction result for each segment. These learning instances compose a hybrid instance set to predict the records of a test dataset. When comparing the results of the traditional approach without segmentation of the training and validation bases with those of the proposed model, an improvement in the metrics used was obtained by 29.52% and 18.4% for the years 2021 and 2022, respectively. And when comparing the results of the current model of the tax body that supported this research, the proposed model improved the metrics by 60.34% and 51.9% for the years 2021 and 2022, respectively. The improvement in the assertiveness metric used suggests that the model is promising in providing information to support decision-making in public management. Furthermore, the model can assist in solutions to combat tax evasion, analyze impacts resulting from changes in the tax system, and address other related challenges.

Keywords: regression, CNN, MLP, segmented learning, tax revenue forecasting.

Lista de Ilustrações

Lista de Tabelas

Lista de abreviaturas e siglas

ANN	*Artificial Neural Network*
CNN	*Convolutional Neural Networks*
CNPJ	Cadastro Nacional de Pessoas Jurídicas
HIS	*Hybrid Instance Set*
IA	Inteligência Artificial
ICMS	Imposto sobre Circulação de Mercadorias e Serviços de Transporte Interestadual e Intermunicipal e de Comunicação
LR	*Linear Regression*
LSTM	*Long Short-Term Memory*
MAE	*Mean Absolute Error*
MLP	*Multilayer Perceptron*
NFCE	Nota Fiscal de Consumidor Eletrônica
NFE	Nota Fiscal Eletrônica
NSLI	*Non-Segmented Learn Instance*
RBF	*Radial Basis Function*
RF	*Random Forest*
RNA	Rede Neural Artificial

SEFAZ-MA	Secretaria de Estado da Fazenda do Maranhão
SLI	*Segmented Learn Instance*
SVM	*Support Vector Machine*
SVR	*Support Vector Regression*
UF	Unidade Federativa

Sumário

1 Introdução

No Brasil, as administrações públicas em todos os níveis de governo devem obedecer aos princípios e regras que definem limites e estabelecem formalidades na elaboração do orçamento público, que é o planejamento para cumprir, durante determinado período, os planos e programas de trabalho a serem desenvolvidos. Em resumo, este plano é feito através de uma previsão de receitas e, a partir destas, da definição das despesas a realizar. Assim como superávit de receita permite o aumento das despesas, o déficit impõe ao administrador a preocupação de contingenciar os gastos e, obviamente, quanto mais cedo se tem a informação, melhor será o planejamento para a execução orçamentária.

Em termos de representatividade na arrecadação das unidades federativas (UF) do Brasil, a principal fonte é o imposto sobre as operações relativas à circulação de mercadorias, sobre a prestação de serviços de transporte interestadual e intermunicipal e de comunicação, chamado de ICMS. Em alguns estados, como no caso do Maranhão, esse imposto responde por aproximadamente 90% da arrecadação mensal. Pela sua relevância para a autonomia estadual, a legislação tributária impõe aos contribuintes do ICMS a obrigação de prestar informações aos órgãos fiscalizadores governamentais por meio da emissão de notas fiscais. Essas informações, disponibilizadas em formato eletrônico, são recebidas no mês em que ocorre a comercialização do produto ou serviço, e servem para apurar, ao final de um período mensal, o valor a ser pago no vigésimo dia do mês subsequente, conforme sistemática de fixada na lei.

Embora existam regras para estimar a receita, esse processo ocorre antes dos acontecimentos econômicos. Considerando que eventos imprevisíveis podem afetar fortemente a economia, como foi o caso durante a crise da COVID-19, revisar a informação de receita estimada após a realização das operações econômicas de cada período mensal fornecerá uma base mais assertiva para o gestor público que executa a despesa. Para isso, a órgão fazendário responsável pela arrecadação dos tributos adota um modelo estatístico para reavaliar a estimativa mensal de receitas, considerando as operações realizadas no mês anterior.

1.1 Objetivos

Nesse contexto, esta pesquisa idealiza um modelo de aprendizagem de máquina para estimar a receita com base em informações fiscais dos contribuintes que estão nas notas fiscais eletrônicas emitidas. O modelo é denominado Conjunto de Instâncias Híbridas (*HIS - Hybrid Instance Set*).

1.1.1 Objetivos Específicos

Esse trabalho pretende avaliar a capacidade de os modelos de aprendizagem de máquina se adequarem a heterogeneidade de dados utilizando bases de dados que abarquem informações fiscais de contribuintes de diversos nichos econômicos, seguindo a abordagem tradicional de treinamento.

Também pretende que o modelo de aprendizagem proposto consiga melhorar a precisão preditiva em comparação à abordagem tradicional não segmentada dos modelos de aprendizado de máquina, e ao modelo atualmente utilizado pela Secretaria da Fazenda do Maranhão (SEFAZ-MA).

Como uma melhoria em relação às abordagens atuais na literatura, o modelo visa não apenas um baixo desvio entre a arrecadação total estimada e a efetivamente realizada, mas também alcançar esse baixo desvio para cada contribuinte em cada período de apuração mensal. Portanto, esse modelo de aprendizagem possibilita uma visão analítica das informações resultantes da previsão, com a indicação das variações positivas e negativas de receita para cada contribuinte de ICMS em uma janela temporal customizável.

1.1.2 Hipóteses da Pesquisa

O modelo baseia-se em duas hipóteses para melhorar a métrica de assertividade da receita estimada. A primeira hipótese envolve a divisão do conjunto de dados de treinamento e validação em clusters (segmentos) de informações fiscais geradas pelos contribuintes que atendam a critérios específicos de similaridade. Com base nisso, a hipótese é que treinar um modelo de aprendizagem em cada segmento resultará em previsões mais precisas. A segunda hipótese sugere que aplicar diferentes modelos de aprendizagem para o mesmo segmento e selecionar o modelo de aprendizagem que alcance o melhor resultado preditivo pode permitir uma melhor adaptação às peculiaridades do segmento e melhorar a precisão da estimativa final.

Adicionalmente, essas duas hipóteses foram testadas em cenários para identificar a influência da extensão temporal do conjunto de dados de treinamento e a influência da proximidade temporal do conjunto de dados de treinamento e validação ao momento de predição.

1.2 Contribuições

O presente estudo busca criar uma ferramenta para estimar as receitas fiscais do ICMS, considerando a atividade econômica que efetivamente servirá como base de apuração dos contribuintes desse tributo, servindo como instrumento de apoio a tomada de decisão do gestor público. Dessa forma, pode-se elencar como importantes contribuições:

- Construção de um modelo que, por se basear na atividade econômica que serve de base para a apuração da receita de ICMS, é capaz de se adequar à situações excepcionais que afetam a economia, como foi o caso durante a COVID;
- Possibilidade de análise da representatividade de cada contribuinte na arrecadação, e a respectiva variação na linha do tempo;
- Estrutura um conjunto de dados com alta coesão para a formação do montante de arrecadação, de acordo com a atual sistemática estabelecida na legislação tributária;
- Escalabilidade para outros nichos de análise em um contexto fiscal e em contextos similares.

1.3 Organização do Texto

Este trabalho está estruturado da seguinte forma:

- O capítulo 2 introduz uma fundamentação teórica dos algoritmos e técnicas utilizados na pesquisa, além de contextualizar a sistemática de apuração do ICMS.
- O capítulo 3 lista trabalhos relacionados com o tema de predição de receita fiscal, estrutura os modelos mais utilizados e diferencia alguns aspectos considerados nessa pesquisa.

- O capítulo 4 descreve as etapas da metodologia, que inicia com a extração e transformação, segmentação dos dados, aprendizagem segmentada, seleção das instâncias de aprendizagem e, por fim, o processo de predição.
- O capítulo 5 discorre acerca dos resultados obtidos, comparando-os aos resultados de outros modelos, inclusive o atualmente utilizado no órgão fazendário.
- O capítulo 6 apresenta as conclusões e algumas considerações finais relacionadas a limitações do modelo e possibilidade de aplicações em trabalhos futuros.

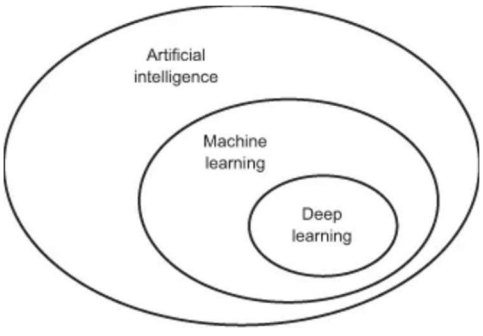
2 Fundamentação Teórica

O foco central deste trabalho é, em síntese, fazer uma análise de regressão entre as receitas fiscais do tributo ICMS e as informações econômicas atreladas ao fato gerador de cada operação realizada pelos contribuintes desse imposto. Sem a pretensão de aprofundar, mas com o objetivo de esclarecer e embasar as discussões que serão trazidas ao longo deste trabalho, esta seção abordará conceitos relacionados às técnicas de aprendizagem de máquina utilizadas no processo de regressão, quais sejam*: Linear Regression* (LR) com transformação polinomial; *Support Vector Machine* (SVM) com kernel *Radial Basis Function* (RBF); *Random Forest* (RF); *Artificial Neural Network – Multilayer Perceptron* (MLP); e *Convolutional Neural Network* (CNN). Em seguida, far-se-á a contextualização necessária para o entendimento sobre as regras que regem a sistemática de apuração do ICMS.

2.1 Aprendizagem de máquina

Antes efetivamente conceituar os algoritmos utilizados, traremos breve explanação de três termos que são bastante citados nos tempos atuais: inteligência artificial (em inglês, artificial intelligence), aprendizagem de máquina (em inglês, machine learning) e aprendizado profundo (em inglês, deep learning). A figura 1 traz uma representação visual que nos permite uma ideia da abrangência de cada termo:

Figura 1. Inteligência artificial, aprendizagem de máquina e aprendizado profundo



Fonte: Adaptado de (Chollet, 2017, 1.1. Artificial intelligence, machine learning, and deep learning section)

A Inteligência Artificial é um campo geral que abrange a aprendizagem automática e a aprendizagem profunda, mas que também inclui muitas outras

abordagens que não envolvem qualquer aprendizagem. Durante bastante tempo, muitos especialistas acreditaram que a inteligência artificial de nível humano poderia ser alcançada fazendo com que os programadores criassem à mão um conjunto suficientemente grande de regras explícitas para a manipulação do conhecimento. Esta abordagem é conhecida como IA simbólica e foi o paradigma dominante na IA desde a década de 1950 até ao final da década de 1980. Embora a IA simbólica tenha se mostrado adequada para resolver problemas lógicos e bem definidos, revelou-se difícil descobrir regras explícitas para resolver problemas mais complexos e confusos, como classificação de imagens, reconhecimento de fala e tradução de idiomas. Uma nova abordagem surgiu para ocupar o lugar da IA simbólica: o aprendizado de máquina (Chollet, 2017, 1.1.1. Artificial intelligence section).

O aprendizado de máquina introduziu um novo paradigma de programação. Na programação clássica, o paradigma da IA simbólica, os seres humanos inserem regras (um programa) e dados a serem processados de acordo com essas regras, e produzem respostas. Com o aprendizado de máquina, os humanos inserem dados, bem como as respostas esperadas dos dados, e criam as regras. Estas regras podem então ser aplicadas a novos dados para produzir respostas originais (Chollet, 2017, 1.1.2. Machine learning section).

O aprendizado profundo é um subcampo específico do aprendizado de máquina: uma nova abordagem no aprendizado de representações a partir de dados que enfatiza o aprendizado de camadas sucessivas de representações cada vez mais significativas. A profundidade na aprendizagem profunda não é uma referência a qualquer tipo de compreensão mais profunda alcançada pela abordagem; em vez disso, representa esta ideia de camadas sucessivas de representações (Chollet, 2017, 1.1.4. The “deep” in deep learning section).

Como já citado, a aplicação da aprendizagem de máquina no caso descrito nessa pesquisa servirá para ajustar uma modelo de regressão que melhor represente a arrecadação fiscal decorrente das operações econômicas declaradas nas notas fiscais emitidas em cada período de apuração, ou seja, faremos uma análise de regressão, que é uma técnica de investigação que utiliza modelagem estatística para representar a relação entre variáveis (Montgomery, Peck and Vining, 2021, 1).

Convém definir brevemente o conceito de modelo, que é uma representação simplificada de sistemas, objetos ou teorias que nos permite entender melhor as coisas.

Segundo Martin (2022,2) há três tipos de modelos: visual, determinístico e estatístico. Um clássico exemplo de modelo visual é a representação de um imóvel em uma planta-baixa feita por um engenheiro, que servirá para seu cliente imaginar como a construção ficará após ser concluída. Outros modelos são representados na forma de equações matemáticas, como por exemplo a equação que mede a pressão em uma dada profundidade oceânica.

Modelos estatísticos também são representados na forma de equações matemáticas, assim como modelos determinísticos. A diferença entre os dois é como eles lidam com as diferenças entre o que o modelo prevê sobre a realidade e as observações da própria realidade. Tal modelo é chamado de determinístico porque, de acordo com o modelo, a profundidade determina a pressão com precisão. Em contraste, os modelos estatísticos são usados em situações em que há uma incerteza considerável sobre a precisão das previsões do modelo. Este é quase sempre o caso nas ciências sociais, porque os humanos, e as sociedades que constroem, são complexos, complicados e não previsíveis com tanta precisão como alguns fenômenos naturais, como a relação entre a profundidade e a pressão da água (Martin, 2022, 3).

2.1.1 Linear Regression (LR)

A regressão linear serve para traçar a reta que melhor representa os dados de um dado conjunto. Essa reta é chamada de linha de regressão e correlaciona uma ou mais variáveis independentes (também referenciadas como preditivas ou de regressão) com uma variável dependente (também referenciada como variável de resposta ou alvo) (Yan and Su, 2009).

O modelo de regressão linear pode ser de dois tipos a depender da quantidade de variáveis independentes que a equação representa. Na regressão linear múltipla, a equação matemática representa duas ou mais variáveis independentes, conforme a seguir:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon,$$

onde:

- y é a variável dependente.
- β_0 é o intercepto da linha com o eixo y.
- $\beta_1,\ \beta_2,\ \ldots,\ \beta_n$ são os coeficientes que representam a inclinação da linha em relação a cada variável independente $x_1,\ x_2, \ldots, x_n$

- $x_1,\ x_2, \ldots, x_n$ são as variáveis independentes.
- ε é o termo de erro ou ruído que representa a variação nos dados que não pode ser explicada pelas variáveis independentes.

A regressão linear simples envolve apenas uma variável independente, ou seja, sua equação é simplificada para:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

O modelo de regressão linear possui algumas premissas (Martin, 2022, 52), quais sejam:

1. Linearidade: A relação entre o preditor e o resultado é linear. Se esta suposição não for atendida, o modelo linear representa erroneamente a verdadeira forma do relacionamento entre o preditor e o resultado.
2. Normalidade dos erros: Os erros seguem uma distribuição normal. Se os erros não forem distribuídos normalmente, os testes de hipóteses e os intervalos de confiança sobre os coeficientes e previsões do modelo podem produzir resultados enganosos.
3. Homocedasticidade dos erros: A variância dos erros em torno da linha de regressão é a mesma em todos os pontos da linha de regressão.
4. Independência dos erros: Os erros são independentes um do outro. Se houver dependência entre alguns ou todos os erros, os testes de hipóteses e os intervalos de confiança em torno dos coeficientes e previsões do modelo podem produzir resultados enganosos.
5. Aleatoriedade dos erros: Os erros são o resultado de um processo aleatório. Se os erros não forem de facto aleatórios, então as inferências da amostra para uma população ou processo podem ser ilusórias
6. O preditor é medido sem erros: o modelo de regressão linear não tem termo de erro para a variável preditora e, portanto, assume implicitamente que a variável preditora não tem erro de medição
7. Ausência de observações extremamente influentes: Não há valores discrepantes extremos nos dados que distorçam a relação observada entre o preditor e o resultado. Se esta suposição não for satisfeita, as estimativas dos coeficientes, bem como os testes de hipóteses e os intervalos de confiança, podem ser enganosos.

Para melhor ajustar o modelo de regressão linear às relações não-lineares entre as variáveis independentes e a variável independente, utilizou-se a transformação polinomial. Uma transformação polinomial envolve adicionar termos polinomiais das variáveis independentes ao conjunto de dados. Por exemplo, para uma variável independente x, uma transformação polinomial de grau d adiciona x^d ao modelo:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_d x^d + \varepsilon$$

2.1.2 Support Vector Machine (SVM)

Support Vector Machine (SVM) é uma técnica de aprendizado de máquina utilizada principalmente para problemas de classificação, embora também podem ser empregados com eficiência para problemas de regressão (Bonaccorso, 2018, Support Vector Regression section), como é o caso desta pesquisa.

O SVM pode ser de dois tipos: Linear ou com Kernel. O primeiro tipo é utilizado no caso de dados linearmente separáveis, quando um único hiperplano pode separar claramente as duas classes. O segundo tipo é utilizado quando os dados não são linearmente separáveis, caso em que o modelo mapeia os dados para um espaço de alta dimensão usando uma função de kernel para encontrar um hiperplano que possa separar as classes no novo espaço transformado. O dataset dessa pesquisa enquadra-se nas circunstâncias atendidas pelo SVM com kernel, onde foi utilizada a função de base radial (Radial Basis Function – RBF), por isso, vamos adentrar um pouco no detalhamento dessa técnica.

A RBF é baseada na seguinte função:

$$K\left(\bar{x}_i, \bar{x}_j\right) = e^{-\gamma|\bar{x}_i - \bar{x}_j|^2}$$

O parâmetro γ determina a amplitude da função, que não é influenciada pela direção, mas apenas pela distância da origem. Este kernel é particularmente útil quando os conjuntos são côncavos e se cruzam, por exemplo, quando um subconjunto pertencente a uma classe está rodeado por outro pertencente a outra classe (Bonaccorso, 2018, Radial Basis Function section).

Diferentemente do modelo de regressão linear, nesse caso não houve transformação polinomial pois o kernel RBF já transforma os dados de forma não-linear de maneira eficaz.

2.1.3 Random Forest (RF)

Antes de introduzir o modelo de Random Forest, convém abordar o modelo Decision Trees, uma vez que este serve como base para aquele. Na mineração de dados, uma árvore de decisão é um modelo preditivo que pode ser usado para representar classificadores e modelos de regressão (Rokach and Maimon, 2014, 10). Quando uma árvore de decisão é usada para tarefas de classificação, ela é mais comumente chamada de árvore de classificação. Quando é usado para tarefas de regressão, é chamado de árvore de regressão.

