INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DA PARAÍBA
CAMPUS CAJAZEIRAS
CURSO SUPERIOR DE TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

UTILIZAÇÃO DE REDES NEURAIS DE MEMÓRIA DE LONGO E CURTO PRAZO (LSTM) PARA PREVISÃO DA ARRECADAÇÃO MENSAL DE RECEITA ORÇAMENTÁRIA EM CAJAZEIRAS, ESTADO DA PARAÍBA

ALLYSON OLIVEIRA DE ABREU

Cajazeiras
2024

ALLYSON OLIVEIRA DE ABREU

UTILIZAÇÃO DE REDES NEURAIS DE MEMÓRIA DE LONGO E CURTO PRAZO (LSTM) PARA PREVISÃO DA ARRECADAÇÃO MENSAL DE RECEITA ORÇAMENTÁRIA EM CAJAZEIRAS, ESTADO DA PARAÍBA

Trabalho de Conclusão de Curso apresentado junto ao Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia da Paraíba - Campus Cajazeiras, como requisito à obtenção do título de Tecnólogo em Análise e Desenvolvimento de Sistemas.

Orientador

Prof. Me. Francisco Paulo de Freitas Neto.

Cajazeiras

2024

IFPB / Campus Cajazeiras
Coordenação de Biblioteca
Biblioteca Prof. Ribamar da Silva
Catalogação na fonte: Cícero Luciano Félix CRB-15/750

A162u Abreu, Allyson Oliveira de.
Utilização de redes neurais de memória de longo e curto prazo (LSTM) para previsão da arrecadação mensal de receita orçamentária em Cajazeiras, estado da Paraíba / Allyson Oliveira de Abreu.– 2024.

40f. : il.

Trabalho de Conclusão de Curso (Tecnólogo em Análise e Desenvolvimento de Sistemas) - Instituto Federal de Educação, Ciência e Tecnologia da Paraíba, Cajazeiras, 2024.

Orientador(a): Prof. Me. Francisco Paulo de Freitas Neto.

1. Desenvolvimento de sistemas. 2. Sistema de gestão pública. 3. Controle de receitas orçamentárias. 4. *Machine Learning*. I. Instituto Federal de Educação, Ciência e Tecnologia da Paraíba. II. Título.

IFPB/CZ CDU: 004.4:336(043.2)

MINISTÉRIO DA EDUCAÇÃO
SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DA PARAÍBA

ALLYSON OLIVEIRA DE ABREU

UTILIZAÇÃO DE REDES NEURAIS DE MEMÓRIA DE LONGO E CURTO PRAZO (LSTM) PARA PREVISÃO DA ARRECADAÇÃO MENSAL DE RECEITA ORÇAMENTÁRIA EM CAJAZEIRAS, ESTADO DA PARAÍBA

Trabalho de Conclusão de Curso apresentado junto ao Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia da Paraíba - Campus Cajazeiras, como requisito à obtenção do título de Tecnólogo em Análise e Desenvolvimento de Sistemas.

Orientador

Prof. Me. Francisco Paulo de Freitas Neto

Aprovada em: **29 de Fevereiro de 2024.**

Prof. Me. Francisco Paulo de Freitas Neto - Orientador

Prof. Me. João Paulo Freitas de Oliveira - Avaliador

IFPB - Campus Cajazeiras

Prof. Dr. Fabio Gomes de Andrade - Avaliador

IFPB - Campus Cajazeiras

Documento assinado eletronicamente por:

- **Francisco Paulo de Freitas Neto**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 29/02/2024 20:36:23.
- **Joao Paulo Freitas de Oliveira**, **PROF ENS BAS TEC TECNOLOGICO-SUBSTITUTO**, em 01/03/2024 07:46:07.
- **Fabio Gomes de Andrade**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 01/03/2024 16:16:03.

Este documento foi emitido pelo SUAP em 29/02/2024. Para comprovar sua autenticidade, faça a leitura do QRCode ao lado ou acesse https://suap.ifpb.edu.br/autenticar-documento/ e forneça os dados abaixo:

Código 539265
Verificador: 70c3e1ab45
Código de Autenticação:

Aos colegas desenvolvedores, que labutam diariamente na esperança de uma mundo conectado, subsidiado na diversidade e respeito entre os povos.

AGRADECIMENTOS

A todos que estiveram ao meu lado nos momentos mais difíceis:

Em especial à minha família, que por maior seja a distância que estejam, sempre me apoiaram e me deram todo o suporte para continuar nesta jornada;

A meu pai, que sempre se colocou como um exemplo a ser seguido e mostrar-me qual caminho percorrer, apesar dos pesares;

À minha mãe, pelos palavras de motivação e apoio, apesar dos pesares;

Aos meus amigos, que em meus dias mais nebulosos me colocavam sobre meus pés e incentivavam a encarar mais uma lide;

Meus mais sinceros agradecimentos.

”God machine
Malfunctioned as it grew
And the circuits blew
Falling down on you
Now you’re free
Unplug from the source
No more underscores
Open up the door”

Greta Van Fleet - Age of Machine

”In God we trust;
All the other must bring data.”

W. Edwads Deming

RESUMO

A previsão de receitas orçamentárias tornou-se crucial para a gestão eficaz dos órgãos públicos, especialmente devido às incertezas econômicas e políticas que impactam diretamente esses dados ao longo do tempo. Este estudo apresenta uma iniciativa de investigação que utiliza técnicas avançadas de Machine Learning, especificamente o modelo *Long Short-Term Memory (LSTM)*, para prever as receitas orçamentárias. A aplicação desse modelo demonstrou resultados promissores, apresentando um erro relativo de previsão pouco superior a 1%, indicando uma eficácia na predição das receitas. Essa abordagem moderna e precisa destaca-se como uma potencial ferramenta para os órgãos governamentais, proporcionando uma visão mais clara e preditiva das receitas, mesmo em contextos marcados por volatilidade econômica e incertezas políticas.

Palavras-chave: Receita Orçamentária. *Machine Learning*. Previsão.

ABSTRACT

Forecasting budget revenues has become crucial for effective management of public agencies, especially due to the economic and political uncertainties that directly impact these data over time. This study presents a research initiative that utilizes advanced Machine Learning techniques, specifically the Long Short-Term Memory (LSTM) model, to predict budget revenues. The application of this model has shown promising results, with a relative forecasting error just over 1%, indicating effectiveness in revenue prediction. This modern and accurate approach stands out as a potential tool for government agencies, providing a clearer and predictive insight into revenues, even in contexts marked by economic volatility and political uncertainties.

Keywords:Budgetary Revenue. Forecasting. Machine Learning.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CSV	*Comma Separated Values*
CRFB/88	Constituição da República Federativa do Brasil de 1988
DS	*Data Science*
LAI	Lei de Acesso à Informação
LGPD	Lei Geral de Proteção de Dados
SGBD	Sistemas de Gerenciamento de Banco de Dados
TICs	Tecnologias das Informações e Computacionais
RNRs	Redes Neurais Recorrentes
RMLP	Rede de Memória Longa de Curto Prazo

SUMÁRIO

1 INTRODUÇÃO

A conjuntura atual da Administração Pública no Brasil reflete a pressão incessante por uma significativa melhoria na prestação de seus serviços, seja para as empresas ou para os cidadãos, visando atender de maneira mais eficaz às demandas sociais.

A Administração Púbica lidera uma nova abordagem no aprimoramento do conjunto de ferramentas destinado à gestão pública. Denominada por alguns autores como *"New Public Service"*(Denhardt e Denhardt (2000)), *"Post-New Public Management"*(Dunleavy et al. (2006)), e também *"New Public Governance"*(Osborne (2006)), essa nova abordagem apresenta três visões distintas para a AP, com ideais e estruturas totalmente diferentes, mas convergentes quando se trata de visões alternativas sobre melhorias nas prestações de serviços públicos.

De acordo com Susar e Aquaro (2019) e Araújo e Souza (2011), este modelo de gestão pública coloca o cidadão, em contraste com o modelo clássico, como ponto focal dos investimentos em serviços públicos. Ao inverter os papéis, altera-se a forma e a função organizacional, direcionando o foco para o serviço público, com esforços voltados para atender às expectativas do cidadão e promover a prestação de contas à população. Este papel atua como um suporte à transparência do orçamento público, inerente ao interesse público.

Diante do exposto, este capítulo apresentará a definição do problema que deu origem aos questionamentos fundamentais para este projeto, a solução proposta para essa problemática, os objetivos gerais e específicos perseguidos, a metodologia utilizada para atingir esses objetivos, e o cronograma adotado para a organização e alcance das metas da implementação do projeto.

