Design, Integration, and Evaluation of a Demand Forecasting Service in the Context of Primary Healthcare

Luis F. Alves Pereira
luis-filipe.pereira@ufape.edu.br
Universidade Federal do Agreste de Pernambuco
Garanhuns, Pernambuco, Brasil

Luann Bento Ferreira
Universidade Federal do Agreste de Pernambuco
Garanhuns, Pernambuco, Brasil

Dimas Cassimiro Nascimento
Universidade Federal do Agreste de Pernambuco
Garanhuns, Pernambuco, Brasil

Igor Medeiros Vanderlei
Universidade Federal do Agreste de Pernambuco
Garanhuns, Pernambuco, Brasil

Daliton Silva
Universidade Federal do Agreste de Pernambuco
Garanhuns, Pernambuco, Brasil

Veruska Borges Santos
Universidade Federal de Campina Grande
Campina Grande, Paraíba, Brasil

Mágno Sillas Gomes
Universidade Federal do Agreste de Pernambuco
Garanhuns, Pernambuco, Brasil

Ines Alessandra Melo
Universidade Federal do Agreste de Pernambuco
Garanhuns, Pernambuco, Brasil

Abstract

Context: Efficient management of Primary Healthcare (PHC) within Brazil's Unified Health System is essential for ensuring quality healthcare services, particularly in addressing challenges related to resource allocation and workforce planning. Ensuring an equitable and sustainable distribution of healthcare resources is crucial to meeting the diverse needs of Brazil's population. **Problem:** A key challenge in PHC is accurately forecasting demand, which is vital for optimizing resource allocation and preventing facility overcrowding. In the absence of precise projections, healthcare services may experience overburdened facilities and prolonged wait times, ultimately compromising the quality of care. **Proposed Solution:** This study presents a demand forecasting service tailored to Brazil's PHC. By leveraging regression techniques, the system facilitates data-driven decision-making in resource allocation and workforce planning. Predictive analytics enable public health managers to anticipate demand more effectively, ensuring a better-prepared healthcare system capable of delivering effective services. **Information Systems (IS) Theory:** This research is grounded in the Theory of Socio-Technical Systems, as the proposed solution integrates diverse datasets and predictive models into an information system that actively interacts with existing humans workflows in PHC. By influencing decision-making and resource allocation, the system fosters a seamless integration between technological automation and human expertise, enhancing the adaptability and efficiency of healthcare services. **Method:** Following a prescriptive approach, the study evaluates the forecasting service through a detailed case study. A quantitative analysis assesses its accuracy and practical applicability, ensuring its effectiveness in supporting decision-making in PHC. **Summary of Results:** Experimental results demonstrate that the proposed service accurately predicts PHC demand by leveraging real-world data, which can lead to more timely and efficient resource allocation. **Contributions and Impact on IS:** This research contributes to the Information Systems field by offering a scalable and adaptable tool for Brazil's public healthcare context, enabling efficient healthcare planning and enhancing resource management.

CCS Concepts

• **Information systems** → **Data analytics**.

Keywords

Primary Health Care, Demand Forecasting, Regression, Artificial Intelligence

1 Introdução

A Atenção Primária à Saúde (APS) desempenha um papel fundamental na estrutura do Sistema Único de Saúde (SUS) no Brasil, sendo responsável pela prevenção, diagnóstico e tratamento de grande parte das doenças e condições crônicas que afetam a população. No entanto, a gestão eficiente dos recursos e o dimensionamento adequado de profissionais de saúde na APS permanecem desafios críticos. Nesse contexto, o planejamento com base em estimativas futuras se apresenta como uma ferramenta essencial na organização do serviço para atender à demanda na APS [3]. Uma estratégia promissora para enfrentar esses desafios envolve a utilização de técnicas de previsão de demandas, que permitem aos gestores antecipar as necessidades de atendimento e, assim, otimizar a alocação de recursos.

Um dos problemas mais críticos na gestão da APS é a previsão precisa da demanda por serviços de saúde. Um centro de saúde deve idealmente disponibilizar recursos humanos e de infraestrutura com base na demanda dos pacientes. Nesse sentido, ao saber com antecedência o número de pacientes que irão ao centro de saúde, torna-se possível preparar os recursos necessários para prestar um serviço de qualidade [11]. A falta de uma projeção confiável sobre a quantidade de atendimentos necessários gera dificuldades para gestores locais na alocação de recursos e dimensionamento de equipes. Esse problema organizacional pode resultar em unidades de saúde sobrecarregadas, tempos de espera elevados, e, em última instância,

na redução da qualidade dos serviços oferecidos aos usuários do SUS. Como a demanda por atendimentos na APS é influenciada por diversos fatores, incluindo mudanças populacionais, sazonalidade de doenças e políticas de saúde, uma solução eficaz precisa integrar esses fatores de maneira dinâmica e adaptável.

Técnicas de previsão de demanda têm se mostrado ferramentas promissoras para resolver este problema ao antecipar as necessidades de atendimento. Estudos na área de Sistemas de Informação têm se debruçado sobre a relação entre pessoas, procedimentos e tecnologias para desenvolver soluções que ajudem organizações de saúde a lidar com esse tipo de desafio. No contexto da APS, a previsão de demanda envolve, portanto, não apenas o desenvolvimento de modelos matemáticos, mas também a integração de sistemas que dialoguem com os fluxos organizacionais e possam ser usados em processos de tomada de decisão de forma acessível aos gestores da saúde pública.

Neste artigo, apresentamos a concepção, integração e avaliação de um serviço voltado para a previsão de demandas no contexto da APS. Nosso sistema se diferencia por sua capacidade de prever com precisão a quantidade de atendimentos futuros, utilizando modelos de regressão disponíveis no estado da arte. Tal previsão não apenas facilita a tomada de decisões estratégicas, como também potencializa a criação de políticas públicas mais assertivas, ajudando a mitigar a sobrecarga de unidades de saúde e garantir que os recursos estejam disponíveis quando e onde são mais necessários.

O estado da arte na aplicação de modelos preditivos para demanda na APS revela uma ampla gama de abordagens que empregam aprendizado de máquina e técnicas de séries temporais. Um estudo no Brasil utilizou o modelo ARIMA para projetar visitas domiciliares na APS, indicando que a demanda por esses serviços pode ser prevista com base em dados históricos, o que facilita o planejamento de recursos em diferentes unidades de saúde e otimiza a distribuição de equipe em resposta a variações na demanda entre unidades de saúde com equipes de tamanho semelhante[4]. Além disso, o modelo Random Forest tem sido empregado para prever a demanda de medicamentos essenciais para o tratamento de doenças crônicas, como no caso de estudo em Ruanda, onde o modelo apresentou alta acurácia na previsão da necessidade de medicamentos para doenças não transmissíveis [20]. Outro estudo em Taiwan desenvolveu modelos especializados, utilizando técnicas como XGBoost e análise discriminante quadrática, para prever demandas específicas de serviços como assistência domiciliar e cuidados diários, levando em consideração variáveis demográficas e tipos de câncer [7]. Entretanto, muitas dessas pesquisas se concentram em contextos internacionais ou em áreas específicas da saúde, como pacientes com câncer e para doenças não transmissíveis.

