

FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE MATEMÁTICA E INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA APLICADA

DISSERTAÇÃO DE MESTRADO

Aplicação de Machine Learning na Análise de Factores Associados as Uniões Prematuras em Moçambique

Bertina Marília do Carmo Cuco Tlhemo

Maputo
2026

UNIVERSIDADE EDUARDO MONDLANE

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE MATEMÁTICA E INFORMÁTICA

PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA APLICADA

DISSERTAÇÃO DE MESTRADO

Aplicação de Machine Learning na Análise de Factores Associados as Uniões Prematuras em Moçambique

Bertina Marília do Carmo Cuco Tlhemo

Dissertação submetida ao Programa de Pós-graduação em Estatística Aplicada da Universidade Eduardo Mondlane, como parte dos requisitos para a obtenção do grau de Mestre em Estatística Aplicada.

Orientador: Prof. Doutor Cachimo Combo Assane

Maputo

2026

FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE MATEMÁTICA E INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA APLICADA

DISSERTAÇÃO DE MESTRADO

Aplicação de Machine Learning na Análise de Factores Associados as Uniões Prematuras em Moçambique

Bertina Marília do Carmo Cuco Tlhemo

Dissertação submetida ao Programa de Pós-graduação em Estatística Aplicada da Universidade Eduardo Mondlane, como parte dos requisitos para a obtenção do grau de Mestre em Estatística Aplicada.

Aprovado em ——— de——————— de 2026, por:

_____

Doutor Adelino José Chingore Juga
(Arguente - Universidade Eduardo Mondlane-DMI)

_____

Prof. Doutor Alberto Chicafo Mulenga
(Presidente - Universidade Eduardo Mondlane-DMI)

_____

Prof. Doutor Cachimo Combo Assane
(Orientador - Universidade Eduardo Mondlane-DMI)

Maputo
2026

DEDICATÓRIA

Este trabalho é dedicado à minha família, ao meu esposo Felisberto Tlhemo e aos meus filhos Mayisha e Thabo, pelo amor, apoio incondicional e paciência ao longo desta caminhada.

AGRADECIMENTOS

A Deus, por me conceder vida, saúde, perseverança e força para concluir esta etapa tão significativa da minha vida. A Ele toda honra, glória e gratidão, pois sem a Sua presença e graça nada disto teria sido possível.

Aos meus queridos pais, Marta Mangue e Paulo Cuco, pelo amor incondicional, pelos valores transmitidos e pelo apoio constante ao longo da minha vida. Aos meus irmãos, pela amizade, incentivo e presença constante, mesmo nos momentos mais difíceis.

Ao meu orientador, Prof. Doutor Cachimo Combo Assane, pela orientação dedicada, disponibilidade, paciência e valiosos ensinamentos transmitidos ao longo da realização desta investigação. O seu rigor científico, apoio constante e incentivo foram fundamentais para a concretização deste trabalho.

A todos os docentes do Mestrado em Estatística Aplicada, pelos conhecimentos partilhados, pela dedicação ao ensino e pelo contributo inestimável para a minha formação académica e profissional.

Aos meus colegas de turma, agradeço pela partilha de conhecimentos, pela entreajuda e pelos momentos vividos ao longo deste percurso académico. Um agradecimento especial a Muemed Cassimo, Samuel Jamisse e Falume Chale pelo companheirismo, amizade e apoio inestimável.

Resumo

As uniões prematuras constituem um fenómeno de relevância social e de saúde pública em Moçambique, com implicações negativas no desenvolvimento educacional e económico e no bem-estar das raparigas. A presente investigação teve como objectivo aplicar técnicas de *Machine Learning* na análise dos factores associados a este fenómeno, complementando-as com métodos estatísticos clássicos. Foram investigados os dados do Inquérito Demográfico e de Saúde de Moçambique (IDS, 2022–23), tendo-se procedido a uma caracterização descritiva e a uma análise bivariada para examinar as associações entre as variáveis explicativas e a ocorrência de uniões prematuras. Em seguida, ajustou-se um modelo de Regressão Logística, considerando o desenho amostral complexo, e aplicou-se o algoritmo *Extreme Gradient Boosting* (XGBoost), com vista a estimar a probabilidade de ocorrência do fenómeno e a comparar o desempenho preditivo das abordagens. O modelo XGBoost foi optimizado por ajuste hiperparamétrico com pesquisa em grelha e validação cruzada. A Regressão Logística apresentou uma área sob a curva ROC (AUC) de 0,82, revelando boa capacidade discriminatória, enquanto o XGBoost registou uma AUC média de 0,832 na validação cruzada e de 0,817 no conjunto de teste, evidenciando desempenho global semelhante, com ligeira vantagem na fase de validação. A análise da importância das variáveis identificou a idade na primeira relação sexual, o estado civil, o nível de escolaridade e os factores socioeconómicos como os principais determinantes, sugerindo que o fenómeno resulta da interacção de factores demográficos, sociais e comportamentais. Conclui-se que a aplicação de técnicas de *Machine Learning*, articulada com métodos estatísticos clássicos, se mostrou adequada para analisar os factores associados às uniões prematuras e avaliar a capacidade preditiva dos modelos, fornecendo evidência empírica relevante para apoiar a formulação de políticas públicas e de estratégias de intervenção orientadas para a prevenção do fenómeno em Moçambique.

Palavras-chave: Inquérito Demográfico e de Saúde; *Machine Learning*; Moçambique; Regressão Logística; Uniões prematuras; XGBoost.

Abstract

Early unions constitute a phenomenon of social and public-health relevance in Mozambique, with negative implications for the educational and economic development and the overall well-being of girls. This research aimed to apply *Machine Learning* techniques to the analysis of the factors associated with this phenomenon, complementing them with classical statistical methods. Data from the Mozambique Demographic and Health Survey (DHS, 2022–23) were investigated, with a descriptive characterization and a bivariate analysis conducted to examine associations between the explanatory variables and the occurrence of early unions. A Logistic Regression model was then fitted, accounting for the complex survey design, and the *Extreme Gradient Boosting* (XGBoost) algorithm was applied to estimate the probability of occurrence and to compare the predictive performance of the approaches. The XGBoost model was optimized through hyperparameter tuning using grid search and cross-validation. Logistic Regression achieved an area under the ROC curve (AUC) of 0.82, showing good discriminative ability, while XGBoost achieved a mean AUC of 0.832 in cross-validation and 0.817 on the test set, indicating similar overall performance, with a slight advantage during validation. The variable-importance analysis identified age at first sexual intercourse, marital status, level of education, and socioeconomic factors as the main determinants, suggesting that the phenomenon results from the interaction of demographic, social, and behavioral factors. It is concluded that the application of *Machine Learning* techniques, combined with classical statistical methods, proved adequate for analyzing the factors associated with early unions and assessing the predictive performance of the models, providing relevant empirical evidence to support the design of public policies and intervention strategies aimed at preventing this phenomenon in Mozambique.

Keywords: Demographic and Health Survey; Early unions; Logistic Regression; *Machine Learning*; Mozambique; XGBoost.

Lista de Figuras

2.1 Curva ROC (Receiver Operating Characteristic). 25

3.1 Fluxograma do processo de análise estatística e modelação preditiva 37

4.1 Distribuição das mulheres em idade reprodutiva segundo a União Prematura. . . 44

4.2 Distribuição das mulheres em idade reprodutiva por quintil de riqueza 45

4.3 Distribuição das mulheres em idade reprodutiva por área de residência 46

4.4 Distribuição das mulheres em idade reprodutiva por nível de escolaridade . . . 47

4.5 Diagnóstico de Observações Influentes através da Distância de Cook 51

4.6 Curva ROC do Modelo 56

4.7 Curva ROC de regressão logística com Validação Cruzada 59

4.8 Curva ROC do modelo XGBoost 61

4.9 Curva ROC do modelo XGBoost para a predição da união prematura 64

4.10 Importância das variáveis no modelo XGBoost 68

Lista de Tabelas

2.1 Estrutura da matriz de confusão utilizada em modelos de classificação binária . 25

2.2 Critérios de interpretação da área sob a curva ROC 26

3.1 Descrição das variáveis do estudo extraídas do IDS 2022–2023 34

4.1 Teste de Associação de Rao & Scott entre as Variáveis Explicativas e a Variável Resposta 48

4.2 Diagnóstico de Multicolinearidade das Variáveis Explicativas 50

4.3 Modelo de Regressão Logística para os Factores Associados à União Prematura 52

4.4 Desempenho do Modelo de Regressão Logística considerando o desenho amostral 55

4.5 Medidas de desempenho do modelo de Regressão Logística 58

4.6 Medidas de desempenho preditivo do modelo XGBoost 62

4.7 Medidas de desempenho do modelo XGBoost obtidas por validação cruzada . . 63

4.8 Resumo do ajuste hiperparamétrico do modelo XGBoost 65

4.9 Desempenho do modelo XGBoost na validação cruzada e no conjunto de teste . 66

4.10 Importância das variáveis no modelo XGBoost 67

4.11 Comparação do desempenho preditivo entre Regressão Logística e XGBoost . . 70

Conteúdo

1 INTRODUÇÃO **1**

1.1 Motivação 2

1.2 Problematização 3

1.3 Objectivos 4

1.3.1 Objectivo Geral 4

1.3.2 Objectivos Específicos 4

1.4 Relevância 4

1.5 Estrutura da Dissertação 5

2 REVISÃO DA LITERATURA **6**

2.1 Síntese conceptual e empírica das uniões prematuras 6

2.1.1 Enquadramento institucional e normativo das uniões prematuras em Moçambique 7

2.1.2 Contexto internacional das uniões prematuras 8

2.1.3 Factores associados às uniões prematuras 9

2.1.4 Consequências das uniões prematuras 10

2.1.5 Abordagens metodológicas em estudos sobre uniões prematuras 11

2.1.6 Aprendizado de Máquina 12

2.1.7 Aprendizagem Supervisionada 13

2.1.8 Síntese crítica da literatura 14

2.2 Modelo de Regressão Logística 16
2.2.1 Formulação Matemática do Modelo de Regressão Logística 17
2.2.2 Função de Ligação Logit 17
2.2.3 Chance e Razão de Chances 17
2.2.4 Estimação dos Parâmetros 18
2.3 Testes de Significância dos Parâmetros 19
2.3.1 Teste de Razão de Verossimilhança 19
2.3.2 Teste de Wald 20
2.3.3 Teste Escore de Rao 21
2.4 Avaliação do Ajuste e do Desempenho Preditivo dos Modelos 22
2.4.1 Pseudo R^2 22
2.4.2 Teste de Hosmer–Lemeshow 23
2.4.3 Curva ROC e Medidas de Desempenho do Modelo 24
2.4.4 Matriz de Confusão 25
2.5 Sensibilidade e Especificidade 26
2.6 Área Abaixo da Curva ROC 26
2.7 Algoritmo XGBoost 27
2.8 Trabalhos Relacionados 31

3 METODOLOGIA **33**

3.1 Material 33
3.2 Variáveis em Estudo 33
3.3 Classificação da Pesquisa 35
3.4 Métodos 35
3.4.1 Etapas do Processo de Análise 36
3.4.2 Análise Exploratória de Dados 37

3.4.3 Divisão da Base de Dados 39
3.4.4 Treinamento dos Modelos 40
3.4.5 Ajuste Hiperparamétrico 41
3.4.6 Técnicas de Validação 41
3.4.7 Avaliação do Desempenho dos Modelos 42