1.1 DEFINIÇÃO DO PROBLEMA

No contexto de um mundo amplamente conectado, onde interações diárias, expressões de opiniões e compartilhamentos de inquietações ocorrem incessantemente nas redes sociais, o volume expressivo de audiência em plataformas de *streaming*, transações financeiras online e o constante fluxo de informações armazenadas por empresas são notáveis. Esse cenário contemporâneo, marcado por termos guarda-chuvas como *data science* e *big data*, reflete a era da conectividade e da geração massiva de dados. Contudo, a pandemia de COVID-19, que assolou recentemente,

desencadeou uma situação peculiar, alterando significativamente a dinâmica social e profissional.

Diante dessa necessidade de mudança, o Poder Público teve que se adaptar, implementando medidas sanitárias rigorosas e reorganizando a prestação de serviços essenciais. No entanto, o aumento exponencial da transmissão de dados durante esse período evidenciou lacunas na capacidade do setor público em disseminar efetivamente informações sobre as medidas adotadas e os gastos públicos realizados. Mesmo com aprimoramentos nos sistemas de transparência, a realidade brasileira, especialmente nos portais da transparência, apresenta desafios consideráveis. Termos técnicos complexos e tabelas extensas dificultam a compreensão dos cidadãos, tornando a interpretação dos dados uma tarefa árdua.

Autores como Araújo e Souza (2011) destacam há mais de uma década a necessidade de uma abordagem mais acessível na apresentação de dados públicos, visando transformá-los em informações úteis e compreensíveis. O desafio reside na inabilidade de converter dados abertos em conhecimento acessível e aplicável.

Administrar recursos e consolidar uma estratégia sem uma base autossustentável e sólida que sirva como ponto de partida para a tomada de decisões dos gestores são ferramentas essenciais para modernizar a aplicação desses recursos e avaliar a eficácia dos programas governamentais e estaduais financiados pela máquina pública.

Ademais, a administração pública brasileira enfrenta obstáculos na gestão estratégica de recursos, marcada pela burocracia excessiva, morosidade nos processos administrativos e carência na capacitação específica dos servidores. Esta realidade contrasta com o potencial das ferramentas tecnológicas disponíveis, que, mesmo sendo parte integrante do cotidiano da população, não são plenamente exploradas em algumas repartições públicas.

Inserir e estimular o uso de aplicações que utilizam princípios de aprendizado de máquina a partir de dados cotidianos na máquina pública surge como uma alternativa para um planejamento orçamentário mais consolidado, que deixa de lado a mera especulação sobre o que pode ocorrer no ano financeiro vindouro e passa a utilizar acontecimentos passados, fixando sazonalidades que são comuns a determinados espaços de tempo e que podem voltar a se repetir.

Neste contexto, surge a necessidade de uma atuação estratégica da gestão, na fase inicial de planejamento da utilização dos recursos por meio da implementação de ferramentas de *machine learning*. Este trabalho propõe a análise e previsão de receitas

orçamentárias, utilizando dados disponibilizados pelo setor público referentes aos gastos e receitas geridos pela Prefeitura Municipal de Cajazeiras, Estado da Paraíba, no período de 2013 a 2022[1]. Este enfoque busca promover uma gestão mais eficiente e transparente, alinhada aos desafios contemporâneos e às demandas da sociedade.

1.2 DESCRIÇÃO DA SOLUÇÃO

No contexto de uma sociedade organizada em um território conduzido por um governo eleito, a dinamicidade do Estado moderno, tanto no âmbito privado quanto no público, redefiniu a abordagem necessária para fiscalizar as contas públicas. Essa mudança para um Poder Público mais ágil e incisivo demandou uma revisão nas práticas de transparência e gestão financeira.

Ao tratar da temática legal dos orçamentos públicos, dois diplomas foram fundamentais: a Constituição da República Federativa do Brasil e a Lei n°. 4.320, de 17 de Março de 1964, que estabeleceu as normas gerais de Direito Financeiro Brasileiro. Além disso, em relação à obtenção e distribuição de dados no Brasil, destacam-se o Marco Civil da Internet, a Lei Geral de Proteção de Dados (LGPD), e a Lei de Acesso à Informação (LAI). Esses marcos legais refletiram os avanços tecnológicos e a necessidade de clareza na disponibilização de dados governamentais para a sociedade.

As mudanças tecnológicas influenciaram movimentos que demandaram métodos mais transparentes de compartilhamento de informações. Autores como Funk et al. (2008) e Valle-Cruz et al. (2019) ressaltaram que essa transição redefiniu a comunicação e o compartilhamento de informações na sociedade brasileira.

O avanço tecnológico, especialmente no campo da informação e comunicação, transformou não apenas a vida cotidiana dos cidadãos, mas também a atuação diária do Poder Público. As Tecnologias da Informação e Comunicação (TICs) não só adiantaram e otimizaram tarefas diárias, mas também introduziram desafios, encargos e questões políticas e administrativas inéditas nos negócios públicos. Cristóvam Lucas Bossoni Saikali (2020) destacaram a necessidade de as instituições públicas se adaptarem rapidamente às mudanças proporcionadas pelas TICs, não apenas em termos institucionais, mas também na percepção dos impactos que a não adaptação poderia gerar. A inovação no setor público, conforme Cubuk et al. (2019), foi um processo que envolveu a gestão de ferramentas para a prestação de serviços públicos por meio de processos originais, buscando respostas inovadoras através de novas ideias, serviços e formas de aplicá-los.

[1] Site utilizado na coleta dos dados: <https://www.cajazeiras.pb.gov.br/acessoainformacao.php?id=3&emed=1>, acesso em 04 de março de 2024

Um modelo de *machine learning* dedicado à previsão de receitas orçamentárias poderia conferir inúmeras vantagens à administração pública municipal. Primeiramente, a utilização de tal modelo permitiria uma abordagem mais precisa e preditiva no que diz respeito à arrecadação, oferecendo *insights* valiosos sobre tendências e padrões históricos. Isso possibilitaria uma tomada de decisão embasada em dados concretos, contribuindo para a eficiência na gestão financeira.

Além disso, ao antecipar possíveis variações nas receitas, o modelo de *machine learning* proporcionaria uma ferramenta estratégica para o planejamento orçamentário. A administração pública municipal poderia se preparar de forma proativa para eventuais períodos de instabilidade econômica, ajustando suas prioridades e direcionando recursos de maneira mais eficaz.

A transparência também seria fortalecida, uma vez que a previsão de receitas por meio de *machine learning* permitiria uma comunicação mais clara e compreensível para a população. Ao disponibilizar informações detalhadas sobre as projeções financeiras, a administração promoveria uma prestação de contas mais robusta, favorecendo a confiança dos cidadãos na gestão pública. Ademais, a automação desse processo reduziria a carga burocrática e o tempo dedicado a tarefas manuais, liberando recursos humanos para atividades mais analíticas e estratégicas. Dessa forma, a administração pública poderia aprimorar sua eficiência operacional e focar em iniciativas inovadoras.

Em suma, a introdução de um modelo de *machine learning* na previsão de receitas orçamentárias representou uma evolução significativa na gestão municipal, proporcionando maior precisão, planejamento estratégico, transparência e eficiência operacional. Essas vantagens contribuíram para uma administração mais ágil, adaptável e alinhada às demandas dinâmicas da sociedade contemporânea.

1.3 OBJETIVOS

1.3.1 Objetivos gerais

Desenvolver e avaliar um modelo preditivo utilizando aprendizagem de máquina para prever a arrecadação de receita orçamentária, empregando dados históricos do município de Cajazeiras, Estado da Paraíba.

1.3.2 Objetivos específicos

- Análise detalhada dos dados históricos de arrecadação financeira, identificando padrões, tendências e *insights* relevantes que possam influenciar nas previsões

futuras;

- Criação e configuração de Redes Neurais *LSTM* capazes de prever séries temporais de arrecadação financeira com base em dados históricos, utilizando diversas arquiteturas e configurações ideais para garantir a eficácia e precisão das previsões;
- Realizar comparações entre os modelos *LSTM* desenvolvidos e métodos tradicionais de previsão de séries temporais, destacando a superioridade dos modelos *LSTM* em termos de precisão e eficácia na previsão de arrecadação financeira;
- Investigr como os hiperparâmetros, como o número de neurônios, tamanho do lote (*batch size*) e taxa de *dropout*, influenciam na precisão das previsões dos modelos *LSTM*, identificando as configurações mais adequadas para otimizar o desempenho dos modelos;
- Avaliar como os modelos *LSTM* desenvolvidos se comportam em cenários reais de curto prazo em diferentes contextos financeiros, determinando a adaptabilidade dos modelos, fornecendo *insights* sobre sua utilidade e eficácia na prática.

1.4 METODOLOGIA

Visando ao melhor entendimento do leitor sobre como se deu a divisão dos processos que levaram à construção deste documento, abaixo segue as principais fases e metodologias utilizadas pelo autor para tal.

