

Modelagem preditiva da malária em crianças de Uganda com árvores de decisão interpretáveis, balanceamento de dados e agrupamento

Cleiane Gonçalves Oliveira
Prog. de Pós-Grad. em Ciências da Saúde
UNIMONTES
Januária, Brasil
cleiane.oliveira@ifnmg.edu.br

Marcos Flávio S. V. D'angelo
Dep. de Ciência da Computação
UNIMONTES
Montes Claros, Brasil
marcos.dangelo@unimontes.br

João B. Mendes
Dep. de Ciência da Computação
UNIMONTES
Montes Claros, Brasil
joao.mendes@unimontes.br

Marcelo Perim Baldo
Dep. de Fisiopatologia
UNIMONTES
Montes Claros, Brasil
marcelo.baldo@unimontes.br

Hasheem Mannan
School of Nursing, Midwifery and Health Systems
University College Dublin
Dublin, Ireland
hasheem.mannan@ucd.ie

Abstract—A malária permanece como um dos principais desafios de saúde pública em países africanos, como Uganda. Este estudo propõe e analisa diferentes abordagens de aprendizagem de máquina para a predição de malária em crianças menores de cinco anos, utilizando dados sociodemográficos do Uganda Malaria Indicator Survey 2018-2019. Foram aplicados modelos de árvores de decisão com diferentes profundidades, avaliando-se também o impacto de técnicas de balanceamento de dados (undersampling e oversampling). Além disso, propõe-se uma nova abordagem que combina agrupamento de indivíduos por similaridade com a aplicação de árvores de decisão específicas para cada grupo. Os resultados mostram que técnicas de balanceamento melhoram o desempenho na detecção de casos positivos, especialmente com árvores mais rasas, e que o agrupamento possibilita a identificação de perfis distintos da população. Conclui-se que dados sociodemográficos tem potencial de contribuição na predição de malária, e que a interpretabilidade e o balanceamento dos modelos são aspectos essenciais para aplicações em saúde pública.

Index Terms—Malária, aprendizagem de máquina, dados sociodemográficos.

I. INTRODUÇÃO

A malária continua a ser uma das principais causas de mortalidade e morbidade em todo o mundo. Até 2022 foram registrados cerca de 249 milhões de casos de malária em 85 países onde a malária é endêmica. A região do continente africano soma 95,4% das mortes a nível mundial, sendo que aproximadamente 78,1% foram crianças com menos de 5 anos de idade [1].

Uganda está entre os países africanos responsáveis por quase metade de todos os casos mundiais de malária. Os poderes públicos já implementaram várias intervenções que incluem mosquiteiros tratados com inseticidas de longa duração, testes rápidos e pulverização residual, no sentido de tentar controlar e eliminar este problema de saúde [2]. Entretanto, há

gargalos nas ações por conta dos custos de testes rápidos, a necessidade de pessoal treinado para sua aplicação e dificuldade de disponibilizar tais recursos em zonas rurais e remotas, onde residem a maior parte das crianças com dados coletados. Neste contexto, há a necessidade de outras ferramentas eficazes para intervir no diagnóstico e prevenção [3].

A área de aprendizagem de máquina é uma abordagem que tem se mostrado eficaz na tomada de decisões e previsões com base na grande quantidade de dados produzidos pelo setor de saúde [3]. Em relação à predição de malária muitos trabalhos têm sido realizados como a aplicação de redes bayesianas em dados de unidades de saúde [4]; comparação da eficiência de algoritmos de aprendizagem de máquina em dados de sintomas e características dos pacientes [5]; uso de modelos de aprendizagem profunda após a clusterização de cidades do estado do Amazonas, no Brasil [6]. Há muitos outros trabalhos que aplicam algoritmos de aprendizagem de máquina avaliando variáveis diversas como a partir de imagens de exames de sangue para identificação do vírus [7], ou que consideram variáveis climáticas para predizer surtos [8].

Outra abordagem que vem sendo explorada é a utilização de dados sociodemográficos para a predição de doenças [3]–[5]. Tais dados permitem identificar quais fatores externos aos sintomas da doença favorecem o seu surgimento. Tal informação possibilita a identificação de importantes fatores que estão frequentemente presentes no surgimento da doença, permitindo a definição de políticas públicas para a prevenção e o enfrentamento de surtos.

Em 2018 e 2019 foi realizada a terceira (e mais recente) Pesquisa de Indicadores de Malária (*Malaria Indicator Survey* - MIS) em Uganda. O MIS faz parte do programa *Demographic Health Surveys* (DHS) que promove assistência técnica para mais de 90 países a partir de pesquisas de larga

escala com o objetivo de promover a compreensão global das tendências da saúde e da população nos países em desenvolvimento [9]. O questionário MIS têm muitos pontos fortes que apoiam a tomada de decisões em saúde. São coletados dados sócio-demográficos dos moradores das residências (gênero, idade), características das residências (como o principal material das paredes, chão e telhado), e algumas dinâmicas do domicílio (principal fonte de água, posses de animais, posses de terras e eletrodomésticos). Para as crianças menores de 5 anos também são coletados resultados de testes de malária e anemia. A base de dados de Uganda conta com o registro de 45767 indivíduos, sendo 7780 crianças com o resultado para o teste de malária [10].