4 RESULTADOS E DISCUSSÃO 43

4.1 Estatística Descritiva 43
4.1.1 Variável resposta 43
4.1.2 Distribuição das mulheres em idade reprodutiva por quintil de riqueza . 44
4.1.3 Distribuição das mulheres em idade reprodutiva por área de residência . 45
4.1.4 Distribuição das mulheres em idade reprodutiva por nível de escolaridade 46
4.2 Análise bivariada da associação entre as variáveis explicativas e as uniões prematuras 47
4.3 Modelação Estatística 49
4.3.1 Diagnóstico de Multicolinearidade 49
4.3.2 Estimação do Modelo de Regressão Logística 51
4.3.3 Avaliação do Desempenho do Modelo de Regressão Logística 55
4.3.4 Curva ROC e Área sob a Curva 56
4.3.5 Validação Cruzada do Modelo de Regressão Logística 57
4.3.6 Curva ROC do Modelo de Regressão Logística 58
4.3.7 Comparação do Desempenho entre o Modelo Inicial e o Modelo com Validação Cruzada 59
4.4 Modelo XGBoost 60
4.4.1 Avaliação do Desempenho do Modelo XGBoost 60
4.4.2 Avaliação do Desempenho do Modelo XGBoost por Validação Cruzada 62
4.4.3 Avaliação do Desempenho do Modelo XGBoost por meio da Curva ROC 63

4.4.4 Ajuste Hiperparamétrico do Modelo XGBoost 64
4.4.5 Avaliação das Combinações de Hiperparâmetros e Identificação da Melhor Configuração do Modelo XGBoost 65
4.4.6 Avaliação do Desempenho do Modelo XGBoost na Validação Cruzada e no Conjunto de Teste 65
4.4.7 Importância das Variáveis no Modelo 66
4.4.8 Comparação entre Regressão Logística e XGBoost 68
4.5 Discussão 70

5 CONCLUSÕES E RECOMENDAÇÕES 73

5.1 Conclusões 73
5.2 Recomendações 74

REFERÊNCIAS BIBLIOGRÁFICAS 76

Capítulo 1

INTRODUÇÃO

As uniões prematuras afectam de forma desigual as raparigas em diversas regiões do mundo, constituindo uma violação dos direitos humanos e limitando significativamente as suas oportunidades de desenvolvimento social, educacional e económico. Esta prática compromete a trajectória de vida de milhões de raparigas, com implicações que transcendem fronteiras culturais e geográficas, configurando-se como um obstáculo persistente ao alcance da igualdade de género e do desenvolvimento sustentável, conforme o Fundo das Nações Unidas para a Infância (United Nations International Children's Emergency Fund [UNICEF], 2022).

Estima-se que, todos os anos, aproximadamente 12 milhões de raparigas no mundo entrem em uniões antes dos 18 anos de idade, sendo que mais de 650 milhões de mulheres actualmente vivas foram casadas durante a infância. Estes números evidenciam a magnitude e a persistência do fenómeno, apesar dos avanços registados nas últimas décadas em matéria de políticas públicas e de salvaguarda dos direitos da criança e do adolescente.

No contexto africano, o fenómeno assume particular gravidade, uma vez que uma proporção significativa dos países com as maiores taxas de uniões prematuras se encontra neste continente. Entre estes destacam-se Moçambique, Níger, Chade, Mali, Burkina Faso, Sudão do Sul, Nigéria, Guiné e República Centro-Africana. Nestes países, a persistência desta prática está fortemente associada a factores estruturais, como desigualdades de género, níveis elevados de pobreza, normas socioculturais e limitações no acesso à educação. Para além de comprometer as oportunidades de desenvolvimento, as uniões prematuras estão também relacionadas com piores indicadores de saúde, incluindo maior risco de mortalidade materna e infantil, desnutrição crónica e maior exposição à violência doméstica (Bassiano & de Lima, 2018).

Em Moçambique, as uniões prematuras continuam a assumir uma expressão significativa, apesar dos esforços desenvolvidos pelo Governo e por organizações nacionais e internacionais no combate ao fenómeno. Estima-se que cerca de 48% de mulheres com idades compreendidas entre 20 e 24 anos tenham entrado em união antes dos 18 anos, evidenciando a persistência de desigualdades socioeconómicas e educacionais, sobretudo em contextos marcados por elevada

vulnerabilidade social. Estes dados evidenciam a necessidade de análises empíricas mais aprofundadas que permitam compreender os factores associados à ocorrência das uniões prematuras no contexto moçambicano.

Neste enquadramento, a utilização de métodos estatísticos e de técnicas de *Machine Learning* tem vindo a ganhar destaque na análise de fenómenos sociais complexos, como as uniões prematuras. Estas abordagens permitem identificar padrões não lineares, interacções complexas entre variáveis e factores de risco associados ao fenómeno, contribuindo para uma compreensão mais abrangente e robusta da sua dinâmica. Estudos recentes realizados em diferentes contextos demonstram o potencial de Machine Learning na identificação de determinantes relevantes e na estimação da probabilidade de ocorrência de uniões prematuras, fornecendo evidência empírica importantes para a formulação de políticas públicas baseadas em evidência.

Alem et al. (2020), num estudo realizado no Bangladesh, aplicaram algoritmos de *Machine Learning* para investigar os determinantes e a distribuição espacial das uniões prematuras, identificando factores como o nível educacional, a autonomia das mulheres e a região de residência como estatisticamente relevantes. De forma semelhante, Mathabatha e Tsawe (2023) recorreram a técnicas de *Machine Learning* na África do Sul para analisar os factores associados às uniões prematuras entre mulheres com idades compreendidas entre 20 e 49 anos, evidenciando a influência do nível de escolaridade, da idade na primeira relação sexual, bem como das condição socioeconómica e da riqueza familiar.

No contexto moçambicano, Hodges (2015) analisou dados relacionados com uniões prematuras e gravidez precoce, evidenciando que uma proporção significativa de raparigas entra em união antes dos 18 anos, frequentemente associada à ocorrência de gravidez em idades muito jovens. Estes resultados reforçam a pertinência de abordagens analíticas que permitam explorar, de forma integrada, os factores associados ao fenómeno.

Neste sentido, o presente estudo propõe-se utilizar métodos estatísticos e técnicas de *Machine Learning* para desenvolver modelos analíticos e preditivos capazes de identificar os factores associados às uniões prematuras e de estimar a probabilidade da sua ocorrência em Moçambique, com informações provenientes do Inquérito Demográfico e de Saúde 2022–2023. A análise realizada visa fornecer evidência empírica robusta que sustente o desenho de estratégias de prevenção e mitigação mais adequadas às particularidades do contexto moçambicano.

1.1 Motivação

O presente estudo pretende constituir mais uma ferramenta de análise que auxilie na compreensão dos factores que impulsionam as uniões prematuras em Moçambique. Tratando-se de um problema que vem sendo combatido e que, por si só, configura um crime, a compreensão

desta problemática reveste-se de extrema importância para os decisores, na medida em que lhes proporciona instrumentos adicionais para a implementação de políticas que contribuam para o combate eficaz do fenómeno, permitindo às menores submetidas a tais actos prosseguir o seu processo normal de crescimento até atingirem a maioridade, momento a partir do qual passam a poder tomar as suas próprias decisões. A investigação desenvolvida poderá lançar nova luz sobre este fenómeno, ao possibilitar uma melhor compreensão do seu historial, das suas consequências e das possíveis formas de mitigação. Constitui igualmente uma oportunidade para evidenciar a importância do recurso a ferramentas estatísticas na identificação e resolução de problemas que afectam significativamente a vida de uma sociedade. Por fim, tratando-se de um trabalho de fim de curso, acresce a motivação de saber que permitirá a obtenção de mais um título académico e que representará um contributo para futuros trabalhos a desenvolver em torno desta temática.

1.2 Problematização

Apesar dos esforços desenvolvidos pelo Governo e por diversas instituições internacionais, as uniões prematuras persistem em Moçambique, continuando a comprometer o desenvolvimento e o bem-estar das raparigas. Os dados do Inquérito Demográfico e de Saúde 2022–2023 evidenciam a magnitude do problema: entre as mulheres dos 25 aos 49 anos, a proporção das que tiveram a primeira relação sexual antes dos 18 anos tem aumentado de forma consistente desde 1997, passando de 75% para 81% em 2022–2023, sendo que cerca de 46% se casaram pela primeira vez aos 18 anos ou menos.

As consequências deste fenómeno são profundas e prolongam-se ao longo da vida das raparigas. A entrada precoce em união encontra-se associada a um risco acrescido de gravidez precoce, abortos inseguros, interrupção da educação, infecções sexualmente transmissíveis e complicações obstétricas, a que se somam um menor estatuto económico das mulheres, taxas elevadas de divórcio e um risco aumentado de mortalidade materna e infantil. A estas repercussões acresce a maior exposição à violência doméstica, nas suas dimensões física, sexual e psicológica, ou mesmo ao abandono, o que coloca em causa a segurança e a integridade das raparigas.

Neste contexto, a procura de soluções eficazes para a erradicação das uniões prematuras assume-se como condição essencial para assegurar um futuro promissor às raparigas moçambicanas. Compreender os factores que contribuem para a ocorrência deste fenómeno, com recurso a modelos de *Machine Learning*, revela-se fundamental para a definição de medidas adequadas de prevenção e combate, de modo a garantir o pleno desenvolvimento e o bem-estar de um grupo cuja qualidade de vida e trajectória pessoal são significativamente afectadas por esta prática.

Diante do exposto, importa perceber quais são os factores associados às uniões prematuras em Moçambique.

1.3 Objectivos

1.3.1 Objectivo Geral

Aplicar técnicas de *Machine Learning* na análise dos factores associados às uniões prematuras em Moçambique.

1.3.2 Objectivos Específicos

- Descrever o perfil sociodemográfico associado à ocorrência de uniões prematuras entre as raparigas moçambicanas;
- Estimar a probabilidade de ocorrência de uniões prematuras através dos modelos de Regressão Logística e *XGBoost*, avaliando o seu desempenho preditivo com base em medidas de validação apropriadas;
- Comparar os modelos de Regressão Logística e *XGBoost*, com vista à identificação do modelo que apresenta melhor desempenho com os dados disponíveis.

1.4 Relevância

As uniões prematuras representam um desafio significativo em Moçambique, onde as implicações sociais, económicas e de saúde são profundamente sentidas, particularmente pelas raparigas e pelas mulheres. Em muitas comunidades moçambicanas, as uniões prematuras estão frequentemente associadas a práticas tradicionais, pobreza, falta de acesso à educação e serviços de saúde inadequados. Estas uniões muitas vezes privam as raparigas de oportunidades educacionais e de desenvolvimento, limitando as suas perspectivas de futuro e perpetuando o ciclo de pobreza. Existem poucos estudos que abordam as questões de uniões prematuras em Moçambique, o que dificulta a compreensão dos factores específicos que contribuem para este grande problema. O presente estudo é um contributo para o desenvolvimento de políticas e intervenções eficazes para a resolução deste grande mal, pois envolve uma análise abrangente das dinâmicas sociais, culturais e económicas dentro das comunidades, levando em consideração questões como uniões prematuras, práticas tradicionais, desigualdade de género, limitações no acesso à informação sobre saúde sexual e reprodutiva, assim como situações de pobreza e vulnerabilidade socioeconómica.

