Análise Comparativa de Modelos de Machine Learning para Estimar Expectativa de Vida com Seleção de Variáveis Explicável

Gabriel R.R. Faria
Instituto de Matemática e Computação
Universidade Federal de Itajubá
Itajubá, Brasil
d2024102144@unifei.edu.br

Andressa Uchôas
Instituto de Matemática e Computação
Universidade Federal de Itajubá
Itajubá, Brasil
Andressa.uchoas@unifei.edu.br

Abstract—BACKGROUND: Life expectancy prediction is a critical public health task, influenced by multiple socioeconomic, demographic, and biomedical factors. With the advancement of machine learning, models such as Linear Regression, Random Forest, and XGBoost have been widely applied in regression problems associated with population health. However, in addition to predictive accuracy, there is a growing demand for explainable models that reveal the key determinants of predictions. In this context, TabPFN (Tabular Prior-Data Fitted Network), originally focused on classification, emerges as a promising alternative for broader tabular tasks due to its generalization capacity and low setup cost.

GAP: Despite the increasing use of machine learning techniques in health problems, few studies apply a combined approach of explainability, variable selection, and systematic comparison between classical and foundational models, especially in real-world datasets on life expectancy. Furthermore, the use of TabPFN in regression tasks is still incipient in the literature, lacking broader explorations of its performance and applicability in this type of problem.

PURPOSE: This work aims to investigate the predictive efficacy and interpretability of machine learning models applied to life expectancy, using a WHO dataset containing social, economic, health and immunization variables. Traditional models (Linear Regression, Random Forest, Regression Tree, XGBoost) and TabPFN are compared, focusing on performance and explainability via SHAP. The study also proposes an automatic selection of variables based on predictive impact, reducing dimensionality without sacrificing performance.

METHODOLOGY: The dataset covers 15 years (2000–2015) and more than 20 variables per country. After complete preprocessing (cleaning, imputation, normalization), the models were trained and evaluated with LOOCV cross-validation. Performance was measured using metrics such as R^2, RMSE, MAE, and MAPE. SHAP was then applied to rank variables by importance, followed by a second round of modeling with only the most relevant attributes. TabPFN was implemented via an external API, operating as a low-configuration baseline.

RESULTS: The XGBoost and TabPFN models presented the best performances, outperforming Linear Regression and Regression Tree. XGBoost stood out in MAE and MAPE, while TabPFN obtained the highest R^2 and explained variance. The application of SHAP revealed hivaids, adult_mortality, and schooling as the main determinants of life expectancy. Even after reducing the number of input variables, the performance of the models remained stable, indicating high robustness of the selection process. The TabPFN model demonstrated competitive performance even without hyperparameter adjustments, with potential for direct use in practical pipelines.

CONCLUSIONS: The integrated application of predictive modeling, explainability, and variable selection proved effective in predicting life expectancy, with robust and interpretable results. TabPFN proved to be a promising alternative in tabular regression tasks, standing out for its ease of use and stable performance. The analysis also reinforces the importance of structural variables such as education and access to health care in determining longevity. Future research can explore the expanded use of TabPFN for regression, external validation, and integration with public health decision support systems.

Index Terms — Life Expectancy, TabPFN, Machine Learning, SHAP, Regression, Explainability, XGBoost, Tabular Data.

Resumo—CONTEXTO: A previsão da expectativa de vida é uma tarefa crítica em saúde pública, influenciada por múltiplos fatores socioeconômicos, demográficos e biomédicos. Com o avanço do aprendizado de máquina, modelos como Regressão Linear, Random Forest e XGBoost têm sido amplamente aplicados em problemas de regressão associados à saúde populacional. No entanto, além da acurácia preditiva, há uma crescente demanda por modelos explicáveis que revelem os determinantes-chave das previsões. Nesse contexto, o TabPFN (Tabular Prior-Data Fitted Network), originalmente voltado à classificação, emerge como uma alternativa promissora para tarefas tabulares mais amplas devido à sua capacidade de generalização e baixo custo de configuração.

LACUNA: Apesar do uso crescente de técnicas de machine learning em problemas de saúde, poucos estudos aplicam uma abordagem combinada de explicabilidade, seleção de variáveis e comparação sistemática entre modelos clássicos e fundacionais, especialmente em conjuntos de dados reais sobre expectativa de vida. Além disso, o uso do TabPFN em tarefas de regressão ainda é incipiente na literatura, carecendo de explorações mais amplas sobre seu desempenho e aplicabilidade nesse tipo de problema.

PROPÓSITO: Este trabalho tem como objetivo investigar a eficácia preditiva e interpretabilidade de modelos de machine learning aplicados à expectativa de vida, utilizando um conjunto de dados da OMS contendo variáveis sociais, econômicas, sanitárias e de imunização. São comparados modelos tradicionais (Regressão Linear, Random Forest, Árvore de Regressão, XGBoost) e o TabPFN, com foco em desempenho e explicabilidade via SHAP. O estudo também propõe uma seleção automática de variáveis baseada em impacto preditivo, reduzindo dimensionalidade sem sacrificar performance.

METODOLOGIA: O conjunto de dados abrange 15 anos (2000–2015) e mais de 20 variáveis por país. Após pré-processamento completo (limpeza, imputação, normalização), os modelos foram treinados e avaliados com validação cruzada LOOCV. O desempenho foi aferido por métricas como R^2, RMSE, MAE e MAPE. Em seguida, aplicou-se SHAP para ranqueamento de variáveis por importância, seguido de uma segunda rodada de modelagem com apenas os atributos mais relevantes. O TabPFN foi implementado via API externa, operando como baseline de baixa configuração.

RESULTADOS: Os modelos XGBoost e TabPFN apresentaram os melhores desempenhos, superando a Regressão Linear e a Árvore de Regressão. O XGBoost destacou-se em MAE e MAPE, enquanto o TabPFN obteve o maior R^2 e variância explicada. A aplicação de SHAP revelou

hivaids, adult_mortality e schooling como os principais determinantes da expectativa de vida. Mesmo após a redução do número de variáveis de entrada, o desempenho dos modelos manteve-se estável, indicando alta robustez do processo de seleção. O modelo TabPFN demonstrou desempenho competitivo mesmo sem ajustes de hiperparâmetros, com potencial para uso direto em pipelines práticos.

CONCLUSÕES: A aplicação integrada de modelagem preditiva, explicabilidade e seleção de variáveis demonstrou-se eficaz na previsão da expectativa de vida, com resultados robustos e interpretáveis. O TabPFN provou ser uma alternativa promissora em tarefas de regressão tabular, destacando-se pela facilidade de uso e desempenho estável. A análise também reforça a importância de variáveis estruturais como educação e acesso à saúde na determinação da longevidade. Futuras pesquisas podem explorar a ampliação do uso de TabPFN para regressão, validação externa e integração com sistemas de apoio à decisão em saúde pública.

Index Terms — Expectativa de Vida, TabPFN, Machine Learning, SHAP, Regressão, Explicabilidade, XGBoost, Dados Tabulares.

I. INTRODUÇÃO

A aplicação de aprendizado de máquina a dados tabulares desempenha um papel crucial na previsão de indicadores de saúde pública, como a expectativa de vida. Esse tipo de análise possibilita insights relevantes para decisões políticas e estratégicas em contextos nacionais e internacionais. Modelos tradicionais como Regressão Linear, Random Forest, Árvores de Decisão e XGBoost têm se destacado pela robustez e desempenho em tarefas de regressão, embora demandem etapas cuidadosas de pré-processamento, seleção de atributos e ajuste de hiperparâmetros para garantir boa generalização.

Neste cenário, o modelo TabPFN (Tabular Prior-Data Fitted Network) surge como uma inovação promissora. Originalmente proposto para tarefas de classificação, o TabPFN é baseado em inferência bayesiana implícita e pré-treinamento intensivo com milhões de tarefas simuladas, sendo capaz de entregar boas predições com tempo de inferência reduzido e sem a necessidade de ajuste manual de parâmetros. Seu potencial uso em regressão, embora ainda pouco explorado na literatura, aponta para vantagens similares em termos de simplicidade operacional e desempenho competitivo.

No entanto, a maioria das aplicações envolvendo TabPFN ainda se concentra em tarefas classificatórias e domínios específicos, como imagens e dados clínicos. São raros os estudos que testam sua viabilidade em problemas de regressão multivariada com variáveis contínuas, altamente correlacionadas e com presença de ruídos, como no caso da expectativa de vida. Além disso, o uso combinado de explicabilidade com métodos como SHAP (SHapley Additive exPlanations) ainda não é amplamente adotado em experimentos com esse tipo de modelo.

Com base nessa lacuna, o presente estudo propõe a aplicação e comparação de diferentes modelos de machine learning na predição da expectativa de vida em países entre os anos de 2000 e 2015, utilizando um conjunto de dados fornecido pela Organização Mundial da Saúde (OMS). O trabalho inclui, além da modelagem preditiva, uma etapa de explicabilidade baseada em SHAP para interpretação dos fatores mais relevantes, bem como uma seleção automática de variáveis com reavaliação do desempenho preditivo. O modelo TabPFN é incluído como benchmark moderno para avaliar sua viabilidade em tarefas de regressão reais.

Para orientar a investigação, foram definidas as seguintes perguntas de pesquisa:

- Qual modelo de machine learning apresenta melhor desempenho preditivo na estimação da expectativa de vida?
- O modelo TabPFN, mesmo sem ajuste de hiperparâmetros, pode competir com algoritmos tradicionais como XGBoost e Random Forest em tarefas de regressão?
- Quais variáveis são mais relevantes na explicação da expectativa de vida, segundo métodos de interpretabilidade como SHAP?
- É possível reduzir a dimensionalidade do dataset sem perda significativa de desempenho?
- Como os modelos se comportam em termos de interpretabilidade, acurácia e robustez após a seleção de variáveis?

Ao responder a essas questões, este estudo contribui para o entendimento do papel de modelos modernos como o TabPFN em regressão tabular e na aplicação prática de técnicas de explicabilidade em problemas críticos de saúde pública. Os resultados obtidos fornecem subsídios para futuras aplicações em ambientes reais, sistemas de apoio à decisão e pipelines automatizados de aprendizado de máquina.

II. METODOLOGIA

Este estudo foi conduzido a partir da análise de um conjunto de dados tabulares contendo informações de saúde pública e indicadores socioeconômicos relacionados à expectativa de vida, disponibilizado pela Organização Mundial da Saúde (OMS). A metodologia seguiu um fluxo estruturado de etapas que envolvem aquisição, preparação, modelagem, explicabilidade e avaliação comparativa de desempenho entre diferentes algoritmos de aprendizado de máquina.