Os modelos de regressão mapeiam o espaço de entrada em um domínho de valor real. Em uma explanação formal, temos que o objetivo é examinar $y|X$ para uma resposta y e um conjunto de preditores X (Rokach and Maimon, 2014, 85).

Adentrando na conceituação do Random Forest, temos que ele é um algoritmo de aprendizado de máquina que pertence à família dos métodos de ensemble. A ideia principal de uma metodologia ensemble é combinar um conjunto de modelos, cada um dos quais resolve a mesma tarefa original, a fim de obter um melhor modelo global composto, com estimativas ou decisões mais precisas e confiáveis do que as que podem ser obtidas usando um único modelo (Polikar ,2006). Dessa forma, a ideia central do modelo Random Forest é combinar múltiplas árvores de decisão para formar uma "floresta" de árvores e fazer previsões mais robustas e precisas.

Ao combinar várias árvores de decisão, o Random Forest tende a ser menos propenso ao *overfitting* do que uma única árvore de decisão, especialmente em conjuntos de dados grandes e complexos. Outra característica importante da floresta aleatória é que ela é rápida (Rokach and Maimon, 2014, 126).

2.1.4 Artificial Neural Network – Multilayer Perceptron (MLP)

As redes neurais artificiais RNA (em inglês ANN) são, como o próprio nome indica, redes computacionais que tentam simular o processo de decisão em redes de células nervosas (neurônios) do sistema nervoso central biológico (Graupe, 2019, 1). Uma RNA ou simplesmente uma rede neural é uma estrutura computacional direcionada ou recorrente que conecta uma camada de entrada a uma de saída. Normalmente, todas as operações são diferenciáveis, e a função vetorial geral pode ser facilmente escrita da seguinte forma:

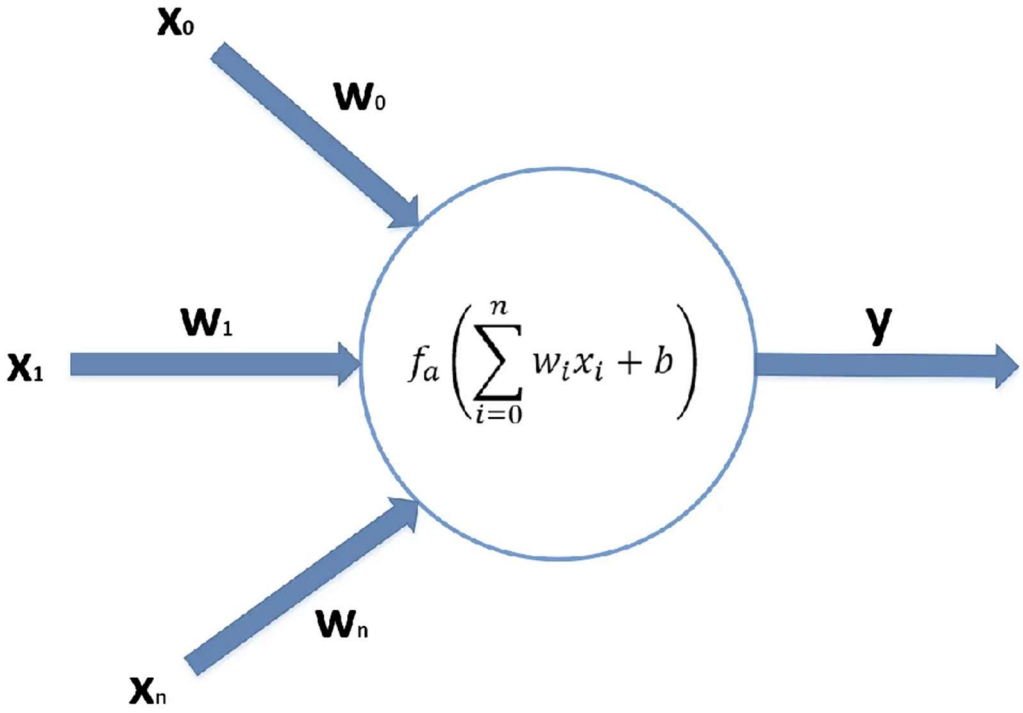
$$\bar{y} = f(\bar{x})$$

A função transforma um vetor em outro aplicando o mesmo operador elemento a elemento. Portanto, assumimos o seguinte:

$$\bar{x} = (x_1, x_2, \dots, x_n) \text{ e } \bar{y} = (y_1, y_2, \dots, y_m)$$

O adjetivo neural provém de dois elementos importantes: a estrutura interna de uma unidade computacional básica e as interconexões entre eles. No diagrama representado na figura 2, há uma representação esquemática de um neurônio artificial.

Figura 2. Estrutura de um neurônio artificial



Fonte: Adaptado de (Bonaccorso, 2018)

A função matemática que descreve um neurônio artificial é dada pela equação:

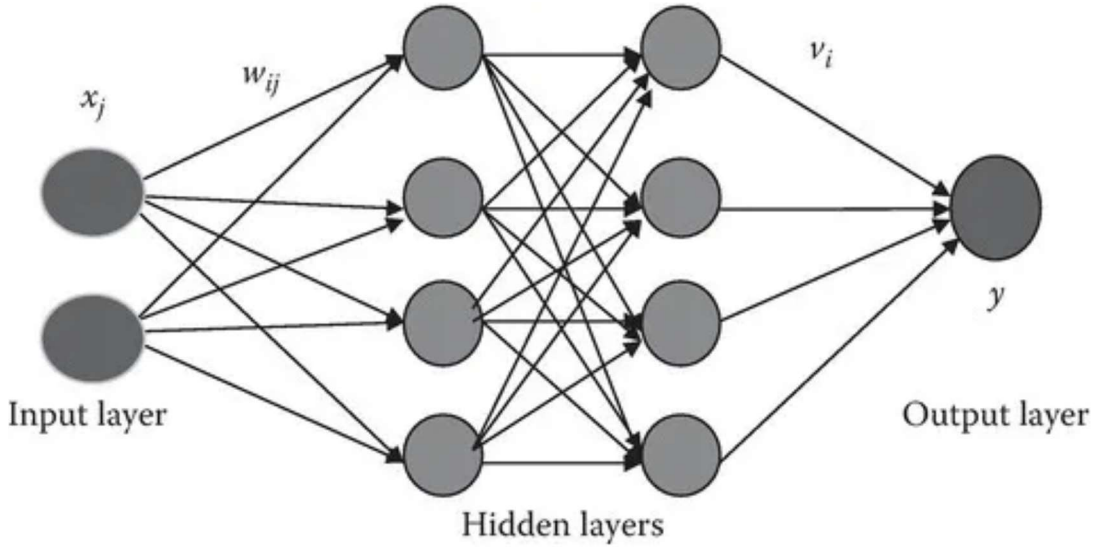
$$y = f_a\left(\sum_{i=0}^{n} w_i x_i + b\right),$$

onde x_i representa um sinal de entrada do neurônio, w_i representa o peso sináptico associado a entrada i, b é o *bias* e f_a é a função de ativação, que produzirá a saída y.

As redes neurais são organizadas em camadas. As camadas são compostas por vários neurônios interconectados que contêm funções de ativação. A RNA processa informações através de neurônios de maneira paralela para resolver problemas específicos. A RNA adquire conhecimento por meio do aprendizado, e esse conhecimento é armazenado na força das conexões interneurais, que é expressa por valores numéricos chamados pesos. Esses pesos são usados para calcular valores de

sinal de saída para um novo valor de sinal de entrada de teste. Os padrões são apresentados à rede através da camada de entrada, que se comunica com uma ou mais camadas ocultas, onde o processamento real é feito através de um sistema de conexões ponderadas. As camadas ocultas então se vinculam a uma camada de saída, onde a resposta é produzida (Chakraverty and Mall, 2017), conforme figura 3.

Figura 3. Estrutura de uma Rede Neural Artificial



Fonte: Adaptado de (Chakraverty and Mall, 2017)

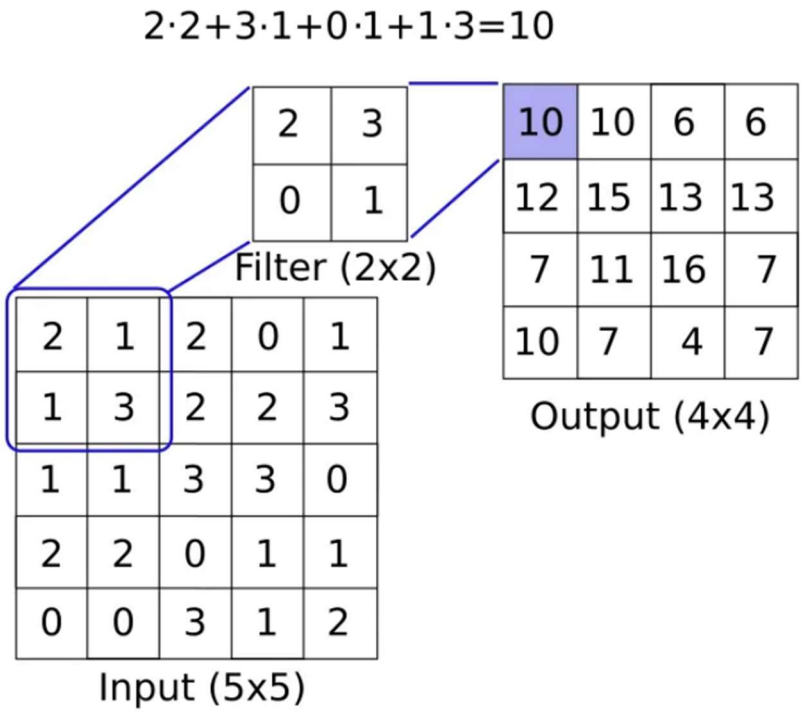
Da figura temos que x_j são nós de entrada, w_{ij} são pesos da camada de entrada para a camada oculta, v_i denotam os pesos da camada oculta para a camada de saída e y o nó de saída.

O processamento de informações em alta velocidade, capacidade de roteamento, tolerância a falhas, adaptabilidade, generalização e robustez são excelentes características das RNAs (Elsheikh and Elaziz, 2022, 1).

2.1.5 Convolutional Neural Network (CNN)

Matematicamente, a convolução é definida como um tipo especializado de operação linear. Na prática, é uma função aplicada à saída de outra função, conforme figura 4. Se analisarmos a CNN, podemos ver que a CNN é semelhante às redes neurais que utilizam operação de convolução (correlação cruzada) entre pesos e entradas no lugar da multiplicação de matrizes em pelo menos uma de suas camadas (McClure, 2017, 8. Convolutional Neural Networks section).

Figura 4. Operação de convolução



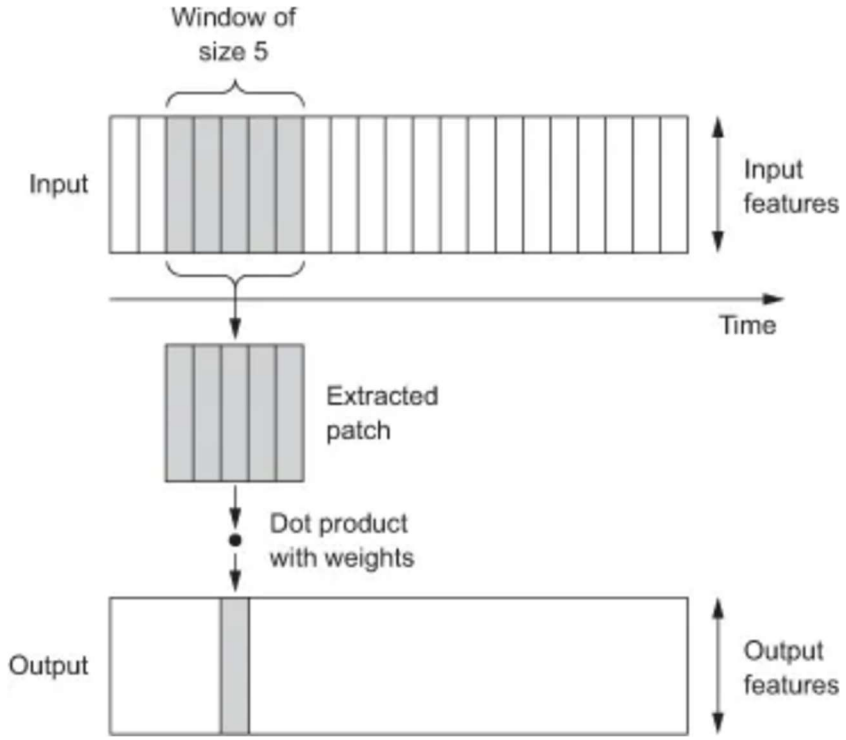
Fonte: adaptado de (McClure, 2017, 8. Convolutional Neural Networks section)

Uma CNN é um tipo de FeedForward Neural Network no qual o padrão de conectividade entre os neurônios dentro e através das camadas é organizado de tal forma que eles podem responder a regiões sobrepostas que ocupam o campo visual (Srinivasan and Laplante, 2023, 166).

A CNN é amplamente utilizada para dados que possuem topologia semelhante a uma grade. Um exemplo inclui dados 1D, como dados de séries temporais que podem ser analisados como uma grade 1D amostrada em intervalos de tempo regulares, e imagens multidimensionais, que podem ser visualizadas como uma grade 2D de pixels empilhados em camadas (Singh, 2020).

É possível usar convoluções 1D, extraindo patches 1D locais (subsequências) de sequências, conforme figura 5. Essas camadas de convolução 1D podem reconhecer padrões locais em uma sequência. Como a mesma transformação de entrada é realizada em cada patch, um padrão aprendido em uma determinada posição em uma frase pode posteriormente ser reconhecido em uma posição diferente, tornando a tradução de convnets 1D invariante (para traduções temporais).

Figura 5. Como funciona a convolução 1D: cada passo de tempo de saída é obtido a partir de um patch temporal na sequência de entrada.



Fonte: adaptado de (Chollet, 2017, 6.4.1. Understanding 1D convolution for sequence data section)

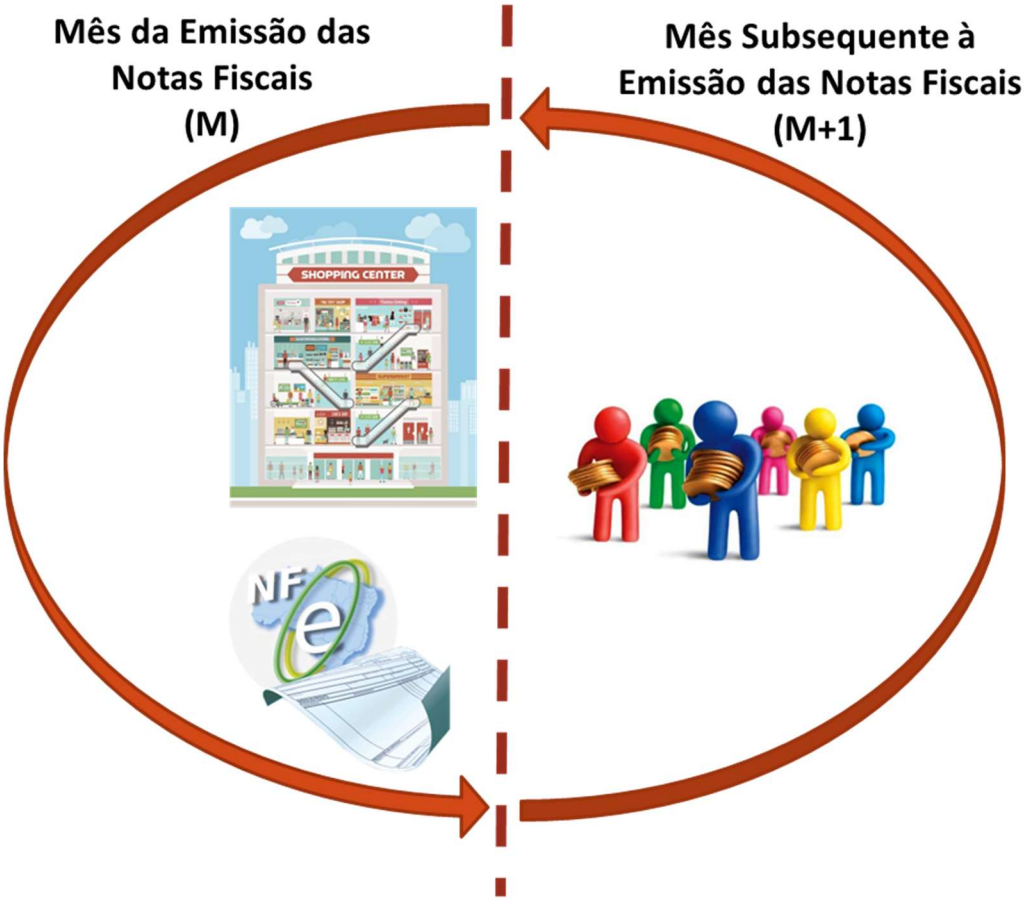
2.2 Introdução ao ICMS

O Estado, enquanto governo de um povo, tem para com estes a obrigação de realizar as prestações positivadas na legislação que rege a vontade coletiva. Para isso, no entanto, são necessárias fontes de receitas para que haja disponibilidades de recursos ao administrador público para a implementação das políticas públicas sociais. Conforme afirma Mello (2022, p.22): "É cristalino que o Estado enquanto instituição tem como uma das fontes de receita a tributação, para que possa sustentar sua administração". Sem a pretensão de aprofundar, mas com o objetivo de fundamentar e servir como base para as discussões que serão trazidas no decorrer desse trabalho, faz-se por diligência uma breve conceituação do direito tributário nas palavras de Harada (2017, p.338): direito tributário é "o direito que disciplina o processo de retirada compulsória, pelo Estado, da parcela de riquezas de seus súditos, mediante a observância dos princípios reveladores do Estado de Direito".

