Além da Distância: Uma Abordagem Multidimensional para Avaliar a Acessibilidade à Saúde via Transporte Público

Cauê Rodrigues de Aguiar[1], Ana Carolina Xavier Castro[1], Euler da Silva Lima[1] Francisco Airton Silva[2], Geraldo Pereira Rocha Filho[1]

[1] Universidade Estadual do Sudoeste da Bahia (UESB)

[2]Universidade Federal do Piauí (UFPI)

cauaguiar@outlook.com.br, {anacarolina.acxc, euler410}@gmail.com, faps@ufpi.edu.br, geraldo.rocha@uesb.edu.br

Abstract. *This work proposes a multidimensional, machine learning-based approach to assess accessibility to healthcare services via public transportation in Salvador, Bahia, Brazil. Unlike traditional approaches, which are limited to geographic proximity, this proposal integrates aspects of transportation network connectivity, temporal reliability, and socioeconomic factors. To this end, a vulnerability score is constructed using Shannon entropy, synthesizing multiple dimensions of access and subsequently used as a target variable in supervised models, including Random Forest and Gradient Boosting. Furthermore, a temporal segmentation of neighborhoods is performed based on travel time to healthcare facilities. The results show that the quality and connectivity of the transportation system have a greater influence on accessibility than geographic distance, and reveal two distinct access profiles (fast and slow), associated with inequalities between central and peripheral regions.*

Resumo. *Este trabalho propõe uma abordagem multidimensional baseada em aprendizado de máquina para avaliar a acessibilidade a serviços de saúde via transporte público em Salvador/BA. Diferentemente de abordagens tradicionais, que se restringem à proximidade geográfica, a proposta incorpora, de forma integrada, aspectos de conectividade da rede de transporte, confiabilidade temporal e fatores socioeconômicos. Para isso, é construído um score de vulnerabilidade por meio da entropia de Shannon, que sintetiza múltiplas dimensões do acesso e é posteriormente utilizado como variável-alvo em modelos supervisionados, incluindo Random Forest e Gradient Boosting. Além disso, é realizada uma segmentação temporal dos bairros com base no tempo de deslocamento até unidades de saúde. Os resultados evidenciam que a qualidade e a conectividade do sistema de transporte possuem maior influência sobre a acessibilidade do que a distância geográfica, além de revelar dois perfis distintos de acesso (rápido e demorado), associados às desigualdades entre regiões centrais e periféricas.*

1. Introdução

O acesso à saúde é um direito fundamental [Brasil 1988] e um dos principais desafios nos grandes centros urbanos, especialmente em cidades marcadas por elevada desigualdade socioespacial. Para uma parcela significativa da população de baixa renda, o transporte

público constitui o principal meio de deslocamento até unidades do Sistema Único de Saúde (SUS), de modo que sua cobertura, qualidade e confiabilidade influenciam diretamente a capacidade dos cidadãos de acessar atendimento médico.

Salvador é a quinta cidade mais populosa do Brasil, com aproximadamente 2,5 milhões de habitantes segundo o Censo 2022 [IBGE 2022]. Assim como em outras grandes capitais brasileiras, a distribuição dos serviços de saúde e da rede de transporte público não é uniforme entre os bairros. No entanto, o grau em que essas assimetrias afetam o acesso ao SUS ainda é pouco quantificado em escala intra-urbana.

Apesar da relevância do tema, a avaliação da acessibilidade à saúde via transporte público ainda apresenta limitações metodológicas importantes. Abordagens tradicionais tendem a se restringir à proximidade geográfica entre a população e as unidades de saúde, desconsiderando aspectos fundamentais como a conectividade da rede de transporte, a confiabilidade temporal do serviço e as condições socioeconômicas dos territórios. Além disso, as dimensões geoespacial, temporal e demográfica são frequentemente tratadas de forma isolada, o que dificulta a captura da natureza multifatorial da vulnerabilidade no acesso à saúde [Sarker 2021].

Diversos estudos têm investigado a acessibilidade a serviços de saúde via transporte coletivo, combinando análise de rede e indicadores geoespaciais para identificar disparidades territoriais [Yuen et al. 2018, Liu et al. 2022, Tomasiello et al. 2024]. Embora esses trabalhos revelem padrões relevantes de desigualdade, a maioria ainda se concentra na proximidade geográfica, sem incorporar de forma integrada a conectividade da rede, a confiabilidade temporal e os fatores socioeconômicos. Abordagens mais recentes utilizam dados no formato *General Transit Feed Specification* (GTFS) e variáveis temporais de demanda [Fayyaz S. et al. 2017, Liu et al. 2023], porém continuam tratando essas dimensões de maneira isolada, sem integrá-las em *pipelines* de aprendizado de máquina. Nesse contexto, e considerando a escassez de estudos aplicados à cidade de Salvador [Instituto de Pesquisa Econômica Aplicada 2024], observa-se uma lacuna na literatura quanto ao uso de abordagens integradas que combinem *scoring* multidimensional, classificação supervisionada e segmentação temporal para avaliar a acessibilidade ao SUS em cenários urbanos com forte assimetria socioespacial.

Diante desse cenário, este trabalho propõe uma abordagem baseada em aprendizado de máquina para avaliar a acessibilidade a unidades de saúde pública via transporte coletivo em Salvador/BA, indo além da análise baseada exclusivamente em distância. A proposta integra cinco fontes de dados em um *dataset* em nível de bairro, constrói um *score* de vulnerabilidade por entropia de Shannon e aplica modelos supervisionados, incluindo modelos *baselines* (Regressão Logística (RL) e Árvore de Decisão (AD)) e métodos *ensemble* (Random Forest (RF) e Gradient Boosting (GB)), com separação entre variáveis utilizadas no *score* e variáveis preditoras. Adicionalmente, emprega-se *clustering* temporal para caracterizar perfis de deslocamento até unidades de saúde. Em conjunto, essas etapas permitem analisar, de forma integrada, a relação entre condições socioespaciais, estrutura da rede de transporte e padrões de acesso à saúde. No âmbito do índice de vulnerabilidade proposto, atributos de qualidade e conectividade do sistema de transporte mostraram-se mais informativos do que a proximidade geográfica às unidades de saúde, o que reforça a relevância de incorporar a dimensão da mobilidade em avaliações de acessibilidade.