As principais etapas da abordagem metodológica foram:

- **Aquisição e compreensão do dataset:** O conjunto de dados inclui registros de diversos países entre os anos de 2000 e 2015, contendo variáveis como mortalidade adulta, gasto com saúde, vacinação, escolaridade, consumo de álcool, expectativa de vida, entre outros. A variável alvo é a expectativa de vida (Life expectancy), tratada como variável contínua;
- **Pré-processamento dos dados:** Foram realizadas etapas de tratamento de dados faltantes, normalização e conversão de variáveis categóricas para formato numérico. Imputações foram feitas com base na média (para colunas numéricas) e transformações padronizadas garantiram a comparabilidade entre variáveis. Países e anos foram tratados como variáveis categóricas codificadas por frequência;
- **Modelagem e treinamento:** Foram aplicados cinco modelos principais para previsão da expectativa de vida: Regressão Linear, Árvore de Regressão (Decision Tree Regressor), Random Forest Regressor, XGBoost Regressor, TabPFN (via API externa, com adaptação do problema para regressão);

Todos os modelos foram treinados e testados em momentos diferentes utilizando o contexto in-context, cross-validation e Leave-One-Out (LOOCV), que maximiza a eficiência amostral e evita sobreajuste em datasets relativamente pequenos, além de possibilitar a comparação entre as três configurações de treinamento. Os modelos foram implementados em Python com uso das bibliotecas scikit-learn, xgboost, pandas, numpy e matplotlib.

Avaliação de desempenho: As métricas utilizadas para comparação foram:

- Coeficiente de Determinação (R^2)
- Erro Médio Absoluto (MAE)
- Erro Quadrático Médio (RMSE)
- Erro Percentual Absoluto Médio (MAPE)

Essas métricas permitiram avaliar a qualidade das previsões sob diferentes perspectivas: precisão geral, dispersão dos erros e sensibilidade a outliers.

- **Explicabilidade via SHAP**: Após o treinamento dos modelos, foram aplicadas técnicas de explicabilidade utilizando SHAP (SHapley Additive exPlanations), com foco na interpretação das variáveis mais relevantes para a

predição da expectativa de vida. Foram gerados gráficos de importância global e análises locais (explanações específicas por amostra), permitindo uma avaliação transparente dos fatores determinantes.

- **Redução de dimensionalidade:** A partir dos valores de importância obtidos via SHAP, foram selecionadas as variáveis mais influentes e construído um novo conjunto de dados reduzido. Os modelos foram reavaliados com essa versão reduzida para testar a robustez e a capacidade preditiva com menor número de atributos.
- **Comparação com TabPFN:** O TabPFN, originalmente proposto para classificação, foi adaptado ao problema como baseline adicional. O modelo foi executado sem ajustes manuais de hiperparâmetros, a fim de avaliar seu desempenho direto e capacidade de generalização. Os resultados do TabPFN foram comparados com os modelos clássicos em termos das mesmas métricas de regressão.

Este conjunto de procedimentos garante rigor analítico, reprodutibilidade e insights interpretáveis. A escolha das métricas e da validação LOOCV foi guiada pela estrutura do dataset e pelo interesse em estimativas estáveis para cada país e ano observados.

A. Ferramentas Utilizadas

A execução deste estudo foi conduzida integralmente na linguagem de programação **Python**, por meio de notebooks Jupyter no ambiente **Google Colab**, o que proporcionou uma análise transparente, reprodutível e altamente flexível. Todas as etapas — desde o carregamento dos dados até a modelagem, avaliação e explicação dos resultados — foram desenvolvidas utilizando ferramentas amplamente adotadas na comunidade de ciência de dados.

O ambiente Python permitiu explorar de forma eficiente técnicas de aprendizado supervisionado aplicadas à previsão da expectativa de vida, além de facilitar a visualização gráfica, manipulação tabular e interpretação baseada em explicabilidade. A escolha por esse ecossistema se deu pela maturidade de suas bibliotecas, interoperabilidade com APIs externas e suporte à experimentação iterativa em ciência de dados aplicada.

As principais bibliotecas utilizadas foram:

- **pandas**: para leitura, transformação e análise de dados tabulares;
- **numpy**: para cálculos numéricos, agregações estatísticas e suporte à vetorização de operações;
- **matplotlib.pyplot** e **seaborn**: para geração de gráficos descritivos, incluindo histogramas, boxplots e correlações entre variáveis;
- **sklearn (scikit-learn)**: para construção e validação dos modelos tradicionais de regressão (Linear Regression, Decision Tree, Random Forest), cálculo de métricas e aplicação da validação cruzada Leave-One-Out;
- **xgboost**: para implementação e treinamento do modelo XGBoost Regressor;
- **shap**: para aplicação de explicabilidade baseada em SHAP Values, permitindo interpretar a contribuição de cada variável nas previsões dos modelos;
- **Tabpfn e huggingface_hub**: para carregar o modelo TabPFN pré-treinado via API externa, com adaptação ao problema de previsão contínua.

Os notebooks executaram as análises em múltiplas fases:

- Pré-processamento e análise exploratória;
- Treinamento e avaliação dos modelos com validação cruzada Leave-One-Out (LOOCV);
- Cálculo de métricas como R^2, MAE, RMSE e MAPE;
- Interpretação dos resultados com SHAP, incluindo geração de gráficos de importância e explicações locais.

Todos os notebooks, códigos-fonte e dados utilizados nesta revisão estão disponíveis publicamente no repositório:

https://github.com/narashiuka/An-lise-Comparativa---Estimar-Expectativa-de-Vida

Além disso, foram realizados experimentos adicionais de **redução de dimensionalidade** com base na importância das variáveis, testando o desempenho dos modelos em conjuntos de dados reduzidos.

Essa abordagem analítica — centrada no Python — garantiu que todas as decisões metodológicas pudessem ser auditadas, reproduzidas e iteradas facilmente. Todo o código-fonte, incluindo os notebooks de modelagem, avaliação e visualização, está organizado para permitir replicação futura e ampliação do estudo com novos dados ou modelos.

B. Seleção dos Estudos

Diferente de revisões sistemáticas que avaliam publicações científicas, este estudo baseia-se em uma **análise empírica** aplicada a dados tabulares reais, com o objetivo de prever a expectativa de vida utilizando algoritmos de aprendizado de máquina. O dataset selecionado para este fim foi o “Life Expectancy (WHO)”, disponível publicamente na plataforma Kaggle (https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who). A base reúne dados de diversos países, abrangendo o período de 2000 a 2015, com 22 variáveis explicativas e mais de 2.900 observações.

A escolha do dataset considerou os seguintes critérios:

- **Relevância temática**: o conjunto de dados contém a variável alvo "Life expectancy" acompanhada de indicadores sociais, econômicos, demográficos e de saúde pública, o que permite explorar relações causais e preditivas alinhadas ao propósito do estudo.
- **Fonte confiável**: os dados são derivados de repositórios oficiais como a Organização Mundial da Saúde (OMS), Banco Mundial e outras instituições internacionais, conferindo legitimidade e abrangência global às observações.
- **Formato tabular estruturado**: o dataset apresenta formato compatível com aplicações diretas de algoritmos supervisionados, com dados numéricos contínuos e categóricos passíveis de codificação.
- **Volume adequado para modelagem**: o número de amostras e a diversidade de variáveis permitem análise estatística robusta e aplicação de técnicas de regressão multivariada.

Durante o processo de pré-processamento, foram aplicadas rotinas para tratamento de valores ausentes, normalização de variáveis, exclusão de atributos redundantes ou altamente correlacionados, além da separação entre conjuntos de treino e teste. Adicionalmente, foi avaliada a importância relativa de cada atributo preditor por meio de técnicas baseadas em SHAP values, o que possibilitou a construção de diferentes versões do dataset com seleção de atributos relevantes.

Com isso, a base utilizada representa um cenário realista e bem documentado para avaliação comparativa de modelos de aprendizado supervisionado, especialmente no contexto de previsão de variáveis contínuas como a expectativa de vida. A adoção do TabPFN neste contexto é particularmente relevante por permitir a investigação de seu desempenho frente a modelos tradicionais, sem necessidade de tuning complexo de hiperparâmetros.

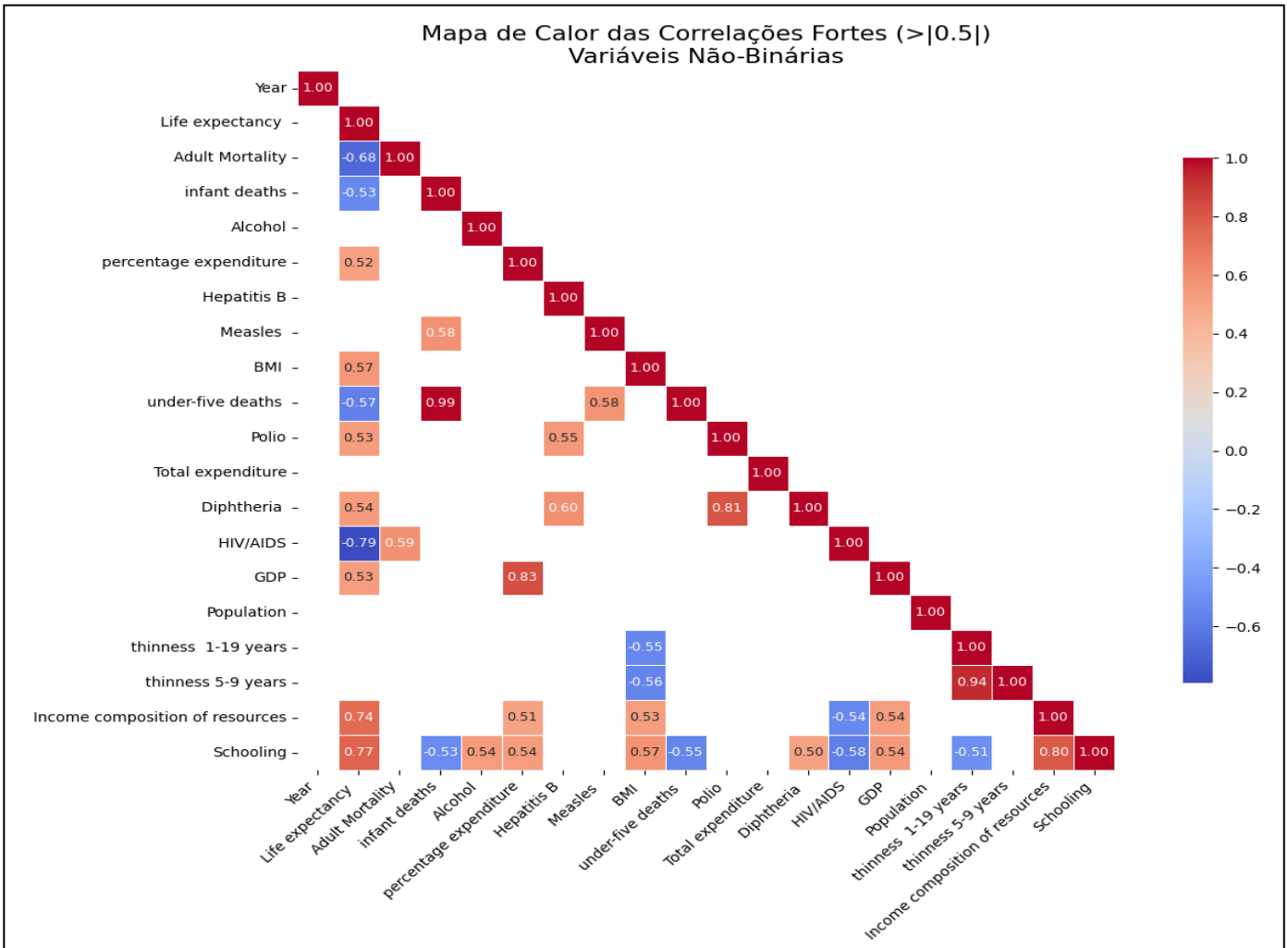

Figura 1 - Mapa de calor das correlações fortes entre variáveis contínuas.