Neste cenário, a aplicação de técnicas de aprendizagem de máquina se torna bastante favorável para a predição de casos de malária. Entretanto, tais técnicas enfrentam alguns desafios. Um deles é a interpretabilidade do modelo gerado. Na área da saúde modelos de aprendizagem de máquina mais interpretáveis são desejáveis. É necessário que os envolvidos na tomada de decisão entendam as variáveis que o modelo considera para classificar um paciente com determinado diagnóstico. Um dos modelos amplamente utilizados é a árvore de decisão que pode ser definida como uma família de métodos estatísticos particularmente adequada para explorar relações potencialmente não lineares entre as variáveis de análise [11]. O modelo de árvore de decisão implica uma regra de predição que define subconjuntos dos dados de forma hierárquica, o que simplifica a interpretação do modelo gerado. Esta técnica é amplamente usada na área da saúde: na análise do uso de árvore de decisão na saúde mental por apresentarem uma abordagem valiosa para a tomada de decisões clínicas devido à sua interpretabilidade, capacidade de identificar pacientes em risco e apoio no planejamento do tratamento [12]; aplicação na área da vigilância da saúde [13]; e comparação de tratamentos a partir da construção de árvores de decisão [14].

Outro desafio enfrentado pelas técnicas de aprendizagem de máquina é o desbalanceamento de dados. Nos dados das 7780 crianças de Uganda, 88,65% dessas crianças não apresentaram resultado positivo para malária, e os restantes 11,35% estavam positivas para malária no momento da coleta de dados. O desbalanceamento dos dados, onde o número de amostras não é igualmente distribuído nas classes, é um fator que dificulta o aprendizado dos modelos. Para isso existem diversas técnicas para balanceamento de dados que tentam, ou criar novos dados, ou reduzir os dados para tentar igualar o número de amostras entre as classes. Essas técnicas colaboram também para diminuir o viés que a classe majoritária pode imputar ao modelo.

Outra discussão é a generalização do modelo. Quando se gera um único modelo para toda uma população características específicas de grupos menores e específicos podem ser perdidas. Assim, técnicas de agrupamento também são aplicadas em conjunto com modelos de aprendizagem de máquina em saúde, inclusive na intervenção contra a malária. Há a tentativa de agrupar os casos de malária por cidades [6], o agrupamento de dados temporais para vigilância e planejamento de

intervensões [15], e a criação de grupos a partir das variáveis de pacientes infectados de forma que aquelas semelhantes à infecção por malária sejam identificadas [16]. Há também o uso de agrupamentos como uma das etapas para técnica de balanceamento de dados [17], [18].

Neste trabalho analisamos o desempenho do modelo interpretável de árvore de decisão para predição de malária em crianças abaixo de 5 anos de Uganda. É realizada a comparação da eficiência do modelo aplicando técnicas de balanceamento e é apresentada uma nova proposta que combina técnicas de agrupamento com árvore de decisão para identificação de diferentes perfis na população estudada. Assim, tem-se como objetivo deste trabalho analisar se os dados socioeconômicos colaboram para a predição de casos de malária considerando a interpretabilidade, o balanceamento dos dados e o agrupamento dos indivíduos.

II. MATERIAIS E MÉTODOS

Todos os códigos realizados foram implementados em linguagem Python com uso das bibliotecas *Scikit-learn*, *Pandas*, *Numpy*, *Statistics* e *Imblearn*.

A. Seleção dos Dados MIS Uganda 2018-19

Os dados usados neste trabalho foram obtidos do DHS Program [19]. Dos questionários disponíveis foi utilizado o módulo *Household Recode*. Este por sua vez é subdividido em *Household's Basic Data*, *Household Characteristics*, *Household Schedule*, *Children's height/weight/hemoglobin variables*, *Malaria Module for Nets Variables*, *Malaria for Household Members*. Uma vez que só há resultados dos testes de malária para crianças menores de 5 anos, e sendo também este o público com maior mortalidade, foram selecionadas somente essas amostras. O conjunto de dados lista 8957 domicílios totalizando 45767 indivíduos, sendo 8125 crianças. Porém, somente 7780 possuem resultado para o teste de malária, sendo este o conjunto de dados utilizado.

Para cada criança foram relacionados os demais dados disponíveis sobre seu domicílio. Das dezenas de dados disponíveis foram selecionados os listados a seguir, obtendo a quantidade de 70 colunas iniciais. O subconjunto *Malaria Module for Nets Variables* não foi utilizado uma vez que os dados apresentados já são contemplados nos outros subconjuntos.

Household's Basic Data - 4 colunas - Número total de membros no domicílio; Número de crianças residentes com menos de 5 anos; Região de residência; Se a residência está em região urbana, rural ou assentamento de refugiados.