Atualmente, as administrações públicas de quaisquer esferas de poder são regidas por princípios e regras que definem os limites e estabelecem as formalidades na elaboração do orçamento público. Segundo Andrade (2002, p. 37), orçamento público "é a função primordial da gestão pública de estimar as receitas e fixar as despesas". Já para Lima e Castro (2000, p.19), orçamento Público "é o planejamento feito pela Administração Pública para atender, durante determinado período, aos planos e programas de trabalho por ela desenvolvidos, por meio da planificação das receitas a serem obtidas e pelos dispêndios a serem efetuados, objetivando a continuidade e a melhoria quantitativa e qualitativa dos serviços prestados à sociedade.". Observa-se, a partir dos conceitos acima, que os recursos para a executar a administração da coisa pública são estimados, ou seja, podem ocorrer diferenças positivas ou negativas quando da efetiva execução orçamentaria.

A Administração Tributária está inserida no âmbito de todos os entes políticos: União, Estados/DF e municípios. Acerca da competência tributária estadual, as fontes de recursos são os impostos, as taxas e contribuições de melhoria, nos termos do Art. 145 da Constituição da República Federativa do Brasil. Entretanto, em termos de representatividade na arrecadação, a principal fonte é o Imposto Sobre Operações Relativas à Circulação de Mercadorias e Sobre Prestações de Serviços de Transporte Interestadual, Intermunicipal e de Comunicação – ICMS. Em algumas unidades da federação, como no caso do Maranhão, esse imposto compõe mais de 90% da arrecadação mensal. Por ser tão relevante para a autonomia estadual, pela representatividade na arrecadação, a legislação tributária impõe aos contribuintes desse imposto a obrigação de gerar informações para os órgãos de fiscalização do governo. Toda essa informação, predominantemente em meio eletrônico, possibilita o controle fiscal do cumprimento da obrigação de pagar. Convém esclarecer que, conforme figura 6, pela sistemática de apuração do montante ICMS a ser recolhido para o Estado, o valor devido pelas operações realizadas em um determinado mês (M) deve ser repassado, em regra, até o vigésimo dia do mês subsequente (M+1).

Figura 6. Sistemática de apuração do tributo ICMS



3 Trabalhos Relacionados

Os algoritmos de aprendizado de máquina têm diversas aplicações no contexto fiscal. Já existem estudos que avaliam vários modelos e aspectos que contribuem para a melhoria da precisão preditiva da receita tributária, que é o foco desta pesquisa. Para sistematizar a análise dos estudos identificados, dividimos esses estudos em três grupos com base nas características e limitações dos dados preditivos utilizados na fase de aprendizado: aqueles que utilizam séries temporais de receita, seja apenas os valores brutos de receita ou valores estratificados em diferentes tipos de receita; aqueles que combinam indicadores econômicos como o produto interno bruto, taxa de inflação, classificação de crédito, índice de preços ao consumidor, entre outros; e aqueles que utilizam informações fiscais de algum tipo de documento fiscal emitido pelo contribuinte. Esta classificação está resumida na coluna **Dados Preditivos** da Tabela 1. Além disso, esta tabela relata os modelos de aprendizado usados em cada estudo (**Modelos Base**) e a capacidade de fornecer uma visão analítica detalhada da receita estimada para cada contribuinte em cada período mensal (em **Suporte à Visão Analítica**).

Tabela 1. Comparação resumida de artigos

Autor	Local	Dados Preditivos	Modelos Base	Suporte à Visão Analítica
Bayer (2015)	Dinamarca	Série Temporal da Arrecadação	Linear Regression (LR)	Não
Buxton et al. (2019)	Illinois	Informações Fiscais	Multi-Layer Perceptron (MLP) and a Long Short-Term Memory (LSTM)	Não
Hui (2020)	China	Informações Fiscais	Logistics Regression Model and Support Vector Machine Regression Model (SVR)	Não
Kumar et al. (2023b)	-	Informações Fiscais	combination of decision trees, logistic regression and neural networks	Não

Lahiri & Yang (2022)	Nova Iorque	Indicadores Econômicos	several boosting and LASSO algorithms	Não
Liu et al. (2011)	Hebei	Série Temporal da Arrecadação	gray system GM (1, 1) and IOWA operator	Não
Li-Xia et al. (2011)	China	Série Temporal da Arrecadação	support vector machine (SVM) and particle swarm optimization (PSO)	Não
Lu et al. (2009)	China	Série Temporal da Arrecadação	support vector machine with genetic algorithm (GA-SVM)	Não
Noor et al. (2022)	Malásia	Série Temporal da Arrecadação	Feed-forward Neural Network, Random Forest (RF) and Linear Regression	Não
Ticona et al. (2017)	Brasil	Indicadores Econômicos	Genetic Algorithms (GAs) and Neural Networks (NNs)	Não
Xu et al. (2021)	Companhia Chinesa	Informações Fiscais	Support Vector Machine, Random Forest	Não
Zaw et al. (2020)	Mianmar	Série Temporal da Arrecadação	Autoregressive Moving-Average (ARMA)	Não
This Research	Brasil	Informações Fiscais	MLP; SVR; Convolutional Neural Network (CNN); RF; LR	Sim

Os seguintes estudos utilizaram séries temporais de receita tributária na fase de aprendizado. Noor et al. (2022) aplicaram três técnicas de aprendizado para estimar a receita do Governo Federal da Malásia a partir de uma série histórica de 50 anos, com categorias segmentando os valores ao longo do tempo. Neste estudo, o autor relata que a rede neural feed-forward foi a técnica que obteve os melhores resultados. Além disso, no âmbito da administração pública, Zaw et al. (2020) usaram o modelo Auto-Regressive Moving Average (ARMA) em uma série temporal de receita de 2005 a 2019 e relataram resultados mais precisos do que a metodologia utilizada pelo governo. Para prever a receita tributária da Província de Hebei, Liu et al. (2011) usaram um método

de previsão combinado baseado no operador de média ponderada ordenada induzida. Li-xia et al. (2011) desenvolveram um método de previsão baseado na combinação de Máquina de Vetor de Suporte (SVM) e Otimização por Enxame de Partículas (PSO). Lu et al. (2009) usaram um algoritmo genético para selecionar parâmetros adequados para o SVM e, em seguida, estimar a receita bruta de impostos na China.

O uso de séries temporais de receita tributária pode facilitar o processo de coleta de dados, pois a receita, quando não detalhada por contribuinte, é informação pública. No entanto, desconsiderar os fatos econômicos e fiscais usados para calcular o montante do imposto pode levar a desvios significativos. Lahiri & Yang (2022) relataram que os modelos convencionais de previsão de receita são particularmente deficientes durante recessões que normalmente devastam os orçamentos das unidades federativas, como foi o caso durante a pandemia de COVID-19. Portanto, eles usaram indicadores econômicos e receita tributária para revisões em tempo real conforme novas informações chegavam em frequências mensais. Ticona et al. (2017) desenvolveram um modelo híbrido com base em indicadores econômicos e receita tributária que usou um algoritmo genético, combinado com uma rede neural, para estimar a receita fiscal no Brasil. Bayer (2015) destaca outro aspecto importante ao desafiar a regra de que quanto mais longa for a série temporal, melhores e mais precisos serão os resultados. A pesquisa usou uma longa série temporal de receita tributária da Dinamarca e concluiu que estimativas baseadas em séries temporais mais curtas são ligeiramente precisas.

Outros trabalhos relacionados usaram informações fiscais de empresas contribuintes na fase de aprendizado. Kumar et al. (2023b) propuseram um algoritmo chamado Transfer Adaptive Boosting (TAB), capaz de criar um sistema de previsão de valores tributários, mas usaram dados de apenas um contribuinte específico. Em outro trabalho, Kumar et al. (2023a), os autores relatam o uso de Inteligência Artificial (IA) na implementação do Imposto sobre Bens e Serviços, semelhante ao ICMS. O uso de algoritmos de aprendizado supervisionado, como regressão logística, árvores de decisão, máquinas de vetores de suporte e redes neurais, pode identificar com precisão potenciais inadimplentes fiscais, conforme Kumar et al. (2023c). Buxton et al. (2019) descrevem a aplicação de aprendizado de máquina para prever receitas de impostos sobre vendas em dez nichos de mercado no estado de Illinois. Os autores extraíram dados de recibos de vendas e relataram dificuldades e limitações no conjunto de dados.

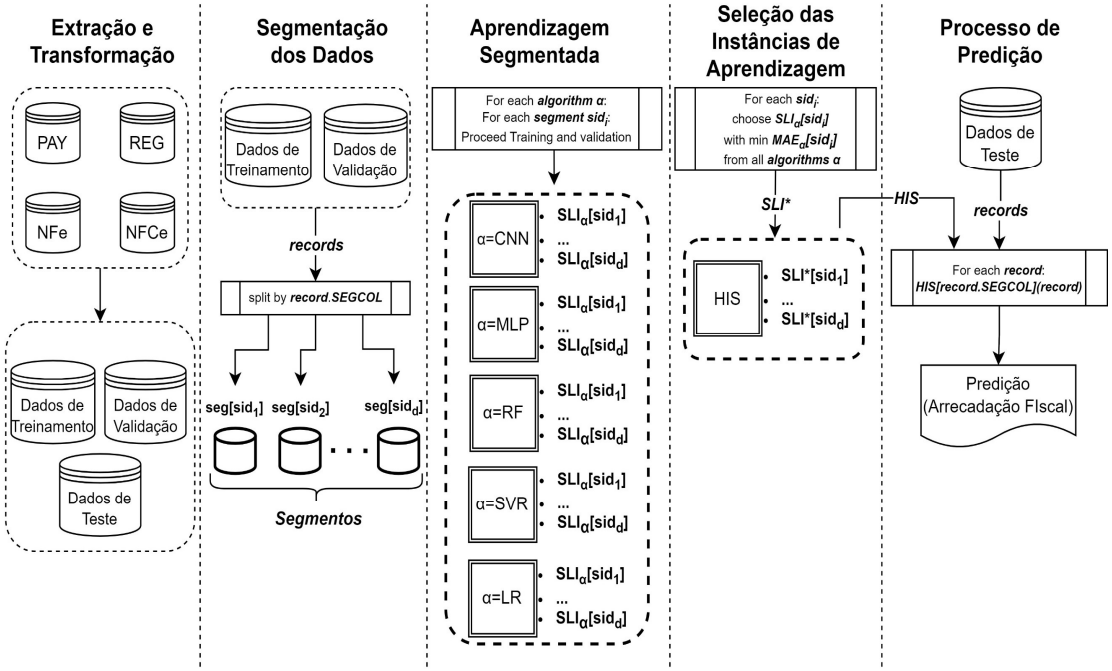
Este estudo compara os resultados produzidos por Perceptron Multicamadas (MLP) e Long Short-Term Memory (LSTM) e conclui que o MLP supera o LSTM. Hui (2020) usou Máquina de Vetores de Suporte (SVM) e Regressão Logística (LR) em um conjunto de dados de apenas uma corporação chinesa e concluiu que ambos os modelos fornecem boas referências para a tomada de decisão; no entanto, o SVM supera a LR. Xu et al. (2021) propuseram um modelo híbrido baseado em Máquina de Vetores de Suporte (SVM) e Floresta Aleatória (Random Forest) para previsão tributária. Eles utilizaram um conjunto de dados público de uma empresa, consistindo em 19 campos, com cada campo contendo 120 pontos de dados. Os autores enfatizam que dados internos específicos da empresa não são totalmente divulgados e que a quantidade de itens de dados é pequena.

Comparado aos trabalhos relacionados, nossa pesquisa apoia uma visão analítica devido à granularidade da previsão, dado que a estimativa da receita tributária ocorre para cada contribuinte e cada período mensal. Além disso, esta pesquisa supera uma limitação destacada em vários estudos que utilizaram informações fiscais: a quantidade e a qualidade dos dados fiscais dos contribuintes. O acesso ao banco de dados da SEFAZ-MA, que contém todas as informações fiscais dos contribuintes no Maranhão, permitiu-nos trabalhar com um conjunto suficiente e atualizado de informações em diversos nichos de mercado, possibilitando que os cenários testados retratassem com maior precisão o desempenho de modelos específicos no contexto fiscal. Finalmente, os experimentos avaliaram o efeito da extensão temporal dos dados de treinamento e como a proximidade temporal dos dados de treinamento e validação em relação aos dados de teste pode impactar a precisão do modelo, problema abordado no trabalho de Bayer (2015).

4 Metodologia

Todo o fluxo do modelo proposto é ilustrado na Figura 7. Ele começa com a etapa de Extração e Transformação, na qual as bases com informações fiscais são preparadas para a criação dos conjuntos de dados de treinamento, validação e teste. Na segunda etapa, Segmentação de Dados, os conjuntos de dados de treinamento e validação são divididos em subconjuntos menores (segmentos) de acordo com um parâmetro categórico definido. Em seguida, na etapa de Aprendizado Segmentado, o processo de aprendizado ajusta vários algoritmos de aprendizado de máquina para cada segmento. Após isso, na Seleção de Instâncias de Aprendizado, o modelo escolhe o algoritmo de aprendizagem (instância de aprendizado) que produz o melhor resultado de previsão para cada segmento. Finalmente, o Processo de Previsão obtém a instância de aprendizado de acordo com o identificador de segmento de cada registro para prever a receita tributária. As seções seguintes fornecerão explicações detalhadas dessas etapas.

Figura 7. Fluxo do modelo de aprendizagem segmentado



4.1 Fontes de dados

A Secretaria da Fazenda do Maranhão (SEFAZ-MA) forneceu as informações fiscais para este estudo. Esses dados são protegidos por sigilo fiscal; portanto, a

pesquisa foi conduzida sob um acordo de confidencialidade que proíbe a divulgação dos dados.

Diversas fontes de dados armazenam informações fiscais. Com base em sua relevância para a receita da unidade federativa, esta pesquisa utilizou quatro subconjuntos das seguintes fontes: Cadastro de Contribuintes (REG), Pagamentos (PAY), Nota Fiscal Eletrônica (NFe) e Nota Fiscal Eletrônica para Consumidor (NFCe).

O banco de dados REG armazena informações relacionadas ao contribuinte e ao setor econômico em que operam. Entre essas informações, a utilizada nesta pesquisa é o número do Cadastro Nacional da Pessoa Jurídica (CNPJ). Por convenção, esse número consiste em 14 dígitos seguindo uma regra de formação específica: os primeiros 8 dígitos (CNPJ RAIZ) identificam o grupo empresarial, que pode ser composto por um ou mais estabelecimentos; os 4 dígitos seguintes, juntamente com o CNPJ RAIZ, identificam um estabelecimento dentro do grupo empresarial; e os últimos 2 dígitos são dígitos verificadores.

O sistema de cálculo estipula que o contribuinte deve calcular o valor do imposto para cada período mensal e pagá-lo até o vigésimo dia do mês subsequente ao cálculo. O banco de dados PAY armazena a data e as informações dos valores de pagamento efetuados por cada estabelecimento para as diversas naturezas fiscais.

O contribuinte fornece informações econômicas e fiscais para cada operação comercial em dois documentos fiscais: a NFe e a NFCe. O primeiro é mais abrangente, distinguindo 6 campos de diferentes naturezas fiscais, embora com mais formalidades no preenchimento. O segundo é mais simples, distinguindo apenas 2 campos de natureza fiscal distinta, e é utilizado para representar transações de menor valor realizadas presencialmente em estabelecimentos de varejo. O banco de dados NFe armazena as notas fiscais eletrônicas, enquanto o banco de dados NFCe armazena as notas fiscais eletrônicas para consumidores.

4.2 Extração e transformação

O processo de criação de um modelo analítico começa com a fase de extração de dados. Nesta etapa, os extratores capturam os dados e realizam limpeza e transformação. A limpeza envolve a remoção de documentos fiscais inválidos com base em critérios que classificam valores como exorbitantes ou eventos que anulam uma

operação específica. A transformação agrega as informações mensalmente de acordo com o sistema de cálculo do ICMS. Como existem sistemas de auditoria fiscal que controlam a entrada de informações na SEFAZ, os dados apresentam poucos problemas estruturais ou de consistência.

Após o processo de extração e transformação do banco de dados REG, o *data mart* para registro resultou em um total de 16.097 estabelecimentos distintos, distribuídos em 995 grupos empresariais. Para o banco de dados PAY, o *data mart* de pagamento resultou em um total de 270.479 registros de pagamento, agregados por contribuinte e período de avaliação mensal. Em relação ao banco de dados NFe, cada campo de natureza fiscal foi agregado pelo contribuinte e período de avaliação mensal. Esses campos selecionados estão diretamente relacionados ao sistema de cálculo do valor a ser pago e geraram o *data mart* de notas fiscais eletrônicas com um total de 580.923 registros representando operações em que os contribuintes estão envolvidos, seja como compradores (213.007) ou vendedores (367.916). Da mesma forma, do banco de dados NFCe, um total de 89.072 registros foram selecionados para compor o *data mart* de notas fiscais eletrônicas para consumidores.

Após a extração, os *data marts* foram unificados. Nesta etapa, o processo gerou uma tabela com informações agregadas por estabelecimento e período mensal, totalizando 435.002 registros na série histórica de janeiro de 2015 a dezembro de 2022.

Para agilizar o tempo de experimentação sem comprometer a confiabilidade dos resultados, limitamos o conjunto de dados a um subconjunto de 70 grupos econômicos com representação significativa na receita fiscal e atividade econômica em vários segmentos de mercado. O subconjunto de dados consiste em 6.201 estabelecimentos distintos e 127.654 registros. Cada registro retrata a receita mensal e os valores econômicos das operações para cada estabelecimento. Convém esclarecer que só estão representados estabelecimentos que realizaram operações econômicas ou efetuaram algum repasse de valores monetários ao fisco, e que há diversos grupos econômicos com grande quantidade de estabelecimentos que operam de maneira eventual, resultando em um total de registros seja inferior a multiplicação da quantidade de estabelecimentos pelo número de meses do período coletado.

A Tabela 2 lista as variáveis e seus respectivos conceitos. Essas variáveis foram selecionadas devido à sua forte correlação com o valor do imposto a ser pago, de acordo com o sistema de cálculo do ICMS estabelecido na legislação tributária.