Nosso trabalho se destaca ao explorar e avaliar modelos de predição do estado da arte no cenário brasileiro, utilizando dados do SUS para prever a demanda por serviços de saúde na APS. A integração de dados heterogêneos e a construção de uma arquitetura que facilita a implementação do sistema no ambiente do SUS constituem um diferencial importante em relação ao estado da arte. Além disso, o serviço de previsão de demandas foi projetado para ser escalável e capaz de se adaptar a diferentes contextos regionais, fornecendo uma solução robusta e de baixo custo para os gestores públicos.

O serviço de previsão de demandas apresentado neste trabalho foi desenvolvido no contexto de um projeto de Pesquisa e Desenvolvimento (P&D) real financiado pelo Departamento de Ciência e Tecnologia da Secretaria de Ciência, Tecnologia, Inovação e Complexo da Saúde do Ministério da Saúde, e será futuramente disponibilizado por meio de um software a ser empregado por gestores públicos que atuam no contexto da saúde pública no Brasil.

O restante deste artigo está organizado da seguinte forma. Na Seção 2, são apresentados os trabalhos relacionados. Na Seção 3, é apresentado o serviço de previsão de demandas, englobando a modelagem do serviço, tecnologias empregadas, detalhes de integração do serviço ao software, tipos de previsão realizadas e a relevância do serviço para a APS. Na Seção 4, é apresentado o planejamento da validação do serviço, abrangendo os modelos de regressão avaliados, design experimental, e métricas de avaliação. Na Seção 5, são apresentados e discutidos os resultados experimentais obtidos. Finalmente, na Seção 6, são apontadas as principais conclusões do trabalho, assim como perspectivas de trabalhos futuros.

2 Trabalhos Relacionados

Nos últimos anos, a aplicação de modelos de aprendizagem de máquina e técnicas de análise de dados têm sido explorada para otimizar a alocação e estruturação dos serviços prestados nas unidades de Atenção Primária a Saúde. No Brasil, boa parte das soluções concentram-se, ainda, na predição ou detecção de doenças tratadas pela atenção básica, a exemplo de predição do risco de doenças crônicas não transmissíveis [12], análise preditiva de doenças cerebrovasculares [15] e risco de desenvolvimento de doenças cardiovasculares [22].

Porém, no contexto da predição de demandas nas unidades da Atenção Primária à Saúde, Bolsoni et al. [3] apresentaram uma solução para predizer a quantidade de visitas domiciliares nas unidades da atenção primária nos 20 meses seguintes. A solução, utilizando o modelo Autoregressive Integrated Moving Average (ARIMA), foi aplicada no dados do município de Florianópolis (SC) e em cada uma de suas 49 unidades APS para estimar os atendimentos domiciliares dos 20 meses subsequentes a fevereiro de 2019. Porém, a solução restringe-se à predição de atendimentos classificados como visitas domiciliares, visto que o objetivo dos autores foi propor uma abordagem sustentável para evitar hospitalizações desnecessárias e manter os indivíduos em sua casa o maior tempo possível, dados às projeções de algumas doenças. Além disso, não houve validação dos resultados obtidos com os valores reais, para verificar a eficácia da solução apresentada.

Já Lorenzi et al. [18] exploraram o uso de regressão linear para estimar o volume de atendimentos relacionados a doenças respiratórias na rede pública da cidade de Curitiba, utilizando dados de atendimentos históricos e dados climáticos. Com isso, os autores consideraram 4 variáveis climáticas (temperatura mínima do dia, média de temperatura, média de umidade e média da amplitude térmica dos últimos 7 dias) como sendo a entrada do modelo para predizer o volume de consultas dos próximos dias, baseado nessas variáveis climáticas dos dias anteriores. Segundo os autores, a predição apresentou variação de 10% para mais ou para menos em relação à quantidade de atendimentos realizados. Tal resultado se

deve ao fato das variáveis consideradas apresentarem baixa correlação ($< |0,3|$) com a variável alvo, além de não considerar os dados históricos de atendimentos, que tendem a apresentar padrões, nem outros fatores externos que podem influenciar na demanda por atendimentos, a exemplo de doenças virais e fatores socioeconômicos da região. O modelo não apresenta suporte para os demais tipos de atendimentos prestados nas unidades e também tal avaliação não foi estendida para os dados dos demais centros de saúde do Brasil, que podem apresentar comportamento distintos, devido às características de cada localidade.

Fora do Brasil, os autores Klute et al. [17] analisaram 20 modelos preditivos, categorizados em tradicional, híbrido (tradicional + aprendizagem de máquina) e aprendizagem de máquina, para definir qual modelo tende a ser mais adequado para prever a demanda por consultas ambulatoriais. Para a previsão de demanda nos dois centros de saúde analisados, o modelo XGBoost mostrou-se mais apropriado, visto que apresentou o menor erro. Tal modelo, que é um algoritmo de divisão de espaço, usa a engenharia de variáveis para transformar as entradas das séries temporais em componentes previsíveis e as entradas são formuladas como uma regressão supervisionada para construir suas previsões. Assim, as novas variáveis, combinadas com o uso de dados estacionários, demonstraram que um método normalmente usado para previsões de eventos pode ser usado efetivamente para previsões de séries temporais. Entretanto, é importante destacar que o XGBoost tende a funcionar bem para séries que exibem padrões cíclicos, mas não capturam tendências com a mesma eficácia, a exemplo de aumentos de demandas causados por surtos virais, pandemias, etc.

Noura Al Nuaimi [2] apresentou a aplicação de técnicas de mineração de dados em quatro modelos para auxiliar os gestores públicos na previsão da demanda por serviços de saúde no Emirado de Abu Dhabi. Os modelos Naive Bayes, K Nearest Neighbor (KNN), Support Vector Machine (SVM) e o C4.5 foram treinados para classificar os distritos em subabastecido ou potencialmente superabastecido de hospitais, além de classificar a adequação do tamanho das clínicas de saúde. Porém, tais classificações são para as necessidades atuais e futuras, conforme os dados disponibilizados pela capital. Assim, os resultados obtidos servem para auxiliar os gestores nas tomadas de decisões, mas não fornecem a predição direta da demanda por serviços de saúde.