2. Trabalhos Relacionados

A aplicação de técnicas de aprendizado de máquina a problemas urbanos consolida-se como campo interdisciplinar. Por exemplo, [Oliveira et al. 2024] investigaram padrões espaciais de criminalidade na cidade de Chicago via técnicas de aprendizado de máquina não-supervisionado, empregando Self-Organizing Maps, K-Means e DBSCAN sobre bases públicas de ocorrências. Os autores demonstraram que a combinação de algoritmos de agrupamento permite identificar correlações entre a distribuição geográfica dos crimes e fatores demográficos locais, subsidiando a formulação de políticas públicas de prevenção. De forma similar, o presente trabalho emprega *clustering* e classificação supervisionada sobre dados públicos urbanos para identificar padrões espaciais, neste caso aplicados à acessibilidade à saúde via transporte público em Salvador/BA.

Na acessibilidade a serviços de saúde, pesquisas recentes têm explicitamente integrado sistemas de transporte urbano para analisar as disparidades espaciais no acesso à assistência médica. [Yuen et al. 2018] investigaram a acessibilidade ao atendimento odontológico na cidade de São Paulo considerando a proximidade entre clínicas e infraestrutura de transporte público. Eles demonstraram que áreas de menor renda apresentam baixa disponibilidade de transporte frequente para acessar serviços de saúde. Sob a mesma ótica, [Liu et al. 2022] investigaram a acessibilidade espacial a hospitais em Chongqing, na China, combinando análise de rede e indicadores populacionais para identificar regiões com maior vulnerabilidade de acesso. Assim, esses estudos evidenciam que a integração entre localização de serviços de saúde e estrutura da rede de transporte é fundamental para compreender desigualdades territoriais na provisão de cuidados médicos, perspectiva que o presente trabalho estende ao incorporar, além da proximidade geográfica, indicadores de conectividade, confiabilidade temporal e fatores socioeconômicos.

Uma segunda linha de pesquisa utiliza dados de transporte público, especialmente no formato GTFS, para modelar redes de mobilidade urbana. [Fayyaz S. et al. 2017] propuseram algoritmos para análise dinâmica de acessibilidade, demonstrando que a variação temporal da operação é determinante para avaliações precisas. Paralelamente, [Liu et al. 2023] introduziram o conceito de acessibilidade realizável a partir de dados em tempo real, apontando que medidas baseadas apenas em horários programados superestimam o acesso devido a atrasos, o que motiva a incorporação de variáveis de confiabilidade temporal como o P90 de tempo de espera e o coeficiente de variação diário da demanda.

Embora os trabalhos citados abordem individualmente acessibilidade a serviços de saúde, uso de dados GTFS e aprendizado de máquina em contextos urbanos, nenhum deles propõe um *pipeline* integrado que articule *scoring* por entropia, classificação supervisionada e *clustering* temporal em nível de bairro, tampouco utiliza registros reais de bilhetagem eletrônica como fonte primária de demanda. O presente trabalho preenche essa lacuna ao integrar tais técnicas para avaliar a vulnerabilidade no acesso à saúde com granularidade intra-urbana.

3. Metodologia

Este estudo propõe uma abordagem computacional multidimensional para avaliar a acessibilidade a unidades de saúde pública via transporte coletivo, indo além de métricas baseadas apenas em proximidade geográfica. A metodologia integra cinco bases de dados em um *pipeline* capaz de capturar aspectos geoespaciais, temporais e socioeconômicos do

acesso em nível de bairro. O *pipeline* é composto por três etapas principais. Na primeira, constrói-se um *score* de vulnerabilidade por entropia de Shannon. Na segunda, são aplicados modelos supervisionados para identificar os fatores associados à vulnerabilidade. Por fim, na terceira etapa, realiza-se *clustering* temporal com base no tempo de deslocamento até unidades de saúde, permitindo a identificação de perfis distintos de acesso.

3.1. Fontes de Dados

O estudo integra cinco bases de dados, como apresentadas na Tabela 1. O *feed* GTFS, disponibilizado pelo *dataset* SUNT (*Salvador Urban Network Transportation*), descreve a rede de transporte público municipal por meio de arquivos padronizados contendo paradas, rotas, viagens e horários programados para ônibus convencional e BRT. O SUNT também registra individualmente cada transação de embarque (*boarding*) e desembarque (*alighting*) realizada no sistema de bilhetagem eletrônica, totalizando 205 arquivos de embarque e 221 de desembarque. O CNES (Cadastro Nacional de Estabelecimentos de Saúde), mantido pelo DATASUS/Ministério da Saúde, fornece localização, tipologia e caracterização das unidades de saúde. Os dados demográficos provêm do Censo IBGE, com população por bairro de 2022 e renda média domiciliar de 2010, última edição com desagregação por bairro disponível. Por fim, os Índices de Qualidade Urbana de Salvador (IQUA), elaborados pelo Projeto QUALISalvador, compreendem seis subíndices temáticos: físico-ambiental, infraestrutura e serviços urbanos, socioeconômico, bem-estar, cultura e participação política, e índice geral.

Tabela 1. Fontes de dados utilizadas na pesquisa.

Fonte	Conteúdo	Volume	Período
GTFS [Ferreira et al. 2025]	Paradas, rotas e horários (ônibus e BRT)	2.996 paradas 51.615 viagens	2023–2024
SUNT [Ferreira et al. 2025]	Registros de embarque e desembarque	426 arquivos	Mar/2024 – Mar/2025
CNES [DATASUS 2025]	Unidades de saúde cadastradas no SUS	268 unidades em 123 bairros	Dez/2025
Censo [IBGE 2022]	Dados demográficos	170 bairros	2022 (pop.), 2010 (renda)
IQUA [Santos et al. 2022]	Índices de qualidade urbano-ambiental	163 bairros, 6 sub-índices	2022

3.2. Tratamento dos Dados

Os dados do CNES foram submetidos a sete filtros sequenciais para reter apenas unidades públicas, ativas, de atendimento direto à população, localizadas em Salvador/BA: (i) estado gestor = Bahia; (ii) município gestor = Salvador; (iii) ausência de motivo de desabilitação; (iv) contrato formalizado com o SUS ou sem informação; (v) exclusão de tipos de estabelecimento não-assistenciais (centrais de gestão, farmácias, núcleos de telessaúde, serviço de verificação de óbito); (vi) exclusão de naturezas jurídicas privadas; e (vii) exclusão de unidades não-fixas ou administrativas. O *dataset* resultante contém 268 unidades de saúde distribuídas em 123 bairros.