- **Início das pesquisas**: Nesta fase, fora realizado estudo bibliográfico sobre os temas que cercavam todas as propostas que este TCC estaria voltado a ser construído. Verificar como estavam os dados abertos orçamentários, no que tange à forma de sua disponibilização ao público, e se havia o uso de alguma solução de *machine learning* utilizados pelas entes e órgão públicos. Por se tratar de fase essencial para todo o sustentáculo do projeto, fora revisitada inúmeras vezes durante a execução deste;
- **Análise dos requisitos**: Aqui foi efetuado as tarefas referentes ao processo de análise dos requisitos para a implementação do algoritmo de aprendizado de máquina. As principais atividades realizadas nesta fase foram: análise exploratória dos dados e coleta de dados que não estavam no corpo disponibilizado pela prefeitura - como dados de população, índice de desenvolvimento humano e produto interno bruto;

- **Implementação do algoritmo**: Realizado o processo de coleta, seguimos para a implementação do algoritmo, escoltado pelos requisitos da segunda fase, ocorre as inúmeras rodadas de treinamentos e verificação dos componentes que fazem parte do treinamento do modelo.

1.5 ORGANIZAÇÃO DO DOCUMENTO

O presente trabalho está estruturado em três capítulos, delineando uma abordagem coesa e progressiva na análise e implementação do modelo proposto.

O segundo capítulo concentra-se na fundamentação teórica essencial para a compreensão do trabalho. O terceiro capítulo detalha o passo a passo da implementação do algoritmo de aprendizado de máquina proposto. Desde a seleção e preparação dos dados até a escolha e treinamento do modelo, cada etapa é descrita de maneira a proporcionar uma compreensão clara do processo. São abordadas as ferramentas utilizadas, as técnicas empregadas e as decisões tomadas durante a implementação do modelo preditivo de receitas orçamentárias.

O último capítulo apresenta a síntese dos objetivos atingidos ao longo do trabalho. Destacam-se os resultados obtidos com a implementação do modelo de *machine learning* e a análise crítica dos benefícios alcançados. Além disso, são discutidas as perspectivas para a continuação desta obra, explorando possíveis aprimoramentos, expansões ou áreas de pesquisa futura. Este capítulo encerra o documento proporcionando uma visão consolidada do estudo realizado.

2 REFERENCIAL TEÓRICO

Neste capítulo, serão explorados os conceitos fundamentais que embasaram a construção deste projeto. Será trabalhado o significado das redes neurais e suas diferentes utilidades, incluindo um aprofundamento nas redes neurais recorrentes de longo e curto prazo (RMLP).

2.1 REDES NEURAIS ARTIFICIAIS

Inspiradas no funcionamento do cérebro humano, as redes neurais artificiais são modelos computacionais que procuram reproduzir o comportamento do sistema nervoso natural. São utilizadas para aprender e detectar padrões não lineares, possuindo como características a alta capacidade de aprender, memorizar e generalizar.

Para Haykin (1998, pp. 24), redes neurais artificiais são:

> Uma rede neural é um processador distribuído maciçamente paralelo composto por unidades de processamento simples, que possui uma propensão natural para armazenar conhecimento experiencial e torná-lo disponível para uso. Ela se assemelha ao cérebro em dois aspectos:
>
> 1. O conhecimento é adquirido pela rede a partir de seu ambiente por meio de um processo de aprendizado.
>
> 2. As forças de conexão entre os interneurônios, conhecidas como pesos sinápticos, são utilizadas para armazenar o conhecimento adquirido. (tradução nossa)

Simon Haykin descreve as redes neurais como sistemas de processamento distribuído massivamente paralelos, compostos por unidades de processamento simples. A semelhança com o cérebro humano reside na maneira como adquire conhecimento, ou seja, através do ambiente ao qual estão expostas, e na forma como armazena essa experiência, por meio de conexões sinápticas.

Tal qual o cérebro humano, a rede neural adquire conhecimento do ambiente ao seu redor através de um processo de aprendizado, sendo possível aprender com os dados que lhes são transmitidos, ajustando suas conexões e padrões internos.

Para tanto, as forças ou pesos das conexões entre os neurônios, conhecidos como pesos sinápticos, são modificados para armazenar as informações que foram adquiridas no processo de aprendizagem pela rede neural. Estes pesos são ajustados no decorrer do processo, o que possibilita que possam ser feitas previsões, classifica-

ções e demais tarefas. Todo esse processo é chamado pelo autor de *learning algorithm*, Haykin (1998).

O propósito das redes neurais que utilizam o aprendizado de máquina com inteligência artificial (IA) é desenvolver paradigmas ou algoritmos que orientem as máquinas a executarem tarefas cognitivas nas quais as redes neurais biológicas se destacam atualmente. Essa definição de IA é baseada em Sage (1990), ressaltando que não é a única definição aceita de IA.

Um sistema de IA, conforme este autor, deve ser capaz de realizar três tarefas fundamentais: 1) armazenar conhecimento, 2) aplicar o conhecimento armazenado para resolver problemas e 3) adquirir novos conhecimentos por meio da experiência.

Sage (1990) sugere que um sistema de IA é composto por três componentes principais: representação, raciocínio e aprendizado. Cada um desses componentes desempenha um papel crucial no funcionamento e na eficiência do sistema de IA.

1. Representação. A característica mais marcante da IA é o amplo uso de uma linguagem de estruturas simbólicas para representar conhecimentos gerais e específicos sobre a resolução de problemas. Essas representações simbólicas tornam-se compreensíveis para os usuários humanos, o que torna a comunicação humano-máquina mais eficiente. No contexto dos pesquisadores de IA, 'conhecimento' é essencialmente outro termo para dados, que podem ser declarativos ou procedimentais.

2. Raciocínio. Essencialmente, o raciocínio é a capacidade de resolver problemas. Um sistema de IA precisa ser capaz de expressar e resolver uma ampla gama de problemas, tornar explícitas as informações e implementar um mecanismo de controle para determinar quais operações aplicar a problemas específicos.

3. Aprendizado. Neste componente, um elemento de aprendizado utiliza informações do ambiente para aprimorar uma base de conhecimento. O elemento de desempenho utiliza essa base de conhecimento para realizar tarefas. As informações fornecidas à máquina pelo ambiente são frequentemente imperfeitas, o que requer que o elemento de aprendizado opere por suposições e feedback do elemento de desempenho para avaliar e ajustar suas hipóteses conforme necessário.

Dessa forma, a compreensão e a aplicação desses componentes-chave são essenciais para o desenvolvimento e o avanço dos sistemas de IA baseados em redes

neurais, visando alcançar resultados cada vez mais precisos e eficientes em diversos domínios de aplicação.

2.2 REDE NEURAL RECORRENTE

Redes Neurais Recorrentes (RNRs) são estruturas poderosas que incorporam ciclos em seus algoritmos, permitindo a realização de tarefas com entradas sequenciais de forma eficiente. Essa capacidade as torna altamente eficazes em domínios como processamento de linguagem natural, análise de séries temporais e reconhecimento de padrões em dados sequenciais, como fala e música.

Um uso exemplar das RNRs é na previsão de séries temporais, como as receitas orçamentárias municipais. Ao alimentar o modelo com dados históricos das receitas, combinados com informações relevantes, como dados demográficos e econômicos municipais (como população e PIB), é possível fazer previsões futuras. Imagine definir que o modelo receberá uma porcentagem das receitas municipais como entrada e fornecerá uma previsão para o próximo mês. Sem treinamento, o modelo não terá capacidade preditiva relevante. No entanto, ao ser treinado repetidamente com dados orçamentários passados e previsões correspondentes, espera-se que o modelo se ajuste e melhore sua capacidade de previsão para os meses seguintes.

A arquitetura das RNRs baseia-se na conexão recorrente entre neurônios, onde cada neurônio está conectado aos neurônios das camadas anteriores e subsequentes. Cada neurônio possui pesos associados às suas entradas, aos quais é adicionado um valor conhecido como viés *bias* e, em seguida, é aplicada uma função de ativação, como a função sigmoide, para determinar a ativação do neurônio. Esses pesos representam o conhecimento adquirido durante o treinamento e compõem a memória da rede.

É importante compreender que cada neurônio está sintonizado para reconhecer certos padrões ou contextos de entrada. Quanto mais ativado um neurônio está para uma situação específica, mais relevante é para o problema em questão. Em problemas complexos, são utilizados muitos neurônios e camadas, formando o que é conhecido como *Deep Learning*. Esse tipo de arquitetura neural é empregado em aplicações como reconhecimento de voz, processamento de texto e outros domínios de análise de dados complexos.

É crucial ressaltar que, embora os neurônios em uma rede neural executem operações matemáticas, não devem ser diretamente associados a unidades intuitivas de reconhecimento ou processamento de informações pelo cérebro humano. São unidades computacionais que realizam transformações matemáticas sobre os dados de

entrada para produzir as saídas desejadas, obedecendo a regras e padrões aprendidos durante o treinamento da rede.