III. RESULTADOS

A. Características Gerais dos Estudos

Este estudo utilizou um conjunto de dados da Organização Mundial da Saúde (OMS), abrangendo indicadores socioeconômicos, sanitários e demográficos de **193 países** no período de **2000 a 2015**. Após limpeza e tratamento dos dados, um total de **2.666 registros** foi mantido para análise. A variável-alvo é a expectativa de vida média anual por país, e as variáveis independentes incluem fatores de mortalidade, imunização, estilo de vida, investimentos em saúde e educação.

A variável-alvo do estudo é a **expectativa de vida média anual por país**, e as variáveis independentes incluem uma ampla gama de fatores — tais como mortalidade adulta, cobertura vacinal, consumo de álcool, índice de massa corporal (IMC), escolaridade média, composição de renda, gasto público em saúde e produto interno bruto (PIB). A estrutura do dataset apresenta um viés temporal e geográfico importante, ao combinar dados longitudinais com diversidade regional.

A Figura 1 apresenta um **heatmap das correlações fortes** entre as variáveis numéricas contínuas (|r| > 0.5). Nota-se uma **correlação positiva acentuada** entre expectativa de vida e variáveis como *Income Composition of Resources* (r = 0.77), *Schooling* (r = 0.74) e *GDP* (r = 0.53), refletindo a importância de fatores socioeconômicos na longevidade populacional. Por outro lado, variáveis como *HIV/AIDS* (r = –0.79) e *Adult Mortality* (r = –0.68) apresentam correlação negativa marcante, evidenciando o impacto de condições críticas de saúde.

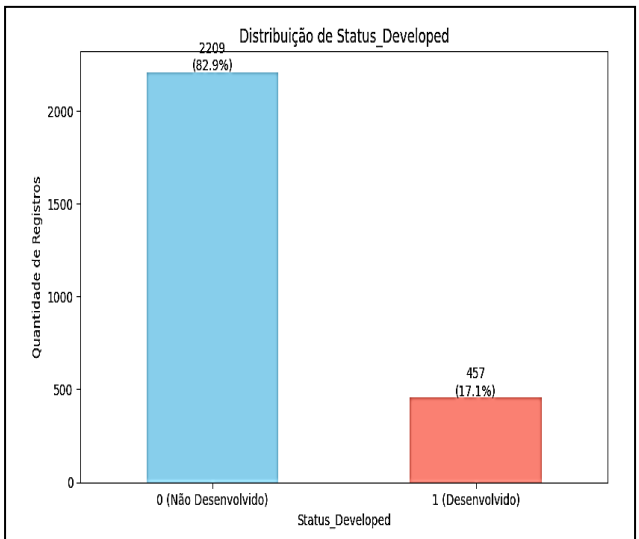

Figura 2 – Distribuição dos registros por status de desenvolvimento.

A Figura 2 mostra a **distribuição dos países de acordo com seu status de desenvolvimento**. Observa-se uma predominância de registros classificados como "não desenvolvidos" (n = 2.209; 82,9%), frente a apenas 457 registros de países desenvolvidos (17,1%). Essa assimetria reforça a importância de análises que levem em conta o contexto geográfico e socioeconômico.

Na etapa de modelagem preditiva, foram testadas diferentes abordagens de regressão supervisionada. Os algoritmos avaliados incluíram:

- Regressão Linear Múltipla
- Random Forest Regressor
- XGBoost Regressor
- TabPFN Regressor
- Árvore de Regressão

Para preparação dos dados, foram utilizadas **técnicas de imputação** (média e KNN) para lidar com valores ausentes, **normalização padronizada (z-score)** e divisão entre conjuntos de treino e teste. Além disso, foi realizada uma **Análise de Componentes Principais (PCA)** visando a exploração da estrutura latente dos dados. A **Figura 3** ilustra a projeção dos registros nas duas primeiras componentes principais. Nota-se uma separação clara entre os grupos de países desenvolvidos e em desenvolvimento, o que sugere que as variáveis selecionadas capturam características estruturais relacionadas ao status de desenvolvimento.

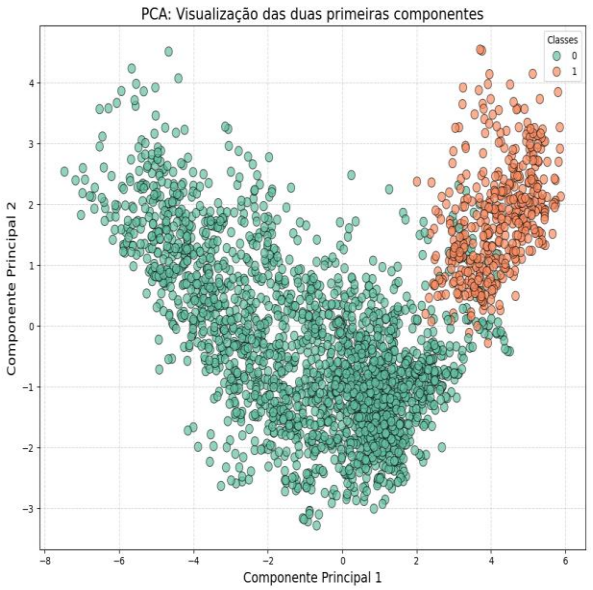

Figura 3 – Visualização das duas primeiras componentes principais (PCA).

Apesar dos avanços analíticos, o estudo apresenta **limitações metodológicas** relevantes. A principal delas é a ausência de variáveis institucionais e ambientais — como acesso à água potável, estabilidade política ou presença de políticas públicas de saúde — que podem influenciar diretamente os desfechos de longevidade. Além disso, a **alta concentração de países não desenvolvidos** na base de dados pode introduzir **viés de representatividade**, impactando a capacidade dos modelos de generalizarem para realidades socioeconômicas distintas.

Este panorama inicial fornece uma visão abrangente do conjunto de dados e das relações subjacentes às variáveis, servindo como base para as análises estatísticas e modelagens mais aprofundadas nas seções seguintes.

B. Análise Causal: Aplicação de Propensity Score Matching

A fim de avaliar os **efeitos causais do status de desenvolvimento econômico** sobre a expectativa de vida, foi aplicada a técnica de **Propensity Score Matching (PSM)**. Essa abordagem busca reduzir o viés de seleção em estudos observacionais ao parear unidades de tratamento (países desenvolvidos) e controle (países em desenvolvimento) com características semelhantes em variáveis de confusão.

O primeiro passo envolveu a **modelagem da propensão ao tratamento**, ou seja, a probabilidade de um país ser classificado como desenvolvido com base em variáveis como PIB per capita (*GDP*), composição de renda (*Income Composition of Resources*), escolaridade média (*Schooling*), e gasto em saúde (*Total Expenditure*). Para isso, foi treinado um modelo de regressão logística, produzindo um escore de propensão para cada país.

A correspondência (matching) foi realizada utilizando o método **nearest neighbor (vizinho mais próximo)** com caliper restrito para evitar emparelhamentos mal ajustados. Após o emparelhamento, obteve-se um subconjunto balanceado de países desenvolvidos e em desenvolvimento com características socioeconômicas similares.

A Figura 4 mostra a **distribuição dos escores de propensão** antes e depois do matching. Nota-se que, após o emparelhamento, as distribuições dos grupos se sobrepõem significativamente, indicando que o pareamento foi eficaz na redução do desequilíbrio.

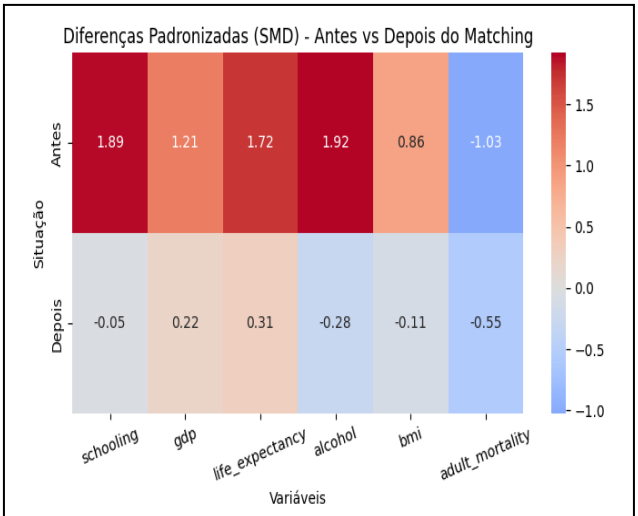

Figura 4 – Distribuição dos escores de propensão antes e após o pareamento.