As coordenadas geográficas de cada unidade foram obtidas via geocodificação pela API do Google Maps Geocoding. Seis tabelas auxiliares do CNES foram integradas por *join* para agregar descrições textuais de atividade, tipo de unidade, clientela e leitos.

O *feed* GTFS foi utilizado como fonte primária de informação sobre a rede de transporte público. O arquivo foi processado para extrair tempos de deslocamento programados entre paradas consecutivas de cada viagem, calculados como a diferença entre horários de chegada. Registros com tempos negativos ou superiores a 180 minutos foram descartados, e apenas pares origem-destino com pelo menos duas observações foram retidos, adotando-se a mediana como estimativa. Cada parada GTFS foi associada ao seu bairro censitário por geocodificação reversa via API do Google Maps e ambiguidades foram tratadas por regras heurísticas baseadas no nome das paradas.

Os dados de demanda foram obtidos do SUNT, compreendendo registros de embarque e desembarque. Já os dados demográficos do Censo 2022 forneceram população total, proporção de idosos e renda média domiciliar por bairro. Duplicatas geradas pelo mapeamento de micro-áreas do Censo foram resolvidas por agregação: para bairros com múltiplas entradas, população e área foram somadas, enquanto proporções e índices foram calculados pela média ponderada pela população. Os IQUA, compostos por seis sub-índices temáticos, também foram integrados ao *dataset* por bairro.

A integração dos *datasets* exigiu normalização dos nomes de bairros, que variam entre fontes. Um dicionário de mapeamento com aproximadamente 50 entradas foi construído manualmente, abrangendo três categorias, sendo elas as variações entre a API Google Maps e o Censo, divergências entre nomenclatura operacional do GTFS e o Censo, e inconsistências de separador nos bairros compostos do IQUA, para padronizar as variações. Bairros insulares (Ilha de Maré, Ilha dos Frades, Ilha de Bom Jesus dos Passos), sem transporte urbano terrestre regular, e bairros de uso institucional (Aeroporto, CAB e Porto Seco Pirajá) cuja ausência de moradores inviabiliza a avaliação de vulnerabilidade no acesso à saúde, foram excluídos da análise, assim como três bairros sem dados de transporte público suficientes para o cálculo das métricas de conectividade (Alto das Pombas, Calabar e Saramandaia). Dos 171 bairros oficiais de Salvador, 170 constam na base do Censo 2022 e após exclusões, 161 compõem o *dataset* final com 38 variáveis.

3.3. Engenharia de *Features*

A partir das cinco fontes de dados descritas, foram construídas variáveis em oito dimensões temáticas (Tabela 2). As métricas geoespaciais são calculadas por parada e agregadas por bairro via média ponderada pelos embarques reais. A acessibilidade gravitacional segue o modelo $G_i = \sum_j e^{-\beta\, d_{ij}}$, onde d_{ij} é a distância haversine em quilômetros e $\beta = 1{,}0\,\mathrm{km}^{-1}$ [Hansen 1959], valor validado por sensibilidade com $\beta \in \{0{,}3;\ 0{,}5;\ 1{,}5;\ 2{,}0\}$ (Spearman ρ entre $0{,}874$ e $0{,}972$), com maior estabilidade em valores intermediários. Para evitar *information leakage*, as variáveis são particionadas em dois conjuntos mutuamente exclusivos: (i) 12 indicadores usados no cálculo do *score* de vulnerabilidade e, portanto, na definição do *target*; e (ii) 21 variáveis candidatas a preditoras nos modelos RF e GB. Nenhuma variável participa de ambos os conjuntos; a correlação de Spearman entre eles indica baixa sobreposição geral (mediana $|\rho| = 0{,}141$, com apenas $4{,}5\%$ dos pares ultrapassando $|\rho| \geq 0{,}70$), embora a proximidade temática entre variáveis contribua parcialmente ao desempenho dos modelos. A seleção das candidatas supervisionadas ocorre em três etapas:

1. **Filtro de multicolinearidade:** pares com correlação de Pearson $|r| > 0{,}80$ são identificados e a variável de menor prioridade temática é removida. Nesta etapa foram excluídas quatro variáveis: usuários únicos ($r = 0{,}96$ com desembarques),

demanda total ($r = 0{,}99$), embarques por dia ($r = 0{,}98$) e proporção pico PM ($r = 0{,}87$ com proporção pico AM).

2. **RFECV (*Recursive Feature Elimination with Cross-Validation*):** eliminação iterativa com RF regularizado com profundidade máxima de 4 e pesos balanceados, StratifiedKFold ($k = 5$), F1-*weighted* e mínimo de 5 variáveis. O procedimento removeu quatro variáveis (presença de unidade local, densidade, razão fim de semana e P90 tempo de viagem), e atingiu melhor F1 = 0,797 com 9 variáveis; porém, como a diferença para as 13 restantes (F1 = 0,781) é inferior ao desvio-padrão (σ = 0,064), mantiveram-se todas para preservar a cobertura temática;
3. **Imputação:** valores ausentes nas variáveis selecionadas são tratados com KNNImputer ($k = 5$ vizinhos), nos dados de treino.

O processo resultou em 13 variáveis supervisionadas, abrangendo conectividade, demanda, padrões temporais e indicadores urbanos, com razão $n/p = 161/13 = 12{,}4$.

Tabela 2. Variáveis construídas por dimensão temática.