1.5 Estrutura da Dissertação

O presente trabalho encontra-se organizado em cinco capítulos.

- **O Capítulo 1**: Apresenta a introdução da investigação, a motivação, a problematização, a relevância da pesquisa, bem como os objectivos gerais e específicos que orientam a investigação.

  • **Capítulo 2**: É dedicado à revisão da literatura, onde se procede ao enquadramento conceptual das uniões prematuras, ao contexto institucional e normativo em Moçambique, ao contexto internacional do fenómeno, aos principais factores associados e às suas consequências, incluindo ainda a discussão das abordagens metodológicas utilizadas em estudos anteriores, com enfoque no modelo de regressão logística e no algoritmo XGBoost baseado em *Machine Learning*.

  • **Capítulo 3**: São descritos os procedimentos metodológicos adoptada no estudo, detalhando a fonte e a natureza dos dados utilizados, a classificação da pesquisa, as variáveis em análise, as etapas do processo analítico, os métodos aplicados e os procedimentos de validação e avaliação do desempenho dos modelos.

  • **Capítulo 4**: Apresenta os resultados obtidos e a respectiva discussão, abrangendo a análise estatística descritiva, a análise bivariada, a regressão logística e os modelos de Machine Learning, com especial enfoque no XGBoost, bem como a comparação do desempenho preditivo entre os modelos considerados.

  • **Capítulo 5**: Apresenta as conclusões do estudo, destacando as principais contribuições da investigação, as implicações no apoio à formulação de políticas públicas, as limitações identificadas e recomendações para pesquisas futuras.

  Esta estrutura busca assegurar a clareza e a organização do conteúdo, facilitando a compreensão dos resultados e contribuições apresentadas.

Capítulo 2

REVISÃO DA LITERATURA

2.1 Síntese conceptual e empírica das uniões prematuras

As uniões prematuras constituem um fenómeno social multidimensional, definido como a entrada em união conjugal, formal ou informal, antes dos 18 anos de idade. Esta realidade é amplamente considerada uma violação dos direitos da criança e da adolescente, na medida que compromete o exercício pleno do direito à educação, à saúde, à protecção e ao exercício da autodeterminação (UNICEF, 2016; UNICEF, 2021). No plano estatístico e demográfico, organismos internacionais como o UNICEF e o Fundo de População das Nações Unidas (United Nations Population Fund [UNFPA]) adoptam o critério etário como definição operacional das uniões prematuras, o que permite a comparabilidade entre estudos baseados em inquéritos populacionais e a análise do fenómeno como uma variável binária em estudos empíricos (UNICEF, 2016; UNFPA, 2016).

No contexto moçambicano, a persistência das uniões prematuras resulta da interacção entre factores legais, culturais e socioeconómicos. Apesar da existência de instrumentos jurídicos e de políticas públicas orientadas para a prevenção e o combate ao fenómeno, a articulação entre o direito formal e as normas tradicionais, associada a limitações na aplicação efectiva da legislação, contribui para a normalização de uniões envolvendo menores de idade, sobretudo em contextos comunitários marcados por vulnerabilidade social (Bassiano & de Lima, 2018). Esta realidade evidencia uma distância entre o enquadramento normativo formal e as práticas sociais vigentes.

A literatura empírica, tanto nacional como internacional, converge no reconhecimento de que as uniões prematuras resultam da combinação de múltiplos factores individuais, familiares, socioeconómicos e contextuais, os quais interagem de forma complexa e reforçam ciclos de desigualdade social e de género (Alem et al., 2020; Belachew et al., 2022; Mathabatha; & Tsawe, 2023).

2.1.1 Enquadramento institucional e normativo das uniões prematuras em Moçambique

A persistência das uniões prematuras em Moçambique resulta da interacção entre factores legais, culturais e socioeconómicos, num contexto marcado pela articulação entre o direito formal e as normas tradicionais. Embora o país disponha de instrumentos jurídicos destinados à protecção dos direitos da criança e da adolescente, a implementação efectiva da legislação enfrenta desafios associados à fraca disseminação do quadro legal, à limitada fiscalização e à forte influência de práticas socioculturais enraizadas, sobretudo em zonas rurais (Bassiano & de Lima, 2018).

No ordenamento jurídico moçambicano, a Lei da Família estabelece o casamento como um vínculo voluntária orientada para a constituição da família. Contudo, a ausência de uma delimitação etária clara e inequívoca no texto legal, aliada à aceitação social de normas tradicionais, contribui para a continuidade de uniões envolvendo menores de idade em determinados contextos comunitários, evidenciando uma lacuna entre o enquadramento normativo formal e a realidade social.

Em resposta a esta problemática, o Governo de Moçambique tem vindo a implementar políticas públicas orientadas para a prevenção e combate às uniões prematuras, com particular atenção à Estratégia Nacional para a Prevenção e Combate às Uniões Prematuras. A referida estratégia adopta uma abordagem multissectorial, envolvendo sectores como a educação, a saúde, a justiça e a acção social, reconhecendo que a eliminação do fenómeno exige intervenções integradas e sustentadas por evidência empírica (Governo de Moçambique, 2016).

Paralelamente, Moçambique é signatário de diversos compromissos internacionais no domínio dos direitos humanos e da equidade de género, que incluem a protecção das raparigas contra práticas nocivas, como as uniões prematuras. Estes compromissos enquadram-se nos Objectivos de Desenvolvimento Sustentável, nomeadamente na meta de eliminação de todas as práticas prejudiciais contra crianças e adolescentes até 2030 (Organização das Nações Unidas [ONU], 2011; Chilimile, 2022).

Apesar dos avanços registados ao nível das políticas públicas e do enquadramento institucional, a literatura evidencia que a persistência das uniões prematuras permanece associada a desigualdades estruturais no acesso à educação, à informação e às oportunidades económicas, reforçando a necessidade de análises empíricas robustas, baseadas em dados representativos, que apoiem a definição de políticas públicas mais adequadas ao contexto moçambicano.

2.1.2 Contexto internacional das uniões prematuras

A nível global, as uniões prematuras continuam a constituir um fenómeno de elevada prevalência, afectando de forma desproporcional raparigas em cenários caracterizados por condições de pobreza, desigualdade de género e acesso limitado à educação. Apesar dos progressos registados nas últimas décadas no domínio dos direitos humanos, milhões de raparigas em todo o mundo continuam a entrar em união antes dos 18 anos de idade, o que evidencia a persistência do fenómeno (UNICEF, 2021).

A literatura internacional indica que a incidência das uniões prematuras é particularmente elevada em zonas da África Subsaariana e da Ásia Meridional, onde factores estruturais, como normas sociais restritivas, práticas culturais tradicionais, insegurança económica e fraco acesso aos serviços de saúde sexual e reprodutiva, desempenham um papel determinante na manutenção desta prática (UNICEF, 2016). Nestes contextos, a união prematura surge frequentemente como mecanismo de subsistência económica ou forma de protecção social em agregados familiares vulneráveis.

Estudos empíricos realizados em países em desenvolvimento têm identificado associações consistentes entre a ocorrência de uniões prematuras e variáveis como o baixo nível de escolaridade, a pobreza, a residência em meio rural e a iniciação sexual precoce. Estes factores interagem de forma complexa, reforçando ciclos de desigualdade intergeracional e limitando as oportunidades de desenvolvimento das raparigas (Alem et al., 2020; Mathabatha & Tsawe, 2023).

Adicionalmente, a evidência internacional demonstra que o adiamento da entrada em união está fortemente associado à permanência das raparigas no sistema educativo e ao acesso à informação, particularmente através dos meios de comunicação social. Países que registaram reduções significativas na prevalência das uniões prematuras tendem a apresentar melhorias nos indicadores de escolarização feminina, empoderamento económico e acesso a serviços de saúde reprodutiva (UNICEF, 2021).

Neste enquadramento, a análise comparativa de experiências internacionais reforça a importância de abordagens baseadas em evidência empírica para a compreensão dos determinantes das uniões prematuras. A utilização de métodos estatísticos e de técnicas de Machine Learning em estudos internacionais tem permitido identificar padrões e factores de risco com maior precisão, contribuindo para o desenvolvimento de políticas públicas mais eficazes e adequadas às especificidades de cada contexto nacional.

2.1.3 Factores associados às uniões prematuras

A literatura científica tem demonstrado de forma consistente que as uniões prematuras resultam da interacção de múltiplos factores individuais, familiares, socioeconómicos e contextuais. Estes factores actuam de forma interdependente, influenciando o momento da entrada em união e moldando as trajectórias de vida das raparigas, particularmente em contextos caracterizados por desigualdades estruturais (Alem et al., 2020; Mathabatha & Tsawe, 2023).

Entre os factores individuais, a idade da mulher e a idade na primeira relação sexual assumem um papel central na explicação da ocorrência das uniões prematuras. A evidência empírica indica que a iniciação sexual precoce está fortemente associada à antecipação da entrada em união, sobretudo em contextos socioculturais onde a gravidez fora da união é socialmente sancionada, o que frequentemente conduz à formalização precoce da convivência conjugal (Alem et al., 2020).

O nível de escolaridade surge de forma recorrente na literatura como um dos principais factores protectores contra as uniões prematuras. A permanência das raparigas no sistema educativo contribui para o adiamento da entrada em união, ao ampliar oportunidades futuras, fortalecer a autonomia individual e promover maior acesso à informação. Em contraste, níveis educacionais mais baixos, particularmente o ensino primário incompleto ou a ausência de escolaridade, estão associados a uma maior probabilidade de ocorrência de uniões prematuras (Mathabatha & Tsawe, 2023).

O estatuto socioeconómico do agregado familiar constitui igualmente um determinante relevante, sendo frequentemente operacionalizado através de indicadores como o quintil de riqueza. Agregados familiares em situação de pobreza ou vulnerabilidade económica tendem a apresentar maior probabilidade de ocorrência de uniões prematuras, quer como estratégia de sobrevivência económica, quer como forma de redução do encargo financeiro associado à permanência das raparigas no agregado familiar (Belachew et al., 2022).

Para além dos factores individuais e socioeconómicos, variáveis contextuais e culturais exercem influência significativa sobre os padrões de união. A área de residência, a província, a religião, o nível de exposição aos meios de comunicação social e a dimensão do agregado familiar têm sido identificadas como factores associados às uniões prematuras, reflectindo diferenças regionais, normas culturais específicas e níveis diferenciados de acesso à informação e a oportunidades sociais (Hodges, 2015).

De forma geral, a evidência empírica sugere que as uniões prematuras não resultam de um único factor isolado, mas sim de uma combinação complexa de determinantes que variam conforme o contexto social, económico e cultural. Esta multiplicidade de factores justifica a utilização de abordagens analíticas multivariadas, capazes de captar simultaneamente os efeitos individuais e contextuais associados à ocorrência das uniões prematuras, como as adoptadas no presente estudo.

2.1.4 Consequências das uniões prematuras

As uniões prematuras produzem efeitos profundos e duradouros na vida das raparigas, com repercussões significativas nos domínios educacional, económico, social e da saúde. A literatura evidencia que a entrada precoce em união compromete o desenvolvimento integral das raparigas, limita as suas oportunidades futuras e reforça ciclos de desigualdade intergeracional (UNICEF, 2016; Gashaw, 2019).

