UNIVERSIDADE FEDERAL DE GOIÁS
INSITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

DOUGLAS VIEIRA DO NASCIMENTO

Previsão de Nascidos Vivos nas Regiões de Saúde do Brasil Através de Modelos de Aprendizado de Máquina Baseados em Árvore

Goiânia
2024

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO (TECA) PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TESES

E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo das Teses e Dissertações disponibilizado na BDTD/UFG é de responsabilidade exclusiva do autor. Ao encaminhar o produto final, o autor(a) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do material bibliográfico

[X] Dissertação [ ] Tese [ ] Outro*:_____

*No caso de mestrado/doutorado profissional, indique o formato do Trabalho de Conclusão de Curso, permitido no documento de área, correspondente ao programa de pós-graduação, orientado pela legislação vigente da CAPES.

Exemplos: Estudo de caso ou Revisão sistemática ou outros formatos.

2. Nome completo do autor

Douglas Vieira do Nascimento

3. Título do trabalho

Previsão de Nascidos Vivos nas Regiões de Saúde do Brasil Através de Modelos de Aprendizado de Máquina Baseados em Árvore

4. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador)

Concorda com a liberação total do documento [X] SIM [ ] NÃO[1]

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante:
a) consulta ao(à) autor(a) e ao(à) orientador(a);
b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo da tese ou dissertação.
O documento não será disponibilizado durante o período de embargo.
Casos de embargo:
- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro;
- Publicação da dissertação/tese em livro.

Obs. Este termo deverá ser assinado no SEI pelo orientador e pelo autor.

Documento assinado eletronicamente por **Arlindo Rodrigues Galvao Filho**, **Professor do Magistério Superior**, em 13/12/2024, às 19:22, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Douglas Vieira Do Nascimento**, **Discente**, em 16/12/2024, às 21:39, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **5039596** e o código CRC **093B2DEE**.

DOUGLAS VIEIRA DO NASCIMENTO

Previsão de Nascidos Vivos nas Regiões de Saúde do Brasil Através de Modelos de Aprendizado de Máquina Baseados em Árvore

Dissertação apresentada ao Programa de Pós–Graduação em Ciência da Computação (PPGCC), do Insituto de Informática da Universidade Federal de Goiás (UFG), como requisito para obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Ciência da Computação.

Linha de Pesquisa: Sistemas Inteligentes e Aplicações

Orientador: Prof. Dr. Arlindo Rodrigues Galvão Filho

Goiânia
2024

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Nascimento, Douglas Vieira Do
Previsão de Nascidos Vivos nas Regiões de Saúde do Brasil Através de Modelos de Aprendizado de Máquina Baseados em Árvore [manuscrito] / Douglas Vieira Do Nascimento. - .
f.

Orientador: Prof. Arlindo Rodrigues Galvão Filho.
Dissertação (Mestrado) - Universidade Federal de Goiás, Instituto de Informática (INF), , Goiânia, .
Bibliografia.
Inclui lista de figuras, lista de tabelas.

1. Previsão de Séries Temporais. 2. Previsão de Longo Prazo. 3. rvores de Decisão. 4. Taxa de Mortalidade Materna. I. Rodrigues Galvão Filho, Arlindo , orient. II. Título.

CDU 004

UNIVERSIDADE FEDERAL DE GOIÁS

INSTITUTO DE INFORMÁTICA

ATA DE DEFESA DE DISSERTAÇÃO

Ata nº **37** da sessão de Defesa de Dissertação de **Douglas Vieira do Nascimento**, que confere o título de Mestre em Ciência da Computação, na área de concentração em Ciência da Computação.

Aos treze dias do mês de novembro de dois mil e vinte e quatro, a partir das dez horas e trinta minutos, via webconferência, realizou-se a sessão pública de Defesa de Dissertação intitulada **"Previsão de Nascidos Vivos nas Regiões de Saúde do Brasil Através de Modelos de Aprendizado de Máquina Baseados em Árvore"**. Os trabalhos foram instalados pelo Orientador, Professor Doutor Arlindo Rodrigues Galvão Filho (INF/UFG) com a participação dos demais membros da Banca Examinadora: Professor Doutor Anderson da Silva Soares (INF/UFG), membro titular interno; Professor Doutor Rafael Teixeira Sousa (ICET/UFMT), membro titular externo. A realização da banca ocorreu por meio de videoconferência. Durante a arguição os membros da banca não fizeram sugestão de alteração do título do trabalho. A Banca Examinadora reuniu-se em sessão secreta a fim de concluir o julgamento da Dissertação/Tese, tendo sido o candidato **aprovado** pelos seus membros. Proclamados os resultados pelo Professor Doutor Arlindo Rodrigues Galvão Filho, Presidente da Banca Examinadora, foram encerrados os trabalhos e, para constar, lavrou-se a presente ata que é assinada pelos Membros da Banca Examinadora, aos treze dias do mês de novembro de dois mil e vinte e quatro.

TÍTULO SUGERIDO PELA BANCA

Documento assinado eletronicamente por **Anderson Da Silva Soares**, **Professor do Magistério Superior**, em 13/11/2024, às 11:29, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Arlindo Rodrigues Galvao Filho**, **Professor do Magistério Superior**, em 13/11/2024, às 11:29, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Douglas Vieira Do Nascimento**, **Discente**, em 13/11/2024, às 11:45, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Rafael Teixeira Sousa**, **Usuário Externo**, em 13/11/2024, às 13:31, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **4944500** e o código CRC **5C5FE801**.

Referência: Processo nº 23070.056580/2024-47 SEI nº 4944500

Resumo

Nascimento, Douglas Vieira Do. **Previsão de Nascidos Vivos nas Regiões de Saúde do Brasil Através de Modelos de Aprendizado de Máquina Baseados em Árvore**. Goiânia, 2024. 38p. Dissertação de Mestrado. Programa de Pós-Graduação em Ciência da Computação , Insituto de Informática, Universidade Federal de Goiás.

Modelos baseados em árvores para previsão são um tipo de técnica de modelagem preditiva que utiliza árvores de decisão para fazer previsões sobre valores ou eventos futuros. Esses modelos são boas escolhas devido à sua capacidade de modelar relações não lineares, por isso foram aplicados à previsão de nascimentos vivos com múltiplas covariáveis. O estudo utiliza dados do Ministério da Saúde do Brasil para treinar e avaliar modelos de previsão, seguindo as diretrizes das expectativas e necessidades do Ministério para o planejamento de políticas públicas. O estudo utiliza dados de todas as 450 microrregiões do Brasil com registros entre os anos de 2000 e 2020. O objetivo é treinar um modelo baseado em árvore com todos os meses entre 2000 e 2018 para avaliar o desempenho da previsão do número de nascimentos ao longo dos anos de 2019 e 2020. LightGBM, XGBoost e Catboost foram avaliados e comparados com AutoARIMA e regressão linear simples. O LightGBM teve um desempenho ligeiramente melhor do que outros modelos avaliados, alcançando um MAPE de 0.0797, com um desempenho mais consistente ao longo dos 24 meses do horizonte de previsão. Os resultados mostram que os modelos baseados em árvores são confiáveis para lidar com múltiplas covariáveis e podem ser uma ferramenta útil para o planejamento de políticas públicas.

Palavras–chave

<Previsão de Séries Temporais, Previsão de Longo Prazo, Árvores de Decisão, Taxa de Mortalidade Materna. >

Abstract

Nascimento, Douglas Vieira Do. **<Live Birth Forecasting in Brazillian Health Regions with Tree-based Machine Learning Models>**. Goiânia, 2024. 38p. MSc. Dissertation. Programa de Pós-Graduação em Ciência da Computação , Insituto de Informática, Universidade Federal de Goiás.

Forecasting tree-based models are a type of predictive modeling technique that uses decision trees to make predictions about future values or events. These models are good choices due to their ability to model non-linear relationships, which is why they were applied to predicting live births with multiple covariates. The study uses data from the Brazilian Ministry of Health to train and evaluate forecasting models, following the guidelines of the Ministry's expectations and needs for public policy planning. The study uses data from all 450 microregions in Brazil with records between the years 2000 and 2020. The objective is to train a tree-based model with all months between 2000 and 2018 to evaluate the performance of predicting the number of births over of the years 2019 and 2020. LightGBM, XGBoost and Catboost were evaluated and compared with AutoARIMA and simple linear regression. LightGBM performed slightly better than other evaluated models, achieving a MAPE of 0.0797, with a more consistent performance over the 24-month forecast horizon. The results show that tree-based models are reliable for handling multiple covariates and can be a useful tool for public policy planning.