Dimensão	Variáveis	Fonte
Proximidade geoespacial	Distância mínima haversine até saúde, unidades em raios de 1 km e 2 km, acessibilidade gravitacional	CNES, GTFS
Conectividade GTFS	Total de conexões, frequência por conexão, bairros com saúde acessíveis, frequência a bairros com saúde, presença de unidade local	GTFS
Oferta de saúde	Quantidade de unidades no bairro, unidades por 1 000 habitantes	CNES, Censo
Demanda (bilhetagem)	Embarques, desembarques, usuários únicos, demanda total, embarques por dia	SUNT
Tempo de deslocamento	Tempo mínimo/mediano/amplitude GTFS até saúde, tempo médio de viagem e de espera, distância de caminhada em transferências	GTFS, SUNT
Padrões temporais	Proporção pico AM/PM, CV diário dos embarques, razão fim de semana, P90 tempo de viagem e espera, CV tempo de viagem	SUNT
Demográficas	População total, densidade, proporção de idosos, renda média	Censo
Qualidade urbana	Sub-índices: infraestrutura, socioeconômico, físico-ambiental, bem-estar, cultura e participação política	IQUA

3.4. *Score* de Vulnerabilidade

Para cada bairro foi calculado um *score* composto de vulnerabilidade no acesso a saúde, empregando o método de pesos por entropia de Shannon EWM (*Entropy Weight Method*) que foi adotado por derivar os pesos diretamente da dispersão observada nos dados, sem necessidade de julgamento subjetivo do pesquisador, o que é especialmente adequado em contextos multidimensionais onde a relevância relativa dos indicadores não é conhecida a priori. O *score* integra 12 indicadores organizados em quatro dimensões:

- **Geoespacial:** distância mínima à unidade de saúde mais próxima, quantidade de unidades em raio de 2 km e índice gravitacional;

- **Mobilidade:** frequência de conexões a bairros com saúde, bairros com saúde acessíveis por transporte direto, tempos médios de viagem e de espera, e distância média de caminhada em transferências entre linhas;
- **Demográfica:** proporção de idosos e renda média domiciliar;
- **Qualidade urbana:** índices IQUA de bem-estar e físico-ambiental.

Previamente ao cálculo, aplicou-se a transformação *shift-to-positive* canônica ($x'_{ij} = x_{ij} - \min_i x_{ij} + \epsilon$) para garantir valores estritamente positivos. Para harmonizar a polaridade, indicadores de benefício (e.g., renda, frequência, índices IQUA) tiveram o sinal invertido após padronização por *z-score*, de modo que valores maiores indicam maior vulnerabilidade em todos os indicadores. A formulação do EWM segue três etapas. Primeiro, cada indicador j é normalizado por proporção:

$$p_{ij} = \frac{x'_{ij}}{\sum_{i=1}^{n} x'_{ij}} \tag{1}$$

Em seguida, calcula-se a entropia de cada indicador:

$$E_j = -\frac{1}{\ln n} \sum_{i=1}^{n} p_{ij} \ln(p_{ij}) \tag{2}$$

onde $n = 161$ (número de bairros) e $m = 12$ (número de indicadores). Indicadores com menor entropia apresentam maior dispersão entre bairros e, portanto, maior poder discriminante. O peso de cada indicador é obtido pela divergência normalizada:

$$W_j = \frac{1 - E_j}{\sum_{k=1}^{m} (1 - E_k)} \tag{3}$$

O *score* final por bairro é dado por $S_i = \sum_{j=1}^{m} W_j \cdot \bar{x}_{ij}$, onde $\bar{x}_{ij}$ é o *z-score* orientado do indicador j no bairro i ($x_{ij} \to x'_{ij} \to \bar{x}_{ij}$). O resultado é normalizado em $[0, 1]$ via *min-max*, onde 1 indica máxima vulnerabilidade. O *score* contínuo foi discretizado em três classes, BAIXA, MODERADA e CRÍTICA, pelo método Natural Breaks, que identifica descontinuidades naturais na distribuição minimizando a variância intraclasse e maximizando a separação interclasses. Essa classificação define a variável-alvo da etapa supervisionada.

3.5. Classificação Supervisionada

A etapa de classificação supervisionada visa predizer a categoria de vulnerabilidade de acesso à saúde de cada bairro a partir de variáveis que não participam do cálculo do *score*. Quatro modelos são empregados: RL, AD, RF e GB. Os dois primeiros são utilizados como modelos *baselines* interpretáveis, permitindo uma referência de desempenho e análise comparativa. Já RF e GB se justificam pela capacidade de lidar com relações não-lineares, variáveis de escalas distintas e amostras reduzidas, além de oferecerem estimativas de importância de variáveis nativamente. O desbalanceamento entre classes é tratado com pesos de classe ajustados automaticamente, que ponderam cada classe inversamente à sua frequência na amostra.

A otimização de hiperparâmetros é feita via GridSearchCV com StratifiedKFold ($k = 5$) e F1-*weighted*, conforme apresentado na Tabela 3. A busca exaustiva foi preferida à aleatorizada por ser viável dado o espaço de hiperparâmetros, garantindo que a configuração ótima seja identificada sem risco de amostragem incompleta.

Tabela 3. Espaços de busca do GridSearchCV por modelo.

Hiperparâmetro	RL	AD	RF	GB
C	{1e-4, …, 150} (30 vals.)	—	—	—
penalty	{l2, l1, elasticnet}	—	—	—
solver	{lbfgs, newton-cg, sag, saga}	—	—	—
l1_ratio	{0.05, …, 0.95} (13 vals.)	—	—	—
criterion	—	{gini, entropy}	—	—
max_depth	—	{2, 3, 4, 5, 7, None}	{2, 3, 4, 5, None}	{1, 2, 3}
min_samples_leaf	—	{1, 3, 5, 8, 12}	{3, 5, 8, 12}	{5, 8, 12}
min_samples_split	—	{2, 5, 10, 15}	{5, 10, 15}	{5, 10}
max_features	—	{sqrt, log2, None}	{sqrt, log2, None}	{sqrt, None}
n_estimators	—	—	{50, 100, 200}	{100, 200, 500}
learning_rate	—	—	—	{0.01, 0.05, 0.1}
subsample	—	—	—	{0.5, 0.7, 0.8}
Combinações	540	720	540	972

A avaliação dos modelos ainda emprega RepeatedStratifiedKFold com 5 *folds* e 3 repetições (15 avaliações), com o F1-*weighted* como métrica principal, reduzindo a variância da estimativa. Adicionalmente, reportam-se a diferença entre F1 de treino e teste como diagnóstico de *overfitting* e a matriz de confusão das predições *out-of-fold*.