No domínio educacional, as uniões prematuras estão fortemente associadas ao abandono escolar precoce, uma vez que a entrada em união implica, frequentemente, a assunção de responsabilidades domésticas e reprodutivas incompatíveis com a continuidade dos estudos. A interrupção do percurso escolar reduz as possibilidades de qualificação profissional e restringe o acesso ao mercado de trabalho formal, comprometendo a autonomia económica das raparigas ao longo da trajectória de vida (Gashaw, 2019).

No plano económico, as uniões prematuras contribuem para a manutenção da pobreza, ao limitar a capacidade de geração de rendimento e aumentar a dependência financeira em relação ao parceiro ou ao agregado familiar. Mulheres que entram em união precocemente tendem a apresentar menores níveis de inserção no mercado de trabalho e maior vulnerabilidade económica, reforçando situações de desigualdade e exclusão social.

No campo da saúde, diversos estudos indicam que as uniões prematuras estão associadas a um risco acrescido de gravidez precoce, complicações obstétricas e aumento da mortalidade materna e infantil. A gravidez em idades muito jovens encontra-se frequentemente associada a uma maior incidência de partos prematuros, baixo peso à nascença e dificuldades no acesso a cuidados de saúde adequados, sobretudo em contextos marcados por limitações estruturais nos serviços de saúde (UNICEF, 2016).

Adicionalmente, as uniões prematuras tendem a aumentar a vulnerabilidade das raparigas à violência doméstica, psicológica e sexual, bem como a limitar a sua capacidade de tomada de decisão no seio da união. A assimetria de poder frequentemente observada em uniões envolvendo menores de idade favorece a continuidade de assimetrias de género e para a normalização de práticas abusivas, com impactos negativos no bem-estar físico e emocional das raparigas (Hodges, 2015).

De forma geral, as consequências das uniões prematuras ultrapassam a esfera individual, produzindo efeitos adversos ao nível familiar, comunitário e social. Estes impactos reforçam a exigência de políticas adequadas e de intervenções baseadas em evidência empírica, orientadas para a prevenção do fenómeno e para a promoção do desenvolvimento pleno das raparigas e jovens mulheres.

2.1.5 Abordagens metodológicas em estudos sobre uniões prematuras

A análise empírica das uniões prematuras tem recorrido a diferentes abordagens metodológicas, seleccionadas em função dos objectivos do estudo, da natureza dos dados e da complexidade do fenómeno em análise. De forma geral, os estudos nesta área têm utilizado tanto métodos estatísticos clássicos como abordagens mais recentes baseadas em técnicas de Machine Learning.

No âmbito dos métodos estatísticos clássicos, são frequentemente aplicados modelos adequados ao estudo de desfechos categóricos, com particular destaque para a regressão logística. Contudo, a literatura evidencia também o uso de outras abordagens, como modelos de regressão multinível, que permitem considerar a estrutura hierárquica dos dados (por exemplo, indivíduos inseridos em comunidades), bem como técnicas de análise de sobrevivência, aplicadas à análise do tempo até à ocorrência da primeira união (Cox, 1972; Goldstein, 2011). Adicionalmente, modelos lineares generalizados têm sido empregados para explorar relações entre variáveis explicativas e o fenómeno em estudo em diferentes contextos (McCullagh & Nelder, 1989).

Entre os métodos mais utilizados, a regressão logística destaca-se pela sua adequação à modelação de variáveis resposta binárias, viabilizando a estimação da probabilidade de ocorrência de uniões prematuras em função de múltiplos factores explicativos. Este modelo tem sido amplamente aplicado em estudos demográficos e de saúde pública, não só pela sua robustez estatística, mas também pela facilidade de interpretação dos resultados através das razões de chances (*odds ratios*) (Maroco, 2007; Hosmer, Lemeshow & Sturdivant, 2013; Gashaw, 2019).

A regressão logística apresenta vantagens relevantes em estudos observacionais baseados em inquéritos populacionais, uma vez que possibilita o controlo simultâneo de variáveis demográficas, socioeconómicas e contextuais, contribuindo para uma interpretação estatística clara e teoricamente fundamentada dos factores associados às uniões prematuras. Adicionalmente, a incorporação de ajustes para o desenho amostral complexo, comum em inquéritos como o Inquérito Demográfico e de Saúde, permite obter estimativas mais fiáveis e representativas da população em estudo (Heeringa, West & Berglund, 2017).

Mais recentemente, métodos de *Machine Learning* têm vindo a ser incorporadas em estudos sociais e demográficos como complemento aos métodos estatísticos tradicionais. Estas técnicas distinguem-se pela sua capacidade de captar relações não lineares, identificar interacções complexas entre variáveis e revelar padrões ocultos nos dados, frequentemente não detectados por modelos paramétricos clássicos (James et al., 2021).

Entre os algoritmos mais utilizados destacam-se métodos como árvores de decisão, *Random Forest, Support Vector Machines* (SVM) bem como redes neurais artificiais, para além de métodos baseados em *boosting*, como o Extreme Gradient Boosting (XGBoost), amplamente reconhecido pela sua elevada capacidade preditiva e eficiência computacional (Breiman, 2001; Cortes

& Vapnik, 1995; Chen & Guestrin, 2016; Santos et al., 2019; Freitas, 2023).

O XGBoost tem demonstrado elevada eficácia em tarefas de classificação binária, permitindo avaliar a importância relativa das variáveis explicativas e optimizar o desempenho dos modelos através de procedimentos como o ajuste de hiperparâmetros, a validação cruzada e a regularização. A utilização deste algoritmo em estudos sobre uniões prematuras tem demonstrado ganhos na capacidade preditiva dos modelos, contribuindo para uma identificação mais precisa de grupos em maior risco.

Deste modo, a combinação entre métodos explicativos clássicos, como a regressão logística, e algoritmos de Machine Learning permite uma abordagem analítica complementar, na qual a regressão logística assegura interpretabilidade estatística, enquanto as técnicas de Machine Learning reforçam a capacidade preditiva dos modelos. Esta integração metodológica revela-se particularmente adequada em estudos aplicados, ao conciliar rigor estatístico, coerência analítica e desempenho preditivo, conforme adoptado no presente estudo para a análise dos factores associados às uniões prematuras em Moçambique.

2.1.6 Aprendizado de Máquina

O *Aprendizado de Máquina* (AM) constitui um subcampo da Inteligência Artificial voltado ao desenvolvimento de métodos capazes de aprender automaticamente com base em dados, identificando padrões e relações relevantes sem exigir programação explícita para cada tarefa específica, (Geron, 2017). Estas técnicas têm sido amplamente empregadas na resolução de problemas complexos, para os quais abordagens tradicionais baseadas em regras fixas se revelam limitadas ou pouco eficientes (Pouyanfar, 2019).

De forma geral, o processo de aprendizagem em Machine Learning pode ser dividido em duas etapas principais. A etapa inicial, designada *treino*, corresponde à exposição do modelo a um conjunto de dados, permitindo a extracção de padrões e a aquisição de conhecimento a partir das informações disponíveis. A segunda etapa, denominada *validação*, tem como objectivo avaliar o desempenho do modelo com base em medidas quantitativas, verificando a sua capacidade de generalização e a eficácia na execução da tarefa proposta.

As aplicações do Aprendizado de Máquina abrangem diversas áreas, incluindo o reconhecimento de voz e de imagem, a filtragem de mensagens não desejadas (*spam*), os sistemas de recomendação e a previsão de valores em contextos económicos e financeiros (Geron, 2017). Adicionalmente, o AM tem assumido um papel relevante na área da segurança informática, nomeadamente na detecção e prevenção de ataques de negação de serviço (DoS) e ataques distribuídos de negação de serviço (DDoS), nos quais a identificação de padrões recorrentes no tráfego de rede é fundamental, (Berman, 2019; Gormez, 2020).

De acordo com o nível de supervisão adoptado durante o treino, as técnicas de Aprendizado de Máquina podem ser classificados em três categorias principais (Souza, 2020):

- **Aprendizagem Supervisionada**, em que o algoritmo é justado com dados rotulados, isto é, com respostas previamente conhecidas;
- **Aprendizagem Não Supervisionada**, que procura detectar padrões em dados não rotulados, agrupando observações com características semelhantes;
- **Aprendizagem Semi-Supervisionada**, que combina dados rotulados e não rotulados, mostrando-se particularmente útil em contextos caracterizados por grandes volumes de dados e escassez de rótulos.

Nos últimos anos, o campo tem-se expandido com o surgimento de abordagens híbridas, como a aprendizagem auto-supervisionada, que utiliza grandes volumes de dados não rotulados para melhorar a capacidade de generalização dos algoritmos (Pereira, 2023).

Em termos operacionais, os modelos de Aprendizado de Máquina dependem da recolha de dados provenientes de múltiplas fontes, os quais são utilizados na fase de treino para o ajustamento dos parâmetros do modelo, (Cavalcante, 2022). A partir deste processo, o modelo torna-se capaz de representar relações subjacentes aos dados e identificar regularidades ou estruturas matemáticas relevantes. Contudo, o desempenho dos algoritmos está fortemente condicionado pela qualidade dos dados utilizados. Dados incompletos, ruidosos ou enviesados conduzem a representações inadequadas do fenómeno em estudo e, consequentemente, a modelos com menor capacidade preditiva, (Chagas, 2024).

Nesta dissertação, recorre-se a métodos de **aprendizagem supervisionada**, nomeadamente à Regressão Logística e ao algoritmo XGBoost, com vista a analisar os factores associados às uniões prematuras em Moçambique.

2.1.7 Aprendizagem Supervisionada

No âmbito do Aprendizado de Máquina, a aprendizagem supervisionada destaca-se como uma abordagem amplamente utilizada para a construção de modelos preditivos, sobretudo em situações em que a variável de interesse é previamente conhecida. Esta abordagem assenta na utilização de conjuntos de dados rotulados, nos quais cada observação é composta por um conjunto de variáveis explicativas e pela respectiva resposta associada, permitindo ao algoritmo aprender a relação subjacente entre entradas e saídas a partir dos dados disponíveis (Han, 2011).

Em função da natureza da variável resposta, os problemas de aprendizagem supervisionada podem ser organizados em duas categorias principais: *classificação* e *regressão*. Nos problemas

de classificação, o objectivo consiste em atribuir cada observação a uma classe previamente definida, sendo esta abordagem aplicada, por exemplo, na filtragem de mensagens indesejadas (*spam*), no apoio ao diagnóstico médico e na identificação de eventos específicos em bases de dados complexas (Silva, 2023). Por outro lado, os problemas de regressão visam estimar valores numéricos contínuos, modelando a relação funcional existente entre a variável dependente e um conjunto de variáveis explicativas, sendo frequentemente utilizados em tarefas de previsão. Entre os modelos de regressão mais utilizados destacam-se a regressão linear, a regressão polinomial bem como a regressão logística, particularmente adequada quando o desfecho assume natureza binária (Silva, 2023).

O funcionamento da aprendizagem supervisionada desenvolve-se de forma sequencial, abrangendo um conjunto de etapas fundamentais. Inicialmente, procede-se à recolha de dados rotulados, assegurando que cada observação possui uma resposta conhecida. Em seguida, na fase de treino, o algoritmo analisa os padrões existentes entre as variáveis de entrada e a respectiva variável resposta, ajustando os seus parâmetros internos com o objectivo de minimizar os erros de previsão. Posteriormente, o modelo é submetido a processos de validação e teste, recorrendo a dados independentes do treino, com vista a avaliar a sua capacidade de generalização. Por fim, após demonstrar desempenho adequado, o modelo é implementado em contextos reais para efectuar previsões ou classificações de novos casos.