Household Characteristics - 53 colunas - Principal fonte de água para beber; Tipo de instalação sanitária; Se tem: eletricidade; rádio; televisão; refrigerador; bicicleta; motocicleta; carro; telefone; telefone celular; relógio; carroça puxada por animais; barco com motor; computador; cassete/cd/dvd player; mesa; cadeira; conjunto de sofá; cama; armário; relógio; barco sem motor; no mínimo um mosquitoireiro; terra própria para agricultura; Principal material: do chão; das paredes; do telhado; Número de quartos usado para dormir; Relacionamento

da estrutura do domicílio: um adulto, dois adultos de sexo oposto, dois adultos do mesmo sexo, três ou mais adultos relacionados; Se o domicílio divide o banheiro com outros domicílios; Número de domicílios que dividem o banheiro; Tipo de combustível para cozinhar; Se as crianças dormem sob o mosquito; Número de mosquiteiros que o domicílio possui; Número de crianças dormindo sob o mosquito na noite anterior; Quantidade de hectares para quem tem terra; Número de animais que o domicílio possui; Tipos de animais: Cavalos, mulas, burros; Cabras; Ovelhas; Galinhas; Gado exótico / mestiço; Gado local; Porcos; Alguém do domicílio tem conta no banco; Habitação pulverizada contra mosquitos nos últimos 12 meses; Pulverização feita: pelo governo; Companhia privada; NGO; Outro; Índice numérico de fator de riqueza; Há quantos meses atrás a habitação foi pulverizada.

Household Schedule - 4 colunas - Tipo de relacionamento com o/a chefe da família; Se o membro é de jure (residente usual); Se o membro é de facto e dormiu no domicílio na noite anterior; Gênero.

Children's height/weight/hemoglobin variables - 5 colunas - Idade em dias; Nível de hemoglobina; Maior nível educacional da mãe; Intervalo anterior ao nascimento; Número de nascimento.

Malaria for Household Members - 4 colunas - Tipo de mosquito que a pessoa dormiu na última noite; Dorme sempre sob um mosquito tratado; Dorme sob um mosquito LLIN; Resultado do exame de sangue para malária (coluna alvo).

Dentro desse cenário 72% das crianças são da zona rural, 19% são da zona urbana e 9% são refugiadas. Ainda, 88,65% dessas crianças não apresentaram resultado positivo para malária, e os restantes 11,35% estavam positivas para malária no momento da coleta de dados. Ressalta-se a necessidade de modelos que possam ser aplicados em zonas rurais e a necessidade de considerar o balanceamento das classes.

B. Pré-processamento dos dados

Foram aplicadas técnicas de pré-processamentos nos dados que envolvem: preenchimento de dados nulos com valores mais frequentes (colunas categóricas) ou valores médios (colunas numéricas), *one-hot-encoding* para colunas categóricas nominais, *label encoder* para colunas categóricas ordinais e normalização dos dados. Após o pré-processamento obteve-se uma base de 7780 crianças com 171 colunas. Sendo a coluna do resultado do exame de sangue para Malária a coluna alvo.

C. Baseline de comparação

Como *baseline* de comparação para a proposta apresentada, definiu-se a aplicação de uma árvore de decisão simples no conjunto de dados original com os parâmetros padrões da função *DecisionTreeClassifier* da biblioteca *Scikit-learn*. Nesta implementação a árvore é construída com a profundidade que for alcançada pelos métodos estatísticos, podendo ter grandes profundidades o que a torna não interpretável. A essa implementação, dá-se o nome de Árvore 0. Para comparar

com uma árvore mais interpretável, isto é, com menor profundidade, implementou-se uma árvore com profundidade 5, Árvore 5; e outra com profundidade 3, Árvore 3. Entende-se que uma menor profundidade acarreta a perda de eficiência dos modelos, o objetivo, então, é analisar a magnitude dessa perda.

D. Balanceamento de dados

Uma vez que os dados em estudo apresentam desbalanceamento entre as classes, também será analisado o impacto da aplicação de técnicas de balanceamento. As duas categorias principais de técnicas de balanceamento são o *Undersampling* e o *Oversampling*.

O *Undersampling* reduz o número de amostras da classe majoritária para uma quantidade semelhante ao da classe minoritária. As duas implementações que serão analisadas são a *RandomUnderSampler* (RUS), que seleciona de forma aleatória quais dados serão excluídos; e o *NearMiss* (NM), que exclui as amostras mais semelhantes às da classe minoritária a partir do algoritmo KNN.

Por sua vez, as técnicas de *Oversampling* criam dados sintéticos da classe minoritária a partir dos dados já existentes. As duas implementações que serão analisadas são o *RandomOverSampler* (ROS), que duplica amostras escolhidas aleatoriamente e com reposição; e o *SMOTE* (*Synthetic Minority Over-sampling Technique*) [20] que cria novos registros a partir da interpolação entre dois registros.

Vale ressaltar que o balanceamento dos dados foi realizado apenas no conjunto de dados de treinamento em cada iteração da validação cruzada.