3.6. *Clustering* Temporal

A segmentação dos bairros segundo o tempo de deslocamento até unidades de saúde via transporte público foi operacionalizada por meio de *clustering* sobre variáveis derivadas exclusivamente do GTFS. Três variáveis foram computadas a partir da matriz de tempos programados entre pares de bairros: (i) tempo mínimo agendado até o bairro mais próximo com unidade de saúde; (ii) tempo mediano entre todos os pares bairro-origem até bairro-com-saúde acessíveis via GTFS; e (iii) amplitude do tempo de viagem, que captura a variabilidade da oferta de rotas. A escolha de variáveis exclusivamente temporais e específicas ao acesso à saúde seguiu recomendações recentes na literatura [Liu et al. 2023, Lindner et al. 2024]. Dentre nove combinações candidatas avaliadas, incluindo indicadores de frequência e conectividade, esta obteve o melhor desempenho e foi selecionada para a análise.

A variável de tempo mínimo apresentou assimetria positiva (*skewness* = $1{,}68$), violando a premissa de esfericidade assumida pelo K-Means. As demais variáveis temporais (mediana e amplitude) foram avaliadas pelo mesmo critério e não necessitaram de transformação (*skewness* de $-0{,}27$ e $0{,}21$, respectivamente). Aplicou-se o PowerTransformer com o método de Yeo-Johnson [Yeo and Johnson 2000], reduzindo a assimetria para $0{,}25$. Os pesos das variáveis foram otimizados por busca em grade para maximizar a separação entre *clusters*. O valor ótimo de K foi determinado por consenso de cinco métricas de validação, avaliadas para $K \in \{2, 3, 4, 5, 6\}$: Silhouette [Rousseeuw 1987], Davies-Bouldin [Davies and Bouldin 1979], Calinski-Harabasz [Caliński and Harabasz 1974], *Gap Statistic* [Tibshirani et al. 2001] e Bootstrap ARI [Hennig 2007] com $B = 50$ reamostragens. Optamos pela combinação de múltiplas

métricas dado que nenhuma é universalmente superior [Arbelaitz et al. 2013] e, para cada métrica, os valores de K foram ranqueados e o K com menor *rank* médio foi selecionado.

4. Resultados e Discussão

4.1. *Score* de Vulnerabilidade

A renda média domiciliar teve maior peso (0,215), seguida da frequência de serviços de saúde (0,164) e distância mínima à unidade mais próxima (0,120), indicando vulnerabilidade de acesso decorrente de fatores socioeconômicos e oferta geográfica de saúde. A discretização de Natural Breaks resultou nas classes: BAIXA ($n = 32$; 19,9%), MODERADA ($n = 104$; 64,6%) e CRÍTICA ($n = 25$; 15,5%). A maioria dos bairros concentra-se na classe intermediária, com distribuição gradual da vulnerabilidade (Figura 1).

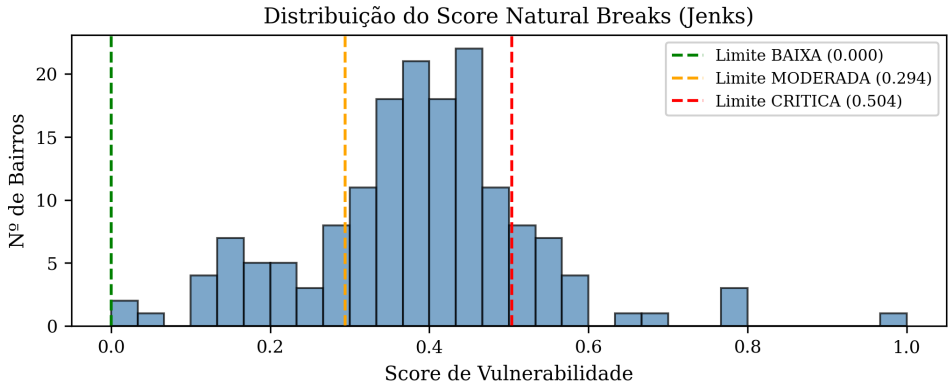

Figura 1. Distribuição do Score de Vulnerabilidade com limites Natural Breaks.

Foi realizada análise de sensibilidade comparando o EWM com pesos uniformes, PCA e AHP. O ranking territorial mostrou alta estabilidade entre métodos (Spearman ρ entre 0,930 e 0,963), indicando robustez da ordenação. A concordância de classificação variou de 57,1% a 64,0%; contudo, 58,0% dos bairros discordantes situam-se a menos de 5% da fronteira Jenks mais próxima vs. 26,4% dos concordantes ($p = 7{,}2 \times 10^{-21}$), sugerindo que a instabilidade é concentrada em casos limítrofes.

4.2. Classificação Supervisionada

A Tabela 4 resume o desempenho dos modelos. O RF obteve o melhor resultado, com F1-*weighted* de 0,781 na validação por RepeatedStratifiedKFold (5×3). O GB atingiu F1 de 0,765. O RF tem *gap* treino-teste de 0,163, indicando boa generalização, enquanto o GB apresenta *gap* de 0,24, indicando *overfitting* moderado, esperado para $n = 161$ com 13 variáveis preditoras, contexto em que *gaps* desse nível são comuns em *datasets* pequenos.

Tabela 4. Desempenho dos modelos supervisionados.