Keywords

<Time Series Forecasting, Long-Horizon Forecasting, Decision Trees, Maternal Mortality Ratio.>

Sumário

Lista de Figuras

Lista de Tabelas

CAPÍTULO 1

Introdução

A mortalidade materna continua sendo um dos grandes desafios para a saúde global, particularmente em países em desenvolvimento. Lideranças mundiais têm buscado melhorar a qualidade de vida das mulheres em fase maternal, conforme destacado no relatório sobre mortalidade materna [Alkema et al. 2016]. Em 2015, a ONU lançou os Objetivos de Desenvolvimento Sustentável (ODS), um conjunto de 17 metas destinadas a enfrentar os principais desafios globais. Entre esses objetivos, a ODS 3 visa atuar com saúde e bem estar da população, incluindo reduzir a taxa de mortalidade materna (TMM) global para menos de 70 mortes por 100.000 nascidos vivos até 2030 [United Nations Development Programme 2015]. Antes das ODS, os Objetivos de Desenvolvimento do Milênio (ODM) já buscavam reduzir a mortalidade materna em 75% entre 1990 e 2015. Embora tenha havido progresso, essa meta não foi plenamente alcançada, resultando na criação de novas metas sob os ODS [Saúde 2024]

A mortalidade materna reflete diretamente como as mulheres são atendidas pelos sistemas de saúde em momentos críticos. Entender esses índices é fundamental para identificar falhas no atendimento e debater soluções e melhorias. Muitas das principais causas de mortalidade, como hemorragias, infecções e pré-eclâmpsia, podem ser prevenidas ou tratadas com atendimento médico adequado. [Fiocruz 2024] A questão da mortalidade materna tem sido uma preocupação central nas políticas de saúde global nas últimas décadas, integrando-se à luta pelos direitos reprodutivos e igualdade de gênero. Apesar de progressos alcançados desde a implementação dos Objetivos de Desenvolvimento do Milênio (ODM), crises recentes, como a pandemia de COVID-19 e conflitos regionais, exacerbaram a vulnerabilidade das mulheres, destacando a urgência de se alcançar as metas propostas pelos Objetivos de Desenvolvimento Sustentável (ODS). A mortalidade materna não afeta apenas as mulheres e suas famílias, mas também representa uma perda significativa para as economias locais e nacionais, contribuindo para ciclos de pobreza e desigualdade. [Pinto et al. 2022] A ausência de cuidados maternos adequados tem impacto direto no desenvolvimento socioeconômico, uma vez que as mulheres são frequentemente responsáveis pela criação de novas gerações e pelo bem-estar das famílias. As desigualdades no acesso a cuidados de saúde materna são um desafio significativo. Em

muitas regiões de baixa e média renda, as disparidades no acesso a serviços médicos de qualidade, especialmente em áreas rurais ou em zonas de conflito, agravam as taxas de mortalidade materna. Mesmo em países com sistemas de saúde desenvolvidos, as mulheres mais pobres, pertencentes a minorias étnicas ou vivendo em áreas remotas continuam a enfrentar riscos elevados. Mulheres em todos os estágios da gravidez necessitam de acompanhamento contínuo para garantir sua própria saúde e a de seus recém-nascidos, uma vez que esses cuidados estão interligados [Ribeiro et al. 2008]

Cumprir a meta do ODS 3 é desafiador em regiões afetadas por conflitos e crises humanitárias, onde os sistemas de saúde encontram-se fragilizados e os recursos, insuficientes. Nessa perspectiva, surgem dois grandes desafios: o primeiro é prever a taxa de nascidos vivos, que pode auxiliar no processo de formulação de políticas públicas, fornecendo aos tomadores de decisão dados cruciais sobre tendências populacionais futuras. Essas informações são essenciais para embasar decisões relacionadas ao planejamento de infraestrutura, alocação de recursos na saúde e educação, além de outras áreas críticas. O segundo desafio envolve a rotatividade de médicos (*churn*) no sistema de saúde, fenômeno que impacta diretamente a continuidade do cuidado, gerando aumento do tempo de espera para consultas e procedimentos, sobrecarga dos profissionais remanescentes e queda na qualidade dos serviços oferecidos [Misra-Hebert, Stoller e Kay 2024]. Embora seja fundamental estabelecer metas claras para a redução da mortalidade materna, a medição precisa dessas mortes continua complexa, com muitas ocorrências não sendo devidamente registradas. Nesse cenário, estratégias que assegurem uma infraestrutura robusta, recursos adequados e a presença de profissionais de saúde qualificados tornam-se vitais para o acompanhamento efetivo das mulheres antes, durante e após a gravidez [Pacagnella et al. 2018].

A mortalidade materna permanece alta devido à falta de infraestrutura, força de trabalho e recursos adequados. Apesar de uma redução de 9% nas taxas de mortalidade nos últimos anos, 295.000 mortes maternas ainda foram registradas em 2017, o que representa cerca de 800 mortes diárias [Organization 2019]. No Brasil, o Ministério da Saúde reportou um aumento na taxa de mortalidade materna entre 2018 e 2020, atingindo 74,4 óbitos por 100.000 nascidos vivos, o dobro da meta estabelecida nos ODS [Health 2020].

Segundo Elena Maria da Silva Duarte [Duarte et al. 2020], a mortalidade materna é considerada uma das mais graves violações dos direitos humanos, sendo evitável em cerca de 92% dos casos. Apesar disso, essa tragédia continua a afetar, de forma desproporcional, os países em desenvolvimento. No Brasil, entre 1996 e 2016, foram registradas 35.546 mortes maternas. A Região Sudeste apresentou o maior número de ocorrências, com 12.686 casos, seguida pela Região Nordeste, com 11.777 mortes. Os dados utilizados neste estudo são fornecidos pelo Ministério da Saúde de Goiás, em parceria com o

Sistema Único de Saúde (SUS), e têm como objetivo compreender as causas, as necessidades e as consequências da mortalidade materna no estado de Goiás, além de identificar estratégias para melhorar o sistema de saúde e reduzir essas ocorrências no futuro. Com semelhante motivação, Abimbola [Abimbola et al. 2015] analisam o cenário nas unidades de atenção básica, quanto a rotatividade de seus profissionais, com objetivo de identificar motivos da evasão e também as consequências que causadas para a sociedade.

Com objetivo de aplicar ações eficazes, a implementação de ferramentas preditivas é essencial, auxiliando na alocação de recursos e no planejamento de intervenções preventivas. A predição envolve a estimativa de eventos futuros com base em dados históricos e padrões identificados. Neste contexto, a previsão de séries temporais utilizando técnicas como árvores de decisão pode desempenhar um papel crucial. Essa metodologia permite analisar dados históricos sobre mortalidade materna, condições socioeconômicas e capacidade dos sistemas de saúde, gerando projeções que ajudam a direcionar recursos e intervenções para áreas e períodos de maior necessidade [Chauhan et al. 2020]. Modelos de aprendizado de máquina baseados em árvores, como árvores de decisão e florestas aleatórias, têm se destacado como opções populares para previsão de séries temporais, principalmente devido à sua capacidade de modelar relações não lineares e lidar com dados ausentes ou incompletos. Esses modelos têm a vantagem de incorporar múltiplas variáveis de entrada, o que os torna particularmente adequados para a previsão de séries temporais com covariáveis [Masini, Medeiros e Mendes 2021]. Ao contrário das técnicas clássicas, os modelos baseados em árvores conseguem explorar relacionamentos complexos entre múltiplos fatores. Ao analisar dados históricos sobre essas tendências, esses modelos podem identificar padrões e capturar interações entre diferentes variáveis que influenciam o crescimento populacional de forma mais precisa, resultando em previsões mais robustas e informadas [Reis 2024]. A metodologia aplicada envolve a comparação dos modelos XGBoost, modelo de aumento de gradiente leve (Light Gradient Boosting Model, LightGBM) para previsão de nascidos vivos e, também, em conjunto o uso de uma Multilayer Perceptron (MLP) para previsão da rotatividade de médicos (*churn*) no sistema de saúde.

As principais contribuições deste trabalho são: (a) avaliar o desempenho e o comportamento de modelos baseados em árvore na previsão do número de nascidos vivos, utilizando múltiplas covariáveis para capturar relações complexas entre fatores demográficos e sociais; e (b) realizar uma análise da rotatividade de médicos(churn) no sistema de saúde, comparando o desempenho de modelos baseados em árvore com o de uma Multilayer Perceptron (MLP), de forma a identificar padrões e prever a saída de profissionais do sistema.

CAPÍTULO 2

Trabalhos Relacionados

Diversos estudos têm explorado a previsão de séries temporais para indicadores de saúde, como taxas de nascidos vivos, óbitos mensais e mortalidade, utilizando tanto abordagens estatísticas tradicionais quanto métodos de aprendizado profundo. Métodos como ARIMA e Suavização Exponencial são amplamente utilizados devido à sua simplicidade e capacidade de capturar padrões sazonais. No entanto, tais abordagens são limitadas pela dedução de linearidade e pela dificuldade em incorporar múltiplas covariáveis, o que pode comprometer a precisão em cenários com maior complexidade.

Por exemplo, Bravo e Coelho (2019) [Bravo e Coelho 2020] aplicaram métodos de previsão sazonais, como o ARIMA Sazonal e o Holt-Winters, para prever nascimentos e mortes mensais em Portugal. Embora esses métodos tenham mostrado eficácia ao lidar com a sazonalidade, eles demonstraram limitações em capturar padrões não lineares e relações interdependentes entre variáveis, o que é crucial para prever com precisão indicadores de saúde em ambientes voláteis, como a evasão de médicos no sistema de saúde.