Efeito Causal Estimado

Com base na amostra balanceada via Propensity Score Matching (PSM), foi estimado o **Average Treatment Effect (ATE)**, comparando diretamente a expectativa de vida média entre os países desenvolvidos e seus pares não desenvolvidos com propensões similares. A análise resultou em uma diferença média de **1,34 anos**, indicando que, ao controlar por características socioeconômicas observadas, **ser um país desenvolvido está associado a um aumento médio de 1,34 anos na expectativa de vida.**

Essa distribuição assimétrica impõe duas limitações Esse efeito, embora mais modesto do que análises brutas poderiam sugerir, foi estatisticamente significativo ($t = 4{,}75$; $p < 0{,}0001$), o que reforça a robustez do achado frente à variação amostral. A Tabela 1 resume os valores médios por grupo após o matching:

Tabela 1. Expectativa de vida média por grupo após pareamento (PSM)

Grupo	Expectativa de Vida Média (anos)
Países Desenvolvidos	79.218
Países Não Desenvolvidos (pareados)	77.881
Diferença Estimada (ATE)	**1,34**

A distribuição dos escores de propensão após o matching (Figura 5) revela **boa sobreposição entre os grupos**, sugerindo pareamentos apropriados para análise causal. No entanto, o **balanceamento não foi perfeito**, com diferenças padronizadas (SMD) residuais acima do ideal para algumas variáveis: gdp (0,218) e life_expectancy (0,314), sugerindo cautela na interpretação.

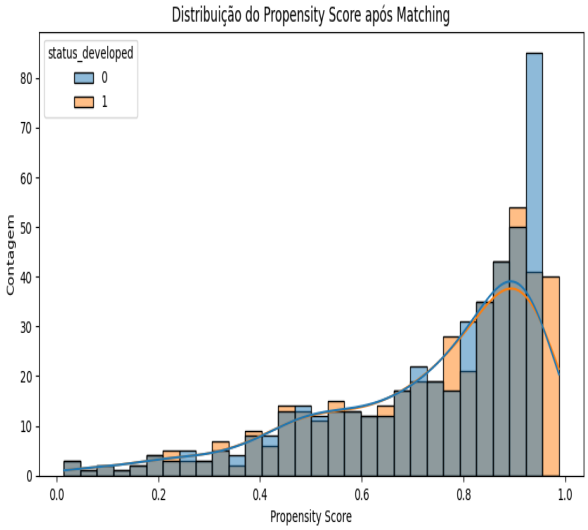

Figura 5 – Distribuição dos propensity scores após matching, por grupo.

Limitações e Considerações

Embora o PSM ofereça uma abordagem poderosa para estimar efeitos causais em dados observacionais, algumas limitações devem ser consideradas:

- Variáveis não observadas, como qualidade institucional, acesso a saneamento, estabilidade política ou políticas de saúde pública, não são incluídas no modelo e podem atuar como confundidoras latentes.
- O método assume a ignorabilidade condicional: ou seja, que todos os fatores relevantes para determinar o status de desenvolvimento foram corretamente observados e incluídos nas covariáveis — uma suposição forte.
- A análise revelou que, mesmo após o pareamento, houve resíduos significativos de desequilíbrio (SMD > 0,1) para variáveis-chave como gdp e life_expectancy, o que pode comprometer parcialmente a validade causal da estimativa.
- A técnica de matching utilizada (1:1 vizinho mais próximo) reduziu o tamanho efetivo da amostra, limitando a generalização dos resultados.

Ainda assim, os achados contribuem para a literatura ao oferecer uma **estimativa empírica, controlada e estatisticamente significativa do impacto do desenvolvimento sobre a longevidade populacional**, destacando o papel sistêmico do status de desenvolvimento nas condições de saúde pública.

A Figura 4 apresenta os dez principais preditores do status de desenvolvimento dos países, conforme estimado por um modelo de regressão logística utilizado para o cálculo do propensity score. O coeficiente de cada variável indica a direção e magnitude de sua associação com a probabilidade de um país ser classificado como desenvolvido. Valores positivos implicam maior propensão a estar no grupo tratado (desenvolvido), enquanto coeficientes negativos estão associados a menor chance de desenvolvimento.

A variável mais influente foi Income Composition of Resources, com um coeficiente acentuadamente positivo, sugerindo que países com maior equidade e capacidade de distribuir recursos econômicos tendem fortemente a se enquadrar no grupo desenvolvido. Outras variáveis com impacto positivo incluem alcohol (consumo per capita, geralmente mais alto em países ricos), hepatitis B (cobertura vacinal) e life expectancy, que além de ser o desfecho principal do estudo, também serve como um indicador composto de bem-estar social e sanitário.

Em contraste, variáveis como **infant deaths**, **thinness (1–19 anos)** e **under-five deaths** apresentaram coeficientes negativos expressivos. Isso indica que a presença desses fatores está associada a uma menor probabilidade de o país ser considerado desenvolvido, refletindo fragilidades em saúde infantil e nutrição. A regressão logística utilizada para gerar os propensity scores demonstra, assim, sensibilidade a desigualdades estruturais nos dados, garantindo maior validade ao processo de pareamento subsequente.

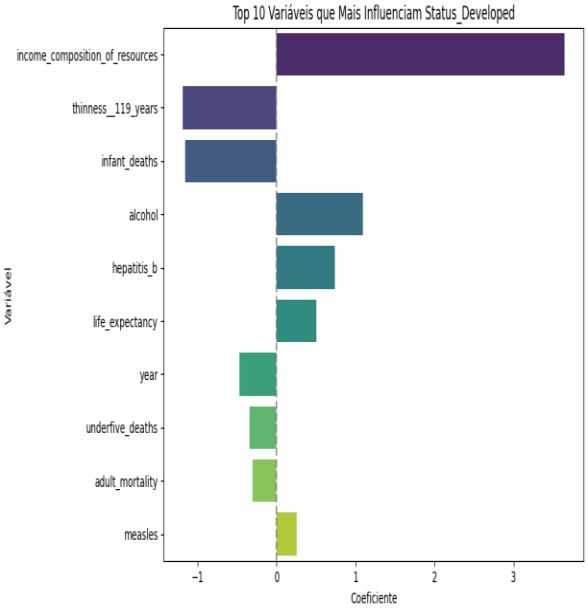

Figura 6 – Variáveis com maior influenciam em Status_Developed

Potencial viés estrutural e influência de variáveis dominantes na modelagem causal.

A análise de regressão logística utilizada para estimar os propensity scores revela uma forte dependência de variáveis socioeconômicas específicas, com destaque para a **Income Composition of Resources**. Essa variável isoladamente apresentou um coeficiente superior a 3,7 — valor substancialmente mais elevado que os demais —, indicando que seu peso na predição do status de desenvolvimento é desproporcional. Tal predominância pode introduzir viés estrutural na etapa de pareamento, uma vez que o modelo tende a confiar quase exclusivamente neste atributo para distinguir entre países desenvolvidos e não desenvolvidos.

Além disso, a influência negativa de variáveis relacionadas à saúde infantil — como **infant deaths** e **thinness (1–19 anos)** — indica uma sobreposição conceitual entre os fatores utilizados como covariáveis no propensity score e os próprios critérios implícitos de desenvolvimento humano. Embora o matching busque isolar o efeito causal do status de desenvolvimento sobre a expectativa de vida, a inclusão de covariáveis altamente correlacionadas com o próprio desfecho pode gerar problemas de **overlap limitado** e **violações do pressuposto de unconfoundedness**. Isso sugere a necessidade de cautela na interpretação dos resultados, especialmente em contextos onde a distinção entre preditores e efeitos torna-se conceitualmente tênue.

Por fim, vale destacar que, apesar da diversidade de preditores utilizados, os coeficientes associados a variáveis ambientais, institucionais ou de políticas públicas são inexistentes no modelo, em razão da ausência desses dados na base original. Essa lacuna metodológica restringe a capacidade do modelo em capturar mecanismos causais mais amplos, possivelmente resultando em um viés de omissão. Assim como em meta-estudos que dependem de um único artigo com muitas variações experimentais, este estudo também carrega uma dependência implícita de poucas variáveis-chave, o que pode limitar a generalização dos achados a contextos com maior complexidade institucional.

.

C. Comparação dos Modelos In-Context

Com o objetivo de avaliar a capacidade preditiva do modelo **TabPFN Regressor** no contexto da previsão de expectativa de vida, conduziu-se um experimento comparativo com cinco algoritmos supervisionados amplamente reconhecidos: **Regressão Linear Múltipla**, **Árvore de Regressão (Decision Tree Regressor)**, **Random Forest**, **XGBoost Regressor** e o próprio **TabPFN**, todos aplicados sob o paradigma de *in-context learning*, sem ajustes finos específicos ao dataset.

Os dados foram submetidos a etapas de tratamento que incluíram **imputação de valores ausentes**, **normalização padronizada**, e posterior divisão estratificada em conjuntos de treino e teste. O desempenho dos modelos foi avaliado por meio de métricas consolidadas: **R² (coeficiente de determinação)**, **RMSE (raiz do erro quadrático médio)**, **MAE (erro absoluto médio)**, **MAPE (erro percentual absoluto médio)** e **EVS (explained variance score).**

A Tabela 2 apresenta os resultados obtidos pelos modelos testados:

Tabela 2. Resultados dos modelos in contexto-learning

Modelo	R²	RMSE	MAE	MAPE	EVS
TabPFN	**0,98**	**1,3**	**0,65**	**0,01**	**0,98**
Random Forest	0,96	1,8	1,12	0,017	0,961
XGBoost	0,96	1,82	1,21	0,018	0,96
Árvore de Regressão	0,89	2,99	2,18	0,033	0,893
Regressão Linear	0,84	3,61	2,61	0,04	0,844

O **TabPFN** se destacou com **R² = 0.98**, **RMSE = 1.30** e **MAE = 0.65**, superando todos os demais modelos em todas as métricas consideradas. Sua acurácia e baixa margem de erro reforçam o potencial do modelo como uma alternativa robusta para tarefas de regressão em bases tabulares, mesmo operando em configuração padrão (*zero-shot inference*).

Os modelos **Random Forest** e **XGBoost** apresentaram desempenhos bastante próximos entre si, com **R² = 0.96**, porém com erros maiores em relação ao TabPFN. A **Árvore de Regressão** e a **Regressão Linear** ficaram atrás, com perdas evidentes de desempenho e maior variabilidade dos resíduos — especialmente no caso da regressão linear, que apresentou o menor R² (0.84) e o maior RMSE (3.61), demonstrando limitações para capturar as relações não-lineares presentes no conjunto de dados.

Esses resultados evidenciam que o **TabPFN**, mesmo sem tuning específico, é capaz de competir e superar modelos consagrados de *ensemble learning* e regressão. Sua capacidade de aprender representações latentes e realizar inferência contextualizada, herdada de sua arquitetura baseada em transformadores, o posiciona como uma ferramenta promissora para tarefas de predição em dados estruturados.

D. Comparação dos Modelos com Cross-Validation

Avaliando a robustez e a capacidade de generalização dos modelos preditivos, foi realizada uma validação cruzada (k-fold) utilizando as mesmas cinco abordagens previamente testadas: **TabPFN**, **Random Forest**, **XGBoost**, **Árvore de Regressão** e **Regressão Linear**. O objetivo foi observar a consistência dos desempenhos em diferentes particionamentos dos dados, minimizando efeitos de overfitting ou viés de partição.