Por outro lado, Skordis-Worrall et al. [24] estimaram a demanda por serviços de saúde em quatro distritos na Cidade do Cabo, África do Sul. Os autores apresentaram dois modelos sobre a demanda por assistência médica: um modelo probit de efeitos aleatórios para estimar a probabilidade de usar qualquer serviço e um modelo binomial negativo para modelar o número de visitas entre os usuários. Entretanto, o trabalho restringe-se apenas em analisar e estimar as necessidades de quatro distritos específicos, considerando apenas suas realidades e características locais.

Ainda, Ramgopal et al. [21] desenvolveram e validaram um modelo preditivo, utilizando algoritmos de aprendizagem de máquina em conjunto (ensemble), para prever a demanda a cada hora por serviços médicos de emergência em Nova Iorque, Estados Unidos. Apesar dos autores terem demonstrado o ganho na eficácia ao utilizar a abordagem de agregação de modelos, a solução não tem foco ou validação na previsão de demandas nos serviços da atenção primária, os quais apresentam comportamentos intrínsecos diferentes do contexto emergencial, a exemplo de acompanhamento contínuo, prevenção de doenças, diagnósticos primários, etc. [23].

Além destes, outros autores também exploraram o tema de predição por serviços no sistema de saúde [13] [1] [5] [6] [9] [10] [19] [25] [26], mas todos com avaliações internacionais apenas. Dessa forma, as principais diferenças entre a estimativa de demanda por serviços no SUS e nos centros de saúde no exterior, diz respeito ao tipo dos dados utilizados, que varia consideravelmente de país para país, às características locais e sazonais de cada região, bem como à divisão de atendimentos de cada local.

Em contrapartida, o presente estudo apresenta uma solução para realizar a predição de todos os atendimentos realizados nas unidades de APS, no contexto do Sistema Único de Saúde no Brasil. Visando atingir melhor generalização da solução, bem como a eficácia nos resultados, para que seja uma solução aplicável em todas as unidades APS do Brasil, foram analisados diversos modelos de Aprendizagem de Máquina, com diferentes combinações de arquiteturas, além da validação em dados reais. A solução apresentada demonstra-se efetiva para realizar a predição do total de atendimentos das unidades APS, facilitando a tomada de decisões dos gestores em saúde pública ao tomar conhecimento sobre a previsão da demanda real por serviços das unidades nos próximos meses e anos.

3 Serviço de Previsão de Demandas

Os modelos de previsão de demanda apresentados neste trabalho foram integrados em um sistema mais amplo, projetado para oferecer ao usuário gestor de saúde uma experiência intuitiva e de fácil utilização, com visualizações de dados por meio de mapas, gráficos e tabelas. Este sistema, denominado **SmartAPSUS**, também conta com algoritmos para a importação de dados provenientes de diversas fontes, como o IBGE, DATASUS e e-SUS, que podem estar em diferentes formatos. A importação desses dados é essencial para assegurar o bom funcionamento dos modelos e a precisão das previsões. A seguir, serão apresentados detalhes da implementação do sistema.

3.1 Arquitetura e Modelagem

O desenvolvimento do sistema foi organizado em duas partes especializadas. A primeira, denominada **Serviço de Dados**, foca na captura e tratamento de dados, além da integração com os modelos de previsão. A segunda parte, denominada **Serviço de Software**, refere-se ao sistema de informação web, que disponibiliza aos usuários, entre outras funcionalidades, a capacidade de interagir facilmente com os modelos desenvolvidos.

Cada parte foi implementada como um serviço autônomo, que se comunicam de forma assíncrona. Essa abordagem permitiu a escolha das tecnologias mais adequadas para cada serviço, considerando suas necessidades específicas. Optou-se pelo modelo cliente-servidor, onde a aplicação web, como cliente, é responsável pela interação com o usuário e exibição dos resultados dos modelos. O servidor contém os dois serviços mencionados, além dos componentes de infraestrutura necessários para seu funcionamento.

A arquitetura com serviços independentes oferece a vantagem de uma gestão mais eficiente dos recursos de hardware, especialmente em ambientes de produção. Isso permite duplicar instâncias dos

serviços mais demandados, garantindo escalabilidade e otimização de desempenho em momentos de carga elevada. A comunicação assíncrona entre os serviços não apenas atende às necessidades de tempo de execução mais longo dos modelos de previsão, mas também adiciona robustez ao sistema, pois os serviços operam de maneira independente e não bloqueante, evitando o encadeamento de falhas.

3.2 Tecnologias Empregadas

A Figura 1 representa os principais componentes da arquitetura do sistema e as tecnologias utilizadas na sua implementação. A seguir, detalhamos as escolhas tecnológicas e as razões para essas opções:

- **Next.js e React (Frontend):** Esses frameworks foram escolhidos por sua capacidade de criar interfaces altamente interativas e pela facilidade de integração com bibliotecas de exibição de gráficos e mapas, melhorando a experiência do usuário. O Next.js também oferece renderização do lado do servidor, o que otimiza a performance e a SEO da aplicação.
- **Java e Spring Boot (Serviço de Software):** A combinação de Java com Spring Boot foi escolhida devido à robustez e escalabilidade que essas tecnologias oferecem, características essenciais para um serviço que lida com alto volume de requisições e operações complexas. O Spring Boot facilita a criação de serviços com configuração mínima, promovendo agilidade no desenvolvimento.
- **Python (Serviço de Dados):** Python foi selecionada pela sua ampla gama de bibliotecas para manipulação e análise de dados, além de ser amplamente utilizada no desenvolvimento de modelos de machine learning e previsão.
- **PostgreSQL com módulo PostGIS (Armazenamento de Dados):** O PostgreSQL, um SGBD relacional de código aberto, foi escolhido pela sua confiabilidade e conformidade com os padrões SQL. O módulo PostGIS foi integrado para oferecer suporte geoespacial.
- **RabbitMQ (Comunicação entre Serviços):** O RabbitMQ foi escolhido como broker de mensagens pela sua eficácia em gerenciar comunicações assíncronas. Ele oferece confiabilidade e suporte para filas de mensagens persistentes, garantindo que mensagens não sejam perdidas em caso de falhas.
- **Técnicas de ETL (Extração, Transformação e Carga):** As técnicas de ETL foram empregadas para integrar dados de diversas fontes e formatos em um modelo interno consistente, garantindo que os dados sejam padronizados antes de serem usados nos modelos de previsão.

Todos os componentes do back-end são executados como containers Docker. Essa escolha oferece maior flexibilidade na implantação, escalabilidade e gerenciamento dos componentes do sistema. Os containers isolam cada serviço e recurso, facilitando a implementação consistente em diferentes ambientes, seja em desenvolvimento, teste ou produção. O uso de Docker não apenas melhora a administração dos recursos, mas também contribui para a agilidade na implantação e gestão do sistema, permitindo a replicação fácil em diversos cenários.