2.3 REDE DE MEMÓRIA LONGA DE CURTO PRAZO

O modelo das Redes de Memória de Longo Prazo (RMLP), conhecido por Long Short Term Memory (LSTM), é uma extensão das Redes Neurais Recorrentes (RNRs) que emprega unidades específicas denominadas células RMLP. Proposto inicialmente por Sepp Hochreiter e Jürgen Schmidhuber em 1997, este modelo foi projetado para lidar com o desafio de capturar e processar informações em sequências temporais de forma mais eficaz do que as RNRs convencionais.

A arquitetura RMLP é composta por células que incorporam estruturas internas complexas, incluindo três portões principais: *forget gate, input gate e output gate*.

O *forget gate*, ou portão do esquecimento, desempenha o papel de avaliar e decidir quais informações passadas são irrelevantes ou não necessárias para a célula. Este portão permite que a RMLP descarte informações menos importantes, focando nas mais relevantes para a tarefa em questão.

Por outro lado, o *input gate*, ou portão de entrada, tem a função de determinar quais novas informações devem ser adicionadas à célula. Composto por duas camadas com diferentes funções de ativação, as entradas geram novos valores e controlam a quantidade que serão incorporadas ao estado atualizado da célula.

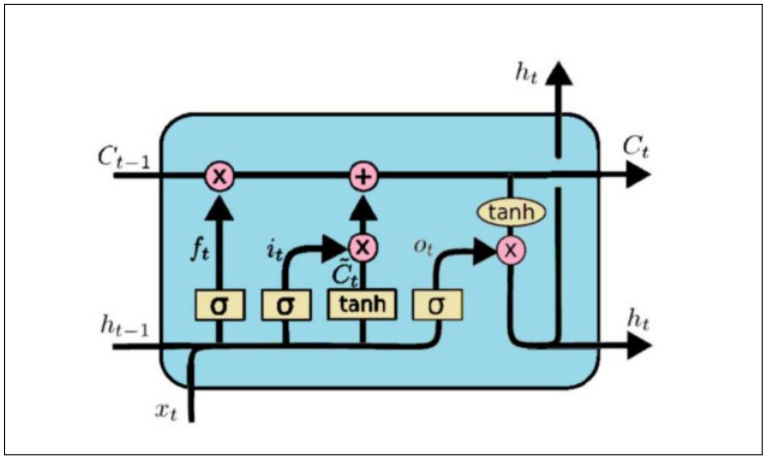
A combinação desses novos valores calculados pelo *input gate* contribui para a atualização do estado da célula, mantendo uma representação interna atualizada das informações relevantes para o contexto da sequência temporal.

Por fim, o *output gate*, ou portão de saída, determina quais partes do estado atualizado da célula serão usadas para calcular a saída da RMLP, gerando um novo estado da célula que incorpora as informações consideradas mais significativas para a tarefa em questão.

As RMLPs, inseridas no campo do aprendizado profundo, oferecem vantagens significativas, como lidar com entradas multivariadas, resistência ao ruído nos dados, previsão de saídas multivariadas, extração automática de características relevantes nos dados e habilidade de modelar relações complexas em sequências temporais. Consequentemente, elas têm apresentado desempenho notável em tarefas de previsão de séries temporais e em diversas outras aplicações no processamento de dados sequenciais e temporais.

A figura 1 apresenta o modelo gráfico do funcionamento de um célula RMLP.

Figura 1 – Célula RMLP



Fonte: Graves e Jaitly (2014)

Em um neurônio RMLP (célula), uma sequência de entrada xt é processada, onde cada gate dentro da célula utiliza unidades de ativação para controlar seu comportamento, permitindo que a informação flua pela célula e atualize seu estado. O parâmetro Ct representa o estado da célula no instante t, que retém informações relevantes do passado. O *forget gate (ft)* decide quais informações devem ser esquecidas, o *input gate (it)* decide quais informações novas devem ser adicionadas, e o *output gate (Ot)* determina o que deve ser produzido com base na entrada e no estado atual da célula. Estes valores são manipulados através de operações de concatenação, multiplicação ou adição, conforme ilustrado no circuito RMLP.

2.4 AVALIAÇÃO DO MODELO

Para avaliar a precisão das previsões, foram empregadas as métricas MAPE (*Mean Absolute Percentage Error*) e RMSE (*Root Mean Square Error*), conforme definido por Cankurt e Subasi (2015). O MAPE expressa a porcentagem da diferença entre os valores reais (Y_i) e os valores preditos ($\hat{Y}_i$) em relação ao valor real, utilizando a fórmula:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right| \times 100 \quad (1)$$

Onde n é o número total de observações. Quanto menor o valor do MAPE, mais preciso é o modelo de previsão.

Por sua vez, o RMSE representa a raiz do erro médio quadrático entre as predições e os valores reais, sendo calculado pela fórmula:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \quad (2)$$

O RMSE fornece uma medida da dispersão dos erros em termos das unidades da variável prevista. Quanto mais próximo de zero, maior a qualidade das predições realizadas pelo modelo. Ambas as métricas são essenciais na avaliação de modelos de previsão, sendo o MAPE uma medida de erro percentual médio e o RMSE uma medida da dispersão dos erros.

O reconhecimento e a utilização de métricas consistentes são práticas fundamentais em trabalhos acadêmicos, como evidenciado pelos estudos de Yamak et al. (2019), Yurtsever (2021) e Wan et al. (2020). Embora essas pesquisas tratem de temas distintos, compartilham a implementação de Redes de Memória de Longo Prazo (RMLPs) e a avaliação desses modelos por meio das mesmas métricas. Essa abordagem padronizada na avaliação proporciona uma base sólida para a comparação e análise de resultados entre diferentes estudos.

A escolha das métricas apropriadas desempenha um papel crucial na avaliação de modelos de aprendizado de máquina, pois oferece uma visão objetiva do desempenho e da eficácia do modelo. O uso consistente dessas métricas contribui para a reprodutibilidade e a comparabilidade dos resultados, permitindo uma análise mais precisa das contribuições de cada estudo para o campo de pesquisa.

3 APLICAÇÃO E MODELAGEM

O presente capítulo insere os dados que forma utilizados no treinamento do modelo e explorará a estrutura das redes neurais compostas por duas camadas RMLP, destacando sua capacidade aprimorada de representação e sua tendência ao sobreajuste. Para mitigar esse problema, são empregadas camadas *dropout*, reduzindo a interconexão entre neurônios. Além disso, será abordada a adição de uma camada *dense*, altamente interconectada, para processamento final e aprendizado dos dados.

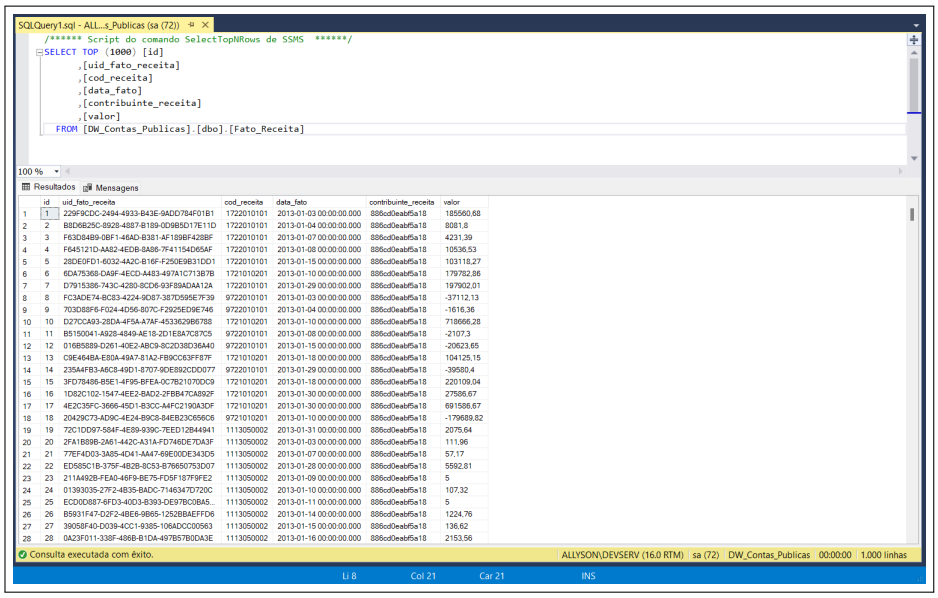
3.1 DADOS UTILIZADOS PARA TREINAMENTO DO MODELO

Os dados foram obtidos do sítio do portal da transparência do município de Cajazeiras[1]. As informações sobre as receitas estão dispostas em formato de tabela, com colunas específicas para a individualização de cada item.

Inicialmente, realizou-se a importação dos dados no formato csv (*comma separated values* ou valores separados por vírgulas) referentes ao período de 2013 a 2022. Após a etapa de extração, foi elaborado um *Data Warehouse* para armazenar e manipular dinamicamente essas informações. Utilizando o Microsoft SQL Server, desenvolveu-se uma solução de *business intelligence* para persistir os dados de forma organizada, incluindo informações sobre o fato gerador da receita, seu tipo e contribuinte, conforme mostra a figura 2.