Tabela 2. Detalhamento das colunas do *dataset*

Variável	Finalidade	Descrição
P1	Preditiva	O valor do imposto resultante da aquisição de bens para comercialização e que serve como crédito para compensar o montante na coluna P2.
P2	Preditiva	O valor do imposto resultante das operações de saída.
P3	Preditiva	O valor do imposto resultante das operações de saída com antecipação de cobrança até o consumo final.
P4	Preditiva	O valor do imposto resultante das operações entre diferentes unidades federativas quando o adquirente não é um contribuinte.
P5	Preditiva	O valor da contribuição destinado ao fundo de combate à pobreza resultante das operações de saída para consumo final.
P6	Preditiva	Antecipação do valor da contribuição destinado ao fundo de combate à pobreza resultante das operações de saída para consumo final.
P7	Preditiva	O valor da contribuição destinado ao fundo de combate à pobreza resultante das operações de saída entre diferentes unidades federativas quando o adquirente não é um contribuinte.
REVENUE	Alvo	Receita paga pelo contribuinte
CNPJ RAIZ	Parâmetro de Segmentação	Identifica o grupo empresarial, que pode incluir 1 ou vários estabelecimentos. É usado para criar segmentos com informações dos estabelecimentos pertencentes ao grupo.
CNPJ	Não usado	Identificador do contribuinte
MES	Delimitação Temporal	Contém o número representando o mês na série histórica. É usado na divisão que forma os conjuntos de dados de treinamento, validação e teste.
ANO	Delimitação Temporal	Contém o número representando o ano na série histórica. É usado na divisão que forma os conjuntos de dados de treinamento, validação e teste.

4.3 Segmentação dos dados

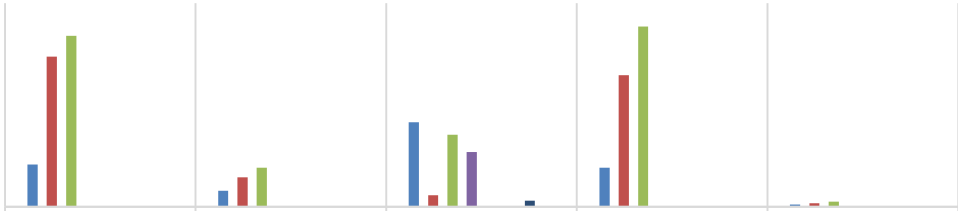
O processo de segmentação antecede a fase de aprendizado e divide os conjuntos de dados de treinamento e validação por meio do agrupamento de registros de acordo com uma coluna categórica informada como parâmetro. Assim, seja $\boldsymbol{DS}$ o conjunto de dados de treinamento ou validação e seja $\boldsymbol{SEGCOL}$ uma determinada coluna categórica. O processo de segmentação divide $\boldsymbol{DS}$ em subconjuntos mutuamente exclusivos $\boldsymbol{seg[sid_i]} \subset \boldsymbol{DS}$ definido como:

$$\boldsymbol{seg[sid_i]} = \{\boldsymbol{record} \in \boldsymbol{DS} | \boldsymbol{record.SEGCOL} = \boldsymbol{sid_i}\},$$

Onde $\boldsymbol{sid_i}$, identificador do segmento, representa os distintos valores de $\boldsymbol{SEGCOL}$, para $\boldsymbol{1 \leq i \leq d}$.

Para os propósitos desta pesquisa, o parâmetro $\boldsymbol{SEGCOL}$ foi o CNPJ RAIZ (Tabela 2). Portanto, considerando que o conjunto de dados possui 70 grupos empresariais, esse número também representa a quantidade d de segmentos distintos. A Figura 8 ilustra os valores (M - milhões; K - milhares) das variáveis preditivas e a Receita agregada para alguns segmentos de diferentes nichos. Nela, é possível observar a heterogeneidade das características tributárias que envolvem a apuração do ICMS, ou seja, não há uma correlação direta entre a arrecadação realizada e as informações de impostos destacados nos documentos fiscais.

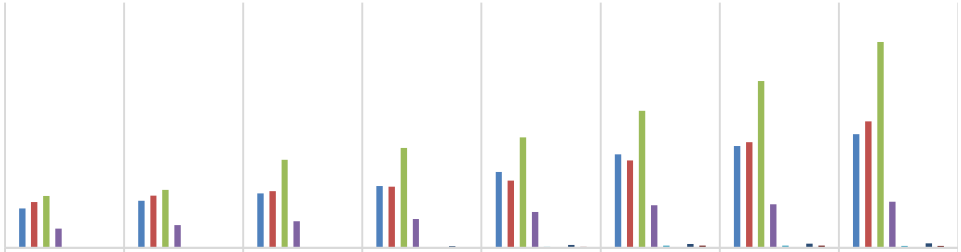
Figura 8. Valores agregados das colunas para 5 segmentos



	seg[51]	seg[08]	seg[12]	seg[25]	seg[55]
REVENUE	R$ 1.245,8M	R$ 466,2M	R$ 2.483,8M	R$ 1.150,0M	R$ 60,6M
P1	R$ 4.426,7M	R$ 862,3M	R$ 336,4M	R$ 3.870,6M	R$ 97,4M
P2	R$ 5.032,6M	R$ 1.144,7M	R$ 2.122,8M	R$ 5.311,6M	R$ 144,6M
P3	0,6K	0,0K	R$ 1.605,9M	R$ 4,9M	0,5K
P4	0,4K	R$ 1,5M	0,0K	0,0K	0,7K
P5	0,0K	0,0K	0,0K	0,0K	0,0K
P6	0,0K	0,0K	R$ 173,6M	0,0K	0,0K
P7	0,4K	R$ 1,5M	0,0K	0,0K	0,7K

A Figura 9 ilustra, para o conjunto dos 70 grupos econômicos selecionados, os valores das variáveis preditivas e a Receita agregada anualmente ao longo da série histórica. Esses valores representam aproximadamente 25% da arrecadação total do Estado do Maranhão.

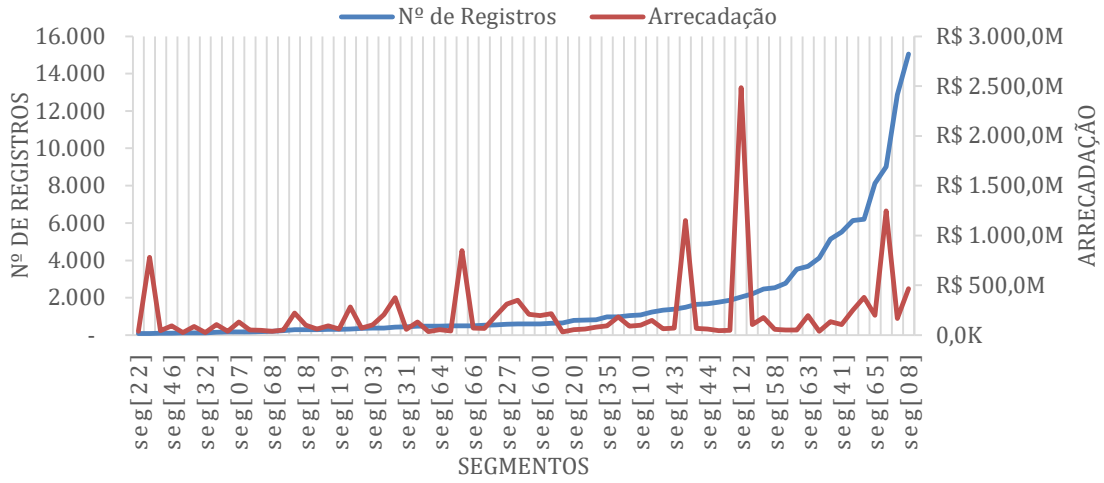
Figura 9. Valores agregados das colunas na série histórica da pesquisa



	2015	2016	2017	2018	2019	2020	2021	2022
REVENUE	R$ 961,8M	R$ 1.150,1M	R$ 1.330,9M	R$ 1.512,1M	R$ 1.857,0M	R$ 2.285,3M	R$ 2.485,8M	R$ 2.779,6M
P1	R$ 1.114,5M	R$ 1.274,5M	R$ 1.386,3M	R$ 1.496,7M	R$ 1.642,7M	R$ 2.137,3M	R$ 2.582,7M	R$ 3.093,9M
P2	R$ 1.265,5M	R$ 1.415,9M	R$ 2.150,8M	R$ 2.438,2M	R$ 2.700,5M	R$ 3.352,5M	R$ 4.075,6M	R$ 5.031,8M
P3	R$ 470,6M	R$ 556,8M	R$ 650,2M	R$ 700,6M	R$ 878,9M	R$ 1.036,6M	R$ 1.059,4M	R$ 1.124,4M
P4	0,0K	R$ 5,6M	R$ 10,1M	R$ 16,2M	R$ 25,0M	R$ 49,7M	R$ 56,2M	R$ 42,3M
P5	0,0K	0,0K	0,0K	0,1K	0,3K	0,3K	0,4K	0,8K
P6	0,0K	0,0K	0,0K	R$ 35,9M	R$ 72,5M	R$ 90,8M	R$ 96,5M	R$ 107,8M
P7	0,0K	R$ 5,6M	R$ 10,1M	R$ 16,2M	R$ 25,0M	R$ 49,7M	R$ 56,2M	R$ 42,3M

A Figura 10 ilustra a quantidade de registros e receita em cada segmento. Cada registro representa as informações econômicas e o valor da receita fiscal de um contribuinte, agregados mensalmente. Nela é possível observar grupos com maiores quantidade de estabelecimentos (maior número de registros) e que não necessariamente possuem uma elevada arrecadação.

Figura 10. Distribuição do número de registros e arrecadação por segmento



A relação entre a carga tributária relatada nos documentos fiscais e o valor do imposto pago apresenta variações significativas. Isso ocorre devido às peculiaridades de cada contribuinte e da legislação tributária aplicada. Alguns desfrutam de benefícios fiscais, enquanto outros têm regras tributárias que aumentam ou diminuem o valor do imposto por várias razões, como o tipo de atividade econômica, o número de empregos gerados pela empresa e outros fatores. Os segmentos retratados mostram as diversas peculiaridades dos contribuintes, tanto na distribuição dos valores da carga tributária em várias variáveis quanto no valor acumulado de cada uma.

4.4 Aprendizagem segmentada

A abordagem de aprendizado segmentado proposta nesta pesquisa envolve etapas de treinamento e validação para cada segmento, e esse processo é repetido com diferentes algoritmos. Este estudo utilizou os seguintes algoritmos: Perceptron Multicamadas (MLP), Máquinas de Vetores de Suporte (SVR), Rede Neural Convolucional (CNN), Floresta Aleatória (RF) e Regressão Linear (LR). A seleção foi baseada em algoritmos validados na literatura e com lógicas de aprendizado diferentes para permitir uma maior adaptabilidade às realidades e peculiaridades dos contribuintes. Considerando que os modelos de aprendizado baseados em redes neurais recorrentes, incluindo LSTM, não ignoram a ordem dos dados e a significância temporal (Huang et al., 2020), eles foram excluídos dos experimentos, uma vez que os dados descritivos das operações econômicas não contêm informações relevantes armazenadas em sequências anteriores, e os padrões subjacentes mudam ao longo do tempo devido a alterações na legislação tributária.

O processo de treinamento aplica as mesmas variações de configurações de hiperparâmetros, de acordo com a arquitetura de cada algoritmo (detalhadas no Anexo I desta pesquisa), aos segmentos criados. Finalmente, o processo de treinamento ajusta uma Instância de Aprendizado (função de regressão) $\boldsymbol{SLI_{\alpha}[sid_i](r)}$ que prevê a receita estimada para cada registro $\boldsymbol{r} \in \boldsymbol{seg[sid_i]}$ de acordo com cada algoritmo $\boldsymbol{\alpha} \in \boldsymbol{\{CNN, MLP, RF, SVR, LR\}}$.

Em relação à métrica usada no processo de validação, considere que $r.REVENUE$ é a receita paga por cada contribuinte em cada período. Dado que nosso objetivo principal é estimar com precisão a receita tributária, ou seja, que

$$|SLI_{\alpha}[sid_i](r) - r.REVENUE| \cong 0$$

a métrica usada nesta pesquisa e que reflete as discrepâncias para todos os registros em um segmento é o Erro Médio Absoluto (MAE), conforme segue:

$$MAE_h = \sum_{r \in DS} \frac{|h(r) - r.REVENUE|}{|DS|},$$

onde $\boldsymbol{h(r)}$ é a função de predição, $\boldsymbol{DS = seg[sid_i]}$ e $\boldsymbol{|DS|}$ é o número total de registros. Quando $\boldsymbol{h(r) = SLI_{\alpha}[sid_i](r)}$, definimos $\boldsymbol{MAE_{\alpha}[sid_i]}$ como:

$$MAE_{\alpha}[sid_i] = MAE_{h= SLI_{\alpha}[sid_i]}$$

4.5 Seleção das instâncias de aprendizagem

A etapa de seleção da instância de aprendizado começa analisando as métricas das instâncias de aprendizado. Para cada identificador de segmento $\boldsymbol{sid_i}$, a instância de aprendizado com o menor $\boldsymbol{MAE_{\alpha}}$ entre todos os algoritmos usados na fase de aprendizado segmentada é selecionada para compor o Conjunto de Instâncias Híbridas (HIS), da seguinte forma:

$$HIS[sid_i] = SLI^*[sid_i],$$

Onde $\boldsymbol{SLI^*[sid_i] = SLI_{\alpha^*}[sid_i]}$, e $\boldsymbol{\alpha^* = \underset{\alpha}{argmin}\, MAE_{\alpha}[sid_i]}$.

4.6 Processo de predição

O processo de previsão de receita requer o parâmetro de segmentação categórica. Assim, para cada registro em um conjunto de dados, o modelo obtém a identificação do segmento e recupera a correspondente $\boldsymbol{SLI^*}$ no $\boldsymbol{HIS}$. Em seguida, a previsão é realizada com as variáveis preditivas do registro. Portanto, devido à granularidade da previsão, o modelo suporta uma visão analítica com drill-down e drill-up para permitir uma análise mais completa do comportamento da arrecadação.

5 Experimentação e Resultados

A abordagem de aprendizado segmentado proposta neste estudo visa um melhor refinamento do aprendizado com base em duas hipóteses. A primeira é que o aprendizado segmentado pode produzir um modelo preditivo mais assertivo. A segunda é que a seleção da instância de aprendizado que produziu o melhor resultado na métrica estabelecida entre as instâncias dos diversos algoritmos usados na fase de aprendizado pode se adaptar melhor à heterogeneidade do conjunto de dados.

Para avaliar a primeira hipótese, ajustamos algoritmos de aprendizado de máquina nos conjuntos de dados de treinamento e validação com as mesmas variações nas configurações de hiperparâmetros duas vezes, de acordo com as abordagens segmentada e não segmentada. A abordagem não segmentada gera apenas uma instância de aprendizado, denotada como $\boldsymbol{NSLI_{\alpha}}$, de acordo com cada algoritmo $\boldsymbol{\alpha}$. A abordagem segmentada gera várias instâncias de aprendizado, denotadas como $\boldsymbol{SLI_{\alpha}[sid_i]}$, uma para cada algoritmo $\boldsymbol{\alpha}$ e cada segmento $\boldsymbol{seg[sid_i]}$, conforme explicado na seção 4. Em seguida, o processo utilizou as instâncias de aprendizado de ambas as abordagens para prever o conjunto de dados de teste para comparar os resultados preditivos.

Para avaliar a segunda hipótese, usamos o Conjunto de Instâncias Híbridas ($\boldsymbol{HIS}$) para prever o conjunto de dados de teste. Finalmente, o processo experimental aplicou o modelo atual da SEFAZ-MA para prever o conjunto de dados de teste. As Tabelas 5 a 12 na seção 5.2 listam os resultados preditivos.

5.1 Detalhamento dos cenários

Dada a mutabilidade da legislação e seus efeitos sobre a relação entre variáveis preditivas e receita mensal, esta pesquisa avaliou a influência da extensão temporal dos dados de treinamento e sua proximidade com o momento da previsão. Portanto, a experimentação adotou a divisão por período de tempo para extrair conjuntos de treinamento, validação e teste. A Tabela 3 lista os seis cenários explorados nos experimentos. Em cada cenário e para cada conjunto de dados ($\boldsymbol{DS}$), a tabela detalha as seguintes características. A **Janela de Tempo** indica o intervalo anual dos registros. O **Nº de Registros** é o tamanho do conjunto de dados ($|\boldsymbol{DS}|$). Os **Nº de Contribuintes**

informam o número de contribuintes distintos. A **Receita Total** representa a soma das receitas de todos os registros $\boldsymbol{r}$ em um conjunto de dados ($\boldsymbol{DS}$), dada por:

$$Receita\ Total\ = \sum_{r \in DS} r.REVENUE$$

Tabela 3. Detalhamento dos cenários dos experimentos

		Dataset de Treinamento	Dataset de Validação	Dataset de Teste
1º Cenário: HIS01	Janela de Tempo	2015 - 2019	2020	2021
	Nº de Registros	59.852	15.222	23.529
	Nº de Contribuintes	3.783	2.493	3.634
	Receita Total	R$ 6.811.922.581,14	R$ 2,285,306,311.70	R$ 2.485.824.096,09
2º Cenário: HIS02	Janela de Tempo	2017-2019	2020	2021
	Nº de Registros	39.243	15,222	23.529
	Nº de Contribuintes	3.512	2,493	3.634
	Receita Total	R$ 4.700.005.897.99	R$ 2.285.306.311,70	R$ 2.485.824.096,09
3º Cenário: HIS01	Janela de Tempo	2015 - 2019	2020	2022
	Nº de Registros	59.852	15.222	25.193
	Nº de Contribuintes	3.783	2.493	4.049
	Receita Total	R$ 6.811.922.581,14	R$ 2.285.306.311,70	R$ 2.779.553.873,09
4º Cenário: HIS02	Janela de Tempo	2017-2019	2020	2022
	Nº de Registros	39.243	15.222	25.193
	Nº de Contribuintes	3.512	2.493	4.049
	Receita Total	R$ 4.700.005.897,99	R$ 2.285.306.311,70	R$ 2.779.553.873,09
5º Cenário: HIS03	Janela de Tempo	2016 - 2020	2021	2022
	Nº de Registros	64.766	23.529	25.193
	Nº de Contribuintes	4.250	3.634	4.049
	Receita Total	R$ 8.135.460.948,66	R$ 2.485.824.096,09	R$ 2.779.553.873,09
6º Cenário: HIS04	Janela de Tempo	2018-2020	2021	2022
	Nº de Registros	43.435	23.529	25.193
	Nº de Contribuintes	3.980	3.634	4.049
	Receita Total	R$ 5.654.416.495,52	R$ 2.485.824.096,09	R$ 2.779.553.873,09

O 1º e o 2º cenários avaliam o efeito da extensão temporal do conjunto de dados de treinamento na previsão de receita para os registros do conjunto de dados de teste no período de 2021. Da mesma forma, os 5º e 6º cenários para o período de 2022. Para avaliar a proximidade temporal dos dados de treinamento e validação com o ano dos registros do conjunto de dados de teste, modelos treinados para estimar receitas em

2021 serão aplicados para estimar as receitas para 2022 nos 3º e 4º cenários. Essas previsões serão comparadas com os resultados dos 5º e 6º cenários, onde o período de tempo de treinamento e validação está mais próximo do ano de 2022.