Por outro lado, Thabang et al. (2020) [Mathonsi e Zyl 2022] introduziram uma abordagem híbrida, combinando métodos estatísticos e de aprendizado profundo, como o Exponential Smoothing Recurrent Neural Network (ES-RNN), que demonstrou desempenho superior aos modelos puramente estatísticos ao lidar com séries temporais multivariadas. Esses modelos são mais eficazes ao lidar com dados ausentes e ao modelar relações não lineares complexas, que frequentemente ocorrem em indicadores de saúde.

Adicionalmente, Rady et al. (2021) [Rady, Fawzy e Fattah 2021] realizaram um estudo aplicando métodos baseados em árvores para prever séries temporais utilizando dados multivariados, comparando o desempenho de Árvores de Decisão, Florestas Aleatórias e Gradient Boosted Trees. O estudo demonstrou que a Floresta Aleatória foi a técnica mais precisa, superando tanto os modelos baseados em árvores quanto o ARIMA em termos de erro quadrático médio. Esses resultados indicam que métodos de aprendizado de máquina baseados em árvores são altamente promissores para previsão de séries temporais em cenários complexos, como os encontrados na saúde pública, onde múltiplas variáveis precisam ser consideradas simultaneamente.

Além disso, o estudo de Predicting COVID-19 mortality risk in Toronto, Canada [Feng, Kephart e Juarez-Colunga 2022] compara o desempenho de métodos baseados em árvores, como Gradient Boosted Trees (GBT) e Random Forest (RF), com métodos de regressão, como Regressão Logística e Regressão Linear, na previsão do risco de mortalidade por COVID-19. Os resultados mostraram que os modelos baseados em árvores superaram os modelos de regressão, capturando melhor as interações entre variáveis complexas, como comorbidades e fatores demográficos, reforçando a superioridade dessas abordagens em cenários de alta complexidade, como a saúde pública.

Scheffle [Scheffler et al. 2008] estimam a procura futura pela necessidade e oferta de médicos por região da OMS para determinar onde ocorrer a escassez para o ano de 2015. Logo, mostraram que, para atender à necessidade global de 80% de cobertura de nascimentos vivos por um profissional qualificado, será necessário aumentar a oferta de médicos em 65% na região africana da OMS.

Embora modelos sazonais como o ARIMA e abordagens híbridas como o ES-RNN tenham alcançado sucesso na previsão de nascimentos e mortes mensais, este trabalho visa expandir essas abordagens ao aplicar modelos de aprendizado de máquina baseados em árvores. Esses modelos são capazes de lidar com múltiplas variáveis de entrada, capturar relações não lineares e proporcionar previsões mais robustas e precisas para indicadores de saúde, como a taxa de nascidos vivos e a evasão de médicos no sistema de saúde.

CAPÍTULO 3

Metodologia

Para melhorar os indicadores de mortalidade materna e aprimorar os serviços de saúde, este estudo utiliza uma abordagem baseada na previsão do número de nascimentos e evasão de médicos no sistema de saúde. A metodologia é aplicada a dados históricos de saúde materna e neonatal, com o objetivo de otimizar a alocação de recursos e melhorar a gestão dos serviços de saúde para mulheres grávidas e recém-nascidos. Este estudo aplica modelos avançados de aprendizado de máquina baseados em árvores para abordar dois problemas distintos no contexto da saúde: (a) prever a taxa de nascidos vivos e (b) analisar a rotatividade de médicos no sistema de saúde. No primeiro caso, os modelos *XGBoost*, *LightGBM* e *CatBoost* serão comparados com abordagens estatísticas, como *AutoARIMA* e Regressão Linear. Já no segundo caso, os modelos XGBoost e LightGBM serão comparados ao desempenho de uma *MLP (Multilayer Perceptron)* para avaliar sua eficácia na previsão da rotatividade de médicos.

3.1 Modelos de aprendizado baseados em árvores

Os modelos aplicados foram escolhidos devido à sua capacidade de lidar com grandes volumes de dados, modelar relações não lineares e capturar interações complexas entre múltiplas variáveis, tornando-os adequados para a previsão de séries temporais e a análise de indicadores de saúde. O *XGBoost* é reconhecido por sua eficiência computacional e por ser capaz de regularizar modelos para evitar o *overfitting*. O *LightGBM* destaca-se por sua alta velocidade e capacidade de lidar com grandes quantidades de dados e muitas categorias, utilizando uma técnica de crescimento de árvore que otimiza a eficiência do treinamento. O *CatBoost* é particularmente eficaz ao lidar com variáveis categóricas e com dados desbalanceados, sendo útil em contextos onde os dados podem conter ruído ou informações incompletas.

3.1.1 *XGBoost*

O *XGBoost* [Chen e Guestrin 2016] é um sistema avançado de aprendizado de máquina que utiliza a técnica de *tree boosting*, onde múltiplas árvores de decisão são treinadas de forma sequencial. Cada nova árvore busca corrigir os erros das árvores anteriores, uma abordagem conhecida como boosting, que se mostrou extremamente eficaz em melhorar a precisão dos modelos preditivos. O XGBoost é projetado para ser altamente eficiente e escalável, tornando-se a escolha preferida em uma variedade de aplicações, incluindo classificação, previsão e detecção de anomalias.

Entre suas principais características, destaca-se a capacidade de lidar com dados esparsos, ou seja, conjuntos de dados que contêm muitos valores ausentes ou zeros. O algoritmo otimiza o processamento desses dados ao atribuir valores padrão para as lacunas, aumentando assim sua eficácia. Além disso, o *XGBoost* incorpora regularização L1 (Lasso) e L2 (Ridge), que ajudam a prevenir o *overfitting*. A regularização L1 promove soluções esparsas, enquanto a L2 reduz o tamanho dos coeficientes, tornando o modelo mais robusto e menos suscetível a flutuações.

Outra inovação significativa do *XGBoost* é seu enfoque em padrões de acesso à memória e uso eficiente de *cache*, o que resulta em um aumento significativo na velocidade de execução. O algoritmo também é otimizado para paralelismo, permitindo que seja executado em múltiplas CPUs e GPUs, o que acelera o treinamento em grandes conjuntos de dados. Essas características fazem do *XGBoost* uma ferramenta poderosa e versátil, amplamente utilizada em competições de aprendizado de máquina e em aplicações do mundo real.

3.1.2 LightGBM

A LightGBM (Light Gradient Boosting Machine) [Ke et al. 2017] é uma implementação eficiente do algoritmo *Gradient Boosting Decision Tree (GBDT)*, desenvolvida pela *Microsoft*. Destaca-se por seu alto desempenho em termos de velocidade e uso de memória, tornando-se uma escolha ideal para lidar com grandes conjuntos de dados e alta dimensionalidade em modelos de previsão em cenários complexos. Um dos principais desafios atuais na modelagem é a necessidade de reduzir o número de instâncias de dados e a quantidade de características sem comprometer a precisão do modelo.

Para abordar esse desafio, o *LightGBM* introduz duas inovações significativas que aumentam sua eficiência e velocidade em comparação com outros algoritmos de *boosting*:

- *Gradient-based One-Side Sampling (GOSS)*: Esta técnica seleciona uma amostra dos dados para treinar o modelo, priorizando instâncias com grandes gradientes. Isso reduz o tempo de treinamento sem sacrificar a precisão das previsões.

- *Exclusive Feature Bundling (EFB)*: Essa abordagem agrupa características mutuamente exclusivas para diminuir a quantidade de variáveis a serem processadas, o que melhora a eficiência do modelo.

Essas otimizações permitem que o *LightGBM* seja significativamente mais rápido do que implementações tradicionais do GBDT, mantendo quase a mesma precisão em tarefas de classificação e regressão. Além disso, ele suporta processamento paralelo e implementações em GPU, tornando-o ideal para grandes conjuntos de dados, como os utilizados na previsão de indicadores de saúde.

3.1.3 CatBoost

O CatBoost [Prokhorenkova et al. 2018] é um algoritmo de aprendizado de máquina que pertence à família de algoritmos de boosting. Seu diferencial reside na capacidade de lidar com dados do mundo real de maneira robusta e eficaz, especialmente com variáveis categóricas. O *CatBoost* oferece um desempenho superior para esse tipo de variável, utilizando uma técnica chamada codificação de impacto, que captura a interação entre os valores das variáveis categóricas e o alvo, fornecendo informações relevantes para o modelo.

Diferentemente de outras implementações de boosting, o *CatBoost* pode aproveitar informações estatísticas de variáveis categóricas sem a necessidade de pré-processamento manual, como rotulagem ou criação de variáveis fictícias (one-*hot encoding*). Além disso, ele lida naturalmente com dados ausentes, eliminando a necessidade de imputação prévia. Entre suas características, destaca-se um mecanismo de otimização integrado chamado *Ordered Boosting*, que melhora a eficiência computacional e permite o treinamento em grandes conjuntos de dados. O *CatBoost* também inclui recursos de regularização, como restrições no tamanho da árvore e força de rotação, que ajudam a evitar o *overfitting* e a melhorar a generalização do modelo.