De modo geral, a eficácia dos modelos de aprendizagem supervisionada depende fortemente da qualidade dos dados, da correcta selecção das variáveis explicativas e da adequação do algoritmo ao problema em estudo. Assim, a definição criteriosa da tarefa de aprendizagem e das etapas do processo constitui um elemento central para a aplicação bem-sucedida destas técnicas em estudos empíricos baseados em dados reais.

Neste estudo, analisam-se os factores associados às uniões prematuras em Moçambique, recorrendo à aprendizagem supervisionada como complemento aos métodos estatísticos tradicionais, de modo a identificar padrões nos dados e estimar a probabilidade de ocorrência da união prematura, bem como comparar o desempenho preditivo dos modelos aplicados.

2.1.8 Síntese crítica da literatura

A literatura sobre uniões prematuras tem recorrido a diferentes abordagens metodológicas para explicar a natureza multifactorial do fenómeno. Segundo Hosmer & Lemeshow (2000), a regressão logística destaca-se entre os métodos mais utilizados na análise de variáveis resposta binárias, dada a sua capacidade de estimar probabilidades condicionais e interpretar os efeitos das variáveis explicativas por meio das razões de chances. Esta perspectiva é reforçada por Menard (2002), que destaca a utilidade do modelo em estudos observacionais, sobretudo quando

se pretende identificar factores associados a eventos dicotómicos.

No estudo conduzido por Gashaw (2019), utilizando dados do *Demographic and Health Survey*, a regressão logística foi aplicada para analisar os determinantes da união prematura. Os resultados reportados pelo autor evidenciaram associação estatisticamente significativa entre baixo nível de escolaridade, estatuto socioeconómico reduzido e maior probabilidade de entrada precoce em união. Contudo, o próprio estudo reconhece que o modelo assume linearidade no logit e independência entre observações, o que pode limitar a captação de interacções complexas ou efeitos contextuais não especificados.

Hair et al. (2009) argumentam que fenómenos sociais complexos podem envolver relações não lineares e estruturas de dependência que desafiam os pressupostos de modelos paramétricos tradicionais. Nesse sentido, alguns estudos têm recorrido a modelos multinível para captar a variabilidade existente entre diferentes contextos sociais e territoriais. Por exemplo, Belachew et al. (2022) analisaram dados demográficos sobre uniões prematuras em países africanos e aplicaram modelos multinível para avaliar simultaneamente factores individuais e contextuais, demonstrando que características comunitárias, como normas sociais e condições socioeconómicas locais, influenciam significativamente a probabilidade de ocorrência do fenómeno. Esta abordagem amplia a compreensão estrutural das uniões prematuras, ao permitir considerar a influência de diferentes níveis de análise. Contudo, a sua aplicação implica maior complexidade analítica e exige uma especificação adequada da estrutura hierárquica dos dados.

Quando o foco incide sobre o momento de ocorrência da primeira união, Kleinbaum & Klein (2012) defendem a aplicação de modelos de sobrevivência estatística, particularmente o modelo de Cox para riscos proporcionais, por permitirem modelar o tempo até à ocorrência do evento. Esta perspectiva temporal acrescenta profundidade analítica ao estudo do fenómeno. Todavia, quando o objectivo consiste em analisar a ocorrência ou não da união antes de determinado limiar etário, a modelação binária mantém-se metodologicamente consistente.

Mais recentemente, técnicas de *Machine Learning* têm sido incorporadas como alternativa ou complemento aos modelos estatísticos tradicionais. De acordo com Géron (2017), algoritmos baseados em árvores possuem elevada capacidade para captar associações de natureza não linear e interacções complexas envolvendo diferentes variáveis explicativas. Alem et al. (2020) aplicaram métodos de aprendizagem supervisionada no Bangladesh e reportaram melhorias no desempenho preditivo quando comparados à regressão logística tradicional, particularmente na capacidade de identificação de casos positivos. De forma semelhante, Mathabatha & Tsawe (2023) conduziram um estudo na África do Sul e observaram maior precisão na classificação de grupos em risco através de modelos baseados em árvores.

Entre os métodos de *boosting*, Chen & Guestrin (2016) introduziram o XGBoost como um algoritmo optimizado para problemas de classificação, incorporando mecanismos de regulari-

zação que reduzem o sobreajustamento e melhoram a generalização do modelo. Os resultados reportados em diferentes aplicações evidenciam ganhos substanciais em desempenho preditivo.

Não obstante, Freitas (2023) e Pereira (2023) alertam que, apesar da elevada capacidade preditiva, muitos algoritmos de aprendizagem automática apresentam menor transparência interpretativa quando comparados a modelos paramétricos. Tal limitação pode comprometer a explicação substantiva dos mecanismos subjacentes ao fenómeno, aspecto particularmente relevante em estudos orientados para a formulação de políticas públicas.

Assim, a literatura revela uma tensão metodológica entre abordagens explicativas, centradas na inferência estatística e interpretação dos efeitos das variáveis explicativas, e abordagens preditivas, orientadas para maximização do desempenho preditivo. Apesar dos avanços registados internacionalmente, permanece limitada, no contexto moçambicano, a integração sistemática destas abordagens e a comparação explícita do seu desempenho na análise das uniões prematuras.

Face a este enquadramento, o presente estudo adopta uma estratégia metodológica complementar, combinando regressão logística e XGBoost, de modo a assegurar simultaneamente interpretabilidade estatística e robustez preditiva na análise dos factores relacionados com as uniões prematuras em Moçambique.

2.2 Modelo de Regressão Logística

A regressão logística corresponde a um método estatístico frequentemente utilizado em situações nas quais a variável resposta apresenta natureza binária, isto é, quando o fenómeno de interesse pode ser representado por dois resultados possíveis, como a ocorrência ou não de um determinado evento. Neste enquadramento, o objectivo do modelo consiste em estimar a probabilidade de ocorrência do fenómeno a partir de um conjunto de variáveis explicativas ou covariáveis, (Batista, 2015).

Diferentemente da regressão linear, cuja variável resposta é de natureza contínua, a regressão logística permite analisar a associação entre variáveis explicativas e uma variável resposta dicotómica, garantindo que as probabilidades previstas permaneçam compreendidas no intervalo $[0, 1]$. Esta característica torna o modelo particularmente adequado para análises preditivas e problemas de classificação binária.

2.2.1 Formulação Matemática do Modelo de Regressão Logística

Considera-se uma variável resposta Y, tal que $Y = 1$ indica a realização do evento em estudo e $Y = 0$ a sua não ocorrência. A regressão logística estima a probabilidade $\Pr(Y = 1 \mid X_1, X_2, \ldots, X_p)$ através da seguinte função logística:

$$\Pr(Y = 1 \mid X_1, \ldots, X_p) = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}. \tag{2.1}$$

Nesta expressão, β_0 representa o intercepto do modelo e $\beta_1, \ldots, \beta_p$ são os coeficientes associados às variáveis explicativas. Cada coeficiente traduz o efeito da respectiva covariável sobre a probabilidade de ocorrência do evento, mantendo as demais constantes (Hosmer & Lemeshow, 2000).

2.2.2 Função de Ligação Logit

A regressão logística utiliza a função de ligação *logit*, que estabelece uma relação linear entre as covariáveis e o logaritmo da razão de probabilidades (odds) do evento. A função logit é definida por:

$$\text{logit}(Y_i) = \ln\left(\frac{\Pr(Y_i = 1)}{1 - \Pr(Y_i = 1)}\right) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi}, \tag{2.2}$$

onde o índice i identifica cada observação da amostra. Esta transformação permite interpretar os coeficientes do modelo em termos de variações logarítmicas nas chances de ocorrência do evento, sendo frequentemente reportados sob a forma de *odds ratios* (Hair et al, 2009).

2.2.3 Chance e Razão de Chances

Na análise de fenómenos binários, como a ocorrência ou não de um determinado evento, a utilização directa da probabilidade pode apresentar limitações, uma vez que modelos lineares podem gerar previsões fora do intervalo $[0, 1]$. Para ultrapassar esta restrição, a regressão logística baseia-se no conceito de *chance* (*odds*), em vez de trabalhar directamente com a probabilidade do evento (Menard, 2002).

A *chance* ocorrência do evento pode ser expressa pela relação entre a probabilidade de ocorrência do sucesso e a probabilidade de insucesso, sendo expressa por:

$$\text{chance} = \frac{P(\text{sucesso})}{1 - P(\text{sucesso})}. \tag{2.3}$$

Enquanto a probabilidade assume valores entre 0 e 1, a chance pode variar entre 0 e $+\infty$, permitindo uma representação mais flexível da intensidade do fenómeno. Valores elevados de chance estão associados a maior propensão para a ocorrência do evento, ao passo que valores reduzidos indicam menor probabilidade (Mesquita, 2014).

Na regressão logística, a interpretação do efeito das variáveis explicativas é feita com base na *razão de chances (RC)*, que permite comparar a ocorrência do evento entre dois grupos ou entre distintas categorias de uma variável explicativa. A razão de chances pode ser expressa por:

$$\text{RC} = \frac{\text{chance do evento ocorrer num grupo}}{\text{chance do evento ocorrer no grupo de referência}}. \tag{2.4}$$

A razão de chances corresponde a uma medida de efeito frequentemente utilizada, permitindo analisar o efeito das variáveis explicativas relativamente à variável dependente. Quando $\text{RC} > 1$ indicam maior probabilidade de ocorrência do evento, enquanto valores de $\text{RC} < 1$ sugerem diminuição dessa probabilidade; valores próximos de 1 apontam para ausência de associação significativa (Hosmer & Lemeshow, 2000).

A regressão logística recorre a função de ligação *logit*, definida como o logaritmo natural da chance, definindo uma associação linear entre as variáveis explicativas e o logaritmo da razão de chances. Esta transformação assegura que as probabilidades previstas permaneçam no intervalo $[0, 1]$ e permite interpretar os coeficientes do modelo de forma directa, uma vez que o exponencial de cada coeficiente corresponde à respectiva razão de chances, (Mesquita, 2014).

Deste modo, a utilização da chance e da razão de chances constitui um elemento central na regressão logística, proporcionando uma interpretação intuitiva e estatisticamente consistente dos efeitos das variáveis explicativas sobre a probabilidade de verificação do evento em estudo.

2.2.4 Estimação dos Parâmetros

Segundo Hosmer e Lemeshow (2000), quando a variável resposta assume natureza dicotómica, o enquadramento da regressão linear deixa de ser adequado, uma vez que o erro associado ao modelo não apresenta distribuição Normal com variância constante. Nestes casos, regressão logística constitui uma alternativa apropriada, considerando que a variável resposta segue uma distribuição Bernoulli, cuja média corresponde a probabilidade de verificação do evento de interesse.

Segundo Cabral (2013), considere-se Y_i como a variável resposta correspondente à i-ésima observação, tal que $Y_i = 1$ indica a ocorrência do evento e $Y_i = 0$ a sua não ocorrência. A probabilidade condicional de sucesso pode ser representada por $\pi(x_i) = \Pr(Y_i = 1 \mid X_i)$, sendo modelada através da função logística. Neste enquadramento, o erro aleatório assume

apenas dois valores possíveis, com média nula e variância $\pi(x_i)[1-\pi(x_i)]$, caracterizando uma distribuição de Bernoulli.