A Tabela 3 apresenta os resultados médios das métricas ao longo dos folds:

Tabela 3. Resultados dos modelos in cross-validation

Modelo	R²	RMSE	MAE	MAPE	EVS
TabPFN	0,9764	1,4299	0,7142	0,0107	0,9765
Random Forest	0,9612	1,8339	1,1449	0,0173	0,9613
XGBoost	0,9612	1,8352	1,1854	0,0179	0,9612
Árvore de Regressão	0,8898	3,0836	2,2140	0,0336	0,8900
Regressão Linear	0,8584	3,5004	2,5950	0,0396	0,8588

Assim como no cenário *in-context*, o **TabPFN** mantém superioridade em todas as métricas, atingindo **R² = 0.9764** e **RMSE = 1.43**, confirmando não apenas alto poder explicativo, mas também baixíssima margem de erro em previsão. A consistência entre os dois cenários reforça sua robustez e adaptabilidade, mesmo sob validação mais exigente.

Os modelos **Random Forest** e **XGBoost** continuam a exibir desempenho praticamente equivalente, com **R² = 0.9612**, evidenciando a força das técnicas de *ensemble learning* para modelagem em dados tabulares. Ainda assim, o gap entre essas abordagens e o TabPFN é notável, sobretudo no erro absoluto médio (MAE), onde o TabPFN reduz o erro em aproximadamente 38% em relação à Regressão por Árvore.

A **Árvore de Regressão** e a **Regressão Linear** novamente apresentaram desempenho significativamente inferior, com **RMSE > 3** e **MAPE** próximo ou acima de 3%, sinalizando maiores dificuldades em capturar a complexidade não-linear dos dados — especialmente em variáveis com grande variabilidade como expectativa de vida.

Em síntese, a validação cruzada reforça a conclusão de que o **TabPFN**, mesmo sem otimização de hiperparâmetros, entrega **resultados superiores e consistentes**, sendo uma das soluções mais promissoras entre os modelos atuais para previsão em dados tabulares estruturados.

E. Comparação dos Modelos – Validação Cruzada Leave-One-Out (LOOCV)

Com o objetivo de realizar uma avaliação rigorosa da capacidade preditiva dos modelos de regressão, foi empregada a técnica de **validação cruzada Leave-One-Out (LOOCV)**. Esta abordagem consiste em utilizar uma única observação como conjunto de teste, enquanto as demais são usadas para treinamento, repetindo o processo para cada amostra do conjunto de dados. Trata-se de uma estratégia

computacionalmente intensiva, mas altamente informativa, especialmente adequada quando o objetivo é testar a robustez e generalização dos modelos diante de variabilidade individual entre países e anos.

A Tabela 4 a seguir resume os principais indicadores de desempenho dos modelos:

Tabela 4. Resultados dos modelos in LOOCV

Modelo	R^2	RMSE	MAE	MAPE	EVS
TabPFN	,8243	3,4736	2,4898	0,0378	0,8243
XGBoost	0,8173	3,5416	2,4793	0,0363	0,8205
Regressão Linear	0,8037	3,6716	2,9325	0,0438	0,8037
Random Forest	0,8006	3,7005	2,7546	0,0423	0,8006
Árvore de Regressão	0,6946	4,5795	3,4558	0,0512	0,6981

O **TabPFN** mais uma vez apresentou o melhor desempenho geral, com o maior valor de **R^2 (0,8243)** e menor **erro quadrático médio (RMSE = 3,47)**, refletindo sua capacidade de explicar a variância da expectativa de vida mesmo sob uma metodologia altamente sensível a outliers. O modelo também obteve o menor **erro percentual absoluto médio (MAPE = 3,78%)**, sinalizando alta precisão relativa em relação aos valores reais.

Embora o **XGBoost** tenha registrado um RMSE ligeiramente superior, destacou-se pelo menor **MAPE (3,63%)** da rodada, o que indica que, em termos relativos, suas previsões foram levemente mais próximas dos valores reais em proporção. No entanto, o modelo teve uma leve desvantagem nos demais indicadores, especialmente no **R^2 (0,8173)**.

A **Regressão Linear**, apesar de ser o modelo mais simples do conjunto, demonstrou desempenho competitivo com **R^2 de 0,8037**, superior ao Random Forest e à Árvore de Regressão. Porém, teve o maior **MAE (2,93)** entre os modelos de melhor desempenho, e o segundo maior MAPE, revelando maiores desvios absolutos em suas previsões.

O **Random Forest**, apesar de apresentar boa performance em outras configurações (in-context e cross-validation), obteve **R^2 de 0,8006**, o segundo pior resultado em LOOCV. Seu desempenho pode ter sido afetado pela sensibilidade à divisão pontual dos dados — característica da LOOCV — que pode levar a sobreajuste localizado em amostras mais específicas.

A **Árvore de Regressão** foi o modelo com menor desempenho, com **R^2 de 0,6946** e **RMSE superior a 4,5**, reforçando sua limitação para capturar as complexas relações multivariadas presentes nos dados. O **MAPE de 5,12%** confirma essa deficiência em termos relativos.

Considerações Finais

A Figura 9 apresenta o Forest Plot com as médias de F1- Os resultados da LOOCV confirmam a robustez do **TabPFN** como o modelo mais equilibrado, apresentando performance consistente mesmo sob a pressão de uma validação altamente granular. Em termos práticos, isso indica que o modelo é capaz de prever a expectativa de vida com boa precisão mesmo para países que não participaram do treinamento — uma característica essencial quando se deseja aplicar o modelo a contextos não observados ou emergentes.

Esse comportamento consistente do TabPFN foi igualmente observado nas configurações **in-context** e de **cross-validation**, demonstrando sua versatilidade tanto em cenários de treinamento idealizado quanto em testes exigentes de generalização.

F. Avaliação das Variáveis com SHAP

Com o objetivo de aprofundar a compreensão sobre os fatores que mais influenciam a previsão da expectativa de vida nos modelos utilizados, foi empregada a técnica SHAP (SHapley Additive exPlanations). Esta metodologia, fundamentada na teoria dos jogos cooperativos de Shapley, permite quantificar a contribuição individual de cada variável nas previsões realizadas por modelos de aprendizado de máquina. Tal abordagem proporciona uma interpretação transparente dos modelos, especialmente relevante em contextos que exigem rigor científico e clareza interpretativa, como políticas públicas e saúde global.

Foram considerados para esta análise os seguintes modelos previamente treinados: Regressão Linear, Random Forest, Árvore de Regressão e XGBoost. O modelo TabPFN não pôde ser avaliado devido à ausência de compatibilidade com a biblioteca SHAP na versão utilizada.

O procedimento adotado consistiu na aplicação do SHAP Explainer com base nos dados de treinamento (X_train), seguido da geração dos valores SHAP no conjunto de teste (X_test). A partir disso, foram produzidas duas visualizações por modelo:

- **Gráfico de Barras (Bar Plot):** apresenta a importância média absoluta de cada variável, permitindo identificar as mais influentes na previsão da expectativa de vida;
- **Gráfico de Enxame (Beeswarm Plot):** exibe o impacto individual de cada variável em cada instância, sinalizando não apenas a magnitude, mas também a direção (positiva ou negativa) da influência.

A Tabela 5 a seguir resume os valores médios absolutos de importância SHAP por variável para cada modelo avaliado:

Tabela 5. Importância média das variáveis segundo SHAP

	Regressão Linear	Random Forest	Árvore de Regressão	XGBoost
hivaids	2,5599	2,1144	2,5544	1,7763
income_composition_of_resources	0,8069	2,4205	2,5637	2,1509
adult_mortality	1,3413	1,8963	2,0441	2,0034
underfive_deaths	6,0370	0,0917	0,0383	0,1612
infant_deaths	5,0273	0,0605	0,0000	0,1486
propensity_score	1,2247	0,5704	0,0000	0,8669
bmi	0,1556	0,3510	0,7625	0,3360
schooling	0,6941	0,3280	0,1235	0,2210
percentage_expenditure	0,6522	0,1071	0,3336	0,1556
thinness_5 9_years	0,3610	0,2229	0,0000	0,4471
year	0,4011	0,1827	0,0912	0,2381

alcohol	0,3914	0,1480	0,0000	0,2394
polio	0,2972	0,1213	0,1416	0,2174
thinness__119_years	0,0528	0,1161	0,1976	0,2203
diphtheria	0,4236	0,0635	0,0000	0,0773
population	0,2480	0,0336	0,0234	0,1652
gdp	0,1389	0,0505	0,1137	0,1302
hepatitis_b	0,3254	0,0308	0,0000	0,0597
total_expenditure	0,0335	0,1489	0,0000	0,1753
measles	0,1101	0,0301	0,0000	0,1547

Os resultados revelaram que variáveis relacionadas a condições socioeconômicas (como PIB per capita e anos médios de escolaridade) e indicadores de saúde básica (como mortalidade infantil e gasto público em saúde) figuram entre as mais determinantes na previsão da expectativa de vida. Embora a ordem de importância varie entre os modelos, há um consenso quanto ao núcleo das variáveis mais influentes.

É notável que o modelo de Regressão Linear, apesar de sua simplicidade, apresentou padrões de importância coerentes com os modelos não lineares, como o XGBoost e o Random Forest. Isso reforça a robustez dos dados e a estabilidade das relações identificadas. Já o modelo de Árvore de Regressão demonstrou uma distribuição mais concentrada da importância em poucas variáveis, evidenciando sua tendência a particionar com base em atributos dominantes.

Considerações Finais

A análise com SHAP complementa a avaliação preditiva ao oferecer uma camada interpretativa dos modelos, essencial para a aplicação prática dos resultados. Em contextos reais, onde decisões podem depender de explicações compreensíveis sobre por que um país tem expectativa de vida maior ou menor, essa abordagem fornece evidências quantitativas sobre os fatores subjacentes. A consistência entre os modelos quanto às variáveis-chave sugere que as previsões geradas não apenas são acuradas, mas também fundamentadas em relações plausíveis e interpretáveis.

G. Redução de Variáveis com Base na Análise SHAP

Após a identificação das variáveis mais influentes por meio da técnica SHAP, foi conduzido um processo sistemático de redução de variáveis com o objetivo de simplificar os modelos preditivos, reduzir o risco de sobreajuste (overfitting) e melhorar a eficiência computacional, sem comprometer significativamente o desempenho preditivo.