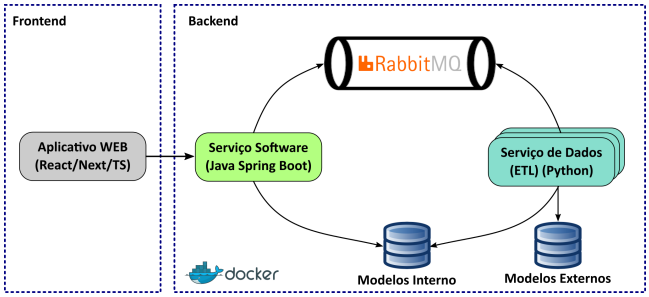

Figura 1: Arquitetura do sistema SmartAPSUS.

3.3 Integração do Serviço

Para exemplificar o processo de integração entre o Serviço Software e o Serviço de Dados, será apresentada, na Figura 2 a sequência de ações do caso de uso "executar predição".

(1) **Início da Requisição pelo Usuário:** A interação é iniciada pelo usuário através da interface front-end web, que realiza uma requisição HTTP para o back-end. Toda a comunicação entre o front-end e o back-end é mediada pelo Serviço Software.
(2) **Captura da Requisição pelo Serviço Software:** O controller do Serviço Software recebe a requisição HTTP, armazena as configurações associadas à solicitação no banco de dados e, em seguida, envia uma mensagem do tipo EXECUTE EVT para a fila do RabbitMQ, incluindo o ID da solicitação de execução recém-criada.
(3) **Confirmação ao Usuário:** Após o envio da mensagem ao RabbitMQ, a requisição HTTP é finalizada e o usuário recebe uma confirmação de que a solicitação foi agendada com sucesso, sendo informado de que o processo será concluído em breve.
(4) **Processamento Assíncrono pelo Serviço de Dados:** Assincronamente, o Serviço de Dados consome a mensagem da fila do RabbitMQ, capturando o ID da execução recém-adicionada.
(5) **Execução do Modelo de Predição:** O Serviço de Dados, em seguida, realiza uma consulta ao banco de dados para recuperar os parâmetros da predição, que estão armazenados em formato JSON. Após a leitura dos parâmetros, é feita uma segunda consulta ao banco de dados para carregar os dados de atendimentos que atendem aos critérios dos parâmetros passados. Os parâmetros decodificados e os dados dos atendimentos são, em seguida, passados para o modelo de predição.
(6) **Armazenamento do Resultado da Predição:** O resultado da execução do algoritmo é armazenado no banco de dados, também no formato JSON, garantindo a consistência e a integridade dos dados para futuras consultas ou análises.
(7) **Notificação de Término:** Após finalizar a execução, o Serviço de Dados adiciona uma mensagem do tipo EXECUTION FINISHED à fila do RabbitMQ, sinalizando que o processo foi concluído com sucesso.
(8) **Notificação ao Usuário e Registro de Logs:** De forma assíncrona, o Serviço Software captura a mensagem de término da execução e notifica o usuário solicitante, seja por

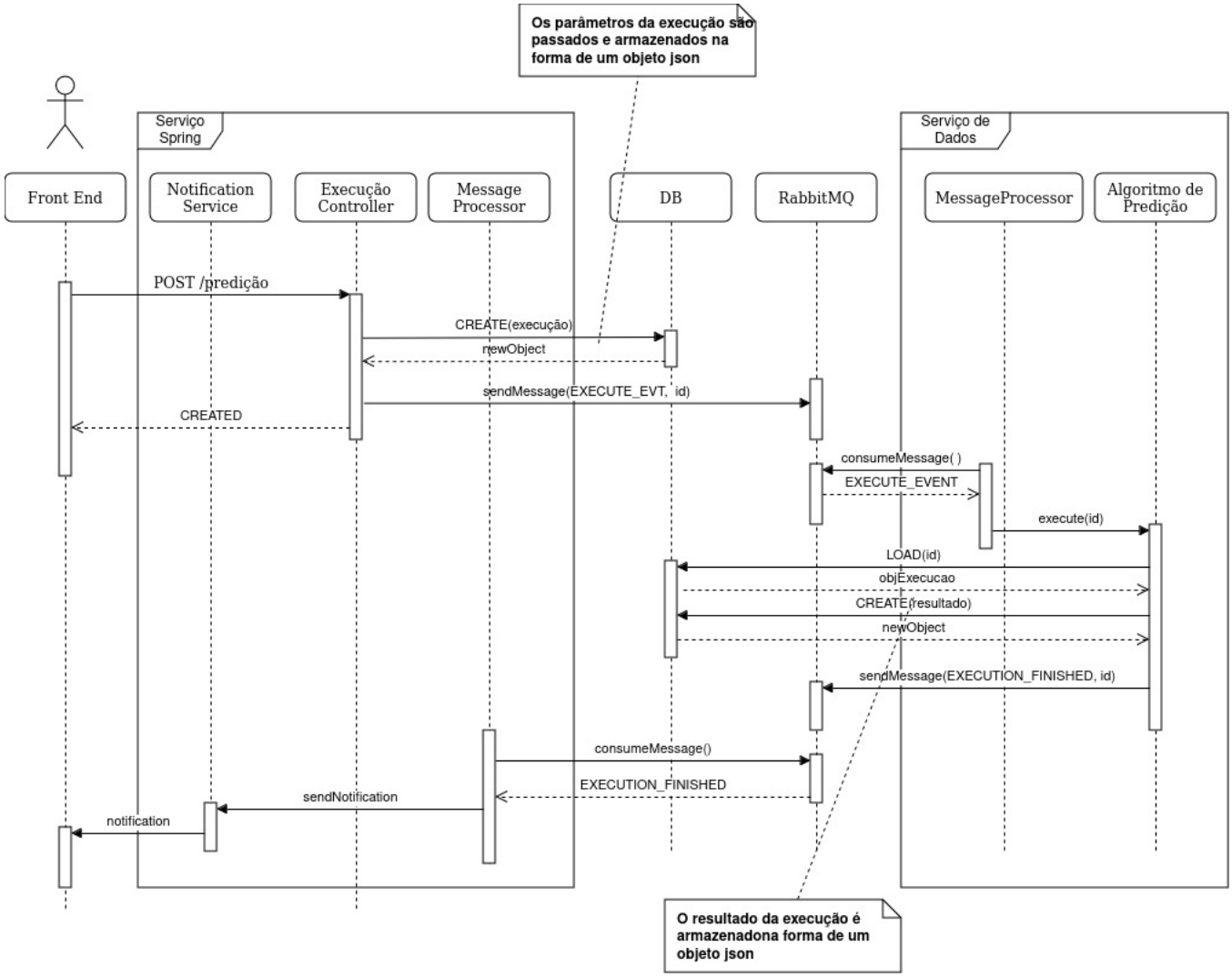

Figura 2: Diagrama de Sequência do caso de uso "executar predição".

e-mail ou SMS, conforme preferido. Para garantir o monitoramento, debugging e auditoria do sistema, todas as operações são registradas em uma tabela de log, permitindo rastrear o andamento e o histórico das requisições.