E. Nova proposta de classificação com agrupamento

Propõe-se uma nova abordagem para classificação dos pacientes com malária agregando o uso de técnicas de agrupamento. Como apresentado na figura 1, os indivíduos são agrupados conforme a sua semelhança em relação a todos os dados selecionados. Para cada grupo é implementada uma árvore de decisão. Para o indivíduo que será classificado, é calculado qual o grupo mais semelhante e aplicada sua respectiva árvore de decisão.

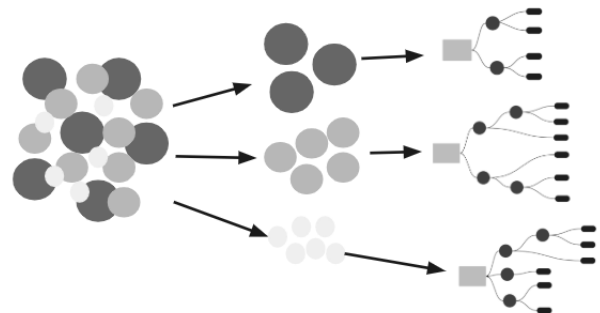


Fig. 1. Nova abordagem de classificação com agrupamento

Para a etapa de agrupamento é aplicado o algoritmo DBSCAN (Density-Based Spatial Clustering of Applications with

Noise) [21] que encontra grupos a partir da densidade de elementos em uma região. Uma região densa é dada pelos parâmetros de quantidade mínima de elementos (minpts) e a distância mínima que esses elementos têm entre si (eps). Dessa forma, o DBScan consegue encontrar grupos de formatos arbitrários, identifica ruídos (elementos que não estão em nenhuma região densa conforme os parâmetros dados) e o número de grupos é encontrado pelo algoritmo.

Contudo, um desafio é definir os parâmetros minpts e eps para identificar regiões densas no conjunto de dados. Para isso, foram calculadas todas as distâncias entre todos os elementos da base original (desconsiderando a coluna alvo) e calculada a média. A partir do valor dessa média aplicou-se o algoritmo DBScan para observar quantos grupos eram identificados. Foram feitas iterações com valores de minpts de 2 e 3. Os parâmetros que resultavam em aproximadamente 2, 7 e 16 grupos foram escolhidos para a realização dos testes seguintes.

Ao tentar classificar um elemento do conjunto de teste é necessário verificar a qual grupo encontrado ele é mais semelhante. Como são esperados grupos de formatos arbitrários, o grupo mais semelhante de um elemento é aquele que a média das distâncias de todos os seus elementos para o item de teste seja a menor.

F. Sequência de testes e Métricas de comparação

Os testes iniciais foram realizados com o *baseline* de comparação nos diferentes tamanhos de árvores. Em seguida, foram aplicadas as quatro técnicas listadas de balanceamento de dados para cada uma das árvores para verificar seu desempenho. Após, foram calculados os parâmetros do DBScan e aplicação da nova proposta para os diferentes tamanhos de árvores. O último teste consiste em juntar o balanceamento de dados e a nova proposta para análise dos resultados em cada tamanho de árvore.

Todos os testes foram realizados com validação cruzada de 10 iterações. Em cada iteração uma parte é usada para teste e as outras nove para treinamento. Ainda, os dados são divididos de forma estratificada para representar a mesma proporção das classes do conjunto de dados original, nas divisões de treinamento e teste.

No contexto destes dados de estudo a classe positiva é a classe onde há resultado positivo para malária, e a classe negativa é o resultado negativo. Os resultados encontrados são classificados então em Verdadeiros Positivos (VP): o classificador previu corretamente os casos positivos de malária; Verdadeiros Negativos (VN): o classificador previu corretamente os casos negativos de malária; Falsos Positivos (FP): o classificador previu o resultado como positivo para malária porém o paciente teve resultado negativo; Falsos Negativos (FN): o classificador identificou o paciente com resultado negativo para malária em situações que o resultado foi positivo.

Para análise das implementações de aprendizagem de máquina foram escolhidas as métricas de acurácia, precisão, recall, F1 e a área sob a curva roc (AUC-ROC). Uma vez que os dados são desbalanceados, optou-se por analisar a precisão, recall e F1 de forma separada para a classe positiva e para a

classe negativa. À precisão, recall e F1 relativa à classe positiva chamamos de P_+ (1), R_+ (2) e $F1_+$ (3) respectivamente. E são dadas pelas fórmulas a seguir.

$$P_+ : VP/(VP + FP) \quad (1)$$

$$R_+ : VP/(VP + FN) \quad (2)$$

$$F1_+ : 2 * ((P_+ * R_+)/ (P_+ + R_+)) \quad (3)$$

O mesmo é feito para a classe negativa, onde se tem respectivamente P_- (4), R_- (5) e $F1_-$ (6).

$$P_- : VN/(VN + FN) \quad (4)$$

$$R_- = VN/(VN + FP) \quad (5)$$

$$F1_- : 2 * ((P_- * R_-)/ (P_- + R_-)) \quad (6)$$

Tem-se como objetivo maior do modelo conseguir prever os casos positivos de malária, ou seja, diminuir a quantidade de falsos negativos. Assim, as análises são concentradas nas medidas da classe positiva. Na área da saúde, classificar um paciente enfermo como saudável (falso negativo) inviabiliza a intervenção e tratamento do mesmo e, dependendo da gravidade da patologia, pode levar o paciente ao óbito.