	GridSearchCV		RepeatedCV (5×3)	
Modelo	F1	*Gap*	F1	*Gap*
Regressão Logística	0,756 ± 0,050	0,049	0,727 ± 0,067	0,078
Árvore de Decisão	0,722 ± 0,075	0,091	0,688 ± 0,065	0,142
Random Forest	**0,811** ± 0,083	0,130	**0,781** ± 0,081	0,163
Gradient Boosting	0,809 ± 0,050	0,191	0,765 ± 0,076	0,236

Além dos modelos principais (RF e GB), são incluídos *baselines* (RL e AD) para análise comparativa; a AD obteve o menor desempenho geral, enquanto a RL, *baseline* mais competitivo, foi adotada como referência para avaliar o ganho dos *ensemble*. Embora a RL apresente menor *gap* treino-teste, o RF supera a RL em todas as métricas globais: F1-*weighted* de 0,781 vs. 0,727 (+0,055), macro-F1 de 0,724 vs. 0,691 e acurácia de 0,789 vs. 0,714, com ganhos concentrados nas classes BAIXA e MODERADA.

As matrizes de confusão (Figura 2), reportadas para RF e GB, principais da análise, mostram que ambos apresentam mínima confusão entre extremos (BAIXA ↔ CRÍTICA): todos os erros são ordinais, isto é, ocorrem entre classes adjacentes, indicando que as fronteiras entre categorias são graduais, coerentes com a natureza contínua do *score* subjacente. O RF atingiu acurácia de 80,7% (130/161), com principal confusão concentrada nos 11 bairros MODERADA classificados como CRÍTICA e nos 9 bairros CRÍTICA classificados como MODERADA, classes que compartilham intervalos próximos do *score*.

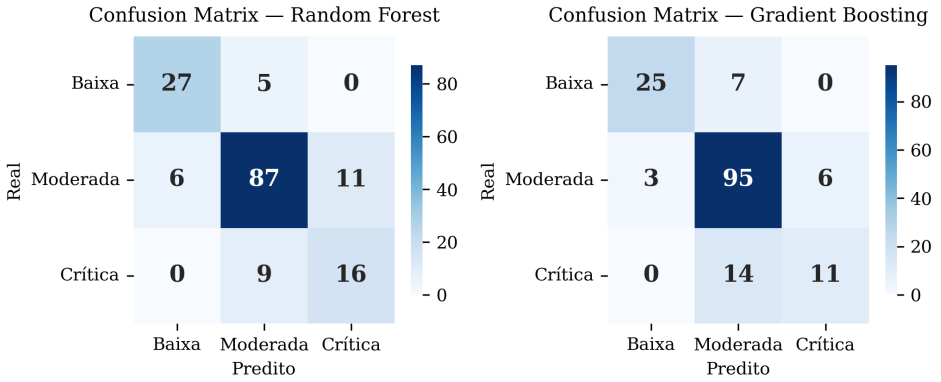

Figura 2. Matrizes de confusão (predições *out-of-fold*).

A análise de importância das variáveis (Figura 3) indica que o total de conexões e a infraestrutura urbana são as mais relevantes, seguidas por frequência de conexões, P90 de tempo de espera, CV do tempo de viagem e desembarques.

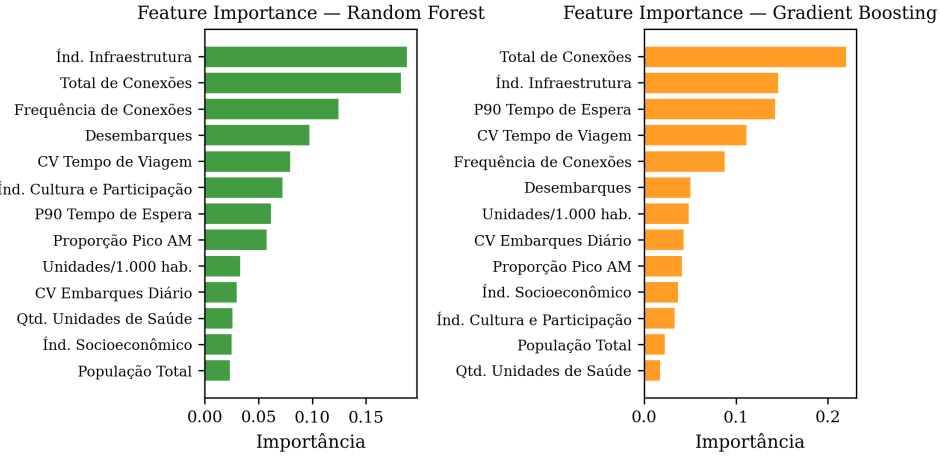

Figura 3. Importância das variáveis para RF e GB.

Cinco das seis principais variáveis estão ligadas à conectividade, confiabilidade temporal e uso do transporte público, sugerindo que a vulnerabilidade de acesso à saúde

é mais determinada pela qualidade do sistema de transporte do que pela proximidade geográfica às unidades. A curva de aprendizado indica saturação do desempenho com aproximadamente 80 amostras de treino (F1 $\approx 0{,}77$), sugerindo limitação de ganhos com $n = 161$.

A Figura 4 apresenta a distribuição espacial das classes sobrepostas às rotas de transporte público. Observa-se que os bairros CRÍTICA tendem a se localizar nas periferias norte e oeste da cidade, onde a malha de transporte é visivelmente mais rarefeita.

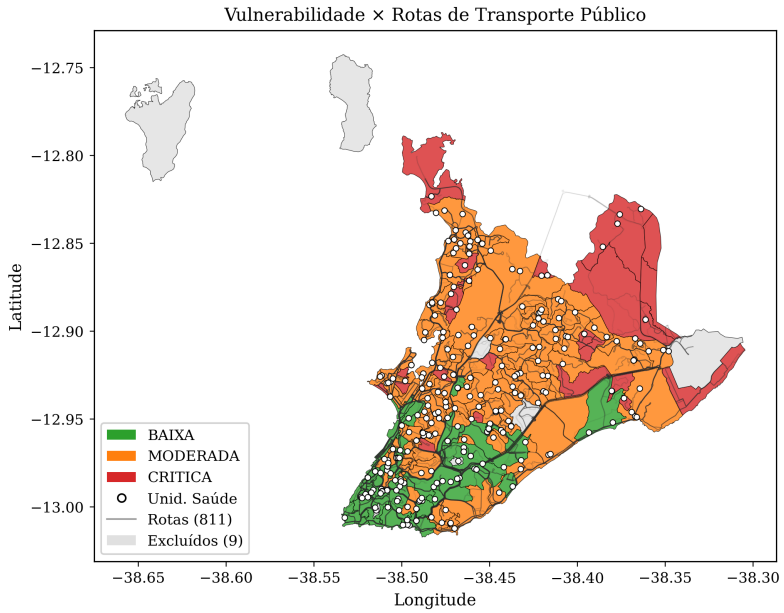

Figura 4. Vulnerabilidade por bairro com rotas de transporte público (cinza) e unidades de saúde (pontos brancos).