A partir dos resultados fornecidos pelo algoritmo de predição, o usuário pode interagir com uma interface gráfica que exibe as estimativas realizadas. A interface inclui:

- **Detalhes da Chamada do Algoritmo:** Informações sobre o tipo de atendimento, o aspecto socioeconômico considerado, a janela histórica utilizada e a data de emissão da previsão.
- **Histórico e Previsões:** Um gráfico que exibe o histórico de atendimentos e as previsões para os próximos meses. Esse gráfico permite uma comparação direta entre o comportamento passado e o comportamento futuro estimado, facilitando a análise das tendências.

A Figura 3 apresenta os resultados obtidos com o algoritmo de previsão de demanda para as unidades de saúde do município de Tacaratu-PE. No gráfico, são representados tanto os valores históricos, provenientes da importação de dados, quanto os valores gerados pelo modelo. Essa visualização facilita a percepção das tendências de demanda. É importante notar que devido às escolhas arquiteturais empregadas, é possível evoluir o modelo utilizado sem a necessidade de alterar a interface do usuário.

3.4 Tipos de Previsões

A versão atual do sistema é capaz de fornecer apenas a **Previsão de Demanda de Atendimentos**, que permite estimar a quantidade de atendimentos previstos para um período futuro, proporcionando uma visão antecipada das necessidades de recursos e facilitando o planejamento estratégico das unidades de saúde. Os principais parâmetros de personalização do modelo são:

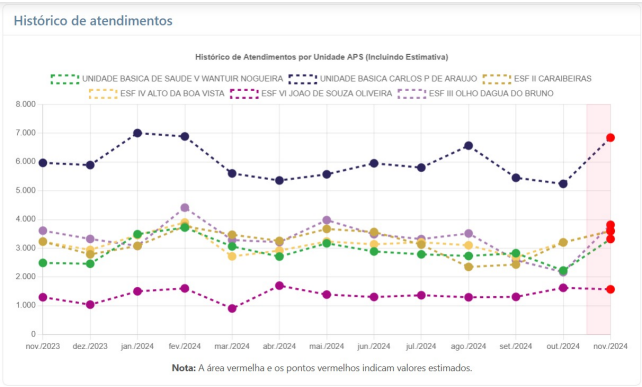

Figura 3: Exemplo de visualização dos resultados.

- **Período de Previsão:** O usuário pode definir o período para o qual deseja realizar a previsão de demanda, podendo escolher entre demanda semanal, quinzenal ou mensal.
- **Tipo de Atendimento:** A previsão de demanda pode ser filtrada de acordo com o tipo de atendimento desejado.
- **Unidade de Saúde:** O sistema possibilita a escolha da unidade de saúde para a qual a previsão será gerada, permitindo a realização de previsões específicas que consideram particularidades locais, como o volume de atendimentos históricos e a capacidade de atendimento.

4 Validação do Serviço

A validação do serviço de previsão de demanda proposto é guiada por duas questões de pesquisa fundamentais: (a) *qual o modelo de rede neural mais adequado para prever a demanda por serviços de saúde no SUS?* e (b) *qual o tamanho ideal da janela temporal de meses a ser observada para realizar a previsão do próximo mês?*

A primeira questão reflete a realidade amplamente reconhecida na literatura de Aprendizado de Máquina de que não é possível determinar a priori qual técnica se adaptará melhor a um conjunto de dados específico. Este fato decorre do *No Free Lunch Theorem* [27], que implica que nenhum algoritmo é universalmente superior em todos os domínios de problema; ao contrário, a eficácia de um algoritmo depende das características específicas do problema ao qual ele é aplicado. Essa incerteza exige a avaliação de diferentes modelos para avaliar a melhor adequação aos dados do SUS.

Já a definição do tamanho da janela temporal é igualmente crucial, uma vez que uma janela muito curta pode ser insuficiente para capturar tendências sazonais ou eventos passados relevantes, comprometendo a acurácia da previsão. Por outro lado, uma janela muito longa pode introduzir ruído, reduzindo a precisão preditiva e aumentando desnecessariamente o custo computacional.

Nas próximas subseções apresentamos as redes neurais avaliadas, o design experimental utilizado, e os resultados obtidos.

4.1 Modelos Avaliados

Redes neurais recorrentes (RNNs, do inglês *Recurrent Neural Networks*) [14] são ideais para processar dados com dependência temporal porque possuem conexões recorrentes que permitem carregar informações entre estados anteriores, capturando padrões sequenciais e dependências de temporais. Neste artigo, além de avaliar a RNN original proposta em 1990 por Jeffrey Elman [14], avaliaremos também duas redes recorrentes mais modernas: a *Long Short-Term Memory* (LSTM) e a *Gated Recurrent Unit* (GRU).

As redes LSTM [16] introduzem uma arquitetura com células de memória e portas de entrada, saída e esquecimento, com o objetivo de capturar dependências de longo prazo de forma mais eficaz em relação às primeiras RNNs. Hoje, as LSTMs são amplamente utilizadas em séries temporais e Processamento de Linguagem Natural devido à sua capacidade de capturar padrões em dados sequenciais longos e complexos.

As redes GRU [8] exploram os mesmos conceitos de portas presentes nas LSTMs, mas simplificam a arquitetura ao integrar as células de memória da LSTM e as conexões recorrentes da RNN em um único estado oculto. Essa simplificação reduz a complexidade computacional, tornando as GRUs mais rápidas para treinar, enquanto ainda são capazes de lidar com dependências de longo prazo e curto prazo em função das operações realizadas nas portas de esquecimento e atualização. As GRUs frequentemente apresentam desempenho competitivo com as LSTMs em muitos problemas.

4.2 Design Experimental

O presente estudo foi conduzido com base em uma série histórica composta por 144 registros mensais de visitas domiciliares realizadas no âmbito do programa governamental de Atenção Primária à Saúde na região metropolitana de Recife [1], abrangendo o período de janeiro de 2004 a dezembro de 2015. A granularidade dos dados é mensal[2], o que permite capturar padrões temporais regulares e tendências de longo prazo. Para garantir uma validação adequada dos modelos, os dados foram divididos em dois conjuntos: o período de janeiro de 2004 a dezembro de 2013 (83%) foi utilizado para o treinamento, enquanto os registros de janeiro de 2014 a dezembro de 2015 (17% do total) foram reservados para validação.