Os cenários detalhados demandaram a criação de quatro Conjuntos de Instâncias Híbridas (***HIS***). Dado que o conjunto de dados possui 70 segmentos distintos, cada ***HIS*** gerado terá 70 ***SLI****. A distribuição das ***SLI**** em cada algoritmo está listada na Tabela 4.

Tabela 4. Distribuição das 70 *SLI em cada algoritmo**

	CNN	MLP	RF	LR	SVR
HIS01	39	20	5	2	4
HIS02	39	24	1	2	4
HIS03	44	14	4	1	7
HIS04	36	22	2	0	10

5.2 Resultados e discussões

Seguindo a sequência de cenários enumerados na Tabela 3, os resultados dos experimentos são apresentados nas tabelas desta seção. Cada linha das Tabelas de 5 a 12 corresponde à previsão de um algoritmo realizada no conjunto de dados de teste. A primeira coluna é o **RANK**, que representa a posição em ordem ascendente pelo valor do MAE. A segunda é o **MODELO**, que se refere à função de regressão do algoritmo usada na previsão. A terceira coluna é o **TOTAL ESTIMADO**, que representa a soma das receitas estimadas de cada registro em um conjunto de dados de teste (DS), dada por:

$$TOTAL\ ESTIMATED_h = \sum_{r \in DS} h(r),$$

Onde $\boldsymbol{h}$ é uma função de regressão dentre as seguintes opções: $\boldsymbol{HIS}[\boldsymbol{sid}_i]$, $\boldsymbol{SLI}_{\boldsymbol{\alpha}}[\boldsymbol{sid}_i]$, $\boldsymbol{NSLI}_{\boldsymbol{\alpha}}$ e o modelo ***SEFAZ-MA***.

A quarta e a quinta colunas são o **ERRO TOTAL** e o **ERRO TOTAL %**, respectivamente, dadas pelas seguintes fórmulas:

$$ERRO\ TOTAL_h = TOTAL\ ESTIMADO_h - RECEITA\ TOTAL$$

$$ERRO\ TOTAL_h\% = |TOTAL\ ESTIMADO_h - RECEITA\ TOTAL| / RECEITA\ TOTAL$$

A sexta coluna é o MAE, o erro médio absoluto das predições. Os resultados produzidos pela metodologia atualmente utilizada na SEFAZ-MA são apresentados em outra tabela após os cenários para cada ano alvo de previsão. Algumas tabelas possuem linhas destacadas porque as análises e discussões as enfatizam.

5.2.1 Resultados preditivos para o ano 2021

Analisando as Tabelas 5 e 6, pode-se observar que, predominantemente, os algoritmos $\boldsymbol{\alpha}$ treinados na abordagem segmentada ($\boldsymbol{SLI_{\alpha}[sid_i]}$) alcançaram melhores resultados do que quando treinados na abordagem tradicional não segmentada ($\boldsymbol{NSLI_{\alpha}}$), com exceção da técnica $\boldsymbol{NSLI_{LR}}$. Também é possível perceber o efeito da extensão temporal do conjunto de dados de treinamento. Para o ano alvo de 2021, o aumento da extensão temporal prejudicou o MAE das técnicas, exceto para $\boldsymbol{NSLI_{MLP}}$ e $\boldsymbol{SLI_{SVR}}$.

As Tabelas 5 e 6 mostram que os modelos nas 4 primeiras posições estimaram a receita com um erro total abaixo de 3%. Segundo esse critério, o melhor resultado foi produzido por $\boldsymbol{SLI_{MLP}}$(Tabela 6), com uma diferença aproximada de menos de 12 milhões, de um total de aproximadamente 2,5 bilhões, representando um desvio de apenas 0,5%. Embora essa medida seja importante para um propósito secundário, um aumento no MAE indica que os desvios para cada previsão são maiores, com um trade-off entre desvios positivos e negativos, prejudicando a análise detalhada buscada com o propósito de uma visão analítica.

Tabela 5. Resultados preditivos do 1º Cenário

Rank	Modelo	Total Estimado	Erro Total	Erro Total %	MAE
1st	HIS	R$ 2.412.503.377,92	-R$ 73.320.718,17	2,9496%	**18.700,26**
2nd	SLI_{MLP}	R$ 2.419.597.264,92	-R$ 66.226.831,17	2,6642%	**19.903,21**
3rd	SLI_{CNN}	R$ 2.416.786.460,13	-R$ 69.037.635,96	2,7773%	**20.206,44**
4th	$NSLI_{CNN}$	R$ 2.420.065.892,48	-R$ 65.758.203,61	2,6453%	**22.865,45**
5th	$NSLI_{MLP}$	R$ 2.278.486.837,64	-R$ 207.337.258,45	8,3408%	**25.400,26**
6th	SLI_{RF}	R$ 2.135.782.328,45	-R$ 350.041.767,64	14,0815%	**28.858,61**
7th	$NSLI_{RF}$	R$ 2.374.899.190,36	-R$ 110.924.905,73	4,4623%	**31.005,88**
8th	SLI_{SVR}	R$ 1.956.580.135,62	-R$ 529.243.960,47	21,2905%	**33.716,44**
9th	$NSLI_{LR}$	R$ 2.541.936.339,71	R$ 56.112.243,62	2,2573%	**37.486,41**

10th	$NSLI_{SVR}$	R$ 2.286.154.872,55	-R$ 199.669.223,54	8,0323%	**40.615,29**
11th	SLI_{LR}	-R$ 6.577.727.414,33	-R$ 9.063.551.510,42	364,6095%	**420.250,32**

Tabela 6. Resultados preditivos do 2º Cenário

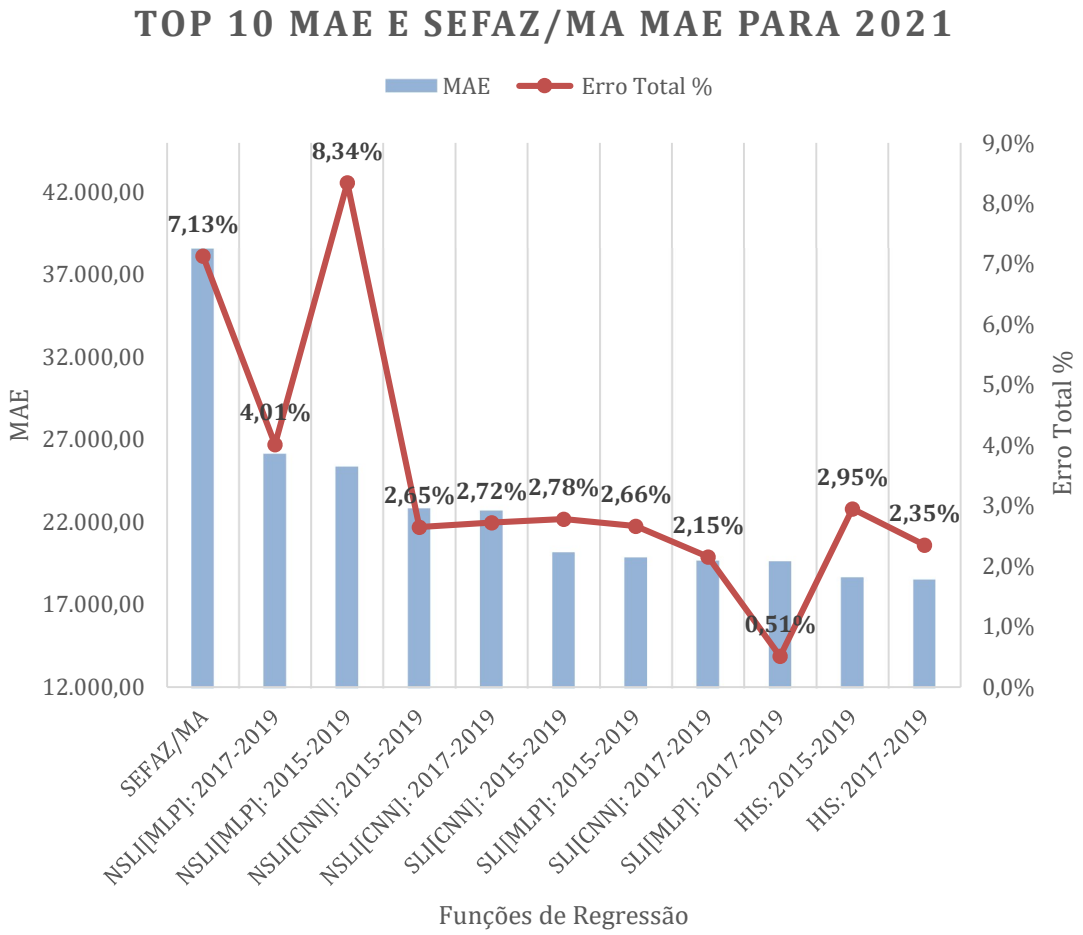
Rank	Modelo	Total Estimado	Erro Total	Erro Total %	MAE
1st	HIS	R$ 2,427,445,855.56	-R$ 58,378,240.53	2.3484%	**18,557.92**
2nd	SLI_{MLP}	R$ 2,473,119,750.78	-R$ 12,704,345.31	0.5111%	**19,674.01**
3rd	SLI_{CNN}	R$ 2,432,309,719.50	-R$ 53,514,376.59	2.1528%	**19,698.42**
4th	$NSLI_{CNN}$	R$ 2,418,229,899.30	-R$ 67,594,196.79	2.7192%	**22,741.39**
5th	$NSLI_{MLP}$	R$ 2,386,143,214.26	-R$ 99,680,881.83	4.0100%	**26,172.07**
6th	SLI_{RF}	R$ 2,164,583,931.59	-R$ 321,240,164.50	12.9229%	**28,160.13**
7th	$NSLI_{RF}$	R$ 2,378,011,265.68	-R$ 107,812,830.41	4.3371%	**29,047.78**
8th	SLI_{SVR}	R$ 1,968,276,750.54	-R$ 517,547,345.55	20.8200%	**33,741.54**
9th	$NSLI_{LR}$	R$ 2,567,475,448.23	R$ 81,651,352.14	3.2847%	**34,733.16**
10th	SLI_{LR}	R$ 2,145,222,297.54	-R$ 340,601,798.55	13.7018%	**39,014.05**
11th	$NSLI_{SVR}$	R$ 2,297,706,861.21	-R$ 188,117,234.88	7.5676%	**39,484.45**

Tabela 7. Resultados preditivos do Modelo SEFAZ em 2021

Modelo	Total Estimado	Erro Total	Erro Total %	MAE
SEFAZ-MA	R$ 2.663.105.656,36	R$ 177.281.560,27	7,1317%	38.578,41

A Figura 11 ilustra os 10 melhores modelos e o modelo SEFAZ-MA para as previsões do ano-alvo de 2021. No eixo X, a disposição dos modelos e o respectivo intervalo de tempo utilizado no treinamento estão em ordem decrescente de MAE. Os resultados mostram que um menor Erro Total % (linha laranja) não indica necessariamente o melhor resultado para uma abordagem analítica, dado pelo menor resultado de MAE (barras azuis).

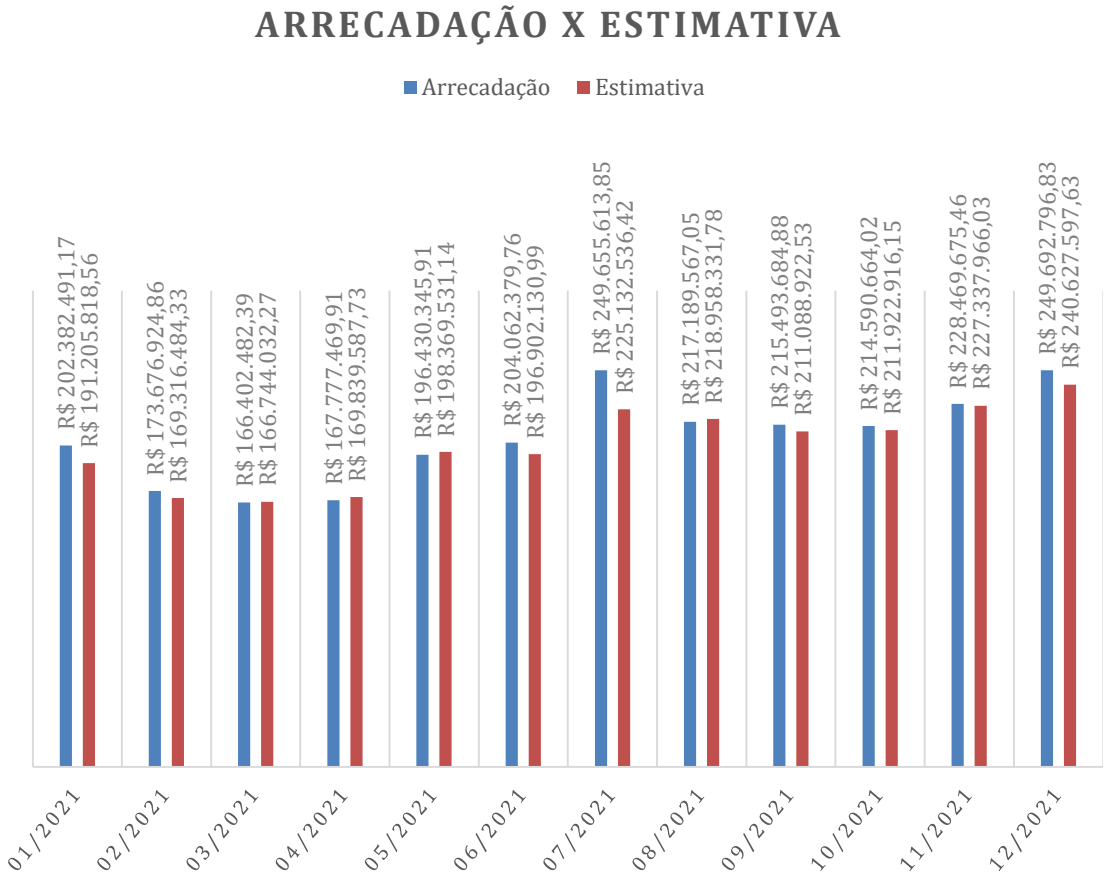
Figura 11. Top 10 MAEs e SEFAZ-MA MAE para estimativas do ano de 2021



O modelo proposto, HIS, foi capaz de gerar o melhor resultado tanto nos cenários da Tabela 5 quanto na Tabela 6. Considerando o melhor resultado da abordagem tradicional, $\boldsymbol{NSLI_{CNN}}$ (sublinhado na Tabela 6), e adicionando à comparação os resultados do modelo atualmente utilizado pela SEFAZ-MA (Tabela 7), o melhor HIS (sublinhado na Tabela 6) reduziu os desvios em 18,4% e 51,9%, respectivamente.

Analisando a figura 12, percebemos as diferenças dos valores estimados e arrecadados das receitas estaduais dos segmentos contidos na massa de dados utilizada por esta pesquisa, em cada período mensal do ano de 2021. Embora tenham ocorrido compensações entre as variações positivas e negativas dos valores estimados e arrecados para os diversos segmentos considerados, observa-se que, mesmo no nível de detalhamento mensal e para um montante que representa aproximadamente 20% da arrecadação total, os valores apresentam uma margem de erro percentual que não supera o 5%.

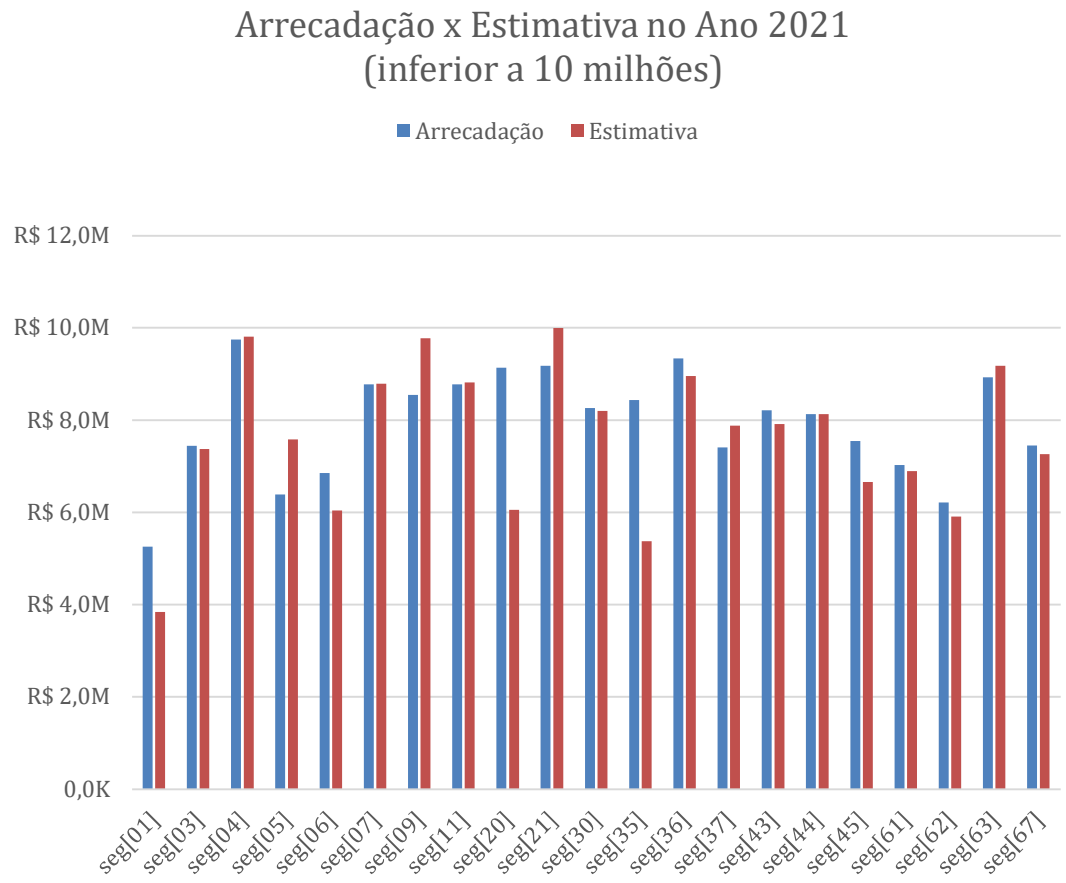
Figura 12. Valores de Arrecadação x Estimativa em períodos mensais do ano 2021



Nas figuras a seguir estão demonstrados os valores de arrecadação e estimativas para cada um dos 70 segmentos utilizados nesta pesquisa. Para melhorar a visualização dos dados nos gráficos, separamos os segmentos em 4 intervalos de acordo com o valor de arrecadação acumulado no ano de 2021: valores inferiores a 10 milhões de reais; valores entre 10 e 30 milhões; valores entre 30 e 100 milhões; e, por fim, valores superiores a 100 milhões.