Assumindo este pressuposto, os parâmetros da regressão logística são obtidos pelo método da máxima verossimilhança. Este procedimento tem como objectivo determinar os valores do vector de parâmetros $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ que maximizam a função de verossimilhança associada aos dados observados, (Hosmer & Lemeshow, 2000).

A função de verossimilhança pode ser expressa como:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_i)^{y_i} \left[1 - \pi(x_i)\right]^{1-y_i}, \tag{2.5}$$

em que $\pi(x_i) = \Pr(Y_i = 1 \mid X_i)$. Para efeitos computacionais, é prática comum maximizar a forma logarítmica da verossimilhança, dada por:

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \Big\{ y_i \log \pi(\mathbf{x}_i) + (1 - y_i) \log[\, 1 - \pi(\mathbf{x}_i)\,] \Big\}. \tag{2.6}$$

Os avanços computacionais tornaram o processo de maximização da verossimilhança eficiente e amplamente acessível, permitindo a aplicação da regressão logística em estudos empíricos com grandes bases de dados, sem perda de precisão estatística (Batista, 2015).

Desta forma, o método da máxima verossimilhança assegura estimativas consistentes e assintoticamente eficientes para os parâmetros do modelo.

2.3 Testes de Significância dos Parâmetros

2.3.1 Teste de Razão de Verossimilhança

Segundo Nelder e Wedderburn (1972), a função associada à verossimilhança constitui a base inferencial dos modelos lineares generalizados, permitindo verificar o ajustamento do modelo e a importância estatística das variáveis explicativas incluídas na análise. Na regressão logística, o teste da razão de verossimilhança é frequentemente utilizado para avaliar a significância global do modelo, assim como para efectuar comparações entre modelos aninhados, ou seja, modelos que diferem apenas pela inclusão ou exclusão de determinadas variáveis.

De acordo com Mesquita (2014), a avaliação da significância de uma variável explicativa pode ser realizada comparando-se dois modelos: um modelo completo, que inclui a variável em análise, e um modelo reduzido, no qual essa variável é excluída. A comparação baseia-se

nos valores da função de verossimilhança associados a cada modelo, medindo-se a perda de ajustamento provocada pela retirada da variável.

A estatística de teste utilizada é definida como:

$$D = -2\ln\left(\frac{\text{verossimilhança do modelo ajustado}}{\text{verossimilhança do modelo saturado}}\right), \tag{2.7}$$

sendo esta quantidade designada por *deviance*. Conforme salientado por Cabral (2013), a deviance desempenha, na regressão logística, um papel análogo ao da soma dos quadrados dos resíduos utilizada na regressão linear, constituindo uma medida do nível de ajustamento do modelo aos dados observados. Menores valores de *deviance* indicam melhor ajustamento do modelo.

Com vista a avaliar a contribuição estatística de uma determinada variável explicativa, considera-se a diferença entre as deviances dos modelos comparados. Esta diferença é expressa pela estatística:

$$G = D_{\text{modelo reduzido}} - D_{\text{modelo completo}}, \tag{2.8}$$

ou, de forma equivalente,

$$G = -2\ln\left(\frac{\text{verossimilhança do modelo sem a variável}}{\text{verossimilhança do modelo com a variável}}\right). \tag{2.9}$$

Segundo Mesquita (2014), sob a hipótese nula de que o coeficiente associado à variável removida é igual a zero, a estatística G segue, assintoticamente, uma distribuição qui-quadrado, cujos graus de liberdade correspondem a diferença entre o número de parâmetros dos modelos em comparação. Assim, a decisão estatística é obtida comparando o valor observado de G com o valor crítico da distribuição χ^2 associado ao nível de significância previamente definido.

De acordo com Cabral (2013), quando a rejeição da hipótese nula indica que a variável em análise contribui significativamente para a explicação da variável resposta. O teste permite verificar a significância global do modelo, contribuindo igualmente para o processo de selecção de variáveis, sendo frequentemente utilizado em conjunto com outros testes, como o teste de Wald, para uma análise mais detalhada dos coeficientes individuais.

2.3.2 Teste de Wald

Segundo Vaz (2020), o teste de Wald constitui um procedimento estatístico amplamente utilizado no contexto da regressão logística para avaliar a significância individual dos coeficientes associados às variáveis explicativas. Este teste baseia-se na comparação entre a estimativa do parâmetro obtida pelo método da máxima verossimilhança e o respectivo erro padrão, permi-

tindo verificar se o coeficiente é estatisticamente diferente de zero.

De acordo com este enquadramento, as hipóteses testadas para cada coeficiente β_j são formuladas como:

$$H_0 : \beta_j = 0 \quad \text{versus} \quad H_1 : \beta_j \neq 0,$$

sendo $j = 1, 2, \ldots, p$.

A estatística do teste de Wald pode ser expressa por:

$$W = \frac{\hat{\beta}_j}{\sqrt{\widehat{\text{Var}}(\hat{\beta}_j)}}, \tag{2.10}$$

onde $\hat{\beta}_j$ representa a estimativa do coeficiente e $\widehat{\text{Var}}(\hat{\beta}_j)$ corresponde à variância estimada do estimador. Sob a hipótese nula, a estatística W segue, assintoticamente, uma distribuição Normal Padrão. De modo equivalente, o quadrado da estatística de Wald segue uma distribuição qui-quadrado com um grau de liberdade.

Conforme salientado por Vaz (2020), o teste de Wald apresenta a vantagem de ser simples de calcular e de fácil interpretação, sendo frequentemente empregue para analisar a relevância individual das variáveis no modelo. No entanto, em determinadas situações, nomeadamente quando os coeficientes assumem valores elevados ou quando as amostras são pequenas, o teste pode apresentar desempenho insatisfatório.

Conforme Hauck e Donner (1977), o teste de Wald pode apresentar falhas na rejeição da hipótese nula, mesmo em situações em que o coeficiente apresenta significância estatística. Deste modo, recomenda-se que os resultados obtidos pelo teste de Wald sejam interpretados em conjunto com outros testes, como o teste da razão de verossimilhança, com vista a garantir uma avaliação mais robusta da significância estatística dos parâmetros do modelo.

2.3.3 Teste Escore de Rao

Segundo Rao (1948), o teste escore, também designado teste do multiplicador de Lagrange, utilizado na análise da significância dos parâmetros em modelos de regressão logística, baseando-se no gradiente da função de log-verossimilhança avaliada sob a hipótese nula. Uma das vantagens deste teste consiste no facto de ou não requer o ajustamento do modelo completo, sendo suficiente estimar o modelo reduzido (Hosmer & Lemeshow, 2000).

A estatística do teste escore pode ser expressa por:

$$S = \mathbf{U}(\boldsymbol{\beta}_0)^\top \mathbf{I}(\boldsymbol{\beta}_0)^{-1} \mathbf{U}(\boldsymbol{\beta}_0), \tag{2.11}$$

em que $\mathbf{U}(\boldsymbol{\beta}_0)$ representa o vector escore e $\mathbf{I}(\boldsymbol{\beta}_0)$ a matriz de informação proposta por Fisher estimada.

Neste enquadramento, a hipótese nula é definida como:

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0,$$

onde $\boldsymbol{\beta}_0$ corresponde ao valor especificado para o parâmetro sob a hipótese nula, sendo frequentemente igual ao vector nulo no contexto da regressão logística.

Assim, tanto o vector escore $\mathbf{U}(\boldsymbol{\beta}_0)$ como a matriz de informação $\mathbf{I}(\boldsymbol{\beta}_0)$ são avaliados no ponto $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, isto é, sob a hipótese nula.

De acordo com Agresti (2013), sob a hipótese nula, a estatística S segue, assintoticamente, uma distribuição qui-quadrado, cujos graus de liberdade correspondem ao número de restrições impostas. Conforme salientado por Mesquita (2014), o teste escore constitui uma alternativa complementar aos testes de Wald e da razão de verossimilhança, sendo particularmente útil em etapas preliminares de selecção de variáveis.

2.4 Avaliação do Ajuste e do Desempenho Preditivo dos Modelos

2.4.1 Pseudo R^2

Nos modelos de regressão logística, a qualidade do ajustamento não pode ser avaliada por meio do coeficiente de determinação tradicional R^2, amplamente utilizado na regressão linear, visto que a variável resposta apresenta natureza dicotómica e os pressupostos desse modelo não são satisfeitos. Por esta razão, foram desenvolvidas medidas alternativas, conhecidas como pseudo R^2, as quais são baseadas na função de verossimilhança do modelo ajustado (Mesquita, 2014; Hosmer & Lemeshow, 2000).

A designação pseudo R^2 decorre do facto de estas medidas apresentarem uma escala semelhante à do R^2 clássico, variando tipicamente entre 0 e 1, em que valores mais elevados indicam melhor ajustamento do modelo aos dados. Contudo, ao contrário do R^2 da regressão linear, os pseudo R^2 não representam a proporção da variabilidade explicada da variável resposta, devendo ser interpretados com cautela e utilizados essencialmente como indicadores comparativos do ajuste entre modelos (Agresti, 2013).

Importa salientar que existem diferentes definições de pseudo R^2, tais como as propostas por Cox e Snell, Nagelkerke e McFadden, as quais podem produzir valores substancialmente dis-

tintos para o mesmo conjunto de dados. Assim, estas medidas devem ser utilizadas de forma complementar a outros critérios de avaliação do modelo, nomeadamente testes de verossimilhança, estatísticas de classificação e medidas de discriminação (Hosmer & Lemeshow, 2000; Agresti, 2013).

2.4.2 Teste de Hosmer–Lemeshow

O teste de Hosmer–Lemeshow corresponde a um dos métodos mais utilizados para avaliar o ajustamento de modelos de regressão logística, baseando-se na comparação entre as probabilidades estimadas pelo modelo e as frequências observadas nos dados. Segundo Hosmer & Lemeshow (2000), este procedimento permite verificar se o modelo estimado representa adequadamente o comportamento da variável resposta nos diferentes níveis de risco estimados.

De acordo com Vaz (2020), o procedimento consiste em ordenar as observações segundo as probabilidades previstas e agrupá-las em g grupos, geralmente $g = 10$, de dimensão aproximadamente igual. Em cada grupo, comparam-se o número observado de ocorrências do evento com o número esperado, obtido a partir das probabilidades estimadas pelo modelo.

As hipóteses do teste podem ser expressas da seguinte forma:

H_0 : Os valores observados e esperados pelo modelo não diferem significativamente;

H_1 : Os valores observados e esperados pelo modelo diferem significativamente.

Neste contexto, o *modelo* refere-se à distribuição teórica (ou função) que se pretende testar face aos dados recolhidos, sendo o investigador quem ajusta esse modelo aos dados observados, e não o inverso.

Estatística do teste é calculada com base na estatística qui-quadrado de Pearson, sendo expressa por:

$$C = \sum_{k=1}^{g} \frac{(o_k - e_k)^2}{e_k \left(1 - \frac{e_k}{n_k}\right)},$$

onde o_k representa o número de ocorrências observadas no k-ésimo grupo, e_k o número esperado de ocorrências e n_k o total de observações pertencentes a esse grupo.

Segundo Hosmer e Lemeshow (2000), sob a hipótese nula e assumindo que o modelo foi correctamente especificado, a estatística C segue aproximadamente uma distribuição qui-quadrado com $(g - 2)$graus de liberdade. Valores elevados da estatística sugerem rejeição da hipótese nula, indicando evidência de mau ajustamento do modelo.