Antes do treinamento das redes neurais, os dados foram normalizados entre 0 e 1, utilizando os valores mínimos e máximos observados no conjunto de treinamento, para facilitar a convergência dos modelos. Para cada rede neural utilizada (RNN, LSTM, e GRU), foram testadas janelas temporais de 3, 6, 9 e 12 meses, que representam o número de recorrências executadas na rede para prever o próximo valor. Todas as redes utilizaram 8 neurônios no estado oculto, enquanto as saídas das recorrências foram processadas por duas camadas totalmente conectadas com 4 e 1 neurônio, respectivamente. A função de ativação *ReLU* foi aplicada em todas as camadas intermediárias, e a função *Sigmoid* foi usada na última camada, garantindo que a predição final como um valor real entre 0 e 1. Esse valor foi posteriormente denormalizado para obter as previsões finais em número absoluto de visitas domiciliares. Todos os modelos foram treinados por 300 épocas, um valor no qual observou-se uma convergência no decaimento das curvas de erro médio quadrático (MSE, do inglês *Mean Squared Error*) durante o treinamento, indicando que o modelo atingiu uma estabilidade no processo de aprendizado.

[1] dados do Sistema de Informação em Saúde para Atenção Básica do governo federal brasileiro, disponível em sisab.saude.gov.br.
[2] nível mais detalhado de granularidade disponível nos dados acessados.

4.3 Resultados

Os resultados obtidos para cada um dos modelos de redes neurais testados (RNN, LSTM e GRU) são apresentados em termos de do Erro Médio Percentual Absoluto (MAPE, do inglês) e coeficiente de determinação R^2.

Considerando o conjunto $\{(\hat{y}^{(1)}, y^{(1)}), \cdots, (\hat{y}^{(N)}, y^{(N)})\}$ contendo N pares contendo a n-ésima predição $\hat{y}^{(n)}$ e o respectivo valor real $y^{(n)}$, onde N é o número de sequências no conjunto de validação, o MAPE é definido como

$$\text{MAPE} = \frac{1}{N}\sum_{n=1}^{N}\left|\frac{y^{(n)} - \hat{y}^{(n)}}{y^{(n)}}\right| \times 100 \tag{1}$$

Por ser dado em termos percentuais, o MAPE permite uma interpretação intuitiva do quão longe, em média, as previsões estão dos valores reais. Por outro lado, o coeficiente de determinação R^2 é definido como

$$R^2 = 1 - \frac{\sum_{n=1}^{N}(y^{(n)} - \hat{y}^{(n)})^2}{\sum_{n=1}^{N}(y^{(n)} - \overline{y})^2} \tag{2}$$

onde $\overline{y}$ representa a média dos valores reais $y^{(n)}, 1 \leq n \leq N$. O valor R^2 próximo de 1 sugere que o modelo se ajusta bem aos dados em termos de captura da dinâmica dos valores. Contudo, se o R^2 for baixo, isso indica que o modelo não captura bem as tendências ou padrões nos dados.

Os resultados obtidos em termos de MAPE em função do tamanho da janela de observação utilizada para prever o valor do próximo mês são apresentados na Tabela 1.

Table 1: MAPEs em função do tamanho da janela em meses para cada modelo de rede neural.

Janela (meses)	RNN	LSTM	GRU
3	7.97	7.86	7.02
6	8.17	8.29	8.55
9	9.78	8.86	9.19
12	9.33	9.97	10.26

A Figura 4 apresenta gráficos de dispersão contendo valores reais versus predições para cada uma das redes neurais quando janelas de observação de tamanho 3, 6, 9 e 12 meses são utilizadas. A linha de identidade $y = x$ - representada pela linha tracejada em vermelho - indica a situação ideal em que todos os valores previstos coincidem exatamente com os valores reais. As retas de ajuste, representadas em verde, azul e amarelo, correspondem às predições das redes RNN, LSTM e GRU, respectivamente. Dessa forma, o coeficiente de determinação R^2 entre as retas de ajuste e a linha de identidade é apresentado nas legendas dos gráficos, fornecendo uma métrica da qualidade do ajuste de cada rede.

5 Discussão

Os resultados apresentados na Tabela 1 indicam que os *setups* utilizando janelas de 3 meses resultaram em um menor MAPE para todos os modelos avaliados. Esse resultado sugere que, para o problema de previsão de demanda na Atenção Primária à Saúde, janelas de observação mais curtas podem capturar de forma mais eficiente as dinâmicas relevantes dos dados.

Considerando especificamente o *setup* com janela de 3 meses, a rede neural que obteve o menor MAPE (7,02%) foi a GRU. No entanto, é importante notar que, apesar do menor MAPE, a GRU não apresentou o maior coeficiente de determinação R^2. O maior valor de R^2 foi obtido pela RNN ($R^2 = 0,68$). Este fato sugere que, a rede mais simples tem maior capacidade de capturar os padrões de variabilidade dos dados em questão.

Ambos os modelos GRU e LSTM apresentaram desempenhos médios melhores, conforme evidenciado pelos valores de MAPE. No entanto, os menores valores de R^2 evidencia falhas em capturar as flutuações ou detalhes dinâmicos nos dados reais. Embora o MAPE sugira que os erros absolutos são baixos, um R^2 baixo indica que os modelos não estão conseguindo explicar bem a estrutura dos dados, o que pode ser crítico para nossa aplicação em que a compreensão da variabilidade temporal é fundamental.

Portanto, a melhor configuração para o sistema de previsão de demanda na Atenção Primária à Saúde é o uso da rede RNN com janelas de observação de 3 meses. Embora a RNN não tenha apresentado o menor erro médio, avaliamos que ela apresenta maior capacidade em capturar a variabilidade nos dados. Este é um aspecto importante para garantir a robustez das previsões em um ambiente complexo e dinâmico como a saúde pública.

6 Desafios e Limitações

A série histórica utilizada neste estudo compreendeu 144 registros mensais, cobrindo o período de janeiro de 2004 a dezembro de 2015. Embora 12 anos de dados sejam, em muitos contextos, considerados um horizonte razoável para a análise de séries temporais, reconhecemos que a robustez das previsões poderia ser aprimorada com um volume maior de dados. Isso se torna particularmente claro ao analisarmos o impacto do tamanho da janela de observação no número de amostras disponíveis para o treinamento dos modelos.

Por exemplo, no experimento em que utilizou-se janelas de 12 meses, cada sequência de entrada era composta por 12 valores históricos para a previsão do 13° mês. Com uma série temporal de 144 registros, o número total de sequências possíveis era dado por $144 - 12 = 132$. Ao reservar 83% dos dados para treinamento, restaram apenas $\lfloor 132 \times 0.83 \rfloor = 109$ instâncias para treinamento. Esse número reduzido pode ter comprometido a capacidade dos modelos em reconhecer padrões em janelas mais longas de dados.

De fato as janelas de 3 meses proporcionaram as previsões mais precisas nos experimentos realizados. Ao mesmo tempo, reconhecemos que certos padrões na demanda por atendimentos de saúde podem envolver ciclos anuais, semestrais ou até mais longos. Sendo assim, o volume reduzido de dados pode ter influenciado o melhor resultado observado nesse estudo, uma vez que pode não haver dados suficientes para modelar relações sazonais mais longas.