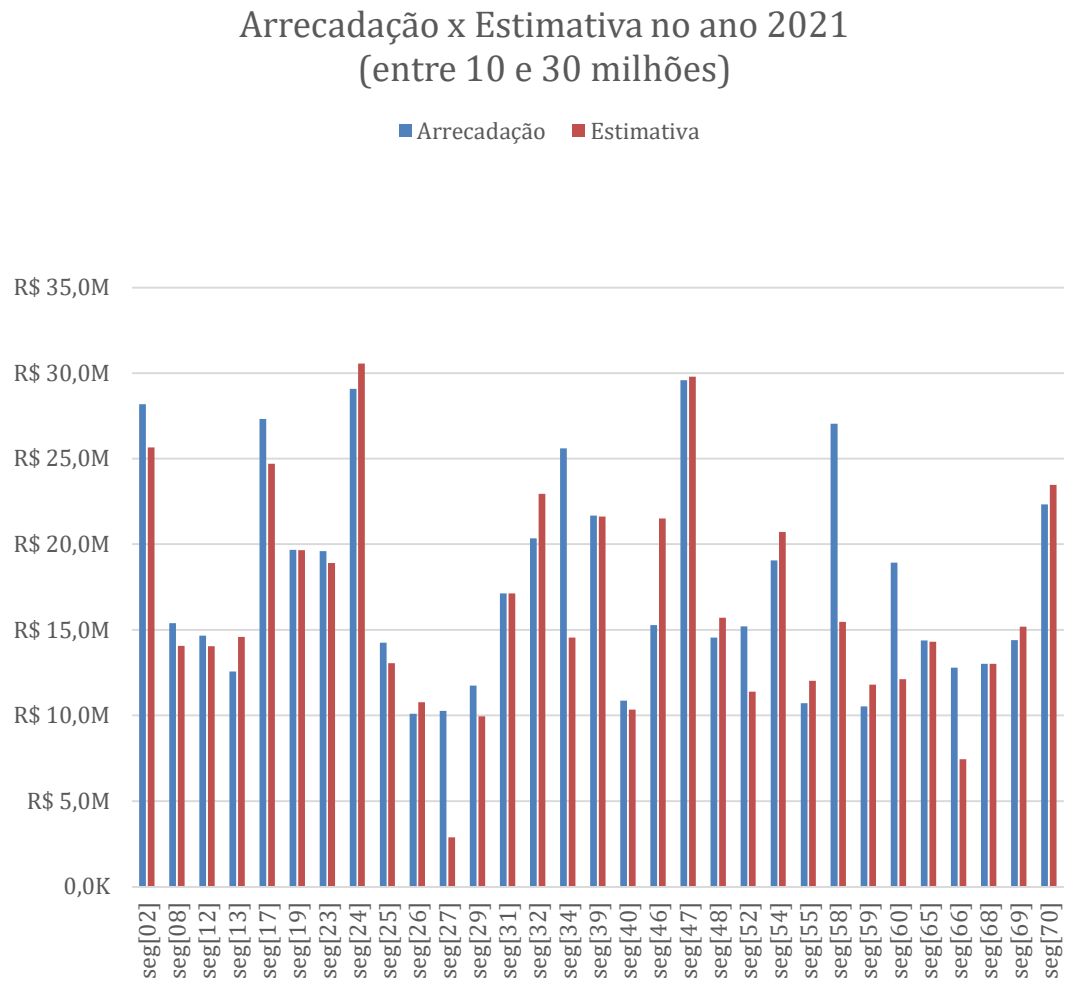
Na figura 13 é possível observar que o modelo estimou um valor próximo ao efetivo para a maioria dos segmentos do 1º intervalo de valores. As maiores diferenças podem ser visualizadas nos segmentos 1, 20 e 35. Em diversos segmentos, é possível observar que o modelo estimou o valor quase equivalente ao arrecadado, como é o caso dos segmentos 3, 4, 7, 11, dentre outros.

Figura 13. Valores de Arrecadação x Estimativa por segmento para o ano 2021 inferiores a 10 milhões



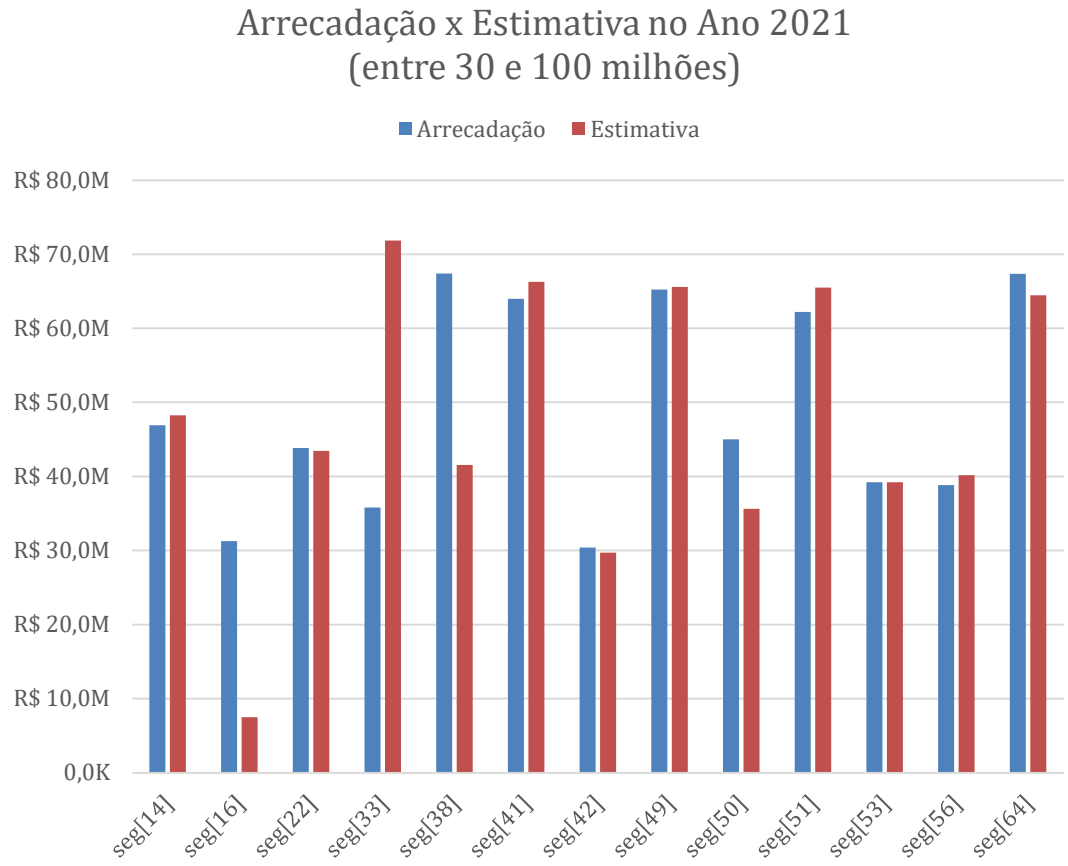
Na figura 14 é possível observar que o modelo também estimou um valor próximo ao efetivo para a maioria dos segmentos do 2º intervalo de valores. As maiores diferenças podem ser visualizadas nos segmentos 27, 34 e 58. Em diversos segmentos, é possível observar que o modelo estimou o valor quase equivalente ao arrecadado, como é o caso dos segmentos 12, 19, 31, 39, dentre outros.

Figura 14. Valores de Arrecadação x Estimativa por segmento para o ano 2021 entre a 10 e 30 milhões



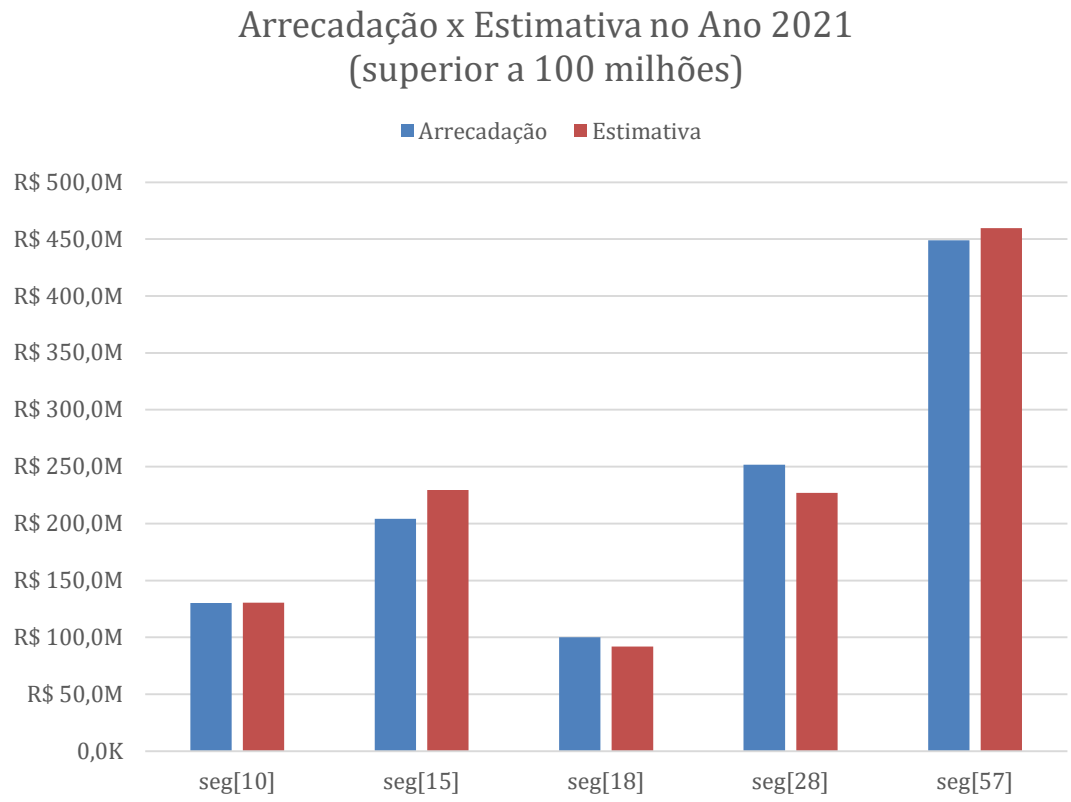
Na figura 15 é possível observar que o modelo também estimou um valor próximo ao efetivo para a maioria dos segmentos do 3º intervalo de valores. As maiores diferenças podem ser visualizadas nos segmentos 16 e 33 e 38. Em diversos segmentos, é possível observar que o modelo estimou o valor quase equivalente ao arrecadado, como é o caso dos segmentos 14, 22, 49, 53, dentre outros.

Figura 15. Valores de Arrecadação x Estimativa por segmento para o ano 2021 entre a 30 e 100 milhões



Por fim, na figura 16, onde estão representadas as maiores arrecadações, segmentos que estão no 4º intervalo, o modelo estimou um valor próximo ao efetivo sem desvios relevantes.

Figura 16. Valores de Arrecadação x Estimativa por segmento para o ano 2021 superiores a 100 milhões



5.2.2 Resultados preditivos para o ano 2022

A partir dos dados tabulados para o ano de 2022, nota-se novamente que, predominantemente, os algoritmos α\alphaα treinados na abordagem segmentada ($\boldsymbol{SLI_{\alpha}[sid_i]}$) obtiveram melhores resultados do que quando treinados na abordagem tradicional não segmentada ($\boldsymbol{NSLI_{\alpha}}$). Além disso, para todos os cenários de 2022, os três melhores resultados foram obtidos utilizando a abordagem segmentada.

Em relação à proximidade temporal dos conjuntos de dados de treinamento e validação ao momento da previsão, comparamos as Tabelas 8 e 10, que têm a mesma extensão temporal de cinco anos do conjunto de dados de treinamento, mas a Tabela 8 apresenta um maior atraso temporal do que a Tabela 10. Uma comparação entre as Tabelas 8 e 10 mostra que, com exceção da $\boldsymbol{NSLI_{LR}}$, todas mostraram melhorias na métrica que mede o desvio. Uma comparação entre as Tabelas 9 e 11 mostra que todos os modelos com conhecimento atualizado apresentaram melhores resultados. Dessa forma, independente da abordagem utilizada, um treinamento conduzido com dados

mais atuais em relação ao momento alvo é capaz de melhorar a estimativa de arrecadação.

Para o aspecto da extensão temporal do conjunto de treinamento, comparando as Tabelas 8 e 9, o aumento prejudicou os MAEs das técnicas, exceto para $\boldsymbol{NSLI_{MLP}}$ e $\boldsymbol{SLI_{SVR}}$. Comparando as Tabelas 10 e 11, o aumento prejudicou o MAE de 5 técnicas ($\boldsymbol{NSLI_{CNN}}$, $\boldsymbol{SLI_{RF}}$, $\boldsymbol{SLI_{LR}}$, $\boldsymbol{NSLI_{LR}}$, $\boldsymbol{SLI_{SVR}}$) e melhorou o de 6 técnicas ($\boldsymbol{HIS}$, $\boldsymbol{SLI_{CNN}}$, $\boldsymbol{SLI_{MLP}}$, $\boldsymbol{NSLI_{MLP}}$, $\boldsymbol{NSLI_{RF}}$, $\boldsymbol{NSLI_{SVR}}$). Mais uma vez, o modelo proposto, $\boldsymbol{HIS}$, foi capaz de gerar os melhores resultados em todos os cenários, conforme mostrado nas Tabelas 8, 9, 10 e 11.

Os resultados para 2022 mostram que os modelos nas quatro primeiras posições estimaram a receita com uma margem de erro total de menos de 3%. Nesse sentido, o melhor resultado foi produzido por $\boldsymbol{SLI_{MLP}}$ (Tabela 9), com uma diferença de apenas 10,5 milhões, em um total aproximado de 2,8 bilhões, representando um desvio de apenas 0,38%. Embora essa medida seja importante para o propósito final, que é estimar a receita, um aumento no MAE indica que os desvios em cada previsão são maiores, com um trade-off entre desvios positivos e negativos, novamente prejudicando a análise detalhada buscada com o propósito de uma visão analítica.

Tabela 8. Resultados preditivos do 3º Cenário

Rank	Modelo	Total Estimado	Erro Total	Erro Total %	MAE
1st	HIS	R$ 2.707.118.071,70	-R$ 72.435.801,39	2,6060%	21.891,30
2nd	SLI_{MLP}	R$ 2.723.803.712,40	-R$ 55.750.160,69	2,0057%	23.144,22
3rd	SLI_{CNN}	R$ 2.724.530.374,95	-R$ 55.023.498,14	1,9796%	23.420,04
4th	$NSLI_{CNN}$	R$ 2.804.165.330,51	R$ 24.611.457,42	0,8854%	25.499,53
5th	$NSLI_{MLP}$	R$ 2.618.514.746,93	-R$ 161.039.126,16	5,7937%	26.770,30
6th	$NSLI_{RF}$	R$ 2.607.168.449,40	-R$ 172.385.423,69	6,2019%	34.857,00
7th	SLI_{RF}	R$ 2.259.855.465,53	-R$ 519.698.407,56	18,6972%	35.879,22
8th	SLI_{SVR}	R$ 2.057.710.712,61	-R$ 721.843.160,48	25,9697%	41.810,82
9th	$NSLI_{SVR}$	R$ 2.563.893.621,27	-R$ 215.660.251,82	7,7588%	42.860,11
10th	$NSLI_{LR}$	R$ 2.302.223.016,95	-R$ 477.330.856,14	17,1729%	61.735,37
11th	SLI_{LR}	-R$ 18.729.128.967,50	-R$ 21.508.682.840,59	773,8178%	882.619,54

Tabela 9. Resultados preditivos do 4º Cenário

Rank	Modelo	Total Estimado	Erro Total	Erro Total %	MAE
1st	HIS	R$ 2.720.489.626,10	-R$ 59.064.246,99	2,1250%	21.679,21
2nd	SLI_{MLP}	R$ 2.790.119.606,27	R$ 10.565.733,18	0,3801%	22.925,36
3rd	SLI_{CNN}	R$ 2.736.253.482,04	-R$ 43.300.391,05	1,5578%	23.333,28
4th	$NSLI_{CNN}$	R$ 2.765.119.098,42	-R$ 14.434.774,67	0,5193%	23.993,50
5th	$NSLI_{MLP}$	R$ 2.757.115.304,28	-R$ 22.438.568,81	0,8073%	27.778,97
6th	$NSLI_{RF}$	R$ 2.613.389.409,84	-R$ 166.164.463,25	5,9781%	32.971,34
7th	SLI_{RF}	R$ 2.278.409.575,51	-R$ 501.144.297,58	18,0297%	35.336,43
8th	$NSLI_{LR}$	R$ 2.908.406.280,51	R$ 128.852.407,42	4,6357%	35.744,09
9th	SLI_{LR}	R$ 2.599.496.516,31	-R$ 180.057.356,78	6,4779%	37.389,25
10th	$NSLI_{SVR}$	R$ 2.547.924.754,07	-R$ 231.629.119,02	8,3333%	42.098,13
11th	SLI_{SVR}	R$ 2.053.775.735,57	-R$ 725.778.137,52	26,1113%	42.358,91