A capacidade do *CatBoost* de processar características categóricas diretamente, sem transformações complexas, e sua eficiência em termos de memória e velocidade são outros pontos fortes. O algoritmo incorpora técnicas que permitem processamento paralelo e implementação em GPU, tornando-o ideal para conjuntos de dados extensos. Estudos demonstram que o *CatBoost* apresenta desempenho superior em comparação com outros algoritmos de *boosting*, como *XGBoost* e *LightGBM*, especialmente em tarefas de classificação.

3.2 Caso de Estudo 1: Previsão da Taxa de Nascidos Vivos

3.2.1 Coleta dos dados

A melhoria dos indicadores de taxa de mortalidade materna é essencial para o avanço dos serviços de saúde pública. Além disso, uma gestão eficiente desempenha um papel crucial nesse processo. A capacidade de prever o número de nascimentos em um determinado período possibilita uma alocação mais eficaz de recursos, permitindo uma preparação antecipada para os cuidados pré-natais e pós-natais. Essa abordagem proativa não só aprimora a qualidade do atendimento médico para as gestantes e os recém-nascidos, como também reduz a pressão sobre os serviços de saúde, melhorando a experiência dos pacientes e otimizando o uso de recursos. A previsão precisa do número de nascimentos pode, assim, ser um componente estratégico fundamental para garantir o bem-estar materno e neonatal, promovendo intervenções preventivas e a melhoria contínua dos serviços de saúde.

Este estudo usa múltiplas variáveis do Sistema de Informação de Nascidos Vivos (SINASC) fornecido pelo Ministério da Saúde. O sistema público de saúde divide geograficamente o Brasil em macrorregiões e microrregiões. Atualmente são 450 microrregiões e 116 macrorregiões, consideradas para a definição de políticas públicas. Os dados extraídos consistem de 17 variáveis de todas as 450 microrregiões, a tabela 3.1 mostra uma breve explicação de cada uma das variáveis. Essas variáveis foram coletadas entre os anos 2000 e 2020, totalizando 252 meses.

Algumas variáveis estão disponíveis à nível diário, como a quantidade de nascidos vivos, outros a nível mensal, como a mortalidade neonatal, e alguns em nível anual como a estimativa populacional. O Ministério da Saúde brasileiro solicitou previsão mensal, os dados foram agregados no nível mensal, então as variáveis diárias foram somadas como valores mensais, e as variáveis anuais foram repetidas em todos os doze meses.

A variável alvo é o número de nascidos vivos, a data mostra de onde vem cada observação. O objetivo é formar um modelo capaz de prever um horizonte de previsão de 24 meses, está estritamente relacionado ao fato de que o Ministério de Saúde geralmente leva um ano inteiro para coletar, organizar e validar todos os dados do ano anterior. Para ser útil a previsão precisa usar dados de um ano atrás, prever a validação de dados período (primeiro ano de previsão), e prever mais um ano à frente (segundo ano da previsão), então as políticas para o próximo ano podem ser planejado usando a previsão com antecedência, favorencendo a otimização e alocação de recursos.

Para utilizar os dados de maneira eficaz e prever 24 meses a frente do que temos é necessário utilizar técnicas de *forescasting*. Essas constituem em obter previsões

Tabela 3.1: Descrições das covariáveis usadas com seu intervalo e distribuição. A distribuição das variáveis contínuas é descrita com uma média (desvio padrão) e as variáveis categóricas com frequência de ocorrência.

Variável	Descrição	Intervalo	Distribuição
Número de nascidos vivos	Variável alvo que quantifica a quantidade de nascidos vivos.	[0 – 18.997]	546,86(1.065,90)
Código da região de saúde	Variável categórica que distinguem todas as 450 regiões de saúde	-	-
Data	Variável que representa a data.	-	-
Mortalidade neonatal	Variável real sobre a taxa de bebês que não sobreviveram.	[0.0048 – 0.1076]	0,0177(0,0052)
Mortes	Número de bebês que não sobreviveram.	[0 – 14.385]	216,73(490,41)
Média de abortos anteriores	Média de abortos anteriores à gravidez atual.	[0 – 99]	1,15(5,28)
Mulheres em idade fértil	Número de mulheres em idade reprodutiva (Inteiro).	[939 – 3.371,491]	19.991,25(162.170,61)
Cuidado pré-natal em sete consultas	Proporção de gestantes que realizaram mais de sete consultas de pré-natal.	[0 – 1,0,32]	0,88(13,87)
Estimativa da população	Estimativa populacional anual por região (Inteiro).	[18.644 – 12.325,232]	43.5623,26(861.949,16)
Seguro de saúde médica	Percentual de mulheres que possuem planos de saúde privados.	[51 – 5,897.685,00]	24.560,05(281.806,62)
Seguro Saúde Odontológica	Percentual de mulheres cobertas por seguro odontológico privado.	[34 – 3,853,367]	15,577.82(184,454.86)
Distribuição de preservativos masculinos	Distribuição de preservativos masculinos.	[0;1]	50.0%
Distribuição de preservativos femininos	Distribuição de preservativos femininos.	[0;1]	99,8%
Pílula anticoncepcional	Distribuição de pílula anticoncepcional.	[0;1]	100,0%
Pílula do dia seguinte	Distribuição de pílula do dia seguinte.	[0;1]	97,5%
Anticoncepcional injetável	Distribuição de injeção anticoncepcional por região	[0;1]	99,8%
Distribuição do diafragma	Distribuição do diafragma anticoncepcional por região	[0;1]	34,0%
Dispositivo ultra-interino	Distribuição do aparelho ultraprovisório por região	[0;1]	49,5%
Outras distribuições	Distribuição de outros métodos contraceptivos	[0;1]	97,5%

sobre eventos futuros através de um processo sistemático, baseado no conhecimento e compreensão de eventos anteriores [Lewis, McGrath e Seidel 2009] . A predição é realizada através de todas as variáveis e, quando possível, também as covariáveis.

3.2.2 Pré Processamento dos dados

Os modelos baseados em árvores não são capazes de extrair automaticamente recursos implícitos dos dados, e são mais eficazes quando trabalham com dados tabulares estruturados. Para adaptar esses modelos ao domínio de séries temporais, foi necessário enriquecer os dados originais com novos recursos. Além das 17 características originais do conjunto de dados, novos recursos derivados foram gerados por meio de uma janela de estatísticas rolantes aplicadas à variável alvo. Esses recursos incluem a extração de média, mediana, variância, desvio padrão, máximo e mínimo, calculados em janelas rolantes de 2, 3, 6, 12 e 24 meses. Isso permitiu que os modelos aprendessem não apenas as magnitudes dos dados, mas também seu comportamento ao longo do tempo. Essa extração de recursos foi realizada utilizando o pacote *Python tsfresh* [Christ et al. 2018] que é amplamente utilizado para a extração automatizada de características de séries temporais.Para adicionar uma capacidade auto-regressiva ao modelo, foram incluídos atrasos (*lags*) dos últimos 24 meses, permitindo que o modelo capturasse dependências temporais ao prever valores futuros. Além disso, as datas foram convertidas em uma

representação cíclica senoidal, utilizando as funções seno e cosseno, criando assim duas novas variáveis que preservam a sazonalidade e padrões temporais de longo prazo.

3.2.3 Modelos comparados as árvores

AutoARIMA é um algoritmo de aprendizado de máquina que pode realizar previsões em séries temporais univariadas. Pertence à família de modelos ARIMA (*Auto-Regressive Integrated Moving Average*) e foi projetado para automatizar o processo de escolha de parâmetros para modelos ARIMA, como: ordem de autorregressão (AR), ordem de diferenciação (I) e ordem de média móvel (MA). O AutoARIMA usa uma extensa técnica de pesquisa para encontrar o conjunto ideal de parâmetros ARIMA que melhor se ajusta aos dados de treinamento. Para cada combinação de parâmetros considerada, um modelo é ajustado aos dados e a qualidade do ajuste é avaliada por meio de métricas como o critério AIC (*Akaike Information Criterion*) e o critério BIC (*Bayesian Information Criterion*). O AutoARIMA seleciona o modelo com a menor métrica para dar ao ARIMA o melhor conjunto de parâmetros. Para avaliar o desempenho de modelos baseados em árvore na previsão de fertilidade, três modelos GBDT populares e modernos são avaliados: *LightGBM*, *XGBoost* e *CatBoost*. Para comparação, o AutoARIMA sem covariáveis é aplicado e, para modelos baseados em árvore, a regressão linear simples é tratada como linha de base com o mesmo conjunto de recursos. Todos os modelos baseados em árvore foram implementados usando pacotes *Python* proprietários e disponíveis publicamente, desenvolvidos pelos autores. Algumas tentativas para otimizar os parâmetros usando pesquisa em grade e pesquisa aleatória não foram eficientes na maioria dos testes resultando em superajuste dos dados de validação. Portanto, optou-se por manter os parâmetros padrão para os três modelos, obtendo um desempenho mais equilibrado ao longo do período de previsão de 24 meses. AutoARIMA é realizado através do pacote *pmdarima* [Smith et al. 2017–] e a regressão linear foi realizada usando Pycaret [Ali 2020] para otimização de parâmetros. Logo, 24 modelos são treinados para produzir uma previsão do tipo regressão para cada 24 meses do período de previsão, cada modelo produzindo um mês de previsões, de modo que um modelo prevê o próximo ano em janeiro, outro em fevereiro e assim por diante. O AutoARIMA foi aplicado por microrregião, portanto, 450 ARIMAs diferentes foram ajustados para prever 24 meses cada. Os códigos de identificação da microrregião não são codificados em todos os modelos baseados em árvore. Isso ocorre porque os modelos baseados em árvore têm sua própria maneira de representar dados categóricos. A codificação *one-hot* foi usada para a regressão linear. Todos os modelos foram avaliados nos últimos 24 meses de dados disponíveis (2019 e 2020), um conjunto de validação de 3 anos foi definido para otimizar o modelo (2016, 2017, 2018) e o restante dos dados foi usado para treinamento (2000-2015). O AutoARIMA é uma

exceção, ele é otimizado usando todos os dados de treinamento e validação, portanto, nenhuma validação é necessária.