[1] Site utilizado: <https://www.cajazeiras.pb.gov.br/acessoainformacao.php?id=3&emed=1>, acesso em 04 de março de 2024

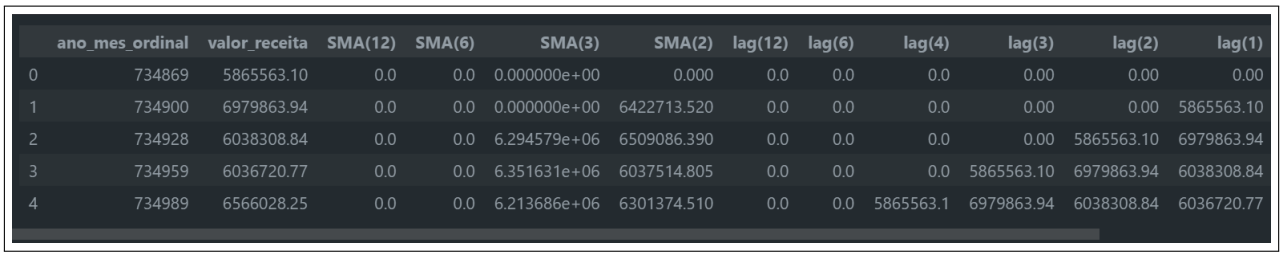
Figura 2 – *Data Warehouse* implementado para persistência das informações colhidas no site.



Fonte: Elaborado pelo autor

Após o processo de extração e transformação dos dados, iniciou-se a análise da qualidade dos mesmos. Por meio de métodos de *feature engineering*, buscou-se reduzir a individualização de cada linha atribuída a uma dada receita orçamentária. Colunas que funcionavam como chaves estrangeiras para outras tabelas auxiliares foram eliminadas, pois carregar um modelo com tanta informação desnecessária poderia induzi-lo a apresentar métricas imprecisas ou produzir resultados incorretos durante sua utilização. Informações adicionais que não estavam disponíveis no sítio da prefeitura foram agregadas à tabela, como dados sobre a população (incluindo sua variação entre os anos e sua taxa de crescimento) e a produção de riqueza (PIB) municipal. Ao final, os dados utilizados no treinamento e teste dos modelos estão representados nas figuras 3 e 4.

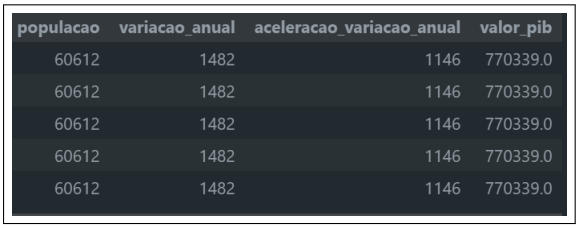
Figura 3 – *Sample* dos dados utilizados para treinamento do modelo, primeira parte.



	ano_mes_ordinal	valor_receita	SMA(12)	SMA(6)	SMA(3)	SMA(2)	lag(12)	lag(6)	lag(4)	lag(3)	lag(2)	lag(1)
0	734869	5865563.10	0.0	0.0	0.000000e+00	0.000	0.0	0.0	0.0	0.00	0.00	0.00
1	734900	6979863.94	0.0	0.0	0.000000e+00	6422713.520	0.0	0.0	0.0	0.00	0.00	5865563.10
2	734928	6038308.84	0.0	0.0	6.294579e+06	6509086.390	0.0	0.0	0.0	0.00	5865563.10	6979863.94
3	734959	6036720.77	0.0	0.0	6.351631e+06	6037514.805	0.0	0.0	0.0	5865563.10	6979863.94	6038308.84
4	734989	6566028.25	0.0	0.0	6.213686e+06	6301374.510	0.0	0.0	5865563.1	6979863.94	6038308.84	6036720.77

Fonte: Elaborado pelo autor

Figura 4 – *Sample* dos dados utilizados para treinamento do modelo, segunda parte.



populacao	variacao_anual	aceleracao_variacao_anual	valor_pib
60612	1482	1146	770339.0
60612	1482	1146	770339.0
60612	1482	1146	770339.0
60612	1482	1146	770339.0
60612	1482	1146	770339.0

Fonte: Elaborado pelo autor

Em suma, a inicial dos dados envolveu a obtenção, importação e transformação destes, disponíveis no portal da transparência. A construção do *Data Warehouse* permitiu o armazenamento organizado e a manipulação dinâmica dessas informações, visando à análise da qualidade dos dados e à preparação para a modelagem. Por meio de técnicas de *feature engineering*, foram realizados ajustes para reduzir a complexidade dos dados, tornando-os mais adequados para a construção de modelos preditivos. Com a inclusão de informações complementares, como dados populacionais e de produção de riqueza municipal, o conjunto de dados foi enriquecido para garantir uma análise mais abrangente.

3.2 PRIMEIRO CONTATO COM OS DADOS E MODELAGEM CONCEITUAL

A estrutura da rede neural, composta por duas camadas RMLP, foi projetada com a intenção de criar um modelo capaz de representar com maior precisão e complexidade as nuances dos dados. Essa configuração reforçada de representação

significa que a rede tem a habilidade de adaptar-se a padrões complexos nos dados de treinamento e, ao mesmo tempo, generalizar esses padrões de maneira mais eficaz para dados não observados. No entanto, a utilização de redes mais profundas traz consigo o desafio do sobreajuste.

O sobreajuste, uma dificuldade comum em modelos de aprendizado de máquina, ocorre quando o modelo se ajusta excessivamente aos dados de treinamento, mas não consegue generalizar bem para dados novos. Para combater esse fenômeno, incluíram-se camadas *dropout* no modelo. Estas camadas funcionam reduzindo aleatoriamente as conexões entre neurônios, permitindo que cada neurônio aprenda de forma mais independente, sem depender excessivamente da colaboração com neurônios vizinhos. Isso ajuda a reduzir a dependência de uma grande quantidade de neurônios interconectados, evitando um ajuste exagerado da rede aos dados de treinamento.

Adicionalmente, uma camada *dense* foi adicionada ao modelo, a qual é altamente conectada, permitindo que cada neurônio dessa camada receba entrada de todos os neurônios da camada anterior. Essa configuração proporciona uma estrutura densa e interconectada, permitindo operações mais refinadas de processamento e aprendizado dos dados.

As redes neurais passam por iterações de treinamento e validação para aprimorar seu desempenho. Durante o treinamento, recalibram-se os pesos da rede a partir do erro entre as previsões e os valores reais dos dados de treinamento. A avaliação do modelo ocorre com a métrica de erro médio quadrático (MSE), a qual é calculada somando as diferenças quadradas entre as previsões e os valores reais.

Cada configuração de arquitetura neural é submetida a múltiplas etapas de treinamento e validação. A determinação dos Hiper Parâmetros, como o número de unidades ocultas na camada RMLP, o uso de *dropout* como método de regularização, e o ajuste do *batch size*, foram conduzidas considerando uma ampla variação entre valores mínimos e máximos. Esse processo busca encontrar faixas de valores que resultem nos menores erros de previsão.

Para selecionar a melhor arquitetura para as redes neurais RMLP neste estudo, foi crucial calcular o erro médio de validação usando MSE em diferentes combinações de parâmetros. O número de unidades ocultas influencia diretamente a complexidade e a capacidade de aprendizado da rede, enquanto o *dropout* atua na regularização para evitar ajustes excessivos aos dados de treinamento, melhorando a capacidade de generalização do modelo para novos dados. O *batch size*, ao limitar o número de amostras processadas durante o treinamento, desempenha um papel crucial na

eficiência e precisão da rede neural.

3.3 PRÉ-PROCESSAMENTO DOS DADOS E APLICAÇÃO

Com o objetivo de analisar minuciosamente os resultados provenientes das Redes Neurais Recorrentes do tipo RMLP, o experimento foi concebido para realizar previsões iterativas de curto prazo. Essas previsões são consideradas de curto prazo devido ao horizonte de tempo abordado, concentrando-se em projeções para um mês adiante (T+1). O caráter iterativo dessas previsões é estabelecido por meio de um processo em *loop*, onde o algoritmo inicialmente prevê um valor para o próximo mês, compara esse valor previsto com a arrecadação real deste mês e, em seguida, registra o erro obtido. O valor previsto é então substituído pelo valor real da arrecadação, permitindo que o algoritmo execute outra previsão, dessa vez para o mês subsequente (T+2). Este padrão iterativo é repetido continuamente ao longo do intervalo de estudo, compreendendo 108 meses de previsões (T+1) de janeiro de 2013 a dezembro de 2022.