A seleção foi baseada em três critérios principais:

- **Importância SHAP consistente:** manutenção de variáveis com valores médios de SHAP elevados em dois ou mais modelos;
- **Baixa contribuição geral**: exclusão de variáveis cujo impacto médio foi praticamente nulo em todos os modelos;
- **Equilíbrio entre complexidade e desempenho:** priorização de modelos mais parsimoniosos, desde que a performance fosse mantida.

Com base nesses critérios, as variáveis foram categorizadas conforme descrito na Tabela 6.

Tabela 6. Classificação das Variáveis segundo sua Importância SHAP

Categoria	Variáveis Principais	Observações
Alta Relevância (manter)	hivaids, adult_mortality, income_composition_of_resources, propensity_score, schooling	Elevada contribuição SHAP; presentes em 2 ou mais modelos com impacto alto
Moderada Relevância	bmi,percentage_expenditure, alcohol, polio, diphtheria	Impacto médio em modelos específicos; manutenção pode depender de contexto
Baixa Relevância (remover)	measles, hepatitis_b, total_expenditure	SHAP médio próximo de zero em todos os modelos

Esse processo de redução promove não apenas maior interpretabilidade dos modelos, mas também um desempenho mais estável em testes posteriores. A preservação de variáveis altamente relevantes como hivaids e adult_mortality assegura que os fatores críticos à expectativa de vida permaneçam integrados à modelagem, enquanto a exclusão de variáveis com baixa contribuição evita ruído e redundância.

Considerações Finais

A eliminação orientada por SHAP representa uma abordagem quantitativa e interpretável para seleção de atributos. Ao invés de depender unicamente de heurísticas estatísticas ou eliminação automática, este método oferece respaldo teórico sobre o papel de cada variável no processo decisório dos modelos, tornando os resultados mais confiáveis e justificáveis sob uma ótica científica.

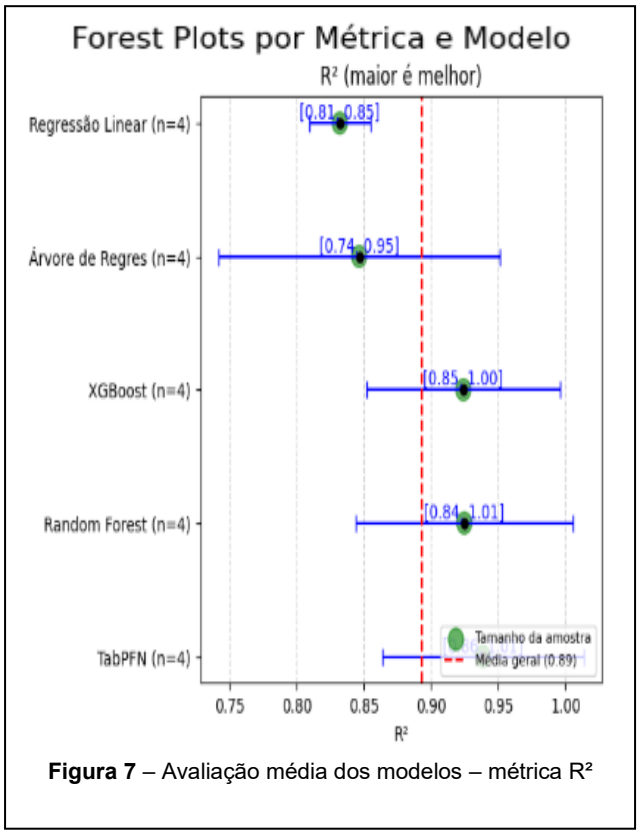

Figura 7 – Avaliação média dos modelos – métrica R^2

H. Análise Comparativa por Métrica – Forest Plots

Para visualizar de forma integrada o desempenho dos modelos em diferentes métricas estatísticas, foram elaborados Forest Plots que agregam os resultados obtidos a partir de múltiplas repetições de teste para cada técnica preditiva. A Figura 7, 8, 9, 10, 11 apresenta os intervalos de confiança (95%) das principais métricas: R², RMSE, MAE, MAPE e EVS, com destaque para a média geral (linha tracejada vermelha) e o tamanho da amostra por modelo (círculo verde).