Ademais, a regressão linear, também aplicada, é um modelo clássico de aprendizado de máquina que cria uma relação linear entre uma variável dependente (alvo) e um conjunto de variáveis independentes (características). O objetivo da regressão linear é encontrar os coeficientes que melhor ajustam uma função linear aos dados, minimizando a soma das diferenças quadradas entre as previsões do modelo e os valores reais [Draper e Smith 1998].

Os modelos de regressão linear são amplamente utilizados devido à sua simplicidade e facilidade de interpretação. Os coeficientes estimados podem ser interpretados como o impacto marginal de cada feição no objetivo. Além disso, a regressão linear é computacionalmente eficiente e fácil de implementar.

3.2.4 Critérios de Avaliação

Para avaliar o desempenho dos modelos aplicados são utilizados as métricas MAPE e MAE. O percentual absoluto médio (MAPE) é uma métrica comumente usada para avaliar a precisão de um modelo de previsão de séries temporais. O objetivo é medir a diferença percentual da média entre os valores previstos e valores reais. A fórmula do MAPE é:

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y_i}}{y_i} \right| \tag{3-1}$$

em que y_i é o valor real, $\hat{y_i}$ é o valor previsto e n é o número de pontos de dados. O valor absoluto é usado na fórmula para garantir que o erro seja sempre um valor positivo. Quanto menor o MAPE, melhor o desempenho do modelo de previsão.

O MAPE é uma boa métrica para usar quando as séries de dados têm um nível relativamente alto de variabilidade, um problema inerente ao conjunto de dados usado neste artigo, pois trata de diferentes regiões do país com populações também diferentes. Possui a desvantagem de ser indefinido quando o valor real é zero [Hyndman et al. 2006], mas esse cenário é extremamente raro em registros de nascidos vivos e até agora há apenas uma ocorrência de uma região sem nascido vivo durante um mês inteiro. O erro médio (MAE) será usado como uma métrica auxiliar.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y_i}| \tag{3-2}$$

em que n é o número de amostras, y_i é o valor real da i-ésima amostra e $\hat{y_i}$ é o valor previsto da i-ésima amostra

3.3 Caso de Estudo 2: Análise da Rotatividade de Médicos no Sistema de Saúde

A compreensão da rotatividade de médicos no Sistema Único de Saúde (SUS) é essencial, dado o impacto negativo que a alta rotatividade pode ter sobre a qualidade dos serviços de saúde. A saída frequente de médicos do sistema gera interrupções na continuidade do atendimento, aumenta o tempo de espera para consultas e procedimentos, sobrecarrega os profissionais remanescentes e, consequentemente, reduz a qualidade do atendimento prestado. A evasão dos médicos para outras áreas ou locais tem se tornado um fenômeno cada vez mais comum e prejudicial ao SUS, comprometendo a eficácia do sistema. O entendimento das razões por trás da baixa retenção desses profissionais é crucial para o desenvolvimento de estratégias de retenção, que busquem mitigar os efeitos da evasão e garantir a estabilidade e o bom funcionamento do sistema de saúde. A retenção médica, portanto, refere-se à capacidade das instituições de evitar o abandono dos profissionais, mantendo-os engajados e atuantes no SUS.

3.3.1 Coleta dos dados

Este estudo utiliza múltiplas variáveis do CIGETS - UFG, fornecido pelo Ministério da Saúde (MS). O sistema público de saúde divide geograficamente o Brasil em regiões. Atualmente 450 regiões de saúde são consideradas para definição de políticas públicas. Para este estudo foram utilizadas 39 regiões na região Centro-Oeste do Brasil. A variável alvo para previsão das informações é a taxa de retenção dos médicos mas para agrupar é utilizado os códigos para cada região e suas datas, usados para agrupá-los e tratá-las por região. São extraídos 18 covariáveis para cada uma das regiões. As covariáveis representam características de observações, que apoiam a análise das previsões das taxas de retenção e contribuem para a identificação de justificativas para o abandono dos cargos pelos profissionais médicos. Os dados usados na predição foram coletados entre 2008 e 2023, totalizando 15 anos. Os anos de 2022 e 2023 foram usados como dados de teste enquanto os demais anos analisados são usados como dados para treino. Algumas informações estão disponíveis em nível bienal, como o Índice de Desenvolvimento da Educação Básica (IDEB), outras em nível mensal, como média de horas trabalhadas, e algumas em nível anual, como estimativa populacional. O MS solicita que os dados sejam padronizados, mantendo todos a nível anual. As covariáveis são somadas (mensalmente) a elas próprias até que contenham as referências anuais. Os dados faltantes são preenchidos com a média dos valores dos anos anteriores. O objetivo é treinar um modelo capaz de prever um horizonte de dois anos.

3.3.2 Pré Processamento dos dados

O pré-processamento do conjunto de dados consiste na normalização das covariáveis para valores entre 1 e -1. As variáveis alvo, taxa de retenção, código da região de saúde e datas não são normalizadas. A otimização dos parâmetros dos modelos é realizada utilizando a busca gulosa em intervalos específicos. Os modelos baseados em árvores são implementados usando as bibliotecas *Python* disponíveis publicamente a MLP é implementado usando a biblioteca *Pytorch Lighting*.

3.3.3 Critérios de Avaliação

Os modelos são avaliados nos últimos 2 anos (2022 e 2023) e os dados restantes são utilizados para treinamento (2008 a 2021). O desempenho dos modelos é avaliado pelo erro quadrático médio (*Mean Square Error*, MSE) e erro médio absoluto (*Mean Absolute Error*, MAE). O MSE é definido pela Equação (3-3)

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3-3}$$

onde n é o número de observações, $\hat{y}_i$ é o valor previsto pelo modelo e y_i é o valor real da i-ésima amostra. Valores próximos a 0 de MSE indica uma boa perfomance do modelo. O MAE é definido pela Equação (3-4).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3-4}$$

onde n é o número de observações, y_i é o valor real da i-ésima amostra, e $\hat{y}_i$ é o valor predito da i-ésima amostra.

3.3.4 Modelo comparado as árvores

Para comparar aos modelos de árvores aplicados (*LightGBM* e *XGBoost*, a Rede Neural Multicamadas (MLP) [Popescu et al. 2009] é utilizada. Este modelo foi escolhido devido à sua capacidade de capturar padrões não lineares complexos, uma vez que o problema de rotatividade envolve múltiplos fatores interdependentes. O MLP é composto por uma arquitetura feed-forward, em que os dados são propagados das camadas de entrada, passando pelas camadas ocultas, até chegar à camada de saída. Para otimizar o desempenho do modelo, foi aplicado o algoritmo de retropropagação do erro, que ajusta os pesos das conexões entre os neurônios com o objetivo de minimizar o erro entre as previsões e os valores reais.

Parâmetros de otimização foram empregados para encontrar a configuração ideal do modelo. Os tamanhos de batch testados foram {16, 32, 64 e 128}, enquanto as *seeds*

aleatórias são utilizadas para garantir a reprodutibilidade dos resultados foram {10, 42, 73, 123 e 3407}. Adicionalmente, o número de neurônios nas camadas ocultas variou entre {32, 64, 128 e 256}, sendo as arquiteturas compostas por 1, 2 ou 3 camadas ocultas. A taxa de aprendizado (*learning rate*) variou de $\{1 \times 10^{-3}$ a $3 \times 10^{-2}\}$, ajustando-se dinamicamente para acelerar o processo de convergência. O Optuna Bayesiana é utilizado como o principal método de otimização para ajustar esses hiperparâmetros. Aplica-se uma abordagem de Tree-structured Parzen Estimator (TPE), explorando o espaço de busca de maneira eficiente e adaptativa, balanceando a exploração de novas combinações de hiperparâmetros e a exploração de regiões promissoras com base em tentativas anteriores. Essa técnica permitiu otimizar os parâmetros da MLP de maneira mais eficiente em termos de tempo e recursos computacionais, resultando em uma melhoria significativa no desempenho preditivo do modelo. O processo de otimização foi conduzido utilizando a biblioteca PyTorch Lightning, que permitiu o gerenciamento eficiente dos experimentos e a execução em escala, facilitando a implementação de técnicas avançadas de treinamento, como aceleração via GPU.