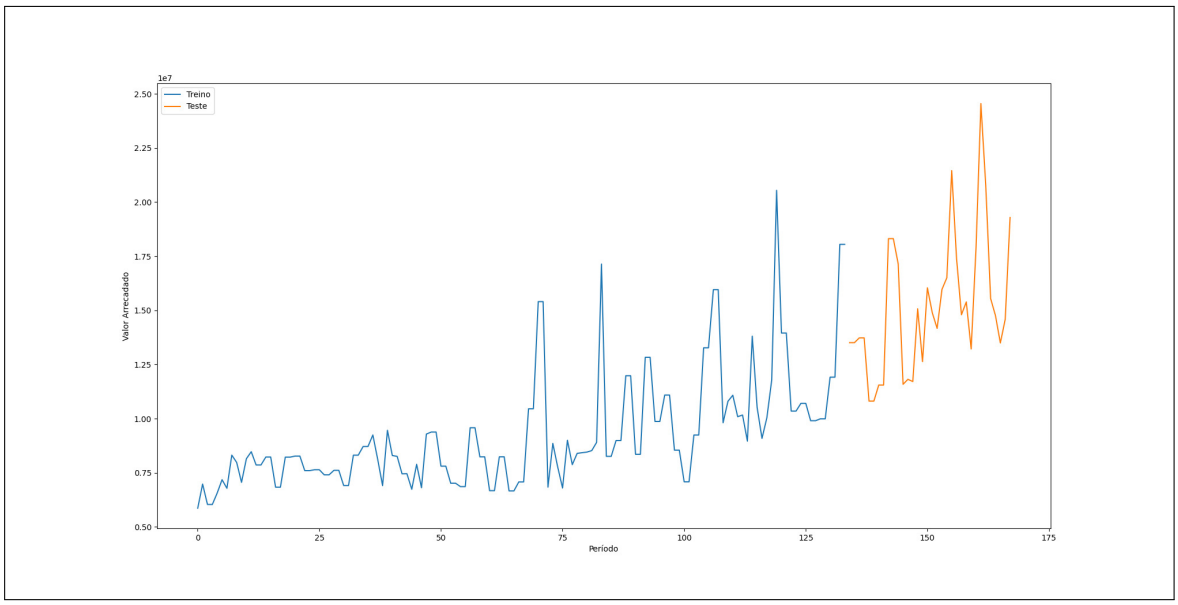
Cada previsão em (T+1) foi realizada usando subamostras distintas dos dados. Para a previsão de janeiro de 2013, por exemplo, utilizou-se uma subamostra composta por dados até dezembro de 2012. Posteriormente, para a previsão de fevereiro de 2013, o modelo foi executado novamente, utilizando uma subamostra que incluía dados até janeiro de 2013. Esse procedimento foi repetido para cada previsão subsequente, avançando mês a mês e utilizando uma nova subamostra que contemplava os dados até o mês imediatamente anterior. Em resumo, a cada nova previsão (T+1), o modelo foi recalibrado e reexecutado, utilizando uma subamostra atualizada dos dados disponíveis até aquele momento. Este processo iterativo permitiu avaliar o desempenho do modelo RMLP em previsões sucessivas, ajustando-se e testando-se constantemente de acordo com os dados mais recentes disponíveis.

Este estudo utilizou uma série temporal derivada dos registros financeiros mensais da cidade de Cajazeiras e seus registros consistiam em dados que abrangiam a arrecadação financeira mensal expressa em valores nominais em reais (R$), começando a partir do mês de janeiro de 2013. A base de dados foi meticulosamente segmentada em duas seções distintas: a base de treinamento, composta pelos 75% iniciais dos dados, e a base de testes, abarcando os 25% restantes e mais recentes, conforme ilustrado na figura 5. Para manipular as informações de data de maneira eficiente, foi empregada uma função da biblioteca *Datetime* em Python, possibilitando a conversão dos registros temporais para o ordinal gregoriano, uma representação numérica de datas. Essa abordagem permitiu uma melhor compreensão temporal dos dados, possibilitando sua utilização efetiva no processo de treinamento e avaliação do

modelo de previsão temporal.

A figura 5 representa a série temporal dos dados utilizados no treinamento do modelo.

Figura 5 – Time series Receitas



Fonte: Elaborado pelo autor

Ambos conjuntos de dados, o de treino e o de teste, foram escalados utilizando a ferramenta *MinMaxScaler*, presente na biblioteca Scikit-learn (ou scikits.learn)[2], uma poderosa ferramenta de aprendizado de máquina de código aberto desenvolvida para a linguagem de programação Python. Essa biblioteca oferece uma variedade de algoritmos para classificação, regressão e agrupamento, como máquinas de vetores de suporte, florestas aleatórias, *gradient boosting*, *k-means* e DBSCAN, e é projetada para interagir de maneira eficiente com as bibliotecas científicas NumPy e SciPy. O processo de treinamento consistiu na normalização dos dados, traduzindo cada característica individualmente para um intervalo entre zero e um, com base nos valores observados no conjunto de treinamento.

Além disso, foi necessário redimensionar as matrizes para três dimensões, uma exigência para a entrada em uma rede RMLP. A primeira dimensão representa o tamanho do lote, a segunda dimensão indica o número de etapas temporais que

[2] Disponível em: https://scikit-learn.org/stable/

alimentam uma sequência e, por fim, a terceira dimensão corresponde ao número de unidades em uma sequência de entrada.

Partindo do conjunto de dados original, foram criadas novas variáveis por meio de manipulações matemáticas, conforme descrito nos estudos de Silva e Figueiredo (2020). Essas manipulações incluem médias móveis e atrasos nos valores originais da série.

Essas variáveis recém-criadas agregam informações valiosas ao conjunto de dados, incluindo médias móveis, atrasos temporais e características demográficas e econômicas do município, tornando-se cruciais para o modelo preditivo, pois fornecem *insights* fundamentais para as previsões de curto prazo.

3.3.1 Otimização dos hiper parâmetros

A seleção criteriosa e otimizada dos parâmetros, aplicada a uma arquitetura de redes neurais, muitas vezes se destaca como a diferença crucial entre redes de desempenho mal otimizadas e aquelas altamente eficazes. O consenso geral é que essa fase desempenha um papel crítico no resultado final do modelo. No entanto, é notável a escassez de literatura científica disponível para avaliar de maneira concreta os impactos reais dessa otimização.

Nesse contexto, visando mitigar a arbitrariedade inerente a essas definições e introduzir uma abordagem mais científica neste estudo, foram estabelecidos para cada hiperparâmetro uma gama de valores potenciais. Em seguida, esses valores foram submetidos a processos iterativos a fim de determinar quais se mostravam mais adequados para a base de dados utilizada. Este método visa trazer mais fundamentação e embasamento científico à seleção de parâmetros, reduzindo a dependência de práticas arbitrárias e promovendo uma escolha mais embasada e efetiva para a configuração dos hiperparâmetros na arquitetura das redes neurais.

A tabela 1 apresenta os parâmetros utilizados para os testes realizados para a escolha no número de neurônios de cada unidade.

Nessa etapa inicial de otimização de hiperparâmetros, foi observado que o intervalo mais propício para o número de neurônios (*units*) que conduziu ao menor erro está em 32 unidades. Em decorrência dessa constatação, foram conduzidos experimentos subsequentes para se aproximar do número ideal de unidades que promovesse não somente o menor erro, mas também uma menor dispersão nos resultados obtidos. Após essa nova bateria de testes, foi constatado que a estrutura com 32 unidades apresentou o desempenho mais otimizado, evidenciando não apenas

Tabela 1 – Resultado de testes com neurônios para LSTM e suas métricas

Parâmetro	Média	Desvio Padrão
32	2.369	1.223
60	4.590	2.778
64	4.507	2.730
65	4.249	2.832
70	3.416	2.542
75	4.127	3.301
80	3.707	1.456
85	4.805	2.424
90	7.183	5.435
128	3.719	2.622

Fonte: Elaborado pelo autor

um menor erro, mas também uma dispersão mais controlada dos resultados, o que sugere uma melhor capacidade de generalização do modelo em relação aos dados não observados.

Para o parâmentro seguinte, qual seja, o *batch size*, ou tamanho do lote de entrada na célula neural, fora observado os valores presentes na tabela 2, a nível de testes e ajustes.

Tabela 2 – Resultado de testes com o tamanho do lote de entrada para LSTM e suas métricas

Parâmetro	Média	Desvio Padrão
1	5.431	2.250
2	2.637	1.239
3	5.940	4.105
4	4.246	5.777
5	2.515	1.612
6	1.911	1.000
7	1.477	0.443
8	1.431	0.638
9	1.231	0.495
10	1.435	0.586
11	1.219	0.267
12	1.246	0.215

Fonte: Elaborado pelo autor

A escolha do valor 11 para o *batch size* foi feita considerando as métricas obtidas nos experimentos realizados. Após uma série de iterações com diferentes valores para o *batch size*, verificou-se que a configuração com tamanho de lote em 11 apresentou a menor média de erro, com um valor de 1.219, e uma dispersão baixaa, expressa pelo desvio padrão em 0.267, indicando uma consistência considerável nos

resultados. Embora alguns lotes menores tenham demonstrado médias ligeiramente mais baixas em alguns casos, o valor 11 se destacou por manter uma média de erro baixa e uma dispersão relativamente reduzida, sugerindo uma tendência a gerar previsões consistentes e com menor variação nos resultados. Essa escolha baseou-se na busca por um equilíbrio entre a precisão das previsões e a estabilidade dos resultados ao longo das iterações, indicando uma adequada capacidade de generalização do modelo para novos dados.