Conforme salientado por Cabral (2013), o teste de Hosmer e Lemeshow analisa essencialmente a capacidade de calibração do modelo, isto é, a concordância entre as probabilidades estimadas e os valores observados, devendo ser interpretado de forma complementar a outras medidas de desempenho, como a capacidade discriminatória e critérios baseados na verossimilhança.

2.4.3 Curva ROC e Medidas de Desempenho do Modelo

A curva ROC (*Receiver Operating Characteristic*) corresponde a uma ferramenta amplamente utilizada para avaliar o desempenho de modelos de classificação binária, como a regressão logística. Esta curva possibilita avaliar a capacidade de discriminação do modelo para distintos valores do ponto de corte, representando graficamente a taxa de sensibilidade (taxa de verdadeiros positivos) em função de 1 e especificidade (taxa de falsos positivos) (Hosmer & Lemeshow, 2013; Agresti, 2013).

Ao variar o ponto de corte aplicado às probabilidades estimadas, cada observação pode ser classificada como evento ou não evento, permitindo avaliar o compromisso entre sensibilidade e especificidade. Em termos práticos, o ponto de corte considerado óptimo é, em geral, aquele mais próximo da região superior esquerda da curva ROC, na qual ambos os indicadores assumem valores elevados (Fawcett, 2006).

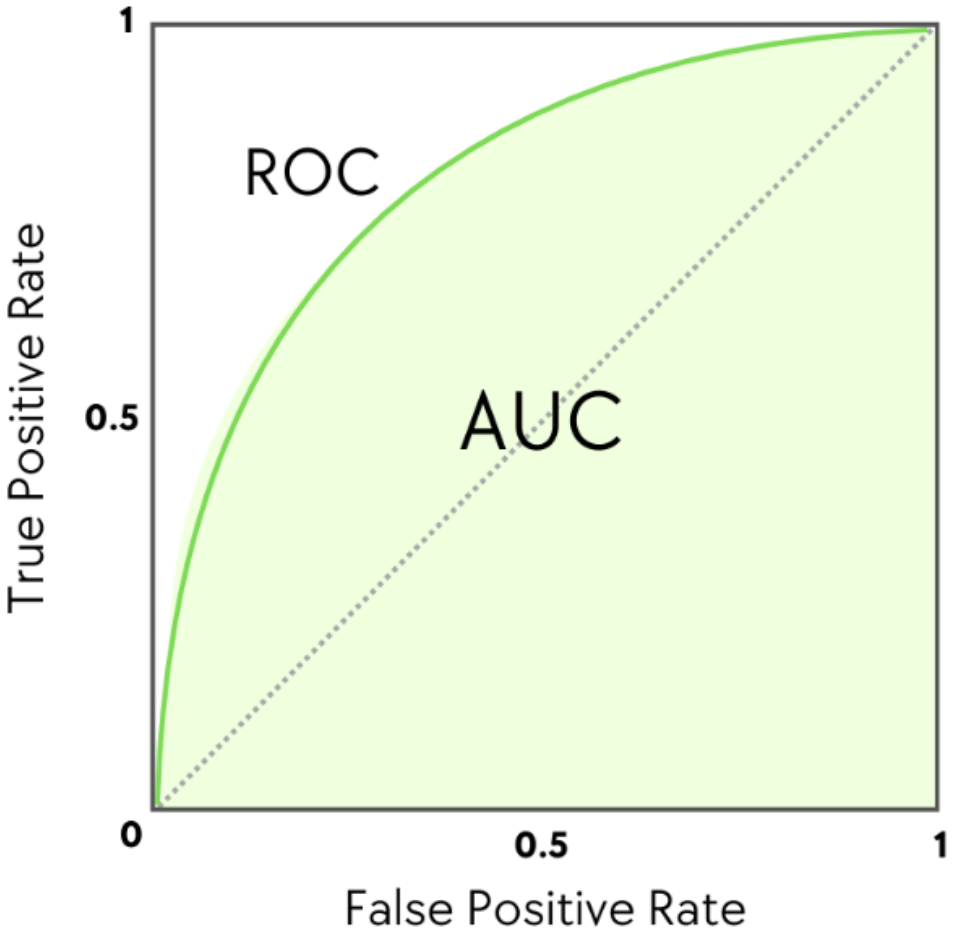


Figura 2.1: Curva ROC (Receiver Operating Characteristic).

Fonte: Adaptado de Sturdivant (2013).

2.4.4 Matriz de Confusão

A base para o cálculo das medidas de desempenho é a matriz de confusão 2×2, conforme ilustra na Tabela 2.1 (Fawcett, 2006).

Tabela 2.1: Estrutura da matriz de confusão utilizada em modelos de classificação binária

Classificação / Realidade	**Evento (+)**	**Não evento (–)**	**Total**
Previsto (+)	VP	FP	VP + FP
Previsto (–)	FN	VN	FN + VN
Total	VP + FN	FP + VN	N

VP: Verdadeiro Positivo; FP: Falso Positivo; FN: Falso Negativo; VN: Verdadeiro Negativo.

2.5 Sensibilidade e Especificidade

Com base na matriz de confusão, definem-se as principais medidas utilizadas para avaliar o desempenho do modelo.

A **sensibilidade** corresponde à probabilidade de o modelo identificar correctamente um indivíduo que apresenta o evento:

$$\text{Sensibilidade} = \frac{VP}{VP + FN}. \tag{2.12}$$

A **especificidade** representa a probabilidade de o modelo identificar correctamente um indivíduo que não apresenta o evento:

$$\text{Especificidade} = \frac{VN}{VN + FP}. \tag{2.13}$$

A taxa de falsos positivos é definida por:

$$1 - \text{Especificidade} = \frac{FP}{FP + VN}. \tag{2.14}$$

Cada ponto representado na curva ROC resulta de um determinado valor específico do ponto de corte aplicado às probabilidades previstas pelo modelo.

2.6 Área Abaixo da Curva ROC

A área sob a curva ROC (AUC) constitui uma medida global da capacidade de discriminação do modelo, assumindo valores entre 0 e 1. Intuitivamente, a AUC corresponde à probabilidade de o modelo atribuir uma probabilidade estimada mais elevada a um indivíduo com evento do que a um indivíduo sem evento (Steyerberg, 2010; Hosmer & Lemeshow, 2013).

A interpretação da AUC segue, de forma geral, as orientações descritas na Tabela 2.2.

Tabela 2.2: Critérios de interpretação da área sob a curva ROC

Intervalo da AUC	**Interpretação**
AUC = 0,5	Ausência de discriminação
$0,5 <$ AUC $< 0,7$	Discriminação fraca
$0,7 \leq$ AUC $< 0,8$	Discriminação aceitável
$0,8 \leq$ AUC $< 0,9$	Discriminação excelente
$0,9 \leq$ AUC $\leq 1,0$	Discriminação excepcional

Em conjunto, a curva ROC, a AUC e as medidas derivadas da matriz de confusão fornecem uma análise robusta da capacidade preditiva, permitindo comparar distintas abordagens e seleccionar aquela com maior capacidade discriminatória e relevância prática para o fenómeno em estudo.

2.7 Algoritmo XGBoost

O *XGBoost* (*eXtreme Gradient Boosting*) consiste num algoritmo de aprendizagem supervisionada pertencente à família dos métodos de *ensemble*, amplamente reconhecido pelo seu elevado desempenho preditivo, escalabilidade e eficiência computacional, particularmente em problemas de regressão e classificação envolvendo grandes volumes de dados. De acordo com Chen (2016), o XGBoost constitui uma implementação optimizada do algoritmo *Gradient Tree Boosting*, originalmente proposto por Friedman (2001), incorporando melhorias significativas ao nível da regularização, paralelização e gestão de memória.

O princípio fundamental do XGBoost baseia-se na construção sequencial de múltiplos modelos fracos, geralmente árvores de decisão do tipo CART (*Classification and Regression Trees*). Cada novo modelo procura corrigir os erros cometidos pelos modelos anteriores através da minimização sucessiva da função de perda, utilizando técnicas baseadas no gradiente descendente (Friedman, 2001; Chen, 2016).

Uma das principais inovações do XGBoost reside na inclusão explícita de termos de regularização na função objectivo, penalizando modelos excessivamente complexos e contribuindo para a redução do risco de sobreajustamento (*overfitting*). Conforme referido por Chen (2016) e Bruce (2019), esta característica confere maior robustez ao algoritmo em contextos caracterizados por elevada dimensionalidade ou multicolinearidade entre variáveis explicativas.

Além disso, o XGBoost foi concebido para explorar eficientemente arquitecturas computacionais modernas, permitindo processamento paralelo e utilização de CPU *multi-threaded*, o que o torna particularmente adequado para aplicações em ciência de dados e *Machine Learning* de larga escala (Chen, 2016).

No contexto do presente estudo, o XGBoost é utilizado como abordagem complementar à regressão logística para estimar a probabilidade de ocorrência do fenómeno em estudo, permitindo explorar relações não lineares e interacções de elevada complexidade entre as variáveis independentes.

Formulação Matemática do Modelo

Considere-se um conjunto de dados composto por n observações e p variáveis independentes, representado por:

$$D = \{(x_i, y_i)\}_{i=1}^{n} \tag{2.15}$$

onde $x_i \in \mathbb{R}^p$ corresponde ao vector de atributos da observação i e y_i representa a variável dependente.

Em problemas de classificação binária, como no presente estudo,

$$y_i \in \{0, 1\} \tag{2.16}$$

De acordo com Chen (2016), o modelo XGBoost constrói um preditor aditivo composto por K árvores de decisão, dado por:

$$F(x_i) = \sum_{k=1}^{K} f_k(x_i), \qquad f_k \in \mathcal{F} \tag{2.17}$$

onde:

- f_k corresponde à função associada à k-ésima árvore de decisão;
- $\mathcal{F}$ corresponde ao espaço das árvores possíveis;
- $F(x_i)$ representa o *score* do modelo para a observação i.

Segundo Friedman (2001), cada árvore adicionada ao modelo procura reduzir os erros residuais das iterações anteriores, permitindo uma aprendizagem progressiva e refinada.

Para problemas de classificação binária, a probabilidade estimada de ocorrência de união prematura pode ser obtida através da função logística:

$$P(Y_i = 1 \mid x_i) = \frac{1}{1 + \exp\left[-F(x_i)\right]} \tag{2.18}$$

onde $P(Y_i = 1 \mid x_i)$ representa a probabilidade estimada de ocorrência de união prematura para a observação i, e $F(x_i)$ corresponde ao *score* do modelo.

Função Objectivo do XGBoost

De acordo com Chen (2016), o processo de optimização do modelo baseia-se na minimização de uma função objectivo regularizada, definida por:

$$\mathcal{L} = \sum_{i=1}^{n} l\left(y_i, F(x_i)\right) + \sum_{k=1}^{K} \Omega(f_k) \tag{2.19}$$

onde $l(y_i, F(x_i))$ representa a função de perda, e $\Omega(f_k)$ corresponde ao termo de regularização associado à árvore f_k.

Conforme apresentado por Chen (2016), o termo de regularização pode ser expresso por:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{2.20}$$

onde

- T representa a quantidade de folhas da árvore;
- w_j representa o peso atribuído à folha j;
- γ penaliza árvores excessivamente complexas;
- λ controla a regularização dos pesos.