CAPÍTULO 4

Resultados

4.1 Case 1: Previsão da Taxa de Nascidos Vivos

Dessa forma, apresenta-se os resultados obtidos a partir da aplicação dos modelos de aprendizado de máquina baseados em árvores para a previsão da taxa de nascidos vivos. Utilizamos os modelos *XGBoost*, *LightGBM* e *CatBoost*, treinados e avaliados com base nos dados fornecidos pelo Ministério da Saúde e pré-processados. Os resultados são comparados com os obtidos por modelos estatísticos tradicionais, como AutoARIMA e Regressão Linear, para avaliar a eficácia dos modelos de árvores em prever séries temporais. O AutoARIMA é uma opção competitiva com uma pequena diferença da métrica MAE, mas um MAPE maior.

Tabela 4.1: Desempenho médio dos modelos nos conjuntos de validação e teste.

Arquitetura	**Validação**		**Teste**	
	MAE	MAPE	MAE	MAPE
LightGBM	**31.872**	**0.0762**	39.429	**0.0797**
XGBoost	32.559	0.0771	37.066	0.0803
CatBoost	32.921	0.0790	41.180	0.0828
AutoARIMA	-	-	**36.147**	0.0835
Linear Regression	79.329	0.1974	75.405	0.1742

Essa diferença é esperada, pois o horizonte de previsão usado (24 meses) é relativamente longo e, como o *LightGBM* é capaz lidar bem com covariáveis e obtém desempenho mais consistente em todos os conjuntos de teste.

Esse comportamento é observado na Fig. 4.1, em que a distribuição de erros em todas as microrregiões e meses do conjunto de teste tem apenas um pequeno aumento ao longo do tempo, mantendo a maior parte das regiões abaixo de 0,10 de erro. Comparando o erro médio do *LightGBM* e do AutoARIMA na Fig. 4.2, é possível ver que o erro MAPE médio do AutoARIMA aumenta especificamente nos últimos 12 meses, que é a parte mais crítica do horizonte devido ao período de validação dos dados do Ministério da Saúde.

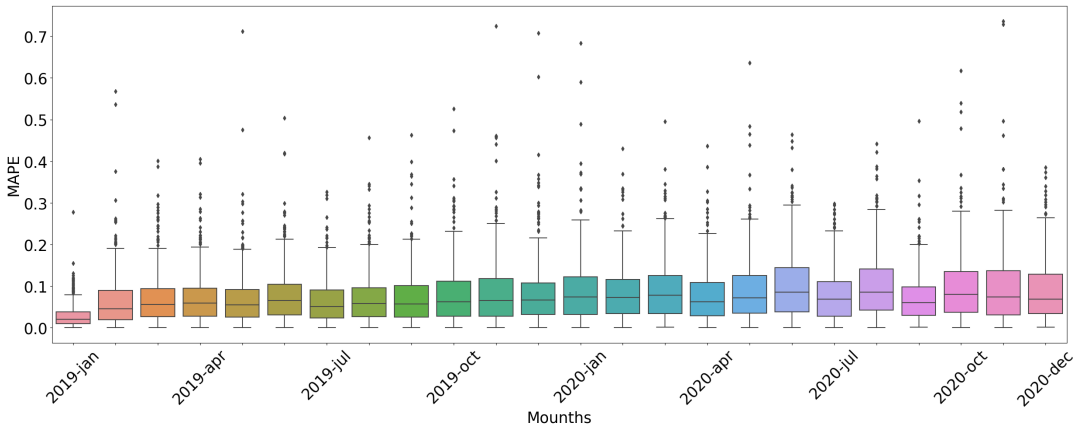


Figura 4.1: Distribuição mensal do erro MAPE da previsão LightGBM ao longo dos 24 meses no conjunto de teste.

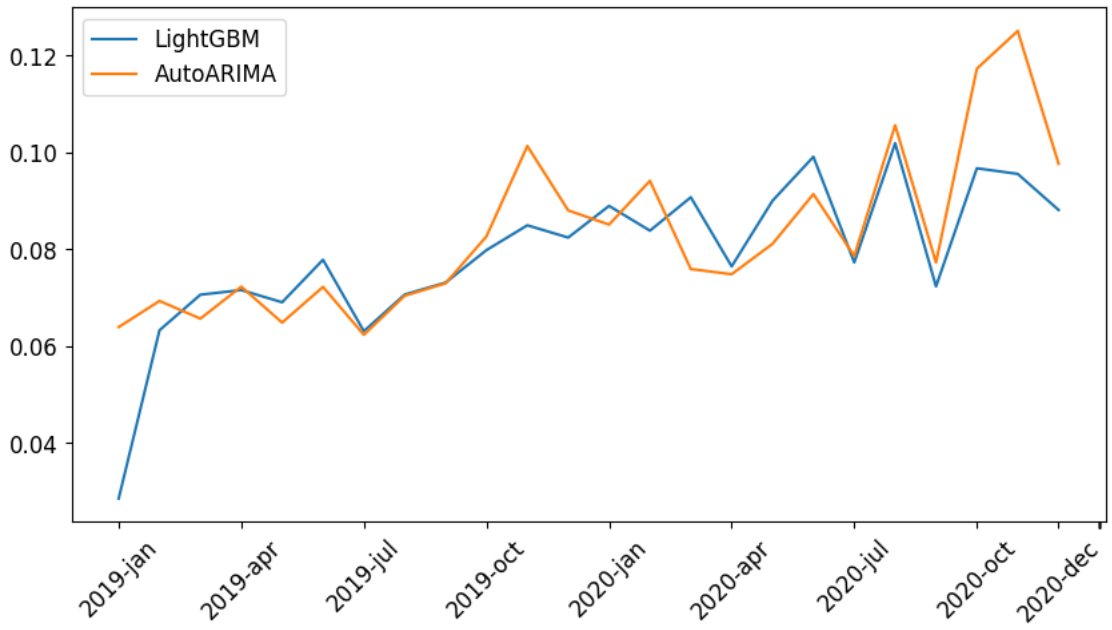


Figura 4.2: Previsão de erro MAPE de AutoARIMA e *LightGBM* em todo o conjunto de teste.

Figura 4.3: Conjunto de teste e previsões das quatro regiões com erro MAPE mais baixo (linha superior) e quatro com erro MAPE mais alto (linha inferior).

Todos os três modelos baseados em árvore levaram menos de 5 minutos para treinar com um processador Intel Core i7-13700k e 32 Gb de RAM, o *LightGBM* foi o mais rápido, treinando em apenas 15 segundos. O AutoARIMA é o modelo de maior tempo, sendo 120 minutos. Os modelos foram treinados em um AMD Ryzen 5 5600G e 24 Gb de RAM, o tempo de treino dos modelos foram próximos à 4 horas.Logo, o modelo de melhor performance é *LightGBM* com realização em 60 minutos, o AutoARIMA finalizou em 20 horas o treino completo, sendo esse o modelo mais caro, pois para cada microrregião um único modelo é treinado.

A Figura 4.3 representa um gráfico das quatro regiões de menores e maiores indices de erros em que é possível visualizar que o modelo *LightGBM* não possui boa performance em regiões de baixa população, ocasionando um MAPE maior especificamente nos últimos 12 meses.

Uma vantagem dos modelos baseados em árvore é a capacidade de medir a importância dos recursos dos modelos ajustados com base na árvore de decisão resultante. A figura 4.4 mostra a importância das covariáveis em diferentes períodos de previsão é possível ver o quão importante as covariáveis se tornam quando o horizonte de previsão é longo, nesse caso 24 observações a frente. Ao prever o próximo mês com base nas observações atuais, o modelo parece estar fazendo uma abordagem auto-regressiva, dando grande importância às informações recentes (desvio padrão dos últimos dois meses, mediana, mínimo e um atraso simples), além de um atraso de doze meses que pode ser imputado a um comportamento sazonal anual. Ao fazer uma previsão de 24 meses à frente, covariáveis que trazem informações gerais da microrregião tornam-se mais relevantes, como estimativa populacional e atendimento pré-natal, e código da própria região de saúde que pode ser usado para identificar um viés de região. O mês seno e cosseno fornece informações sobre quais meses estão sendo previstos, por isso faz sentido ser relevante,

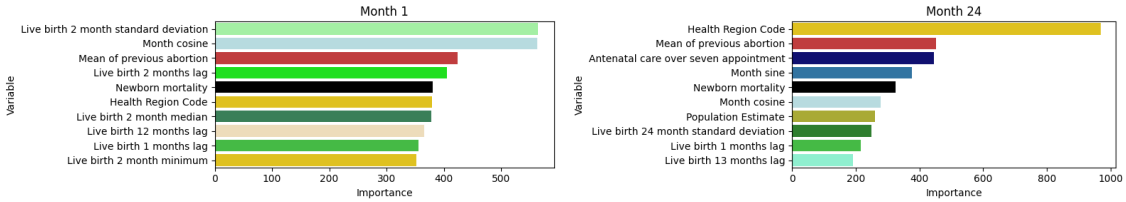


Figura 4.4: Importância das variáveis no primeiro e último mês de previsão com LightGBM.

enquanto as informações sobre aborto e mortalidade neonatal também são relevantes em todos os cenários.