Por fim, ao tratarmos de um valor fixo para o *dropout* de informações, visando um treinamento com as mais variadas porcentagens de saída de dados, fora obtida a tabela 3, a nível de testes e ajustes.

Tabela 3 – Resultado de testes com o *dropout* de dados para LSTM e suas métricas

Parâmetro	Média	Desvio Padrão
0.05	0.029	0.011
0.1	0.035	0.020
0.2	0.026	0.019
0.3	0.013	0.003

Fonte: Elaborado pelo autor

A seleção do valor 0.3 para o parâmetro de *dropout* foi realizada com base nas métricas obtidas durante os experimentos. Após uma análise detalhada dos diferentes valores de *dropout* testados, verificou-se que o valor de 0.3 resultou na menor média de erro, com um valor de 0.013, e uma dispersão mínima, expressa pelo desvio padrão de 0.003. Essa configuração demonstrou consistentemente um desempenho superior em termos de redução do erro médio, mantendo simultaneamente uma estabilidade notável nos resultados ao longo das iterações. Embora valores menores de *dropout* tenham produzido médias ligeiramente inferiores em alguns casos, o *dropout* de 0.3 destacou-se por manter consistentemente um baixo erro médio, juntamente com uma dispersão muito reduzida nos resultados, sugerindo uma capacidade robusta do modelo em generalizar e evitar o sobreajuste. A escolha foi embasada na busca por um equilíbrio entre a minimização do erro médio e a manutenção da estabilidade dos resultados durante o processo de treinamento da rede neural.

Em todas as rodadas de testes e ajustes foram realizadas 10 (dez) rodadas de repetições, utilizando os dados que constam nas figuras 3 e 4, e a variados valores de épocas que, dentro da fase de ajuste dos modelo, atuam como um outro *looping* de testes e validações dos valores de entrada e saída.

3.3.2 Seções de treinamento do modelo

Um modelo de treinamento constitui um conjunto de dados utilizado para instruir um algoritmo de aprendizado de máquina, composto pelos dados de entrada que influenciam diretamente na geração da saída. Esse modelo é empregado para submeter os dados de entrada ao algoritmo, permitindo que a saída processada seja correlacionada com a saída de amostra.

O resultado dessa correlação é essencial para ajustar e modificar o modelo, em um processo iterativo conhecido como *model fitting* ou ajuste do modelo. A qualidade e precisão dos conjuntos de dados utilizados para treinamento e validação exercem um papel crucial na exatidão do modelo final gerado.

O treinamento de um modelo de aprendizagem de máquina representa o procedimento de alimentar um algoritmo de aprendizado de máquina com dados, visando identificar e aprender os melhores valores para todos os atributos envolvidos no processo. Este processo iterativo visa aprimorar a capacidade do modelo em compreender padrões nos dados, a fim de fazer previsões ou tomar decisões mais precisas quando apresentado a novos conjuntos de dados não vistos anteriormente.

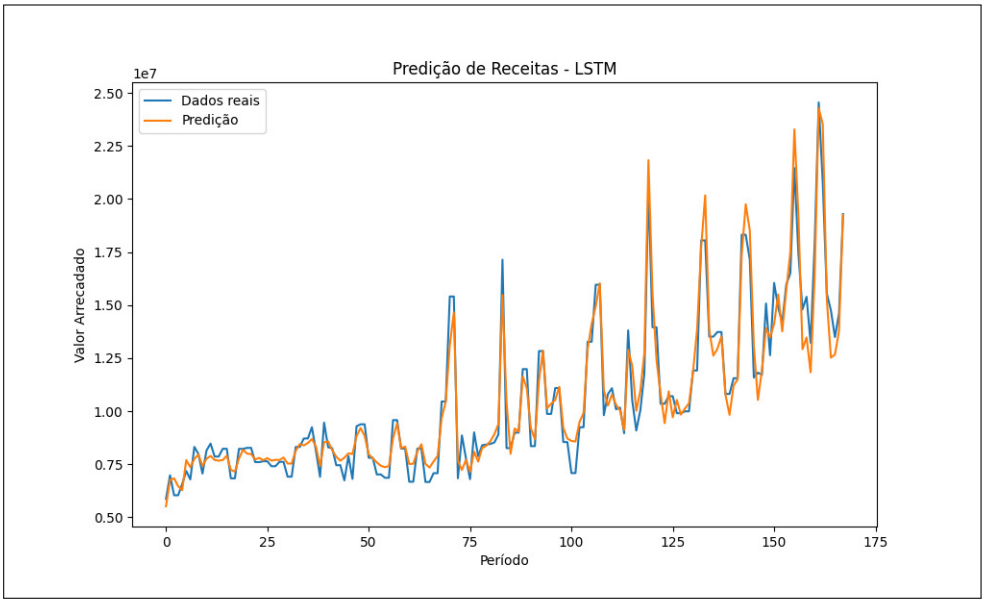
O aprendizado supervisionado, adotado neste estudo, se realiza quando os dados de treinamento incluem tanto os valores de entrada - como as médias móveis e *lags* empregados neste contexto - quanto os resultados de saída, neste caso, a arrecadação das receitas orçamentárias anuais. Cada par de dados que contém as informações de entrada e a respectiva saída desejada é denominado de sinal de supervisão. Durante o treinamento, o modelo é ajustado com base na diferença entre o resultado processado e o resultado documentado quando as entradas são inseridas no modelo.

De maneira mais simplificada e acessível, o treinamento pode ser compreendido como uma etapa em que são fornecidos tanto os dados de entrada quanto os resultados desejados ao algoritmo. Em outras palavras, a máquina tem acesso tanto à informação de entrada quanto ao resultado que se espera dela. Isso permite que a máquina realize seus cálculos de forma independente, mas, ao término desse processo, ela compara o resultado obtido com o resultado esperado, calculando assim o desvio entre eles. Dessa forma, torna-se viável realizar ajustes na sua estrutura matemática com o objetivo de minimizar essa discrepância. No contexto deste estudo, o modelo final foi treinado com camadas RMLP contendo 32 unidades, aplicando um *dropout* de 0.3 e um lote de 11.

A figura 6, apresenta o comparativo entre os dados de testes, referente aos

anos de coleta dos dados utilizados, e as previsões realizadas pelo modelo.

Figura 6 – Comparativo entre a predição e a arrecadação concreta

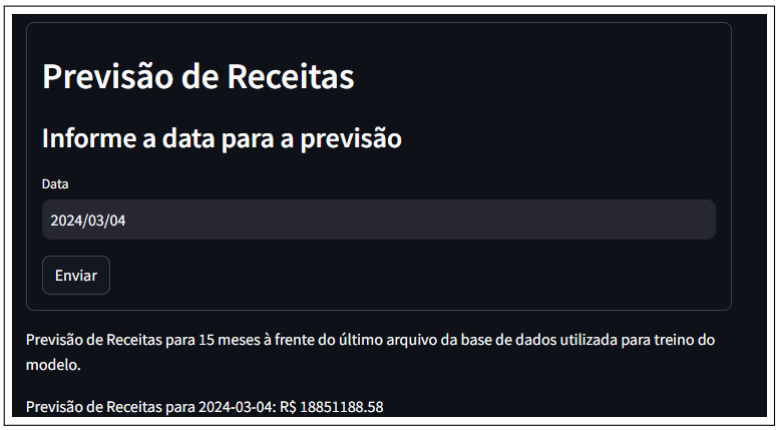


Fonte: Elaborado pelo autor

A análise realizada na figura 6 evidencia que o modelo, a partir da base de treino, conseguiu capturar de maneira eficiente e bem-sucedida o comportamento geral da série temporal. No entanto, é crucial ressaltar que a precisão alcançada neste teste, especialmente nas regiões caracterizadas por uma variação mais expressiva, pode ser atribuída, em parte, à presença de *lags* (retardos temporais) e médias móveis na composição da base de dados utilizada pelo modelo. Essas informações pregressas, incorporadas ao conjunto de dados, contribuem para que o modelo possa reconhecer e se adaptar aos padrões temporais existentes, melhorando sua capacidade de prever variações significativas na série temporal analisada.

A figura 7 mostra a previsão para o mês de março do ano de 2024 do modelo.

Figura 7 – Receita prevista março de 2024



Fonte: Elaborado pelo autor

Essa visualização online oferece uma interface interativa para explorar e compreender as previsões do modelo para o referido período, proporcionando uma ferramenta acessível e transparente para análise e tomada de decisões[3].