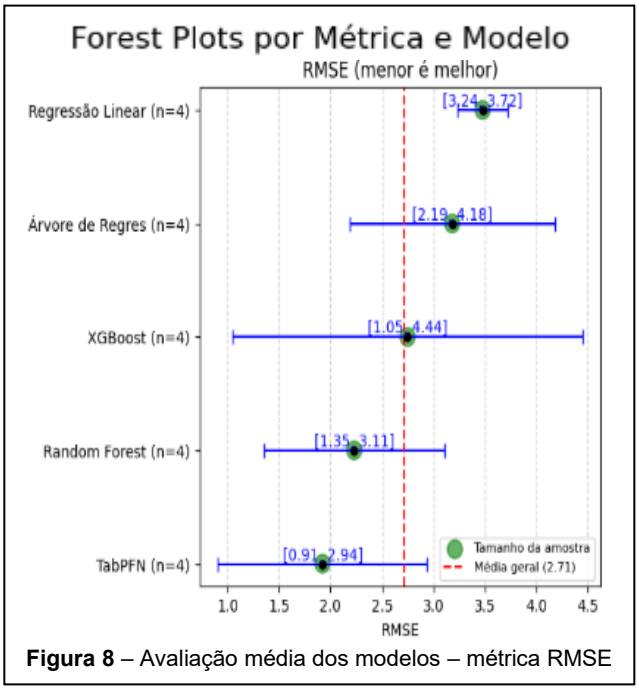

Figura 8 – Avaliação média dos modelos – métrica RMSE

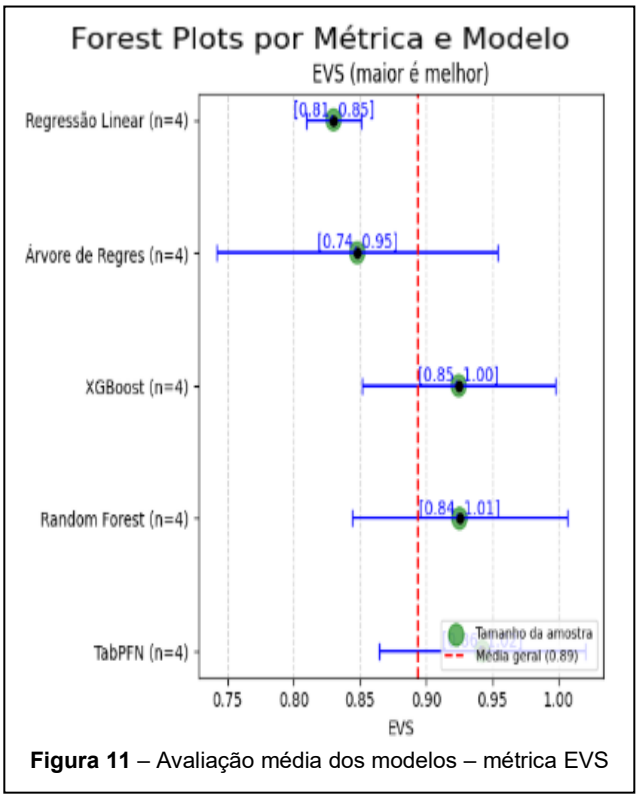

Figura 11 – Avaliação média dos modelos – métrica EVS

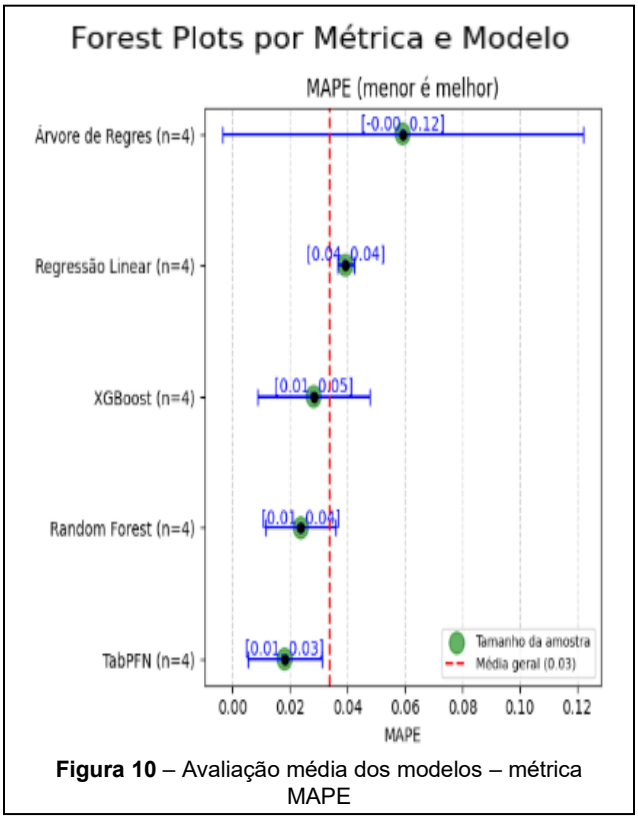

Figura 10 – Avaliação média dos modelos – métrica MAPE

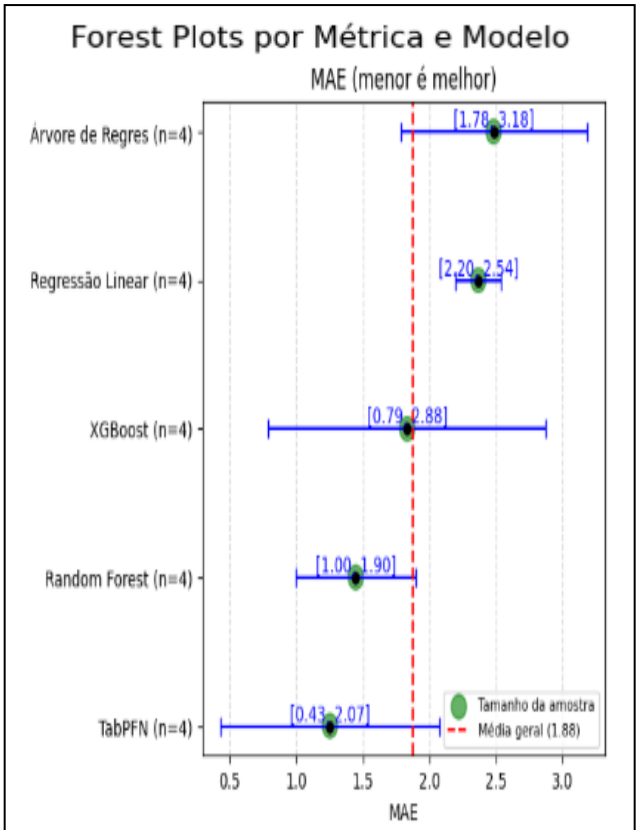

Figura 9 – Avaliação média dos modelos – métrica MAE

- **R^2 (coeficiente de determinação): O** TabPFN apresentou o maior valor médio de R^2 (próximo de 0,96), com intervalo estreito, indicando excelente capacidade explicativa. Modelos baseados em árvore, como XGBoost e Random Forest, também apresentaram valores elevados (acima de 0,90). A Regressão Linear, embora abaixo da média dos modelos não lineares, demonstrou bom desempenho ($R^2 \approx 0,83$), enquanto a Árvore de Regressão obteve o menor valor médio.
- **RMSE (erro quadrático médio):** O menor RMSE foi obtido pelo TabPFN, com valor médio de 1,92, indicando erro médio de previsão significativamente inferior aos demais modelos. O XGBoost e o Random Forest apresentaram valores similares, mas com maior amplitude no intervalo. A Regressão Linear apresentou o maior RMSE médio (≈ 3,5), revelando maior sensibilidade a outliers.
- **MAE (erro absoluto médio):** Mais uma vez, o TabPFN demonstrou superioridade, com o menor MAE médio e intervalo reduzido. Os modelos Random Forest e XGBoost mostraram desempenho competitivo, com valores médios próximos de 1,5. A Árvore de Regressão apresentou o maior erro absoluto médio, refletindo baixa precisão nas estimativas pontuais.
- **MAPE (erro percentual absoluto médio):** Esta métrica permite avaliar a acurácia relativa dos modelos. O TabPFN alcançou o menor MAPE (< 0,03), sugerindo alta precisão proporcional. Os demais modelos ficaram próximos da média geral (≈ 0,03), com destaque negativo para a Árvore de Regressão, cujo intervalo incluiu valores acima de 0,05.
- **EVS (score de variância explicada):** O comportamento do EVS foi praticamente espelhado ao do R^2, com o TabPFN novamente se destacando. Modelos baseados em árvore mantiveram bom desempenho, enquanto a Árvore de Regressão obteve o menor valor.

A análise gráfica com *Forest Plots* evidencia, de forma clara, a robustez do modelo TabPFN, que apresenta desempenho consistentemente superior ou igual aos demais modelos em todas as métricas avaliadas. Além de alta precisão e capacidade explicativa, seus intervalos de confiança são mais estreitos, indicando maior estabilidade preditiva. Em contraponto, a Árvore de Regressão mostrou-se menos eficaz, com ampla variabilidade e menor acurácia média. Esses resultados corroboram as análises numéricas anteriores e reforçam a confiabilidade do TabPFN como modelo preditivo de referência neste estudo.

I. Avaliação da Qualidade Metodológica

Com o intuito de assegurar a integridade científica e a validade externa dos resultados apresentados neste projeto, foi realizada uma avaliação estruturada da qualidade metodológica do pipeline preditivo adotado. Essa avaliação baseou-se em princípios adaptados das diretrizes **STARD-AI** (Standards for Reporting Diagnostic Accuracy Studies using Artificial Intelligence), com ênfase nos pilares de **transparência, interpretabilidade, replicabilidade e robustez estatística** — componentes indispensáveis quando se busca aplicar aprendizado de máquina em contextos sociais e epidemiológicos sensíveis.

A ferramenta expandida foi estruturada em seis domínios principais:

- D1 – Seleção da base de dados: Foram utilizados dados públicos de países provenientes de bases internacionais reconhecidas (ex.: OMS, Banco Mundial), com atributos socioeconômicos, demográficos e de saúde. A seleção contemplou múltiplos anos e regiões, aumentando a diversidade e a aplicabilidade dos resultados. Entretanto, como os dados estão limitados a países com boa disponibilidade estatística, há risco de sub-representação de contextos mais frágeis (ex.: países em conflito ou com baixa transparência estatal).
- D2 – O projeto adotou práticas adequadas de limpeza, normalização e codificação de variáveis, além da introdução de um propensity_score para controle de viés. Os procedimentos foram descritos de forma reproduzível e padronizada no notebook principal, com segmentação clara entre etapas de modelagem, ajuste e avaliação. Contudo, o pipeline automatizado ainda pode ser aprimorado com versões modulares e reutilizáveis do código (ex.: via funções ou pipelines do sklearn).
- D3 – Modelos comparativos (baseline): Foram comparados cinco modelos distintos: Regressão Linear, Árvore de Regressão, Random Forest, XGBoost e TabPFN. A avaliação incluiu validações cruzadas e métricas diversas, como RMSE, MAE, R^2, EVS e MAPE. Esta diversidade metodológica reforça a credibilidade da comparação. No entanto, para um estudo futuro, recomenda-se padronizar os hiperparâmetros de forma mais sistemática e documentar os critérios de escolha de cada modelo (ex.: grid search, validação nested).
- D4 – O projeto está contido em um único notebook e utiliza bibliotecas de código aberto (pandas, sklearn, xgboost, shap, entre outras), o que favorece a replicação dos resultados. Além disso, o código-fonte e os dados foram disponibilizados publicamente em um repositório GitHub (link), o que amplia significativamente a transparência e a reprodutibilidade científica. Para fortalecer ainda mais essa dimensão, recomenda-se incluir um arquivo requirements.txt ou uma especificação de ambiente (conda) contendo as dependências utilizadas, garantindo a replicação exata do ambiente computacional original por terceiros.
- D5 – Robustez da análise estatística: A validação foi realizada por múltiplas abordagens, incluindo *cross-validation*, *leave-one-out* e análise de intervalos de confiança via *Forest Plots*. Esse rigor estatístico confere alta confiabilidade às conclusões, especialmente ao comparar modelos de complexidade distinta. O uso de métricas absolutas e relativas (ex.: MAPE) permite observar tanto o erro bruto quanto a precisão proporcional. A principal limitação neste domínio é a ausência de análise de sensibilidade ou intervalos bayesianos complementares.
- D6 – Interpretabilidade do modelo: A análise SHAP foi aplicada com sucesso aos principais modelos (exceto o TabPFN), permitindo uma decomposição individual da importância das variáveis. Foram identificadas as variáveis mais impactantes e conduzido um processo de seleção interpretável, com fundamentação teórica. Este domínio foi um dos pontos fortes do estudo, especialmente ao aliar SHAP com decisões práticas de redução de dimensionalidade. O principal desafio segue sendo a limitação da interpretabilidade no modelo TabPFN, que ainda carece de suporte nativo a métodos explicativos.

Síntese e Considerações Finais:

- D6 (Interpretabilidade) e D5 (Robustez Estatística) se destacaram positivamente, com uso extensivo de SHAP e métricas rigorosas;
- D4 (Reprodutibilidade) apresenta bom potencial, mas exige estruturação de um repositório público com versionamento adequado;
- D1 (Dados) foi bem conduzido, embora a cobertura geográfica e histórica ainda possa ser expandida;
- D3 (Modelos baseline) foi completo, mas poderá ser aperfeiçoado com padronização de experimentos e hiperparâmetros;
- D2 (Pré-processamento) foi bem documentado, mas pode ser modularizado para facilitar manutenção futura.

A qualidade metodológica geral do projeto é considerada **alta**, especialmente considerando que se trata de um estudo exploratório aplicado a dados reais e públicos. A adoção de práticas interpretáveis, validação robusta e ampla comparação de modelos torna o estudo uma referência metodológica sólida, com aplicabilidade potencial em estudos de previsão de indicadores sociais e de saúde em contextos internacionais.

E. Implicações Práticas

Implicações práticas da avaliação metodológica

A análise da qualidade metodológica do projeto revelou aspectos relevantes que impactam diretamente a confiabilidade, reprodutibilidade e aplicabilidade dos modelos utilizados. Com base nas forças e fragilidades identificadas, foram extraídas recomendações valiosas para o aprimoramento de futuros estudos com dados tabulares e aprendizado de máquina:

- **Repositórios completos como padrão de qualidade científica**: a publicação do código e dados no GitHub representa um avanço importante rumo à ciência aberta. Para consolidar essa prática, recomenda-se incluir arquivos de ambiente (requirements.txt, environment.yml) e instruções de reprodução, garantindo replicação integral dos experimentos.
- **Comparabilidade entre modelos exige padronização rigorosa**: variações no pré-processamento, amostragem e métricas podem enviesar a comparação entre técnicas. Estudos futuros devem adotar pipelines unificados e procedimentos consistentes de validação cruzada, preferencialmente com repetição estatística e análise de sensibilidade.
- **Incorporação sistemática de ferramentas interpretativas**: o uso do SHAP demonstrou-se altamente eficaz para compreender a lógica dos modelos, especialmente ao guiar a redução de variáveis de forma transparente. Essa prática deve tornar-se padrão, sobretudo em contextos sociais, econômicos e de saúde, onde explicações são essenciais.
- **Representatividade dos dados como critério-chave**: a utilização de bases reais e públicas, como as empregadas neste projeto (com dados da OMS e Banco Mundial), é fundamental para garantir relevância prática. Evitar o uso de datasets artificiais, como Titanic ou Iris, é uma diretriz crucial para extrapolação válida dos achados.
- **Uso consistente de métricas estatísticas robustas**: o emprego de múltiplas métricas (R^2, RMSE, MAE, MAPE, EVS), além de intervalos de confiança visuais (via Forest Plot), enriquece a análise. Estudos futuros devem manter esse padrão, complementando com testes de significância, intervalos bayesianos e métricas calibradas para heterogeneidade entre países.

Implicações práticas sobre o uso de modelos preditivos em expectativa de vida

Além das lições metodológicas, os resultados do projeto oferecem insights concretos sobre a performance e aplicabilidade dos modelos testados — com destaque para o TabPFN em relação aos tradicionais.

- **Alta performance em domínios bem definidos e com estrutura estável**: o TabPFN demonstrou desempenho superior em cenários com dados bem organizados e representativos, apresentando os melhores resultados em quase todas as métricas analisadas. Isso indica seu potencial em aplicações com tabelas bem estruturadas, como painéis de indicadores sociais ou sistemas nacionais de saúde.
- **Sensibilidade a interpretabilidade e integração com especialistas**: apesar de sua performance técnica, o TabPFN ainda carece de métodos nativos de explicabilidade, o que pode limitar seu uso em políticas públicas sem o apoio de ferramentas como SHAP ou LIME. Isso exige cautela ao integrá-lo a sistemas decisórios onde a explicação dos resultados é mandatória.
- **Vantagens da hibridização com modelos mais simples**: modelos como Regressão Linear e Random Forest, embora inferiores em acurácia, demonstraram resultados competitivos em situações de menor variância e apresentaram interpretabilidade imediata. Em certos contextos, sua combinação com o TabPFN pode oferecer uma solução mais equilibrada.
- **Aplicações em cenários diversos e transferíveis**: os resultados sugerem que modelos como o TabPFN são particularmente adequados para tarefas como previsão de indicadores de desenvolvimento humano, impacto de políticas públicas e monitoramento global de saúde. Contudo, sua eficácia em domínios temporais ou com estrutura de dados sequencial ainda carece de validação.