4.3. *Clustering* Temporal

O consenso das cinco métricas selecionou $K = 2$ como número ótimo de *clusters* (Tabela 5), por apresentar menor *rank* médio (1,40) e superior em quatro das cinco métricas, enquanto valores crescentes de Davies-Bouldin e decrescentes de Calinski-Harabasz para $K \geq 3$ indicam sobreposição entre grupos sem ganho interpretativo. A solução gerou dois perfis de acesso à saúde: RÁPIDO ($n = 82$; 50,9%) e DEMORADO ($n = 79$; 49,1%).

Tabela 5. Validação do *clustering* temporal (K-Means).

Métrica	$K = 2$	$K = 3$	$K = 4$
Silhouette (↑)	**0,623**	0,456	0,325
Davies-Bouldin (↓)	**0,537**	0,935	1,149
Calinski-Harabasz (↑)	**452,8**	314,7	274,9
Gap Statistic (↑)	0,944	0,984	**1,023**
Bootstrap ARI (↑)	**0,981**	0,890	0,830
Consenso (*avg rank* ↓)	**1,40**	2,00	2,60

O Bootstrap ARI de 0,981 indica boa estabilidade. O grupo RÁPIDO tem tempo mínimo até saúde de 0,13 min ($\pm$ 0,09) e maior amplitude de rotas (38,42 min $\pm$ 7,88),

contra o tempo mínimo $16\times$ superior (2,06 min $\pm$ 1,29) e menor variabilidade (32,43 min $\pm$ 7,21) no grupo DEMORADO, sugerindo oferta concentrada em poucas conexões.

A Figura 5 evidencia que os bairros DEMORADO concentram-se nas regiões periféricas da cidade, onde a malha de transporte é menos densa, padrão coerente com a distribuição da vulnerabilidade.

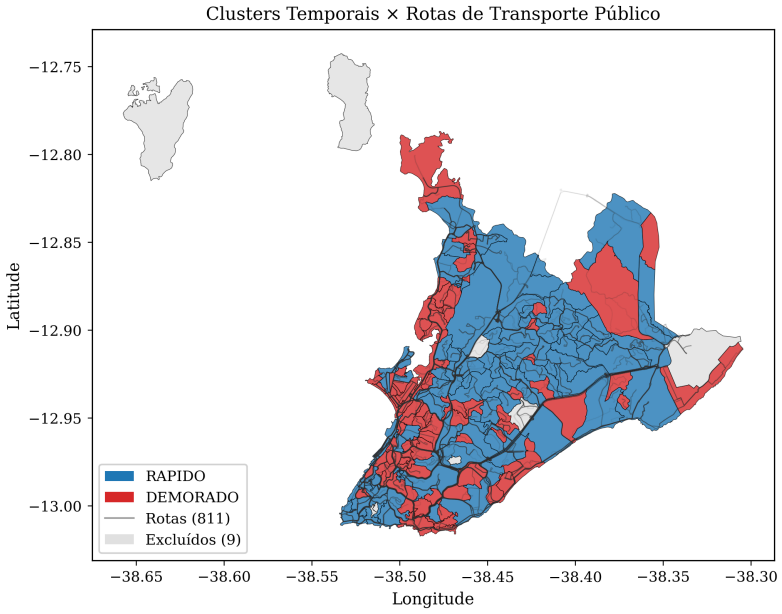

Figura 5. *Clusters* temporais com rotas de transporte público.

4.4. Vulnerabilidade e Tempo de Acesso

A Figura 6 cruza os resultados dos dois eixos de análise. A associação mais frequente (bairros CRÍTICA no *cluster* DEMORADO sendo 14 de 25 (56%)) evidencia coerência entre vulnerabilidade socioespacial e dificuldade temporal de acesso, ou seja, bairros periféricos com menor renda e infraestrutura mais precária tendem também a apresentar os maiores tempos de deslocamento até unidades de saúde. Esse padrão é consistente com o gradiente centro-periferia do *score* e com as desigualdades territoriais descritas por [Tomasiello et al. 2024] para cidades brasileiras.

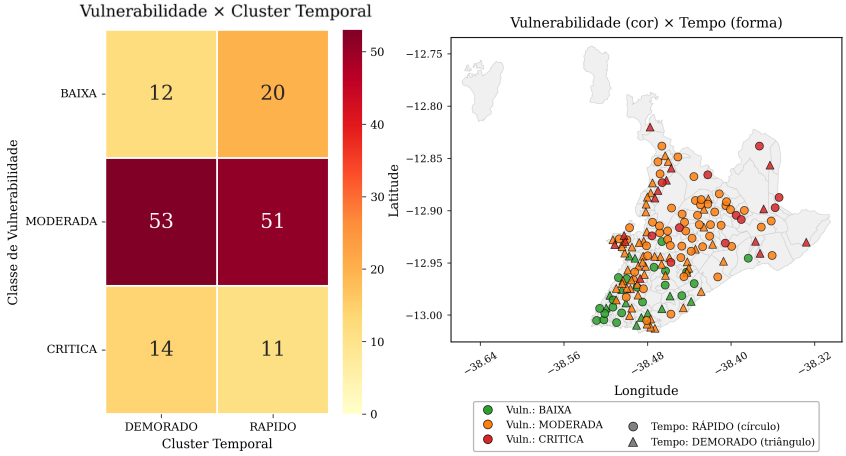

Figura 6. Cruzamento entre vulnerabilidade (cor) e *cluster* temporal (forma).