7 Conclusões e Trabalhos Futuros

Este trabalho apresentou a concepção, integração e avaliação de um serviço de previsão de demandas para a Atenção Primária à Saúde (APS), destacando sua relevância no contexto do Sistema Único de Saúde (SUS) no Brasil. A aplicação de modelos de redes neurais, como RNN, LSTM e GRU, demonstrou resultados promissores,

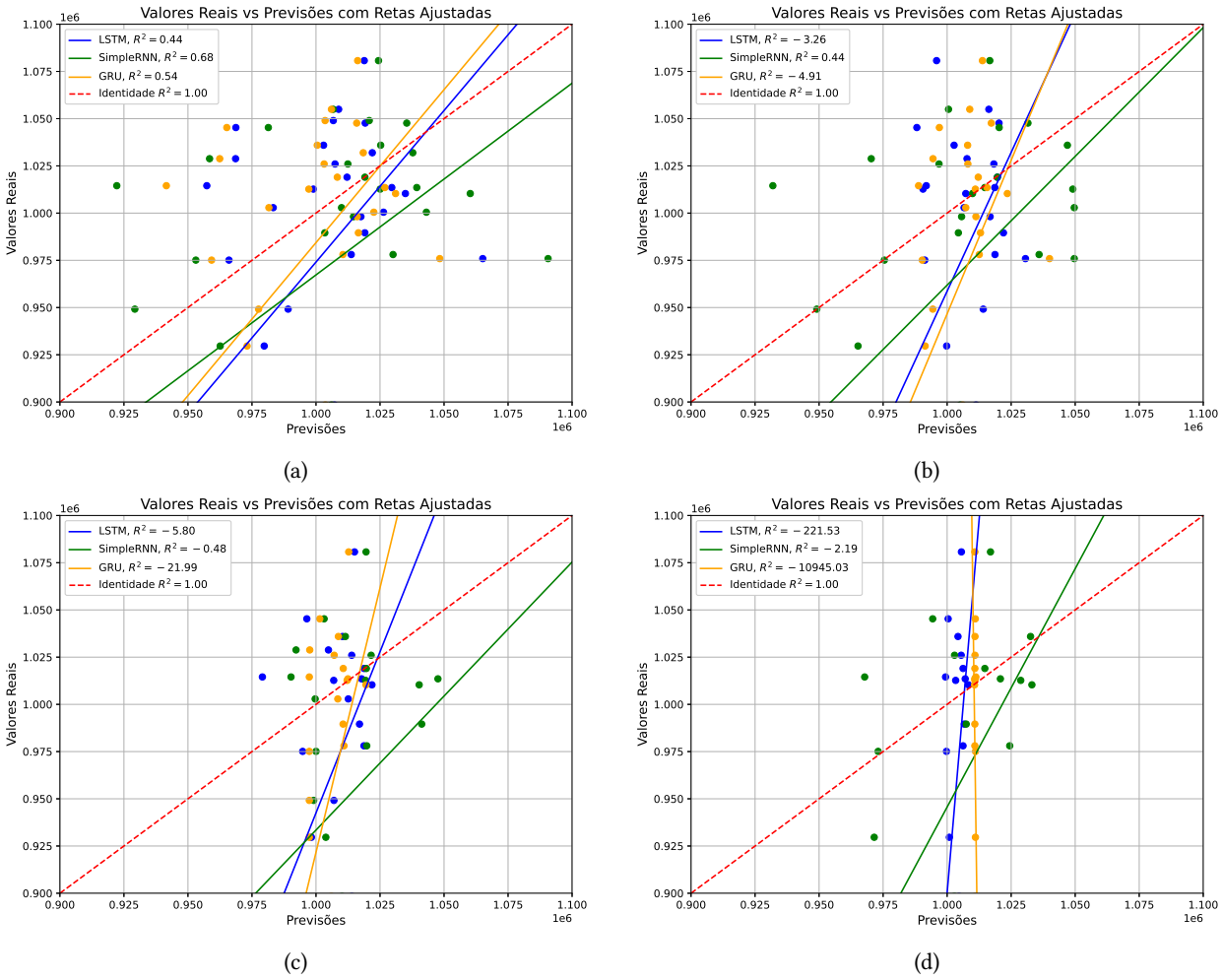

Figura 4: gráficos de dispersão contendo valores reais versus predições para cada uma das redes neurais quando janelas de observação de tamanho 3 (a), 6 (b), 9 (c) e 12 (d) meses são utilizadas.

sendo a RNN com janela de observação de 3 meses identificada como a configuração mais adequada para capturar padrões de variabilidade nos dados e garantir previsões robustas.

Os resultados obtidos confirmam o potencial da solução proposta em contribuir para a melhoria da gestão dos recursos na APS, permitindo a antecipação de demandas e a alocação eficiente de profissionais e insumos. Além disso, a abordagem escalável e adaptável do serviço reforça sua aplicabilidade em diferentes contextos regionais, promovendo maior equidade no acesso aos serviços de saúde. O serviço desenvolvido poderá contribuir significativamente para a otimização dos recursos aplicados por gestores públicos no contexto ad APS, fortalecendo sua capacidade de atender de forma proativa às necessidades de saúde da população brasileira.

Adicionalmente, destaca-se a importância de tornar essa tecnologia acessível a gestores e profissionais que não possuem domínio técnico avançado em Ciência de Dados ou Inteligência Artificial. Para isso, a integração de uma interface amigável e intuitiva ao sistema permite que esses usuários possam explorar os benefícios do serviço de previsão de demandas de forma prática e eficiente. Essa característica contribui para democratizar o uso da tecnologia, ampliando sua adoção e impacto no fortalecimento das políticas públicas de saúde.

Apesar dos resultados promissores, ressalta-se que o modelo treinado com dados de visitas domiciliares na região metropolitana de Recife pode não ser diretamente aplicável a outras localidades, dadas as diferenças nos perfis epidemiológicos e socioeconômicos das regiões. No entanto, os achados deste estudo demonstram a viabilidade do uso de redes neurais para modelar a demanda por serviços de saúde. Estudos futuros poderão explorar a necessidade de treinar modelos específicos para diferentes regiões do país, ajustando-os às particularidades locais.

Como continuidade deste trabalho, pretende-se incorporar séries temporais mais longas e dados adicionais - como indicadores socioeconômicos e epidemiológicos - para refinar a acurácia das previsões. Também planejamos avaliar o desempenho do serviço em diferentes regiões do Brasil, considerando particularidades locais

e desafios específicos. Outrossim, almeja-se investigar outras técnicas de aprendizado de máquina (*e.g.*, ARIMA, Random Forest, e XGBoost) para comparar o desempenho com os modelos avaliados neste estudo.