Tabela 10. Resultados preditivos do 5º Cenário

Rank	Modelo	Total Estimado	Erro Total	Erro Total %	MAE
1st	HIS	R$ 2.743.179.605,38	-R$ 36.374.267,71	1,3086%	16.301,09
2nd	SLI_{CNN}	R$ 2.812.497.687,93	R$ 32.943.814,84	1,1852%	17.350,50
3rd	SLI_{MLP}	R$ 2.818.517.110,77	R$ 38.963.237,68	1,4018%	17.662,29
4th	$NSLI_{CNN}$	R$ 2.751.248.744,24	-R$ 28.305.128,85	1,0183%	23.457,97
5th	$NSLI_{MLP}$	R$ 2.657.734.062,70	-R$ 121.819.810,39	4,3827%	26.092,02
6th	SLI_{RF}	R$ 2.533.300.942,89	-R$ 246.252.930,20	8,8594%	27.597,05
7th	SLI_{LR}	R$ 2.771.645.433,48	-R$ 7.908.439,61	0,2845%	28.841,71
8th	$NSLI_{RF}$	R$ 2.700.302.416,13	-R$ 79.251.456,96	2,8512%	30.245,64
9th	SLI_{SVR}	R$ 2.302.182.338,73	-R$ 477.371.534,36	17,1744%	33.211,60
10th	$NSLI_{SVR}$	R$ 2.524.342.752,12	-R$ 255.211.120,97	9,1817%	38.955,82
11th	$NSLI_{LR}$	R$ 1.479.045.725,84	-R$ 1.300.508.147,25	46,7884%	96.928,12

Tabela 11. Resultados preditivos do 6º Cenário

Rank	Modelo	Total Estimado	Erro Total	Erro Total %	MAE
1st	HIS	R$ 2.755.449.487,67	-R$ 24.104.385,42	0,8672%	17.047,87
2nd	SLI_{CNN}	R$ 2.836.304.112,13	R$ 56.750.239,04	2,0417%	17.751,57
3rd	SLI_{MLP}	R$ 2.831.344.195,11	R$ 51.790.322,02	1,8633%	18.229,78
4th	$NSLI_{CNN}$	R$ 2.803.730.216,45	R$ 24.176.343,36	0,8698%	23.127,43
5th	$NSLI_{MLP}$	R$ 2.682.106.432,61	-R$ 97.447.440,48	3,5059%	26.554,04
6th	SLI_{RF}	R$ 2.553.756.527,02	-R$ 225.797.346,07	8,1235%	27.028,93
7th	SLI_{LR}	R$ 2.824.764.563,15	R$ 45.210.690,06	1,6265%	28.822,17
8th	$NSLI_{RF}$	R$ 2.721.278.671,67	-R$ 58.275.201,42	2,0966%	30.748,52

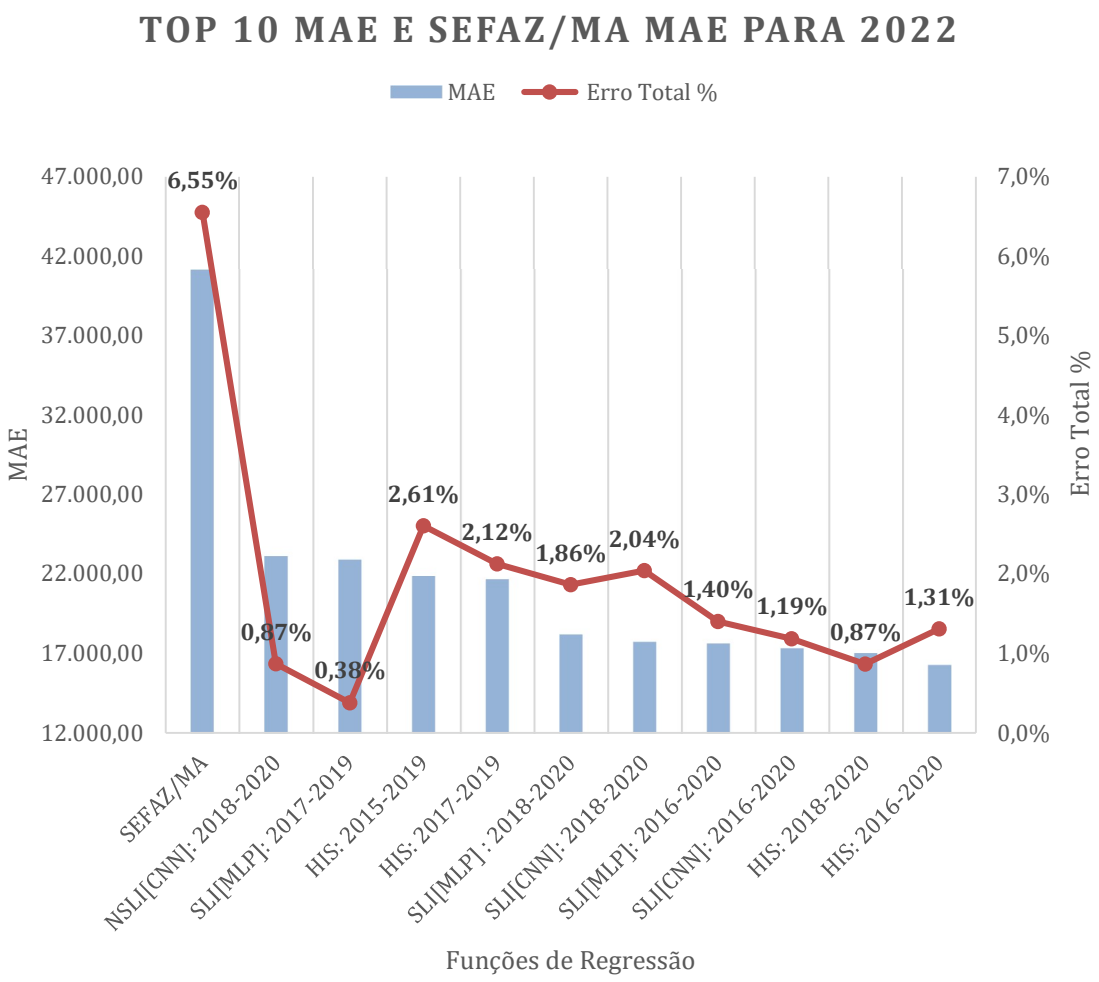
9th	SLI_{SVR}	R$ 2.279.955.297,63	-R$ 499.598.575,46	17,9741%	32.721,61
10th	$NSLI_{LR}$	R$ 2.932.159.984,30	R$ 152.606.111,21	5,4903%	35.133,01
11th	$NSLI_{SVR}$	R$ 2.552.774.058,47	-R$ 226.779.814,62	8,1589%	39.278,44

Tabela 12. Resultados preditivos do Modelo SEFAZ em 2022

Modelo	Total Estimado	Erro Total	Erro Total %	MAE
SEFAZ-MA	R$ 2.961.664.516,94	R$ 182.110.643,85	6,5518%	41.102,82

A Figura 17 ilustra os 10 melhores modelos e o modelo SEFAZ-MA para as previsões de 2022. No eixo X, a disposição dos modelos e o respectivo intervalo de tempo usado no treinamento estão em ordem decrescente do MAE. Mais uma vez, é possível observar que um menor Erro Total % (linha laranja) não necessariamente seleciona o melhor resultado para uma abordagem analítica, dado pelo menor resultado de MAE (barras azuis).

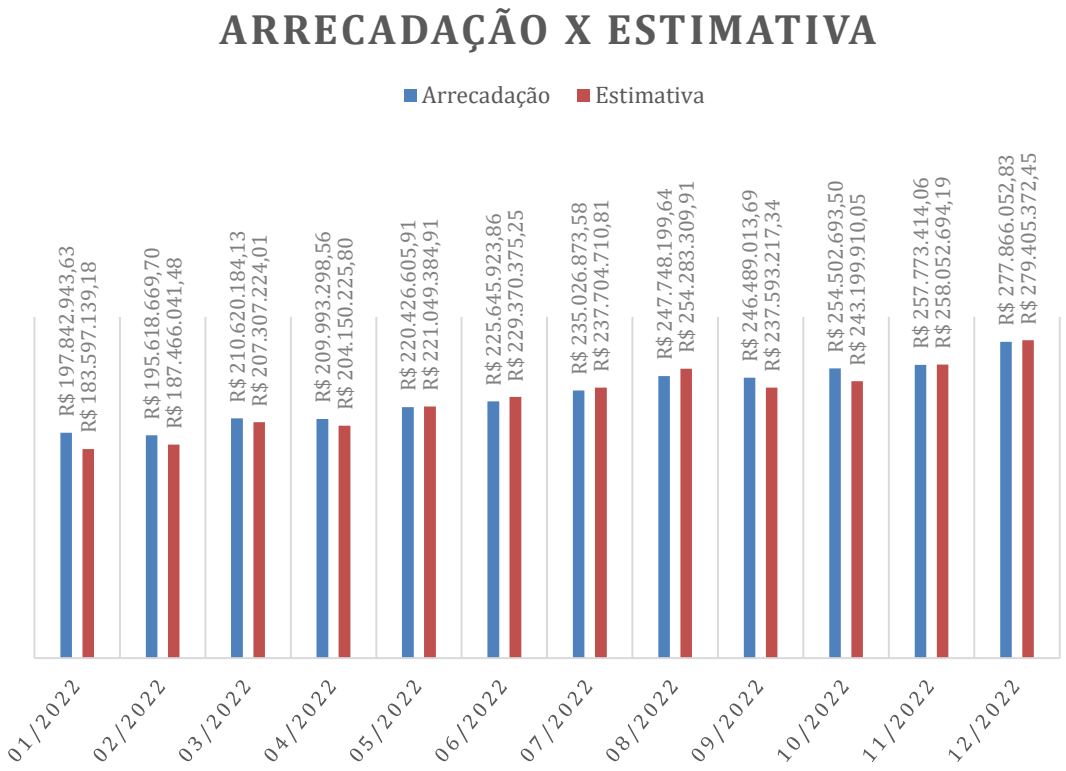
Figura 17. Top 10 MAEs e SEFAZ-MA MAE para estimativas do ano de 2022



Para o ano de 2022, considerando o melhor resultado da abordagem tradicional, $\boldsymbol{NSLI_{CNN}}$ (sublinhado na Tabela 11), e os resultados do modelo atualmente utilizado pela SEFAZ-MA (Tabela 12), o melhor $\boldsymbol{HIS}$ (sublinhado na Tabela 10) reduziu os desvios em 29,52% e 60,34%, respectivamente.

Analisando a figura 18, percebemos as diferenças dos valores estimados e arrecadados das receitas estaduais dos segmentos contidos na massa de dados utilizada por esta pesquisa, em cada período mensal do ano de 2022. Novamente, embora tenham ocorrido compensações entre as variações positivas e negativas dos valores estimados e arrecados para os diversos segmentos considerados, observa-se que, mesmo no nível de detalhamento mensal e para um montante que representa aproximadamente 20% da arrecadação total, os valores apresentam uma margem de erro percentual que não supera o 5%. Nas figuras seguintes apresentaremos as variações acumuladas no ano de 2022 para cada segmento, e discutiremos os erros usando essa visão da informação.

Figura 18. Valores de Arrecadação x Estimativa em períodos mensais do ano 2022

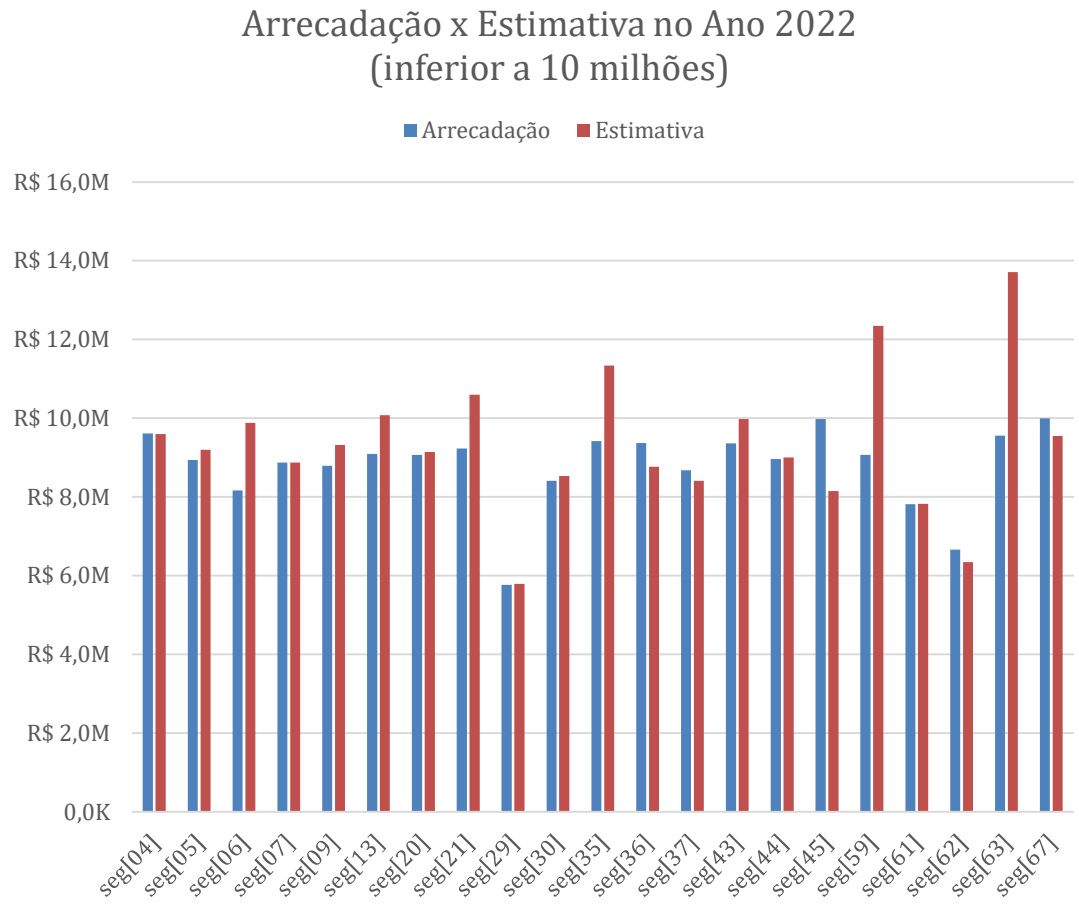


Nas figuras a seguir estão demonstrados os valores de arrecadação e estimativas para cada um dos 70 segmentos utilizados nesta pesquisa. Novamente, separamos os segmentos em 4 intervalos de acordo com o valor de arrecadação

acumulado, só que para o ano de 2022: valores inferiores a 10 milhões de reais; valores entre 10 e 30 milhões; valores entre 30 e 100 milhões; e, por fim, valores superiores a 100 milhões.

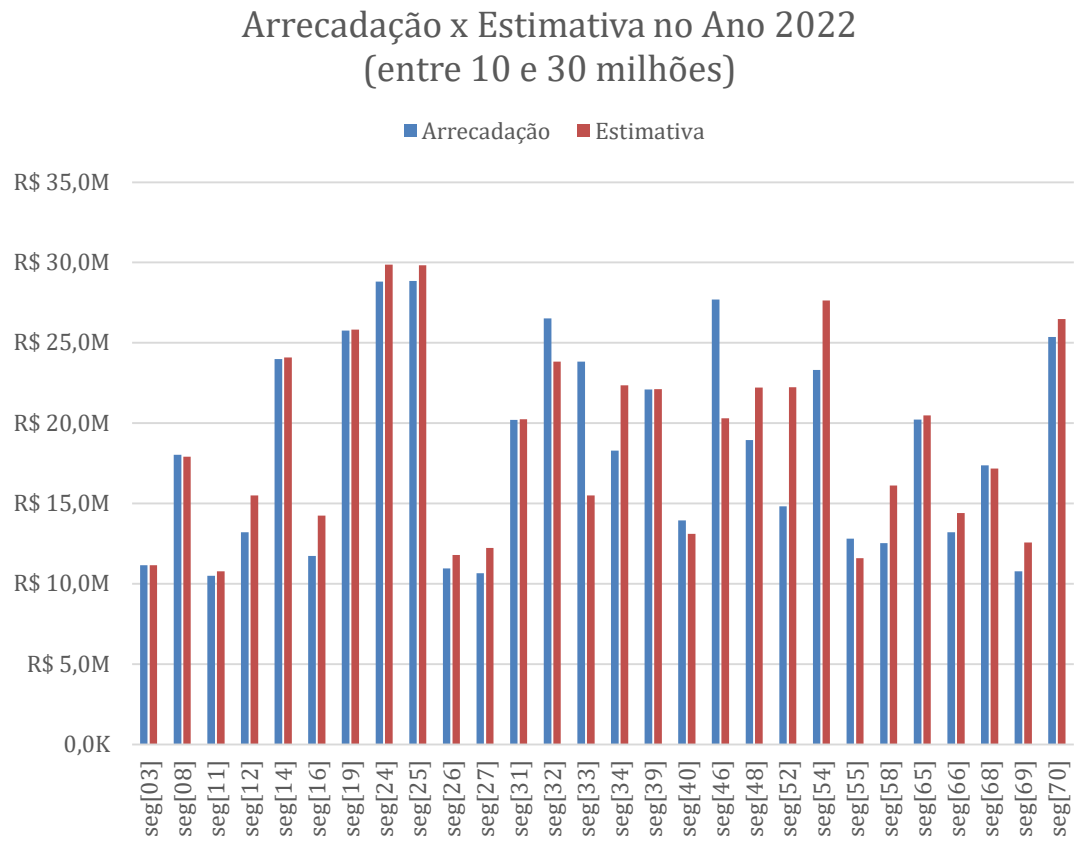
Na figura 19 é possível observar que o modelo estimou um valor próximo ao efetivo para a maioria dos segmentos do 1º intervalo de valores. As maiores diferenças podem ser visualizadas nos segmentos 59 e 63. Em diversos segmentos, é possível observar que o modelo estimou o valor quase equivalente ao arrecadado, como é o caso dos segmentos 04, 07, 44, dentre outros.