4.2 Case de Estudo 2: Análise da Rotatividade de Médicos

No contexto da rotatividade de médicos (churn) no Sistema Único de Saúde (SUS), são avaliados e comparados os desempenhos dos modelos de aprendizado de máquina MLP, LightGBM e XGBoost na previsão dos padrões de evasão médica nos estados de Goiás, Distrito Federal, Mato Grosso e Mato Grosso do Sul. Utilizamos o erro médio absoluto (MAE) como principal métrica de desempenho, com os resultados resumidos na Tabela 4.2

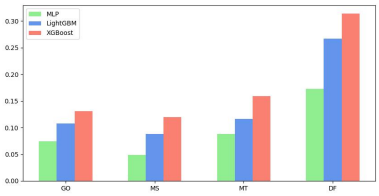
A MLP obteve consistentemente os melhores resultados em termos de MAE para todos os estados, destacando-se como o modelo mais eficaz. Para o estado de Goiás, a MLP alcançou um MAE de 0,008, enquanto o *LightGBM* e o *XGBoost* apresentaram valores de MAE de 0,018 e 0,024, respectivamente. No Distrito Federal, a MLP obteve um MAE de 0,047, superando novamente o *LightGBM* (MAE de 0,071) e o XGBoost, que apresentou o maior erro (MAE de 0,103). Em Mato Grosso, a MLP obteve um MAE de 0,010, enquanto *LightGBM* e *XGBoost* registraram valores de MAE de 0,019 e 0,033, respectivamente. No estado de Mato Grosso do Sul, a MLP também foi superior, com MAE de 0,003, em comparação com os valores de *LightGBM* (MAE de 0,012) e *XGBoost* (MAE de 0,021).

Tabela 4.2: Valores de MAE do conjunto de teste.

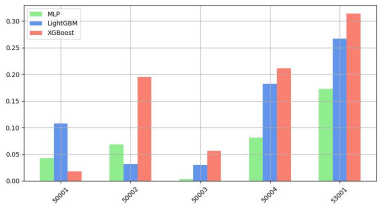
Arquitetura	Estado			
	GO	DF	MT	MS
MLP	**0.008**	**0.047**	**0.010**	**0.003**
LightGBM	0.018	0.071	0.019	0.012
XGBoost	0.024	0.103	0.033	0.021

A Figura 4.5(a) apresenta uma análise visual dos resultados de MAE para cada estado, demonstrando que a MLP obteve os menores erros em todos os cenários. Assim, exibe os resultados gerais para todos os estados, indicando que o *XGBoost* consistentemente apresentou os maiores erros, especialmente em Goiás e Mato Grosso do Sul, onde a MLP obteve um desempenho significativamente superior. No Gráfico 4.5(b), que foca em Mato Grosso do Sul e Distrito Federal, nota-se que o *XGBoost* manteve o pior desempenho no Distrito Federal, onde o MAE foi consideravelmente mais alto.

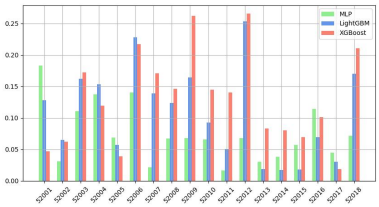
Ao analisar cada estado separadamente, a Figura 4.5(c) detalha os resultados para Goiás, destacando que a MLP obteve os menores erros em várias sub-regiões, enquanto o XGBoost apresentou os maiores valores. No Gráfico 4.5(d), que apresenta os resultados para Mato Grosso, a MLP continua a se destacar com o menor MAE na maioria das regiões, com o *LightGBM* tendo um desempenho intermediário entre a MLP e o *XGBoost*.



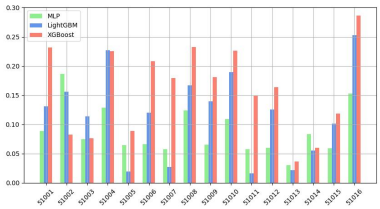
(a) *Todos estados.*



(b) *Mato Grosso do Sul e Distrito Federal*



(c) *Goiás.*



(d) *Mato Grosso*

Figura 4.5: Gráfico do MAE para cada região de saúde por estado/região de saúde.

Esses resultados demonstram o potencial do modelo MLP como uma ferramenta estratégica para o planejamento e gestão de recursos humanos no Sistema Único de Saúde (SUS). A capacidade da MLP de prever padrões de rotatividade médica com alta precisão permite que gestores de saúde antecipem possíveis lacunas no quadro de profissionais em regiões específicas, proporcionando tempo hábil para implementar medidas corretivas, como a contratação de novos profissionais ou o redirecionamento de recursos para áreas mais afetadas. Dessa forma, o uso de previsões precisas sobre a evasão de médicos pode ajudar a manter a continuidade e qualidade do atendimento, evitando a sobrecarga dos profissionais remanescentes e minimizando os impactos negativos para a população atendida. Esse tipo de abordagem preditiva, alinhado a recursos às necessidades reais das diferentes regiões de saúde, contribui para uma gestão de saúde pública eficaz e proativa,

CAPÍTULO 5

Conclusão

Ambos objetivos desse trabalho ilustram como modelos de aprendizado de máquina podem ser aplicados para abordar diferentes desafios no sistema de saúde. A previsão da taxa de nascidos vivos, com modelos baseados em árvores, como *XGBoost*, *LightGBM* e *CatBoost*, mostraram-se ferramentas eficazes para obter padrões temporais e ajudar na alocação mais eficiente de recursos, antecipando necessidades de atendimento materno e neonatal. Assim como, a rotatividade de médicos (*churn*), a aplicação de uma Rede Neural Multicamadas (MLP) demonstrou uma maior capacidade de capturar as complexas interações entre variáveis, proporcionando previsões mais precisas sobre a evasão de médicos no SUS.

5.0.1 Conclusão: Caso de Estudo 1

A mortalidade materna tem sido uma prioridade central nas políticas de saúde pública global, inserida no contexto da luta pelos direitos reprodutivos e pela igualdade de gênero, com impactos diretos na formulação de políticas públicas. Prever a taxa de nascidos vivos é uma ferramenta crítica para apoiar tomadores de decisão, fornecendo dados para o planejamento de infraestrutura, alocação de recursos em saúde e educação, e demais áreas críticas. Este estudo apresenta uma abordagem de modelagem baseada em árvores de decisão, aplicada à previsão do número de nascidos vivos no Brasil para os anos de 2019 e 2020, utilizando dados do Sistema Único de Saúde (SUS). O método proposto incorpora novas variáveis, geradas através de uma janela deslizante aplicada à série temporal, transformando-a em um problema de regressão. O impacto desta modelagem reflete-se na maior precisão preditiva, auxiliando na gestão otimizada de recursos nas microrregiões de saúde, e potencialmente contribuindo para a redução da mortalidade materna ao melhorar o planejamento das intervenções de saúde materna e neonatal.

Considerando os melhores resultados obtidos (valores MAPE e MAE de 0,0797 e 39,429, respetivamente), as previsões do modelo *LightGBM* aproximaram-se dos dados reais do número de nascidos vivos. As baixas pontuações MAPE e MAE indicam que as novas covariáveis ajudaram o modelo a aprender, de modo que previu bem as

pontuações de nascidos vivos e teve um bom desempenho. Portanto, podemos concluir que a abordagem de previsão com 228 meses de treinamento é promissora.

Algumas limitações da abordagem proposta foram identificadas. Por exemplo, áreas de saúde pouco povoadas onde o erro MAPE pode atingir valores maiores que 0,10. O Ministério da Saúde avaliou os resultados e constatou que o problema poderia ser superado, pois as projeções costumam ser mais altas do que as observadas e, portanto, ocorreriam o excedente de profissionais de saúde e não uma escassez.

No futuro, essa metodologia será revisada para os anos de 2021 e 2022. Essa pode ser uma ferramenta para ajudar engenheiros do Ministério da Saúde a prever o número de nascidos vivos, a fim de redistribuir de forma otimizada recursos como médicos, enfermeiros e unidades de terapia intensiva (UTIs) neonatais em diferentes regiões do Brasil. Contribuindo, assim, para o Objetivo de Desenvolvimento Sustentável da ONU em reduzir a mortalidade materna. Ademais, também como trabalhos futuros, será utilizados modelos de churn para prever a oferta de profissionais da saúde em cada microrregião de saúde e caso possa haver algum deficit ou migração desse profissionais pode-se avaliar medidas de alocação de novos profissionais naquela unidade.

5.0.2 Conclusão: Caso de Estudo 2

A metodologia proposta para a predição da rotatividade de médicos (*churn*) em serviços de saúde demonstrou resultados promissores, especialmente com o uso da Rede Neural Multicamadas (MLP), que obteve os melhores desempenhos em termos de erro absoluto médio (MAE) e erro quadrático médio (MSE) nos estados analisados. A MLP, com valores de MAE de 0,07 para Goiás, 0,05 para Mato Grosso do Sul e 0,09 para Mato Grosso, mostrou-se superior aos modelos baseados em árvores, como XGBoost e LightGBM, reforçando sua capacidade de capturar padrões complexos associados à rotatividade de médicos. O XGBoost, por sua vez, apresentou os piores resultados, evidenciando que sua abordagem não é a mais adequada para este problema.