Agradecimentos

Agradecemos o apoio e financiamento do Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq e do Departamento de Ciência e Tecnologia da Secretaria de Ciência, Tecnologia, Inovação e Complexo da Saúde do Ministério da Saúde - Decit/SECTICS/MS no projeto de pesquisa em execução.

Referências

[1] A Okay Akyuz, Mitat Uysal, Berna Atak Bulbul, and M Ozan Uysal. 2017. Ensemble approach for time series analysis in demand forecasting: Ensemble learning. In *2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*. IEEE, 7–12.

[2] Noura Al Nuaimi. 2014. Data mining approaches for predicting demand for healthcare services in Abu Dhabi. In *2014 10th International Conference on Innovations in Information Technology (IIT)*. IEEE, 42–47.

[3] Luiza Bolsoni, Leandro Pereira Garcia, and Daniela Baumgart de Liz Calderón. 2022. Predição de visitas domiciliares na atenção primária: uma abordagem de séries temporais com o modelo Autoregressive Integrated Moving Average. *Revista Brasileira de Medicina de Família e Comunidade* 17, 44 (2022), 3012–3012.

[4] Franciele Guimarães de Brito et al. 2019. Aplicação do modelo ARIMA na previsão de atendimentos em pontos de atenção com alta demanda da Rede de Assistência à Saúde do município de Monte Carmelo, MG. (2019).

[5] Rafael Calegari, Flavio S Fogliatto, Filipe R Lucini, Jeruza Neyeloff, Ricardo S Kuchenbecker, and Beatriz D Schaan. 2016. Forecasting daily volume and acuity of patients in the emergency department. *Computational and mathematical methods in medicine* 2016, 1 (2016), 3863268.

[6] Chen-Yang Cheng, Kuo-Liang Chiang, and Meng-Yin Chen. 2016. Intermittent demand forecasting in a tertiary pediatric intensive care unit. *Journal of medical systems* 40 (2016), 1–12.

[7] Shuo-Chen Chien, Yu-Hung Chang, Chia-Ming Yen, Ying-Erh Chen, Chia-Chun Liu, Yu-Ping Hsiao, Ping-Yen Yang, Hong-Ming Lin, Xing-Hua Lu, I-Chien Wu, et al. 2023. Predicting long-term care service demands for cancer patients: A machine learning approach. *Cancers* 15, 18 (2023), 4598.

[8] Kyunghyun Cho. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[9] Murray J Cote, Marlene A Smith, David R Eitel, and Elif Akçali. 2013. Forecasting emergency department arrivals: a tutorial for emergency department directors. *Hospital topics* 91, 1 (2013), 9–19.

[10] Murray J Cote and Stephen L Tucker. 2001. Four methodologies to improve healthcare demand forecasting. *Healthcare Financial Management* 55, 5 (2001), 54–54.

[11] Juan J Cubillas, María I Ramos, and Francisco R Feito. 2022. Use of Data Mining to Predict the Influx of Patients to Primary Healthcare Centres and Construction of an Expert System. *Applied Sciences* 12, 22 (2022), 11453.

[12] Oberdan Santos da Costa and Luis Borges Gouveia. 2023. Plataforma inteligente de predição do risco de doenças crônicas não transmissíveis de apoio à decisão clínica na atenção primária de saúde, usando Inteligência Artificial. *Revista Fontes Documentais* 6, Ed. Especial (2023), 67–69.

[13] Stephen DeLurgio, Brian Denton, Rosa L Cabanela, Sandra Bruggeman, Arthur R Williams, Sarah Ward, Ned Groves, and John Osborn. 2009. Forecasting weekly outpatient demands at clinics within a large medical center. *Prod Invent Manag J* 45, 2 (2009), 35–46.

[14] Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science* 14, 2 (1990), 179–211.

[15] Jades Fernando Hammes et al. 2018. *Dashboard e um modelo de análise preditiva para doenças cerebrovasculares na atenção primária em saúde.* Masther thesis. Universidade Federal de Santa Catarina, Santa Catarina, Brasil.

[16] S Hochreiter. 1997. Long Short-term Memory. *Neural Computation MIT-Press* (1997).

[17] Brian Klute, Andrew Homb, Wei Chen, and Aaron Stelpflug. 2019. Predicting outpatient appointment demand using machine learning and traditional methods. *Journal of medical systems* 43 (2019), 1–10.

[18] Mayara Regina Lorenzi, Cristiano da Cunha Ribas, and Luiz Gomes Jr. 2018. Predição do volume de atendimentos de saúde na cidade de Curitiba utilizando dados abertos. In *Escola Regional de Banco de Dados (ERBD)*. SBC.

[19] Li Luo, Le Luo, Xinli Zhang, and Xiaoli He. 2017. Hospital daily outpatient visits forecasting using a combinatorial model based on ARIMA and SES models. *BMC health services research* 17 (2017), 1–13.

[20] François Mbonyinshuti, Joseph Nkurunziza, Japhet Niyobuhungiro, and Egide Kayitare. 2022. Application of random forest model to predict the demand of essential med. *Pan African Medical Journal* 42, 1 (2022).

[21] Sriram Ramgopal, Ted Westling, Nalyn Siripong, David D Salcido, and Christian Martin-Gill. 2021. Use of a metalearner to predict emergency medical services demand in an urban setting. *Computer Methods and Programs in Biomedicine* 207 (2021), 106201.

[22] Joaquim Assis Araújo Rangel. 2023. *Risco de desenvolvimento de doenças cardiovasculares em usuários da atenção primária à saúde.* B.S. thesis. Universidade Federal do Rio Grande do Norte, Rio Grande do Norte, Brasil.

[23] MINISTÉRIO DA SAÚDE. 2024. *SAPS - Secretaria de Atenção Primária à Saúde.* Retrieved November 10, 2024 from https://www.gov.br/saude/pt-br/composicao/saps

[24] Jolene Skordis-Worrall, Kara Hanson, and Anne Mills. 2011. Estimating the demand for health services in four poor districts of Cape Town, South Africa. *International health* 3, 1 (2011), 44–49.

[25] Yan Sun, Bee Hoon Heng, Yian Tay Seow, and Eillyne Seow. 2009. Forecasting daily attendances at an emergency department to aid resource planning. *BMC emergency medicine* 9 (2009), 1–9.

[26] Melanie Villani, Arul Earnest, Natalie Nanayakkara, Karen Smith, Barbora De Courten, and Sophia Zoungas. 2017. Time series modelling to forecast prehospital EMS demand for diabetic emergencies. *BMC health services research* 17 (2017), 1–9.

[27] David H Wolpert and William G Macready. 2005. Coevolutionary free lunches. *IEEE Transactions on evolutionary computation* 9, 6 (2005), 721–735.