Contudo, a relação não é determinística: onze bairros CRÍTICA integram o *cluster* RÁPIDO, com alta vulnerabilidade apesar da proximidade temporal às unidades de saúde. Esse grupo requer atenção especial do ponto de vista de política pública, pois a barreira não é o tempo de viagem, mas fatores estruturais como renda baixa, infraestrutura precária e menor bem-estar, que inibem o uso efetivo do serviço mesmo quando o transporte conecta o bairro à unidade em tempo hábil. Essa multidimensionalidade é consistente com o modelo de [Geurs and van Wee 2004], que distingue quatro componentes de acessibilidade (uso do solo, transporte, temporal e individual), e com os determinantes sociais da saúde sistematizados por [Marmot et al. 2008]. Na direção oposta, 20 bairros BAIXA vulnerabilidade integram o *cluster* RÁPIDO, confirmando que bairros centrais acumulam vantagens socioespaciais e temporais, aprofundando assimetrias de acesso urbano.

5. Conclusão e Trabalhos Futuros

Este estudo investigou a acessibilidade a serviços de saúde via transporte público em Salvador/BA com uma abordagem multidimensional baseada em aprendizado de máquina. Ao integrar dados de mobilidade urbana, infraestrutura de transporte e indicadores socioeconômicos, permitiu ir além da análise tradicional baseada exclusivamente em proximidade geográfica, indicando a natureza multifatorial do acesso à saúde em meios urbanos.

Os resultados mostram que cinco das seis variáveis mais importantes nos modelos supervisionados ligam-se a conectividade, confiabilidade temporal e uso do transporte público, indicando que a qualidade da rede influencia a acessibilidade mais que a distância às unidades de saúde. A análise temporal gerou dois perfis de acesso (rápido e demorado), associados a diferenças estruturais entre regiões centrais e periféricas. A combinação do *score* de vulnerabilidade e os *clusters* temporais indica que o acesso à saúde é condicionado não apenas pelo tempo de deslocamento, mas por fatores socioeconômicos e de infraestrutura urbana, sugerindo a adoção de políticas públicas diferenciadas.

Como trabalhos futuros, destaca-se a incorporação de dados de mobilidade em tempo real, como informações de GPS da frota, visando reduzir a discrepância entre tempos programados e observados. Além disso, a aplicação da metodologia em outras capitais brasileiras constitui um passo importante para avaliar sua generalização e ampliar sua contribuição para o planejamento integrado de transporte e saúde pública.

Referências

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., and Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46(1):243–256.

Brasil (1988). Constituição da república federativa do brasil. Promulgada em 5 de outubro de 1988.

Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics – Theory and Methods*, 3(1):1–27.

DATASUS (2025). Cadastro nacional de estabelecimentos de saúde – CNES. Competência dezembro/2025. Disponível em: `https://cnes.datasus.gov.br/`. Acesso em: jan. 2026.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227.

Fayyaz S., S. K., Liu, X. C., and Zhang, G. (2017). An efficient general transit feed specification (gtfs) enabled algorithm for dynamic transit accessibility analysis. *PLOS ONE*, 12(10):e0185333.

Ferreira, M. V., Souza, M., Rios, T. N., Fernandes, I. F. C., Nery, J., Gama, J., Bifet, A., and Rios, R. A. (2025). Salvador urban network transportation (sunt): A landmark spatiotemporal dataset for public transportation. *Scientific Data*, 12(1):1320.

Geurs, K. T. and van Wee, B. (2004). Accessibility evaluation of land-use and transport strategies: review and research directions. *Journal of Transport Geography*, 12(2):127–140.

Hansen, W. G. (1959). How accessibility shapes land use. *Journal of the American Institute of Planners*, 25(2):73–76.

Hennig, C. (2007). Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis*, 52(1):258–271.

IBGE (2022). Censo demográfico 2022: resultados por bairro. Disponível via WFS em: `https://dservices6.arcgis.com/GP5qdNaePRPh2SdT/arcgis/services/censo_2010_e_2022_por_bairro/WFSServer`. Acesso em: jan. 2026.

Instituto de Pesquisa Econômica Aplicada, I. (2024). Projeto acesso a oportunidades. Acesso em: 11 mar. 2024.

Lindner, A., Kühnel, N., Schrömbges, T., and Kuhnimhof, T. (2024). When to measure accessibility? Temporal segmentation and aggregation in location-based public transit accessibility. *Urban Science*, 8(4):165.

Liu, L., Porr, A., and Miller, H. J. (2023). Realizable accessibility: Evaluating the reliability of public transit accessibility using high-resolution real-time data. *Journal of Geographical Systems*, 25:429–451.

Liu, Y., Gu, H., and Shi, Y. (2022). Spatial accessibility analysis of medical facilities based on public transportation networks. *International Journal of Environmental Research and Public Health*, 19(23):16224.

Marmot, M., Friel, S., Bell, R., Houweling, T. A. J., and Taylor, S. (2008). Closing the gap in a generation: health equity through action on the social determinants of health. *The Lancet*, 372(9650):1661–1669.

Oliveira, E. A. d., Gomes, G. L., and Cunha, F. D. d. (2024). Aprendizado de máquina aplicado ao cenário de criminalidade na cidade de chicago. In *Anais do VIII Workshop de Computação Urbana (CoUrb)*, pages 1–14, Niterói/RJ. SBC.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Santos, E., Benevides, T., Borja, P. C., Moraes, L. R. S., De Oliveira, N., Pedrassoli, J. C., Souza, J., Gama, C. M., and Fróes, F. (2022). QUALISalvador: qualidade do ambiente urbano na cidade da Bahia.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2(3):160.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B*, 63(2):411–423.

Tomasiello, D. B., Pereira, R. H. M., Braga, C. K. V., and van Wee, B. (2024). Racial and income inequalities in access to healthcare in Brazilian cities. *Journal of Transport & Health*, 34:101722.

Yeo, I.-K. and Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959.

Yuen, A., Rocha, C. M., Kruger, E., and Tennant, M. (2018). Does public transportation improve the accessibility of primary dental care in são paulo, brazil? *Community Dentistry and Oral Epidemiology*, 46(3):273–278.