III. RESULTADOS E DISCUSSÃO

A. Baseline de comparação

A tabela I apresenta os resultados obtidos para a classe positiva nas três implementações de árvore de decisão proposta. Observa-se que a acurácia e a curva ROC dos modelos são altas enquanto a precisão e o recall para a classe positiva são extremamente baixos. As medidas altas são provocadas pelo desbalanceamento de dados: o modelo acerta muito para a classe negativa que é a majoritária. Ao se tentar alcançar um modelo mais interpretável com uma árvore de profundidade 5 ou 3 observa-se medidas, principalmente de recall, muito baixas. Tais valores inviabilizam o uso desses modelo para predição do resultado de malária.

TABLE I
TABELA BASELINE DE COMPARAÇÃO

	Acurácia	P_+	R_+	$F1_+$	Auc-Roc
Árvore 0	0.84	0.30	0.30	0.30	0.61
Árvore 5	0.89	0.61	0.13	0.21	0.79
Árvore 3	0.89	0.00	0.00	0.00	0.76

TABLE II
BALANCEAMENTO DA ÁRVORE 0

	Acurácia	P_+	R_+	$F1_+$	Auc-Roc
Árvore 0	0.84	0.30	0.30	0.30	0.61
RUS	0.68	0.21	0.68	0.32	0.68
NM	0.44	0.12	0.64	0.21	0.53
ROS	0.84	0.28	0.30	0.29	0.60
SMOTE	0.82	0.28	0.36	0.31	0.62

TABLE III
BALANCEAMENTO DA ÁRVORE 5

	Acurácia	P_+	R_+	$F1_+$	Auc-Roc
Árvore 5	0.89	0.61	0.13	0.21	0.79
RUS	0.73	0.25	0.71	0.37	0.78
NM	0.48	0.13	0.65	0.22	0.54
ROS	0.73	0.26	0.73	0.38	0.79
SMOTE	0.74	0.25	0.65	0.36	0.77

B. Balanceamento de dados

As tabelas II, III e IV apresentam o desempenho das técnicas de balanceamento nas três profundidades de árvores de decisão em relação à classe positiva. Na primeira linha de cada tabela têm-se os valores do baseline de comparação. E, em destaque, os maiores valores de cada métrica.

Na Árvore 0 observa-se que as técnicas conseguem melhorar as medidas de recall porém resultando em uma piora das medidas de precisão. Isto é, há uma diminuição dos falsos negativos, mas um aumento dos falsos positivos, o que também implica na diminuição da acurácia. Na prática, o modelo prediz que muitos estão doentes, sendo que não estão. Isto pode acarretar um gasto maior em relação aos atendimentos de saúde que poderão ser realizados sem necessidade. A mesma situação acontece para Árvore 5, que consegue com a técnica ROS 73% de recall. Já na Árvore de profundidade 3 as técnicas de balanceamento conseguem melhorar as medidas de precisão e recall que eram nulas, mas também com a diminuição da acurácia do modelo.

C. Definição dos parâmetros para o DBSCAN

A tabela V apresenta os valores de minpts e eps encontrados após várias iterações do DBSCAN com todos os dados da base. O valor inicial de eps foi a média das distâncias entre todos os elementos (4.75) e o minpts era 2. São apresentados os valores encontrados para aproximadamente 2, 7, 15 e 20 grupos, variando entre 2, 5 e 10 minpts.

TABLE IV
BALANCEAMENTO DA ÁRVORE 3

	Acurácia	P_+	R_+	$F1_+$	Auc-Roc
Árvore 3	0.89	0.00	0.00	0.00	0.76
RUS	0.67	0.22	0.76	0.34	0.76
NM	0.42	0.12	0.68	0.21	0.58
ROS	0.68	0.23	0.73	0.34	0.76
SMOTE	0.76	0.26	0.58	0.36	0.74

TABLE V
DEFINIÇÃO DOS PARÂMETROS PARA O DBSCAN

Grupos	Eps	Minpts	Ruídos
2	3.86	2	2
7	3.64	2	8
16	3.52	2	23
2	3.38	5	136
7	2.93	5	962
16	2.75	5	1669
2	2.91	10	1410
7	2.62	10	3130
15	2.38	10	4620

Somente a partir do eps 3.86 e minpts 2 é que se consegue encontrar 2 grupos. Quanto menor a distância e maior densidade (minpts), também há uma maior quantidade de ruídos. Esta situação demonstra que os dados não tem grupos tão separados. Quanto maior a distância, mais esparsos são os grupos e eles acabam se juntando formando um número menor de grupos.

Para analisar a eficiência da proposta foram escolhidas as linhas em negrito que variam a quantidade de minpts e de quantidade de grupos para os testes da nova proposta.

D. Nova proposta de classificação com agrupamento

Definidos os parâmetros para o DBSCAN, aplicou-se então a proposta apresentada. As tabelas VI, VII e VIII apresentam os resultados alcançados para as árvores sem profundidade definida, profundidade 5 e 3 respectivamente. Em ambas, a primeira linha apresenta os dados do baseline para comparação.