Em síntese, o treinamento do modelo de aprendizado de máquina revelou-se uma etapa vital na capacitação do algoritmo para compreender e antecipar padrões nas receitas orçamentárias municipais. A eficiência do modelo, evidenciada pela comparação entre dados de teste e previsões, destaca sua capacidade de capturar o comportamento geral da série temporal.

Notavelmente, a incorporação estratégica de *lags* e médias móveis na base de dados contribuiu significativamente para a precisão das previsões, permitindo ao modelo adaptar-se a padrões temporais complexos. A visualização interativa online da previsão para março de 2024 proporciona uma ferramenta acessível para a análise contínua e a tomada de decisões informadas, reforçando a utilidade prática do modelo na gestão das receitas orçamentárias municipais. Este processo de treinamento, aliado à estrutura matemática ajustada do modelo, resultou em uma abordagem eficaz e promissora para aprimorar a previsão e gestão financeira no âmbito municipal.

[3] Disponível em: https://neuralnetworkstudies-6yuujz5nctvqgek8rhiobx.streamlit.app

4 CONSIDERAÇÕES FINAIS

O presente trabalho é delineado por dois objetivos fundamentais que se alinham à exploração das metodologias para a predição de receitas orçamentárias, especificamente da arrecadação mensal do município de Cajazeiras. O primeiro objetivo é voltado para a apresentação de novas abordagens metodológicas, demonstrando a viabilidade de obter resultados consistentes através dessas ferramentas. O modelo desenvolvido obteve um erro médio percentual de 1.15%, evidenciando sua eficácia na previsão de receitas. No entanto, vale ressaltar que apesar dos esforços, o contato com a ouvidoria do município sobre com acontece o processo de previsão de receitas pelo município não obteve resposta.

O segundo objetivo, de extrema relevância, está centrado aprimoramento do processo de planejamento e na abertura para a busca e inserção de novos métodos de previsão de receitas. É notável que a utilização de redes neurais para análises dessa natureza ainda é um campo em fase inicial, sem uma regulamentação técnica e normativa consolidada. Este estudo visa contribuir para o avanço nesse campo, revelando que o uso de redes neurais em previsão de arrecadação é um território em que há muito a ser explorado cientificamente. O vasto campo do aprendizado de máquina oferece inúmeras bibliotecas que podem proporcionar soluções mais robustas para a base de dados utilizada. Apesar da eficácia demonstrada, outras bibliotecas não abordadas neste estudo, como o *Extreme Gradient Boosting (XGBoost)*, devem ser consideradas e confrontadas com as previsões do modelo atual. Ademais, há a necessidade de adaptar a arquitetura da rede para previsões em intervalos mais longos, ampliando sua aplicabilidade.

As limitações do método proposto devem ser consideradas, especialmente sua complexidade e a incerteza frente a mudanças abruptas nos padrões da série, como variações repentinas decorrentes de fatores externos, tais como mudanças legislativas ou eventos inesperados, como crises econômicas, semelhante a gerada pela pandemia. Além disso, a ausência de uma regulamentação formal e a complexidade da escolha da melhor arquitetura para cada problema constituem desafios substanciais. Por fim, é imprescindível mencionar que este trabalho não pretende substituir modelos estatísticos consagrados por algoritmos que utilizam redes neurais. O caminho para uma análise eficiente reside na colaboração entre diferentes metodologias, permitindo uma análise crítica e objetiva dos resultados, visando sempre o aprimoramento contínuo e a abertura para inovações metodológicas.

REFERÊNCIAS

ARAÚJO, L. de R.; SOUZA, J. F. de. Aumentando a transparência do governo por meio da transformação de dados governamentais abertos em dados ligados. **Revista Eletrônica de Sistemas de Informação**, v. 10, n. 1, 2011.

CANKURT, S.; SUBASI, A. Comparasion of linear regression and neural network models forecasting tourist arrivals to turkey. **Eurasian Journal of Science & Engineering**, 2015.

CRISTóVAM LUCAS BOSSONI SAIKALI, T. P. d. S. José Sérgio da S. Governo digital na implementação de serviços públicos para a concretização de direitos sociais no brasil. 2020.

CUBUK, E. B. S.; KARKIN, N.; YAVUZ, N. Public sector innovativeness and public values through information and communication technologies. In: **Proceedings of the 20th Annual International Conference on Digital Government Research**. New York, NY, USA: Association for Computing Machinery, 2019. (dg.o 2019), p. 353–361. ISBN 9781450372046. Disponível em: <https://doi.org/10.1145/3325112.3325215>.

DENHARDT, R. B.; DENHARDT, J. V. The new public service: Serving rather than steering. **Public administration review**, Wiley Online Library, v. 60, n. 6, p. 549–559, 2000.

DUNLEAVY, P.; MARGETTS, H.; BASTOW, S.; TINKLER, J. New public management is dead—long live digital-era governance. **Journal of public administration research and theory**, Oxford University Press, v. 16, n. 3, p. 467–494, 2006.

FUNK, A.; LI, Y.; SAGGION, H.; BONTCHEVA, K.; LEIBOLD, C. Opinion analysis for business intelligence applications. In: **Proceedings of the First International Workshop on Ontology-Supported Business Intelligence**. New York, NY, USA: Association for Computing Machinery, 2008. (OBI '08). ISBN 9781605582191. Disponível em: <https://doi.org/10.1145/1452567.1452570>.

GRAVES, A.; JAITLY, N. Towards end-to-end speech recognition with recurrent neural networks. In: PMLR. **International conference on machine learning**. [S.l.], 2014. p. 1764–1772.

HAYKIN, S. **Neural networks: a comprehensive foundation**. [S.l.]: Prentice Hall PTR, 1998.

OSBORNE, S. P. **The new public governance? 1**. [S.l.]: Taylor & Francis, 2006.

SAGE, A. P. E. **Concise Encyclopedia of Information Processing in Systems and Organizations**. New York: Pergamon, 1990.

SILVA, P.; FIGUEIREDO, K. Aprendizado profundo aplicado na previsão de receita tributária utilizando variáveis endógenas. In: **Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional**. Porto Alegre, RS, Brasil: SBC, 2020. p. 414–425. ISSN 2763-9061. Disponível em: <https://sol.sbc.org.br/index.php/eniac/article/view/12147>.

SUSAR, D.; AQUARO, V. Artificial intelligence: Opportunities and challenges for the public sector. In: **Proceedings of the 12th International Conference on Theory and Practice of Electronic Governance**. New York, NY, USA: Association for Computing Machinery, 2019. (ICEGOV2019), p. 418–426. ISBN 9781450366441. Disponível em: <https://doi.org/10.1145/3326365.3326420>.

VALLE-CRUZ, D.; RUVALCABA-GOMEZ, E. A.; SANDOVAL-ALMAZAN, R.; CRIADO, J. I. A review of artificial intelligence in government and its potential from a public policy perspective. In: **Proceedings of the 20th Annual International Conference on Digital Government Research**. New York, NY, USA: Association for Computing Machinery, 2019. (dg.o 2019), p. 91–99. ISBN 9781450372046. Disponível em: <https://doi.org/10.1145/3325112.3325242>.

WAN, H.; GUO, S.; YIN, K.; LIANG, X.; LIN, Y. Cts-lstm: Lstm-based neural networks for correlatedtime series prediction. **Knowledge-Based Systems**, Elsevier, v. 191, p. 105239, 2020.

YAMAK, P. T.; YUJIAN, L.; GADOSEY, P. K. A comparison between arima, lstm, and gru for time series forecasting. In: **Proceedings of the 2019 2nd international conference on algorithms, computing and artificial intelligence**. [S.l.: s.n.], 2019. p. 49–55.

YURTSEVER, M. Gold price forecasting using lstm, bi-lstm and gru. **Avrupa Bilim ve Teknoloji Dergisi**, Osman SAĞDIÇ, n. 31, p. 341–347, 2021.

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DA PARAÍBA
Campus Cajazeiras - Código INEP: 25008978
Rua José Antônio da Silva, 300, Jardim Oásis, CEP 58.900-000, Cajazeiras (PB)
CNPJ: 10.783.898/0005-07 - Telefone: (83) 3532-4100

Documento Digitalizado Ostensivo (Público)

TCC - Trabalho de conclusão de curso

Assunto:	TCC - Trabalho de conclusão de curso
Assinado por:	Allyson Abreu
Tipo do Documento:	Projeto
Situação:	Finalizado
Nível de Acesso:	Ostensivo (Público)
Tipo do Conferência:	Cópia Simples

Documento assinado eletronicamente por:

- **Allyson Oliveira de Abreu**, **ALUNO (201922010007) DE TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS - CAJAZEIRAS**, em 12/03/2024 08:50:36.

Este documento foi armazenado no SUAP em 12/03/2024. Para comprovar sua integridade, faça a leitura do QRCode ao lado ou acesse https://suap.ifpb.edu.br/verificar-documento-externo/ e forneça os dados abaixo:

Código Verificador: 1111854
Código de Autenticação: e3d5eb2e36