Figura 19. Valores de Arrecadação x Estimativa por segmento para o ano 2022 inferiores a 10 milhões



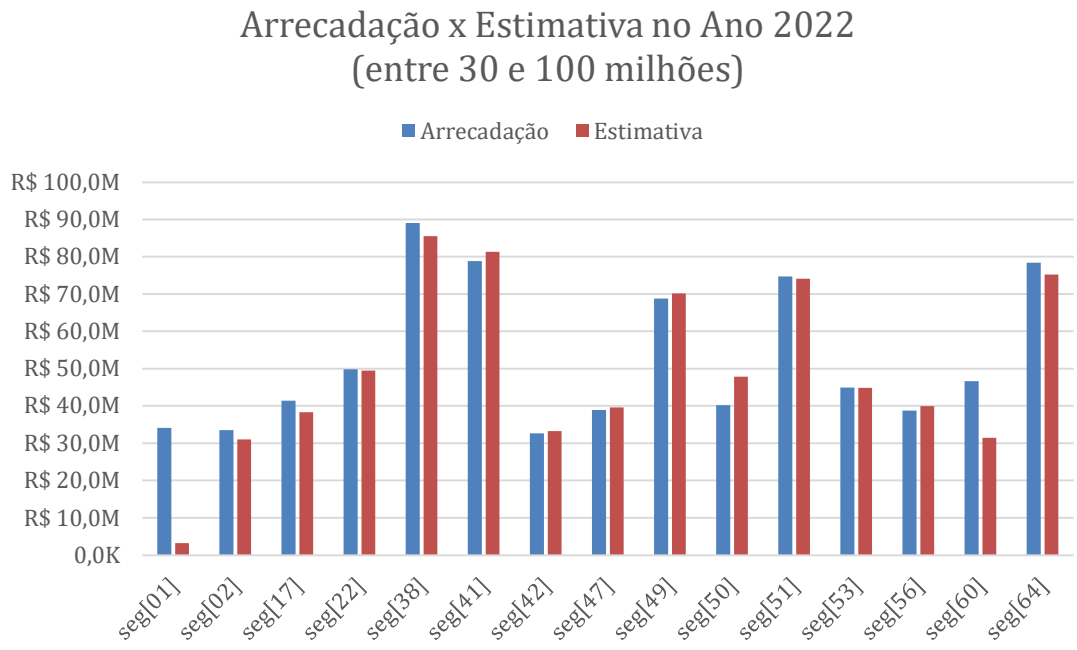
Na figura 20 é possível observar que o modelo também estimou um valor próximo ao efetivo para a maioria dos segmentos do 2º intervalo de valores. As maiores diferenças podem ser visualizadas nos segmentos 33, 46 e 54. Em diversos segmentos, é possível observar que o modelo estimou o valor quase equivalente ao arrecadado, como é o caso dos segmentos 03, 06, 11, dentre outros.

Figura 20. Valores de Arrecadação x Estimativa por segmento para o ano 2022 entre a 10 e 30 milhões



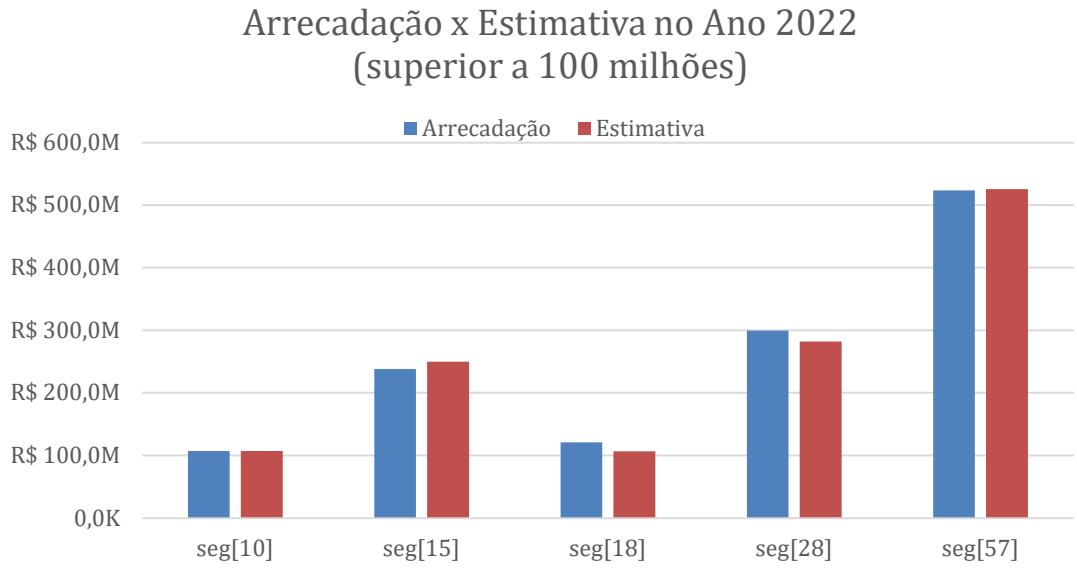
Na figura 21 é possível observar que o modelo também estimou um valor próximo ao arrecadado para a maioria dos segmentos do 3º intervalo de valores. Houve uma diferença relevante no segmento 01. Em diversos segmentos, é possível observar que o modelo estimou o valor quase equivalente ao arrecadado, como é o caso dos segmentos 02, 22, 42, 47, dentre outros.

Figura 21. Valores de Arrecadação x Estimativa por segmento para o ano 2022 entre a 30 e 100 milhões



Por fim, na figura 22, onde estão representadas as maiores arrecadações, segmentos que estão no 4º intervalo, o modelo estimou um valor próximo ao efetivo sem desvios relevantes.

Figura 22. Valores de Arrecadação x Estimativa por segmento para o ano 2022 superiores a 100 milhões



6 Conclusão

Este estudo propõe um modelo de aprendizado segmentado para estimar a receita fiscal com base em informações econômicas utilizadas no sistema para calcular o valor a ser pago. As previsões do modelo de aprendizado segmentado foram comparadas com abordagens tradicionais de aprendizado não segmentadas e com o modelo atualmente aplicado pelo órgão fiscalizador que detém as informações utilizadas nesta pesquisa.

O modelo proposto, o Conjunto de Instâncias Híbridas (HIS), proporcionou as maiores reduções na discrepância medida pela métrica estabelecida entre as receitas estimadas e reais em todos os cenários de experimentação projetados. Conforme ilustrado na Figura 11, ao comparar o método atual e o HIS, houve uma redução no MAE de 38.578,41 para 18.557,92, representando uma melhora na métrica utilizada de 51,9% para as previsões no ano de 2021. Para as previsões no ano de 2022, conforme se observa na Figura 17, o MAE diminuiu de 41.102,82 para 16.301,09, representando uma melhora de 60,3%.

Além disso, os resultados indicam que a atualização do conhecimento preditivo pode melhorar a precisão da previsão. Em relação aos efeitos da extensão temporal do conjunto de dados de treinamento, os resultados não mostram uma tendência de melhoria consistente que cubra todos os algoritmos.

6.1 Limitações do modelo

É importante notar que há uma limitação no modelo em relação à extensão temporal futura da previsão. Como o processo de predição trabalha com variáveis descritivas de um dado período mensal, que determinam o montante de receita a ser recolhido no período mensal subsequente, o modelo requer informações econômicas do período de avaliação anterior ao período alvo da predição, isso limita, em última análise, a capacidade de estimar a receita até o mês corrente. Ou, no máximo, projetar a arrecadação do mês subsequente de acordo com a evolução temporal do mês corrente. No entanto, o modelo pode ser aplicado ao passado para identificar discrepâncias entre os valores estimados e arrecadados para avaliar a regularidade do contribuinte ou até mesmo orientar o planejamento das ações fiscais, diante dos desvios observados.

6.2 Trabalhos futuros

Pesquisas futuras podem avaliar modelos para estimar informações fiscais futuras e usar essas informações no modelo proposto, bem como abordar estratégias de otimização para a etapa de treinamento. Além disso, devido à sua capacidade analítica, é possível combinar variáveis descritivas dependentes da legislação tributária para avaliar a aplicação do modelo em estudos que mensuram o impacto em caso de mudanças no sistema de cálculo de impostos. Também é possível avaliar o modelo como uma ferramenta de apoio para a detecção de indícios de evasão fiscal criando segmentos de contribuintes de um mesmo setor econômico ou que possuem uma correlação forte entre as operações realizadas e a carga tributária devida, possibilitando que a estimativa indique os desvios de comportamentos. Por fim, estudos para validar o modelo em outros contextos não fiscais e até mesmo não governamentais podem avaliar a escalabilidade do modelo proposto.

Referências

Andrade, N. A. (2008), Contabilidade Pública na Gestão Municipal, São Paulo: Atlas, 3ª Edição

Bayer, O. (2015). Relevance of Input Data Time Series for Tax Revenue Forecasting. Procedia Economics and Finance, 25, 518–529. https://doi.org/https://doi.org/10.1016/S2212-5671(15)00765-0

Bonaccorso, G. (2018) Machine Learning Algorithms. 2nd edn. Packt Publishing. Available at: https://www.perlego.com/book/799726 (Accessed: 15 June 2024).

Buxton, E., Kriz, K., Cremeens, M., & Jay, K. (2019). An Auto Regressive Deep Learning Model for Sales Tax Forecasting from Multiple Short Time Series. 2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA), 1359–1364. https://doi.org/10.1109/ICMLA.2019.00221

Chakraverty, S. and Mall, S. (2017) Artificial Neural Networks for Engineers and Scientists. 1st edn. CRC Press. Available at: https://www.perlego.com/book/1521126 (Accessed: 23 June 2024).

Chollet, F. (2017) Deep Learning with Python. [edition unavailable]. Manning Publications. Available at: https://www.perlego.com/book/1469354 (Accessed: 23 June 2024).

Elsheikh, A. H. and Elaziz, M. A. (2022) Artificial Neural Networks for Renewable Energy Systems and Real-World Applications. [edition unavailable]. Academic Press. Available at: https://www.perlego.com/book/3864488 (Accessed: 23 June 2024).

Graupe, D. (2019) Principles of Artificial Neural Networks. 4th edn. WSPC. Available at: https://www.perlego.com/book/979124 (Accessed: 23 June 2024).

Harada, K. (2017), Direito financeiro e tributário, São Paulo: Atlas, 26ª Edição.

Huang, J., Chai, J., & Cho, S. (2020). Deep learning in finance and banking: A literature review and classification. Frontiers of Business Research in China, 14(1), 13. https://doi.org/10.1186/s11782-020-00082-6

Hui, X. (2020). Comparison and Application of Logistic Regression and Support Vector Machine in Tax Forecasting. 2020 International Signal Processing, Communications and Engineering Management Conference (ISPCEM), 48–52. https://doi.org/10.1109/ISPCEM52197.2020.00015

Kumar, N. N., Sridhar, R., Prasanna, U. U., & Priyanka, G. (2023a). Opportunities of Machine Learning Algorithms in the Tax Default Prediction - A Survey. 2023 7th International Conference on Trends in Electronics and Informatics (ICOEI), 1225–1231. https://doi.org/10.1109/ICOEI56765.2023.10125962

Kumar, N. N., Sridhar, R., Prasanna, U. U., & Priyanka, G. (2023b). Tax Management in the Digital Age: A TAB Algorithm-based Approach to Accurate Tax

Prediction and Planning. 2023 International Conference on Inventive Computation Technologies (ICICT), 908–915. https://doi.org/10.1109/ICICT57646.2023.10133949

Kumar, R., Khannna Malholtra, R., & Grover, C. N. (2023c). Review on Artificial Intelligence Role in Implementation of Goods and Services Tax (GST) and Future Scope. 2023 International Conference on Artificial Intelligence and Smart Communication (AISC), 348–351. https://doi.org/10.1109/AISC56616.2023.10085030

Lahiri, K., & Yang, C. (2022). Boosting tax revenues with mixed-frequency data in the aftermath of COVID-19: The case of New York. International Journal of Forecasting, 38(2), 545–566. https://doi.org/https://doi.org/10.1016/j.ijforecast.2021.10.005

Lima, D. V. e Castro, R. G. (2000), Contabilidade Pública: Integrando União, Estados e Municípios (Siafi e Siafem), São Paulo: Atlas, 2ª Edição.

Liu, D., Zhang, R., & Li, J. (2011). Tax Revenue Combination Forecast of Hebei Province Based on the IOWA Operator. 2011 Fourth International Joint Conference on Computational Sciences and Optimization, 516–519. https://doi.org/10.1109/CSO.2011.251

Li-Xia, L., Yi-qi, Z., & Yong Xue-Liu. (2011). Tax forecasting theory and model based on SVM optimized by PSO. Expert Systems with Applications, 38(1), 116–120. https://doi.org/https://doi.org/10.1016/j.eswa.2010.06.022

Lu, S., Jian Zhong-Cai, & bin Xiao-Zhang. (2009). Application of GA-SVM time series prediction in tax forecasting. 2009 2nd IEEE International Conference on Computer Science and Information Technology, 34–36. https://doi.org/10.1109/ICCSIT.2009.5234606

Martin, P. (2022) Linear Regression. 1st edn. SAGE Publications Ltd. Available at: https://www.perlego.com/book/3277496 (Accessed: 14 June 2024).

McClure, N. (2017) TensorFlow Machine Learning Cookbook. 1st edn. Packt Publishing. Available at: https://www.perlego.com/book/118231 (Accessed: 23 June 2024).

Mello, A. G. F. (2022), Tributação, Hipernomia e Medo, Santa Catarina: Emais Editora, 1ª Edição.

Montgomery, D., Peck, E. and Vining, G. (2021) Introduction to Linear Regression Analysis. 6th edn. Wiley. Available at: https://www.perlego.com/book/2237423 (Accessed: 13 June 2024).

Noor, N., Sarlan, A., & Aziz, N. (2022). Revenue Prediction for Malaysian Federal Government Using Machine Learning Technique. 2022 11th International Conference on Software and Computer Applications, 143–148. https://doi.org/10.1145/3524304.3524337

Polikar R., Ensemble based systems in decision making, IEEE Circuits and Systems Magazine 6(3):21–45, 2006.

Ravichandiran, S. (2019) Hands-On Deep Learning Algorithms with Python. 1st edn. Packt Publishing. Available at: https://www.perlego.com/book/1031668 (Accessed: 13 June 2024).

Rokach, L. and Maimon, O. (2014) Data Mining With Decision Trees: Theory And Applications (2nd Edition). 2nd edn. WSPC. Available at: https://www.perlego.com/book/851363 (Accessed: 23 June 2024).

Singh, N. (2020) Fundamentals of Deep Learning and Computer Vision. [edition unavailable]. BPB Publications. Available at: https://www.perlego.com/book/1681480 (Accessed: 13 June 2024).

Srinivasan, S. M. and Laplante, P. (2023) What Every Engineer Should Know About Data-Driven Analytics. 1st edn. CRC Press. Available at: https://www.perlego.com/book/3845688 (Accessed: 13 June 2024).

Ticona, W., Figueiredo, K., & Vellasco, M. (2017), Hybrid model based on genetic algorithms and neural networks to forecast tax collection: Application using endogenous and exogenous variables. 2017 IEEE XXIV International Conference on Electronics, Electrical Engineering and Computing (INTERCON), 1–4. https://doi.org/10.1109/INTERCON.2017.8079660

Xu, H., & Ma, M. (2021). An Improved Hybrid Model base on SVM and Random Forest for the Prediction of Corporate Taxation. 2021 IEEE 3rd International Conference on Civil Aviation Safety and Information Technology (ICCASIT), 1035–1038. https://doi.org/10.1109/ICCASIT53235.2021.9633659

Yan, X. and Su, X. (2009) Linear Regression Analysis: Theory And Computing. [edition unavailable]. World Scientific. Available at: https://www.perlego.com/book/849247 (Accessed: 13 June 2024).

Zaw, T., Kyaw, S. S., & Oo, A. N. (2020), ARMA Model for Revenue Prediction. Proceedings of the 11th International Conference on Advances in Information Technology. https://doi.org/10.1145/3406601.3406617

Anexo I

Hiper Parâmetros dos Algoritmos de Aprendizagem de Máquina

1. Regressão Linear

- A implementação utilizou a classe *LinearRegression* do *sklearn.linear_model*
- Para gerar novas características polinomiais a partir das características originais de um conjunto de dados utilizou-se a transformação polinomial com a classe *PolynomialFeatures* do *sklearn.preprocessing*
    - Parâmetro *degree* variando, em cada ciclo de treinamento, de 1 a 6.

2. Floresta Aleatória

- A implementação utilizou a classe *RandomForestRegressor* do *sklearn. ensemble*
- Variou-se, em cada ciclo de treinamento, o parâmetro que controla a profundidade máxima das árvores individuais na floresta aleatória
    - Parâmetro *max_depth* variando entre os valores 3 a 16.

3. Máquina de Vetor Suporte

- A implementação utilizou a classe *SVR* do *sklearn.svm*
- Fixou-se o kernel "rbf" (Gaussiano/Radial Basis Function)
- Variou-se, em cada ciclo de treinamento, o parâmetro de regularização (que controla a penalidade por erros de treinamento)
    - Parâmetro C assumindo os valores 10, 100, 1000, 10000, 100000, 1000000, 10000000 e 100000000
- O parâmetro *gamma* foi ajustado para calcular automaticamente com base nas características dos dados de entrada
    - Parâmetro *gamma* foi definido como "*scale*"

4. Rede Neural Perceptron Multicamadas

- A implementação utilizou a classe *MLPRegressor* do *sklearn.neural_network*
- Estabeleceu-se um loop de 5 repetições de treinamento, percorrendo todas as variações listadas a seguir.
- Em cada ciclo de treinamento, variou-se
    - os neurônios da camada intermediária
        - Parâmetro *hidden_layer_sizes* variando de 30 a 73.
    - A taxa de aprendizagem
        - Parâmetro *learning_rate_init* variando de 0,0010 a 0,0016
- Fixou-se o nº de épocas
    - Parâmetro *max_iter* foi definido como 3000
- Fixou-se a função de ativação para a camada oculta
    - Parâmetro *activation* foi definido como “relu”
- Fixou-se o algoritmo otimizador usado para treinar a rede neural
    - Parâmetro *solver* foi definido como “adam”

5. Rede Neural Convolucional

- A implementação utilizou as classes
    - *Sequential* do *keras.models*
    - *Dense, Conv1D, Flatten* do *keras.layers*
- Em cada ciclo de treinamento, utilizou-se:
    - Na classe *Conv1D:*
        - Parâmetro *filters* variando de 55 a 59.
        - Parâmetro *kernel_size* variando de 1 a 4
        - Parâmetro activation foi definido como “relu”
    - Na classe *Dense:*
        - Parâmetro *units* variando de 17 a 34
        - Parâmetro activation foi definido como “relu”
- Fixou-se o algoritmo otimizador usado para treinar a rede neural
    - Parâmetro *optimizer* foi definido como “adam”