TABLE VI
PROPOSTA PARA A ÁRVORE 0

	Acurácia	P_+	R_+	$F1_+$	Auc-Roc
Baseline	0.84	0.30	0.30	0.30	0.61
2 grupos	0.84	0.30	0.31	0.31	0.61
7 grupos	0.83	0.24	0.21	0.21	0.55
15 grupos	0.76	0.16	0.24	0.19	0.54

TABLE VII
PROPOSTA PARA A ÁRVORE 5

	Acurácia	P_+	R_+	$F1_+$	Auc-Roc
Baseline	0.89	0.61	0.13	0.21	0.79
2 grupos	0.89	0.61	0.13	0.21	0.79
7 grupos	0.85	0.31	0.17	0.19	0.58
15 grupos	0.77	0.17	0.23	0.19	0.55

Em todas as profundidades, a acurácia da proposta foi igual ou inferior ao baseline. Na Árvore 0 há uma mínima melhora na medida do recall. Na Árvore 5 há uma melhora do recall com o aumento da quantidade de grupos, porém com prejuízo para a precisão. Já na Árvore 3, há a melhora da precisão e do recall a partir de 7 grupos e uma diminuição da acurácia.

TABLE VIII
PROPOSTA PARA A ÁRVORE 3

	Acurácia	P_+	R_+	$F1_+$	Auc-Roc
Baseline	0.89	0.0	0.0	0.0	0.76
2 grupos	0.89	0.0	0.0	0.0	0.77
7 grupos	0.85	0.22	0.12	0.13	0.58
15 grupos	0.77	0.16	0.23	0.19	0.55

E. Nova proposta com balanceamento

Ainda foram realizados os testes com o balanceamento dos dados de treinamento antes da criação dos grupos na proposta apresentada. As tabelas IX, X e XI apresentam os resultados. A primeira linha de cada tabela apresenta o baseline de comparação, a segunda, o resultado para o balanceamento e as demais, os resultados da proposta com balanceamento nas diferentes quantidades de grupos. Para o balanceamento dos dados foi definida a técnica de *RandomUnderSampler* que obteve o melhor recall para o balanceamento da Árvore 3.

TABLE IX
PROPOSTA COM BALANCEAMENTO PARA A ÁRVORE 0

	Acurácia	P_+	R_+	$F1_+$	Auc-Roc
Baseline	0.84	0.30	0.30	0.30	0.61
RUS	0.68	0.21	0.68	0.32	0.68
2 grupos	0.67	0.11	0.29	0.16	0.05
7 grupos	0.71	0.14	0.31	0.19	0.53
15 grupos	0.59	0.12	0.41	0.18	0.51

TABLE X
PROPOSTA COM BALANCEAMENTO PARA A ÁRVORE 5

	Acurácia	P_+	R_+	$F1_+$	Auc-Roc
Baseline	0.89	0.61	0.13	0.21	0.79
RUS	0.73	0.25	0.71	0.37	0.78
2 grupos	0.75	0.12	0.18	0.14	0.50
7 grupos	0.70	0.14	0.30	0.19	0.59
15 grupos	0.56	0.12	0.43	0.18	0.50

TABLE XI
PROPOSTA COM BALANCEAMENTO PARA A ÁRVORE 3

	Acurácia	P_+	R_+	$F1_+$	Auc-Roc
Baseline	0.89	0.00	0.00	0.00	0.76
RUS	0.67	0.22	0.76	0.34	0.76
2 grupos	0.68	0.13	0.29	0.18	0.50
7 grupos	0.71	0.15	0.32	0.20	0.59
15 grupos	0.54	0.12	0.49	0.20	0.51

Observa-se em todas as árvores testadas que o recall diminuiu com o aumento dos grupos. Como a técnica RUS é uma técnica de diminuição de dados, ao se dividir em grupos, tem-se menos elementos em cada grupo, e assim menos representantes. Na Árvore 3 observa-se um aumento do recall em relação aos resultados apresentados na tabela VIII, porém com a diminuição da precisão.

A 2 ilustra a interpretabilidade dos modelos gerados pela Árvore 3, apresentando a árvore gerada após o treinamento com a base completa e com balanceamento de dados RUS.