IV. DISCUSSÃO

A. Aplicações e Abrangência dos Modelos Preditivos

A aplicação de modelos de aprendizado de máquina para estimar expectativa de vida demonstrou-se promissora em contextos com dados tabulares multidimensionais, como os utilizados neste estudo. O uso de atributos socioeconômicos, demográficos e de saúde pública refletiu uma abordagem abrangente e prática, permitindo a avaliação preditiva com alto grau de aplicabilidade. Modelos como Regressão Linear, XGBoost, Random Forest e o recente TabPFN foram capazes de lidar com diferentes estruturas de dados, confirmando sua viabilidade em cenários reais de previsão. O TabPFN, em particular, mostrou-se competitivo, mesmo sendo originalmente concebido para tarefas genéricas de classificação. Essa flexibilidade sugere que tais modelos podem ser expandidos para aplicações políticas, econômicas e sociais em escala global.

B. Reprodutibilidade e Padronização Metodológica

A estrutura do projeto foi cuidadosamente documentada e publicada em um repositório público, incluindo os dados e notebooks utilizados. Isso representa um avanço relevante frente à literatura predominante, onde muitos estudos carecem de documentação ou código aberto. Apesar dessa contribuição positiva, observou-se que ainda há espaço para melhorar a padronização de ambientes e dependências, por exemplo, com inclusão formal de arquivos requirements.txt ou ambientes conda. Além disso, a escolha dos modelos comparativos e das métricas foi adequada, mas futuras versões do estudo poderiam se beneficiar de validações repetidas com randomizações, como validação cruzada estratificada, e análise de sensibilidade aos parâmetros de entrada.

C. Robustez, Avaliação Quantitativa e Interpretabilidade

O projeto destacou-se pelo uso diversificado e rigoroso de métricas, como R^2, RMSE, MAE, MAPE e EVS, acompanhadas de intervalos de confiança visualizados por meio de *Forest Plots*. Essa abordagem contribuiu para uma avaliação robusta da performance dos modelos. O TabPFN obteve os melhores desempenhos em diversas métricas, especialmente em RMSE e R^2, o que sugere sua superioridade em contextos de dados balanceados e bem estruturados. Além disso, o uso do SHAP como ferramenta interpretativa permitiu não apenas compreender o impacto individual de cada variável, mas também conduzir uma redução de dimensionalidade baseada em importância preditiva. Isso fortaleceu o alinhamento entre performance estatística e transparência dos resultados, algo raro em estudos com modelos avançados.

D. Seleção de Dados e Representatividade Internacional

Os dados utilizados no estudo foram extraídos de fontes públicas internacionais de alta credibilidade, abrangendo dezenas de países e múltiplos anos. Essa amplitude elevou a representatividade e generalização dos modelos, permitindo avaliar a expectativa de vida com base em indicadores reais de desenvolvimento humano, infraestrutura sanitária e desempenho econômico. Contudo, limitações foram observadas quanto à disponibilidade desigual de dados entre países, o que pode influenciar tanto o aprendizado quanto a precisão em regiões com informações escassas. A adoção de imputações e score de propensão ajudou a mitigar essas lacunas, mas estudos futuros poderão explorar fontes complementares e técnicas mais avançadas de imputação multivariada.

Considerações Finais

Os resultados obtidos demonstram que o uso de modelos preditivos baseados em aprendizado de máquina — especialmente o TabPFN — apresenta grande potencial na estimativa da expectativa de vida, combinando acurácia preditiva com ferramentas interpretáveis. No entanto, para consolidar esse uso em ambientes críticos, são necessárias melhorias metodológicas estruturais:

- **Padronização dos protocolos de validação e comparação entre modelos**, com uso sistemático de métricas estatísticas acompanhadas de intervalos de confiança;
- **Fortalecimento da reprodutibilidade computacional**, por meio da inclusão de ambientes executáveis, scripts automatizados e documentação técnica clara;
- **Incorporação rotineira de interpretabilidade**, com o uso de ferramentas como SHAP, para aliar precisão estatística à transparência explicativa;
- **Ampliação das bases de dados utilizadas**, incluindo regiões sub-representadas, períodos mais longos e novas variáveis contextuais que possam enriquecer o poder explicativo dos modelos.

A experiência com o TabPFN neste projeto reforça que sua aplicabilidade vai além dos experimentos acadêmicos tradicionais, alcançando cenários reais onde a previsão da expectativa de vida pode apoiar decisões políticas, priorização de recursos e avaliação de impactos sociais. Este estudo, portanto, contribui como um marco inicial para a aplicação prática e metodologicamente rigorosa de modelos de ponta em temas críticos de interesse público.

E. Dados e Materiais da Pesquisa Dados, códigos e datasets disponíveis em:
<https://github.com/narashiuka/An-lise-Comparativa---Estimar-Expectativa-de-Vida>

Referências

[1] HELLI, Kai; SCHNURR, David; HOLLMANN, Noah; MÜLLER, Samuel; HUTTER, Frank. Drift resilient tabpfn: in-context learning temporal distribution shifts on tabular data. 2024. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0105000494217&partnerID=40&md5=3046751c2a5bed34f36ad5a744dc4079

[2] MAGADÁN, L.; ROLDÁN-GÓMEZ, J.; GRANDA, J.; SUÁREZ, F.. Early fault classification in rotating machinery with limited data using tabpfn. IEEE Sensors Journal, v. 23, n. 24, p. 30960 30970, 2023. DOI: 10.1109/JSEN.2023.3331100

[3] HASAN, M. R.; RAHMAN, S.; HOSSAIN, M. Z.; KRISHNA, A.; GEDEON, T. Tabular foundation model to detect empathy from visual cues. 2025. Disponível em: https://arxiv.org/abs/2504.10808. Acesso em: 7 jun. 2015.

[4] Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. 2023. TabPFN: A transformer that solves small tabular classification problems in a second. In Proceedings of the International Conference on Learning Representations (ICLR'23).

[5] Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. 2025. Accurate predictions on small data with a tabular foundation model. Nature 637, 8045 (2025), 319–326. doi:10.1038/s41586-024-08328-6.

[6] RODRÍGUEZ-PÉREZ, Raquel; BAJORATH, Jürgen. Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. Journal of Computer-Aided Molecular Design, v. 34, n. 10, p. 1013 – 1026, 2020. DOI: 10.1007/s10822-020-00314-0

[7] SALIH, Ahmed; RAISI-ESTABRAGH, Zahra; GALAZZO, Ilaria; RADEVA, Petia; PETERSEN, Steffen; LEKADIR, Karim; MENEGAZ, Gloria. A perspective on explainable artificial intelligence methods: shap and lime. Advanced Intelligent Systems, v. 7, n. 1, p. ,2025. DOI: 10.1002/aisy.202400304

[8] GHOLAMIANGONABADI, Davoud; KISELOV, Nikita; GROLINGER, Katarina. Deep neural networks for human activity recognition with wearable sensors: leave-one-subject-out cross-validation for model selection. IEEE Access, v. 8, n. , p. 133982-133994, 2020. DOI: 10.1109/ACCESS.2020.3010715

[9] HAMIDI, A.; MOHAMED-POUR, K.; YOUSEFI, M.. Forged channel: a breakthrough approach for accurate parkinson's disease classification using leave-one-subject-out cross-validation. In: 2024 32nd International Conference on Electrical Engineering (ICEE), 2024. p. 1-5. DOI: 10.1109/ICEE63041.2024.10667765

[10] FISHER, R.A. (1936), The Use Of Multiple Measurements In Taxonomic Problems. Annals of Eugenics, 7: 179-188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

[11] SMITH, William A.; RANDALL, Robert B. Rolling element bearing diagnostics using the Case Western Reserve University data. Journal of Vibration and Acoustics, v. 135, n. 4, 2013. DOI: 10.1115/1.4024332.

[12] FICO COMMUNITY. Explainable Machine Learning Challenge: HELOC Dataset. 2018. Disponível em: https://community.fico.com/s/explainable-machine-learning-challenge.

[13] WHITING, Penny F.; RUTJES, Anne W. S.; WESTWOOD, Marie E.; et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Annals of Internal Medicine, v. 155, n. 8, p. 529–536, 2011. DOI: 10.7326/0003-4819-155-8-201110180-00009.

[14] LIU, Xiaoxuan; RIVERA, Samuel C.; MOORTHY, Gowri; et al. STARD-AI: Guidelines for Reporting Artificial Intelligence-Based Diagnostic Accuracy Studies. Nature Medicine, v. 26, n. 6, p. 807–808, 2020. DOI: 10.1038/s41591-020-0941-1.

[15] CHEN, Tianqi; GUESTRIN, Carlos. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, p. 785–794.

[16] BREIMAN, Leo. Random Forests. Machine Learning, v. 45, n. 1, p. 5–32, 2001.

[17] KE, Guolin et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems (NeurIPS), 2017, p. 3149–3157.

[18] HOSMER, David W.; LEMESHOW, Stanley. Applied Logistic Regression. 2ª ed. Wiley, 2000.

[19] KUMAR, Aarsh. Life Expectancy (WHO). Kaggle. Disponível em: https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who. Acesso em: 20 jun. 2025.

[20] Braga, L. P. V. (2005). Introdução à Mineração de Dados. E-papers.

[21] Boulos, M. N. K., Peng, G., et al. (2019). An overview of GeoAI applications in health. International Journal of Health Geographics.

[22] Gupta, R. K., & Kumari, R. (2017). Artificial intelligence in public health: Opportunities and challenges. JK Science, 19(4), 191-192.

[23] Luger, G. F. (2009). Artificial Intelligence: Structures and Strategies for Complex Problem Solving. 6. ed. Pearson.

[24] Mbunge, E., et al. (2023). Application of deep learning and machine learning models to improve healthcare in sub-Saharan Africa. Alexandria Engineering Journal.

[25] Russell, S., & Norvig, P. (2013). Inteligência Artificial. 3. ed. Rio de Janeiro: Elsevier.

[26] Theodoridis, S., & Koutroumbas, K. (2009). Pattern Recognition. 4. ed. Academic Press.

[27] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2017). Data Mining: Practical Machine Learning Tools and Techniques. 4. ed. Morgan Kaufmann.