Em contribuições, esse trabalho busca além de prever com precisão a evasão de médicos, poder auxiliar equipes técnicas e gestores de saúde na manutenção de uma força de trabalho estável. Ao antecipar os padrões de rotatividade, as instituições podem planejar a substituição de profissionais de maneira proativa, evitando lacunas no atendimento e a sobrecarga dos médicos remanescentes, que poderiam comprometer a qualidade do serviço prestado à população.

A eficácia da MLP ao lidar com as características e covariáveis relacionadas à rotatividade médica sugere que ela pode ser uma ferramenta valiosa no planejamento estratégico das instituições de saúde. A partir desses resultados, conclui-se que a MLP oferece uma base robusta para compreender e prever a evasão de médicos, sendo reco-

mendável expandir a análise, incorporando mais covariáveis e um maior volume de dados para aprimorar o modelo.

Embora os resultados sejam robustos, uma limitação deste estudo foi a restrição dos dados à Região Centro-Oeste, o que resultou em um conjunto de dados de menor volume para o treinamento dos modelos. Como trabalho futuro, propõe-se a ampliação do estudo para cobrir todas as regiões do Brasil, possibilitando a criação de modelos regionais específicos e a identificação de causas subjacentes à evasão de médicos. Com uma análise mais abrangente, será possível implementar estratégias ainda mais eficazes para mitigar a rotatividade e garantir um atendimento de saúde consistente e de alta qualidade em todo o país.

Referências Bibliográficas

[Abimbola et al. 2015]ABIMBOLA, S. et al. How decentralisation influences the retention of primary health care workers in rural nigeria. *Global health action*, Taylor & Francis, v. 8, n. 1, p. 26616, 2015.

[Ali 2020]ALI, M. *PyCaret: An open source, low-code machine learning library in Python*. [S.l.], April 2020. PyCaret version 1.0.0. Disponível em: <https://www.pycaret.org>.

[Alkema et al. 2016]ALKEMA, L. et al. Global, regional, and national levels and trends in maternal mortality between 1990 and 2015, with scenario-based projections to 2030: a systematic analysis by the un maternal mortality estimation inter-agency group. *The Lancet*, v. 387, n. 10017, p. 462–474, 2016. Published Online: November 12, 2015. Disponível em: <http://dx.doi.org/10.1016/S0140-6736(15)00838-7>.

[Bravo e Coelho 2020]BRAVO, J. M.; COELHO, E. Modelling monthly births and deaths using seasonal forecasting methods as an input for population estimates. In: *Demography of Population Health, Aging and Health Expenditures*. Lisboa, Portugal: Springer, 2020. p. 203–222. Accessed: 2024-09-26.

[Chauhan et al. 2020]CHAUHAN, S. et al. Application of machine learning to predict hospital churning. In: *Proceedings of the 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*. [S.l.: s.n.], 2020. p. 14. Accessed: 2024-09-26.

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

[Christ et al. 2018]CHRIST, M. et al. Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package). *Neurocomputing*, Elsevier, v. 307, p. 72–77, 2018.

[Draper e Smith 1998]DRAPER, N. R.; SMITH, H. *Applied regression analysis*. [S.l.]: John Wiley & Sons, 1998.

[Duarte et al. 2020]DUARTE, E. M. d. S. et al. Mortalidade materna e vulnerabilidade social no estado de alagoas no nordeste brasileiro: uma abordagem espaço-temporal. *Revista Brasileira de Saúde Materno Infantil*, v. 20, n. 2, p. 587–598, 2020. Accessed: 2024-09-26. Disponível em: <http://dx.doi.org/10.1590/1806-93042020000200014>.

[Feng, Kephart e Juarez-Colunga 2022]FENG, C.; KEPHART, G.; JUAREZ-COLUNGA, E. Predicting covid-19 mortality risk in toronto, canada: a comparison of tree-based and regression-based machine learning methods. *Medical Research Methodology*, 2022.

[Fiocruz 2024]Fiocruz. *Principais Questões Sobre Prevenção da Mortalidade Materna por Hipertensão*. 2024. Accessed: 2024-09-26. Disponível em: <https://portaldeboaspraticas.iff.fiocruz.br/atencao-mulher/principais-questoes-sobre-prevencao-da-mortalidade-materna-por-hipertensao/>.

[Health 2020]HEALTH, B. M. of. *Epidemiological Bulletin*. 2020. Accessed: 2024-09-26. Disponível em: <https://www.gov.br/saude/pt-br/centrais-de-conteudo/publicacoes/boletins/epidemiologicos/edicoes/2022/boletim-epidemiologico-vol-53-no20/view>.

[Hyndman et al. 2006]HYNDMAN, R. J. et al. Another look at forecast-accuracy metrics for intermittent demand. *Foresight: The International Journal of Applied Forecasting*, International Institute of Forecasters, v. 4, n. 4, p. 43–46, 2006.

[Ke et al. 2017]KE, G. et al. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, v. 30, 2017.

[Lewis, McGrath e Seidel 2009]LEWIS, J. B.; MCGRATH, R. J.; SEIDEL, L. F. *Essentials of applied quantitative methods for health services managers*. [S.l.]: Jones & Bartlett Publishers, 2009.

[Masini, Medeiros e Mendes 2021]MASINI, R. P.; MEDEIROS, M. C.; MENDES, E. F. Machine learning advances for time series forecasting. *Journal of Economic Surveys*, Wiley Online Library, 2021. Accessed: 2024-09-26.

[Mathonsi e Zyl 2022]MATHONSI, T.; ZYL, T. L. van. A statistics and deep learning hybrid method for multivariate time series forecasting and mortality modeling. *Forecasting*, v. 4, n. 1, p. 1–25, 2022. Accessed: 2024-09-26. Disponível em: <https://doi.org/10.3390/forecast4010001>.

[Misra-Hebert, Stoller e Kay 2024]MISRA-HEBERT, A. D.; STOLLER, J. K.; KAY, R. A review of physician turnover: rates, causes, and consequences. *American Journal of Medical Quality*, v. 19, n. 2, p. 56–66, 2024. Accessed: 2024-09-26.

[Organization 2019]ORGANIZATION, W. H. *Trends in Maternal Mortality*. 2019. Accessed: 2024-09-26. Disponível em: <https://apps.who.int/iris/bitstream/handle/10665/327596/WHO-RHR-19.23-eng.pdf>.

[Pacagnella et al. 2018]PACAGNELLA, R. C. et al. Maternal mortality in brazil: proposals and strategies for its reduction. SciELO Brasil, p. 501–506, 2018. Accessed: 2024-09-26.

[Pinto et al. 2022]PINTO, K. B. et al. Panorama de mortalidade materna no brasil por causas obstétricas diretas. *Research, Society and Development*, v. 11, n. 6, p. e17111628753–e17111628753, 2022. Accessed: 2024-09-26.

[Popescu et al. 2009]POPESCU, M.-C. et al. Multilayer perceptron and neural networks. *WSEAS Transactions on Circuits and Systems*, World Scientific and Engineering Academy and Society (WSEAS) Stevens Point . . . , v. 8, n. 7, p. 579–588, 2009.

[Prokhorenkova et al. 2018]PROKHORENKOVA, L. et al. Catboost: unbiased boosting with categorical features. In: BENGIO, S. et al. (Ed.). *Advances in Neural Information Processing Systems*. [S.l.]: Curran Associates, Inc., 2018. v. 31.

[Rady, Fawzy e Fattah 2021]RADY, E. H. A.; FAWZY, H.; FATTAH, A. M. A. Time series forecasting using tree-based methods. *Journal of Statistics Applications & Probability*, Natural Sciences Publishing, v. 10, n. 1, p. 229–244, 2021.

[Reis 2024]REIS, M. M. *Capítulo 4: [Análise de Séries Temporais]*. 2024. Accessed: 2024-09-26. Disponível em: <https://www.inf.ufsc.br/ marcelo.menezes.reis/Cap4.pdf>.

[Ribeiro et al. 2008]RIBEIRO, L. d. A. et al. *Indicadores de mortalidade materna em Goiás no período de 1999 a 2005: implicações para a enfermagem*. 2008. Accessed: 2024-09-26.

[Saúde 2024]SAúDE, O. P.-A. d. *Objetivos de Desenvolvimento Sustentável e a Estratégia Global para a Saúde das Mulheres, das Crianças e dos Adolescentes*. 2024. Accessed: 2024-09-26. Disponível em: <https://www.paho.org/pt/node/63100>.

[Scheffler et al. 2008]SCHEFFLER, R. M. et al. Forecasting the global shortage of physicians: an economic-and needs-based approach. *Bulletin of the World Health Organization*, SciELO Public Health, v. 86, n. 7, p. 516–523B, 2008.

[Smith et al. 2017–]SMITH, T. G. et al. *pmdarima: ARIMA estimators for Python*. 2017–. [Online; accessed <today>]. Disponível em: <http://www.alkaline-ml.com/pmdarima>.

[United Nations Development Programme 2015]United Nations Development Programme. *Human Development Report 2015: Work for Human Development*. 2015. Accessed:

2024-09-26. Disponível em: <https://hdr.undp.org/content/human-development-report-2015>.