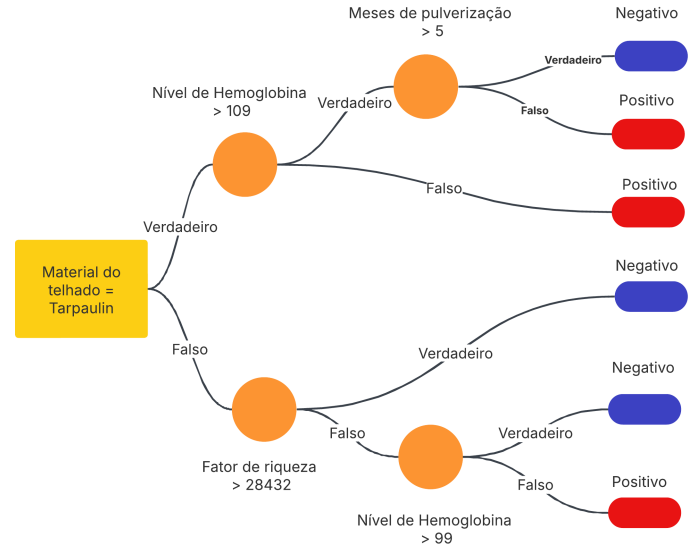


Fig. 2. Árvore de decisão de profundidade 3

IV. CONCLUSÃO

A malária continua sendo um grande problema de saúde pública em diversos países e, em especial, países do continente africano. Identificar fatores que favorecem a proliferação da doença e atuar na sua prevenção e intervenção precoce das suas consequências é urgente. A aplicação de técnicas de aprendizagem de máquina são ferramentas com grandes potenciais de auxílio a essa demanda. Tais técnicas geram modelos que podem ser usados pelos profissionais de saúde para predição de casos de malária e também na formulação de políticas públicas.

Entretanto, tais modelos precisam ser interpretáveis para poderem ser de fato utilizados, inclusive facilitando o uso de um modelo pronto em áreas remotas. Outro desafio para essas técnicas é o desbalanceamento dos dados de saúde, onde há poucos registros sobre quem está acometido pela doença.

Neste trabalho analisamos a aplicação de técnicas de aprendizagem de máquina para predição de malária em crianças menores de 5 anos de Uganda. Como classificadores, foram aplicadas árvores de decisão de diferentes profundidades para buscar modelos mais interpretáveis. Também foram analisadas técnicas de balanceamento para verificar se haveria ganho de eficiência. Também foi proposta uma nova forma de classificação incorporando técnicas de agrupamento para buscar diferentes perfis de indivíduos e obter uma melhor eficiência dos modelos e a sua combinação com técnicas de balanceamento.

Conclui-se, com os testes realizados, que a diminuição da profundidade da árvore faz com que as técnicas de balanceamento alcancem melhorias nas medidas de recall e precisão. A nova proposta apresentada também consegue melhorar essas

medidas de forma modesta e ainda permite a identificação de diferentes perfis dos indivíduos. A variedade de perfis permite mostrar diferentes regras para cada grupo/perfil encontrado, permitindo compreender melhor a doença em estudo. Estas informações podem ser utilizadas para a formulação de políticas públicas ao se entender os principais fatores socioeconômicos relacionados à presença da doença. Como no exemplo de árvore apresentado, a maior frequência de pulverização das casas poderia colaborar no resultado negativo para malária.

Outra discussão é sobre o uso de dados socioeconômicos como predição. Nos testes realizados observa-se que possuem informações que favorece a predição de casos de malária. É necessária a realização de pesquisas futuras para determinar subconjuntos de dados que resultem em melhor eficiência dos modelos e a agregação destes com outros dados que sejam possíveis de coletar.

Observa-se também que os modelos gerados correspondem ao conjunto de dados de Uganda. A aplicação dessas técnicas em dados de outros países irá gerar modelos diferenciados de acordo com suas especificidades. Assim, há a necessidade de análise especializada dos modelos para que políticas públicas sejam formuladas de forma eficiente.

REFERENCES

- [1] World Health Organization, *World Malaria Report 2023*. Geneva: World Health Organization, 2023, acesso em: 28 maio 2025. [Online]. Available: <https://www.who.int/publications/i/item/9789240086173>
- [2] Uganda National Malaria Control Division (NMCD), Uganda Bureau of Statistics (UBOS), and ICF, *2018-19 Uganda Malaria Indicator Survey: Atlas of Key Indicators*. Kampala, Uganda, and Rockville, Maryland, USA: NMCD, UBOS, and ICF, 2019, acesso em: 28 maio 2025. [Online]. Available: <https://dhsprogram.com/pubs/pdf/ATR21/ATR21.pdf>
- [3] M. Mariki, E. Mkoba, and N. M. and, "Combining clinical symptoms and patient features for malaria diagnosis: Machine learning approach," *Applied Artificial Intelligence*, vol. 36, no. 1, p. 2031826, 2022. [Online]. Available: <https://doi.org/10.1080/08839514.2022.2031826>
- [4] M. Semakula, F. Niragire, and C. Faes, "Spatio-temporal bayesian models for malaria risk using survey and health facility routine data in rwanda," *International Journal of Environmental Research and Public Health*, vol. 20, no. 5, 2023. [Online]. Available: <https://www.mdpi.com/1660-4601/20/5/4283>
- [5] H. M. Semakula, S. Liang, P. I. Mukwaya, F. Mugagga, D. Nseka, H. Wasswa, P. Mwendwa, P. Kayima, S. P. Achuu, and J. Nakato, "Bayesian belief network modelling approach for predicting and ranking risk factors for malaria infections among children under 5 years in refugee settlements in Uganda," *Malaria Journal*, vol. 22, no. 1, p. 297, Oct. 2023.
- [6] M. F. X. Barboza, K. H. d. C. Monteiro, I. R. Rodrigues, G. L. Santos, W. M. Monteiro, E. A. G. Figueira, V. d. S. Sampaio, T. Lynn, and P. T. Endo, "Prediction of malaria using deep learning models: A case study on city clusters in the state of amazonas, brazil, from 2003 to 2018," *Revista da Sociedade Brasileira de Medicina Tropical*, vol. 55, p. e0420–2021, 2022. [Online]. Available: <https://doi.org/10.1590/0037-8682-0420-2021>
- [7] W. Silka, M. Wiczorek, J. Silka, and M. Woźniak, "Malaria detection using advanced deep learning architecture," *Sensors*, vol. 23, no. 3, 2023. [Online]. Available: <https://www.mdpi.com/1424-8220/23/3/1501>
- [8] O. Khan, J. O. Ajadi, and M. P. Hossain, "Predicting malaria outbreak in The Gambia using machine learning techniques," *PLOS ONE*, vol. 19, no. 5, pp. 1–22, May 2024.
- [9] H. M. Semakula and F. Mugagga, "Predictors of insecticide-treated net utilization among children under age 5 in refugee settlements in uganda: Analysis of the 2018–19 uganda malaria indicator survey," ICF, Rockville, Maryland, USA, DHS Working Paper No. 206, 2024, acesso em: 28 maio 2025. [Online]. Available: <https://dhsprogram.com/pubs/pdf/WP206/WP206.pdf>
- [10] I. D. Ozodiegwu, M. Ambrose, K. E. Battle, C. Bever, O. Diallo, B. Galatas, M. Runge, and J. Gerardin, "Beyond national indicators: Adapting the Demographic and Health Surveys' sampling strategies and questions to better inform subnational malaria intervention policy," *Malaria Journal*, vol. 20, no. 1, p. 122, Mar. 2021.
- [11] A. Venkatasubramaniam, J. Wolfson, N. Mitchell, T. Barnes, M. JaKa, and S. French, "Decision trees in epidemiological research," *Emerging Themes in Epidemiology*, vol. 14, no. 1, p. 11, Sep. 2017.
- [12] A. B. N, T. Patel, S. Patil, R. L. S, and V. Singh, "Enhancing the quality and efficiency of mental health care using decision trees," in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2023, pp. 1–5.
- [13] K. Battista, L. Diao, K. A. Patte, J. A. Dubin, and S. T. Leatherdale, "Examining the use of decision trees in population health surveillance research: An application to youth mental health survey data in the compass study," *Health Promotion and Chronic Disease Prevention in Canada: Research, Policy and Practice*, vol. 43, no. 2, pp. 73–86, 2023, acesso em: 28 maio 2025. [Online]. Available: <https://www.canada.ca/en/public-health/services/reports-publications/health-promotion-chronic-disease-prevention-canada-research-policy-practice/vol-43-no-2-2023/decision-trees-population-health-surveillance-research-youth-mental-health-compass-study.html>
- [14] N. Swart, A. M. Sinha, A. Bentley, H. Smethurst, G. Spencer, S. Ceder, and M. H. Wilcox, "A cost-utility analysis of two Clostridioides difficile infection guideline treatment pathways," *Clinical Microbiology and Infection*, vol. 29, no. 10, pp. 1291–1297, Oct. 2023.
- [15] E. Legendre, L. Lehot, S. Dieng, S. Rebaudet, A. M. Thu, J. D. Rae, G. Delmas, F. Girond, V. Herbreteau, F. Nosten, J. Landier, and J. Gaudart, "Malaria temporal dynamic clustering for surveillance and intervention planning," *Epidemics*, vol. 43, p. 100682, 2023.
- [16] J. Omoha, O. Omodele, V. Abifade, and A. Najib Isyaku, "Assessment of agglomerative clustering techniques on grouping of malaria infected patients in kaduna state, nigeria," *International Journal of Quantitative and Qualitative Research Methods*, vol. 12, pp. 44–57, 05 2024.
- [17] P. Jiraamphorn and E. Pacharawongsakda, "DB2SM: An efficient resampling technique for imbalanced data classification," in *Recent Advances in Information and Communication Technology 2021*, P. Meesad, D. S. Sodsee, W. Jitsakul, and S. Tangwannawit, Eds. Cham: Springer International Publishing, 2021, pp. 1–12.
- [18] E. Legendre, L. Lehot, S. Dieng, S. Rebaudet, A. M. Thu, J. D. Rae, G. Delmas, F. Girond, V. Herbreteau, F. Nosten, J. Landier, and J. Gaudart, "Malaria temporal dynamic clustering for surveillance and intervention planning," *Epidemics*, vol. 43, p. 100682, 2023.
- [19] The DHS Program, "Available datasets," <https://dhsprogram.com/data/available-datasets.cfm>, 2025, acesso em: 28 maio 2025.
- [20] V. C. Nitesh, "Smote: synthetic minority over-sampling technique," *J Artif Intell Res*, vol. 16, no. 1, p. 321, 2002.
- [21] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, ser. KDD'96. AAAI Press, 1996, p. 226–231.