Segundo Hastie (2009), a inclusão destes termos de regularização permite controlar a complexidade do modelo, contribuindo para melhorar a sua capacidade de generalização.

Etapas de Aprendizagem do Modelo

A aprendizagem do XGBoost é realizada de forma iterativa. Conforme descrito por Chen (2016), na iteração t o score do modelo é actualizado da seguinte forma:

$$F^{(t)}(x_i) = F^{(t-1)}(x_i) + f_t(x_i) \tag{2.21}$$

onde $f_t(x_i)$ é a função (árvore) adicionada na iteração t. Na prática, $f_t(x_i)$ corresponde à diferença entre os scores de duas iterações consecutivas, isto é,

$$f_t(x_i) = F^{(t)}(x_i) - F^{(t-1)}(x_i) = d_i \tag{2.22}$$

onde d_i representa o incremento (ou correcção) que a nova árvore introduz no score do modelo. Assim, cada nova árvore adicionada ao modelo procura minimizar os erros cometidos pelas árvores anteriores.

Segundo Hastie (2009) e Chen (2016), a optimização do modelo baseia-se numa aproximação de segunda ordem da função objectivo, obtida através da expansão em série de Taylor, resultando na seguinte expressão:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2 \right] + \Omega(f_t) \tag{2.23}$$

onde

$$g_i = \frac{\partial l\left(y_i, F^{(t-1)}(x_i)\right)}{\partial F^{(t-1)}(x_i)} \tag{2.24}$$

$$h_i = \frac{\partial^2 l\left(y_i, F^{(t-1)}(x_i)\right)}{\partial \left[F^{(t-1)}(x_i)\right]^2} \tag{2.25}$$

representam respectivamente o gradiente e a hessiana da função de perda.

Ajuste Hiperparamétrico

O desempenho do XGBoost pode ser ajustado através de um conjunto de hiperparâmetros que controlam o nível de complexidade do modelo e a sua capacidade de generalização.

De acordo com Brownlee (2018), dentre os principais hiperparâmetros podem destacar-se:

- **max_depth**: define a profundidade máxima das árvores;
- **eta**: taxa de aprendizagem responsável por controlar a contribuição de cada árvore;
- **min_child_weight**: peso mínimo necessário para criar uma nova divisão;
- **gamma**: ganho mínimo necessário para efectuar uma divisão;
- **subsample**: proporção de observações empregues na construção de cada árvore;
- **colsample_bytree**: proporção de variáveis consideradas na construção de cada árvore;
- **lambda** -**alpha**: parâmetros de regularização L2 e L1.

Conforme referido por Kuhn (2013), o ajuste hiperparamétrico pode ser realizado utilizando técnicas como *Grid Search*, *Random Search* ou optimização Bayesiana, geralmente combinadas com validação cruzada.

2.8 Trabalhos Relacionados

Diversos estudos têm analisado as uniões prematuras e fenómenos sociais associados com base em dados demográficos e de saúde, com a finalidade de identificar factores de risco e contribuir para a definição de políticas públicas. A literatura evidencia que grande parte destes trabalhos recorrem a métodos estatísticos tradicionais, embora estudos mais recentes tenham incorporado técnicas de *Machine Learning* com vista a melhorar a capacidade analítica e preditiva.

Alem et al. (2020) realizaram um estudo no Bangladesh utilizando dados provenientes de inquéritos populacionais, no qual aplicaram técnicas baseadas em *Machine Learning*, nomeadamente *Random Forest*, *Gradient Boosting* e *Support Vector Machines*, para analisar os determinantes e a distribuição espacial das uniões prematuras. Os autores concluíram que o nível de escolaridade, a autonomia feminina e a região de residência constituem factores fortemente associados à ocorrência de uniões prematuras. O estudo evidenciou ainda que os modelos de *Machine Learning* apresentaram melhor desempenho preditivo quando comparados às abordagens estatísticas tradicionais, destacando o potencial destas técnicas na identificação de padrões complexos e relações de natureza não linear entre as variáveis explicativas e o fenómeno em estudo.

Num estudo realizado na África do Sul, analisaram os factores associados às uniões prematuras entre mulheres em idade reprodutiva, recorrendo a métodos estatísticos e modelos preditivos. Os resultados indicaram que a idade na primeira relação sexual, o nível de escolaridade e a condição socioeconómica do agregado familiar estão significativamente associados à ocorrência de uniões prematuras. Os autores destacam que a incorporação de modelos preditivos contribui para uma melhor compreensão do fenómeno, sublinhando, contudo, a importância da interpretação dos resultados para efeitos de políticas públicas (Mathabatha & Tsawe, 2023).

No contexto moçambicano, Hodges (2015) analisou dados relacionados com casamento e gravidez precoce, evidenciando uma elevada prevalência de uniões envolvendo raparigas menores de 18 anos. O estudo identificou factores tais como a pobreza, o fraco acesso à educação e normas socioculturais como determinantes relevantes. Apesar do contributo significativo para a compreensão do fenómeno em Moçambique, o autor recorreu predominantemente a análises descritivas e estatísticas tradicionais, não explorando técnicas avançadas de *Machine Learning*.

No mesmo enquadramento metodológico, Malerba (2023) recorreu a algoritmos de aprendizagem supervisionada, dos quais se destacam a Regressão Logística e as Árvores de Decisão,

Floresta Aleatória, *Gradient Boosting* e XGBoost, para a previsão da evasão universitária. Os resultados indicaram melhor desempenho dos modelos de maior complexidade, em particular XGBoost e *Gradient Boosting*, com valores de *F1-Score* superiores a 90%. Ainda assim, a Floresta Aleatória evidenciou um equilíbrio relevante entre capacidade preditiva e interpretabilidade, atingindo 91% de precisão. O autor conclui que a combinação entre modelos explicativos tradicionais e algoritmos de alta complexidade tende a ser mais eficaz na análise de fenómenos sociais e comportamentais, conclusão que se revela coerente com a orientação metodológica do presente estudo.

De forma convergente, Mendes (2023) observou resultados semelhantes ao aplicar técnicas de *Machine Learning* na previsão de infecções hospitalares, destacando a robustez e consistência dos algoritmos *Random Forest* e XGBoost. No contexto desta dissertação, estas evidências reforçam a pertinência da aprendizagem automática como complemento aos métodos estatísticos clássicos, permitindo aprofundar a compreensão dos factores associados às uniões prematuras em Moçambique, fenómeno de natureza multifactorial e influenciado por determinantes sociodemográficos, económicos e culturais.

Paralelamente, a literatura metodológica tem evidenciado a crescente aplicação de técnicas de *Machine Learning* aplicadas na análise de fenómenos sociais e problemas de saúde pública. Estudos nesta área indicam que algoritmos fundamentados em árvores de decisão associadas a métodos de *boosting*, como o XGBoost, apresentam elevada capacidade de previsão e melhor desempenho em tarefas de classificação de desfechos binários. No entanto, os autores sublinham que estes métodos devem ser utilizados de forma complementar aos modelos estatísticos tradicionais, de modo a assegurar interpretabilidade e robustez analítica.

Deste modo, os trabalhos relacionados evidenciam avanços importantes na aplicação de *Machine Learning* ao estudo das uniões prematuras, sobretudo em contextos internacionais. Contudo, observa-se uma lacuna no contexto moçambicano no que respeita à integração de modelos estatísticos explicativos com técnicas de *Machine Learning* aplicadas a dados recentes e representativos a nível nacional. O presente estudo procura colmatar esta lacuna ao combinar regressão logística e XGBoost na identificação dos factores relacionados com as uniões prematuras em Moçambique, utilizando dados provenientes do Inquérito Demográfico e de Saúde 2022–2023.

Capítulo 3

METODOLOGIA

3.1 Material

Este estudo utilizou dados secundários provenientes do Inquérito Demográfico e de Saúde (IDS) 2022–23, realizado em Moçambique. Os dados encontram-se disponíveis no site oficial do programa DHS: (`https://dhsprogram.com`), sendo o acesso concedido mediante autorização formal, em conformidade com as políticas de uso da base de dados.

A base de dados considerada integra 13 183 observações (linhas) e 6 499 variáveis (colunas), correspondendo a informações recolhidas junto de mulheres em idade reprodutiva, entre outros grupos populacionais abrangidos pelo inquérito.

Para a análise estatística foram utilizados os softwares RStudio (versão 5.31) e **Stata 17**. A redacção e a formatação do trabalho foram realizadas em **LaTeX (versão 4.8)**. Em todas as análises adoptou-se um nível de significância de 5%.

3.2 Variáveis em Estudo

A selecção das variáveis consideradas neste estudo baseou-se na literatura científica sobre uniões prematuras e na disponibilidade e consistência das informações recolhidas no Inquérito Demográfico e de Saúde (IDS) 2022–2023. A literatura consultada permitiu identificar um conjunto de factores frequentemente associados à ocorrência deste fenómeno, os quais abrangem dimensões demográficas, de saúde, socioeconómicas, culturais e contextuais.

Com base nessa revisão, foram definidas variáveis relevantes para a análise, constituídas por variáveis discretas e categóricas, que caracterizam as condições individuais e sociais da população em estudo.

Tabela 3.1: Descrição das variáveis do estudo extraídas do IDS 2022–2023

Variável (Código)	Categorias	Tipo
Características Demográficas e Geográficas		
Idade Actual (v012)	Idade da mulher em anos completos	Discreta
Província (v024)	Niassa, Cabo Delgado, Nampula, Zambézia, Tete, Manica, Sofala, Inhambane, Gaza, Maputo Província, Maputo Cidade	Categórica
Local de Residência (v025)	Urbano; Rural	Categórica
Língua Materna (v034)	Língua principal falada pela inquirida	Categórica
Membros do Agregado (v136)	Número total de pessoas no agregado familiar	Discreta
Características Socioeconómicas		
Nível Educacional (v106)	nenhum; Primária; Secundária; Superior	Categórica
Quintil de Riqueza (v190)	Mais pobre; Pobre; Médio; Rico; Mais rico	Categórica
Religião (v130)	Católico, Islâmico, Sionista, Evangélico, Outro, Sem religião	Categórica
Profissão (v717)	Ocupação actual da inquirida	Categórica
Histórico Reprodutivo e Estado Civil		
Estado Civil (v501)	Nunca casou; Casada; União de facto; Viúva; Divorciada	Categórica
Idade 1ª União (v511)	Idade (em anos) da mulher na primeira união	Discreta
União Prematura (v511a)	Ocorrência da primeira união antes dos 18 anos	Categórica
1ª Relação Sexual (v525/v531)	Idade (em anos) na primeira relação sexual	Discreta
Duração da Gravidez (v214)	Duração da gravidez actual em meses	Discreta
Exposição aos Meios de Comunicação		
Leitura de Jornal (v157)	Frequência de leitura de jornal ou revista	Categórica
Exposição à Mídia (v384a,b,c)	Exposição a mensagens de PF via Rádio, TV ou Jornal	Categórica

Fonte: INE, MISAU & ICF (2023).

3.3 Classificação da Pesquisa

A presente investigação configura-se como um estudo de natureza quantitativa, com carácter observacional, analítica, transversal e aplicada, desenvolvida a partir da análise de dados secundários provenientes do Inquérito Demográfico e de Saúde (IDS) de Moçambique, referente ao período 2022–2023.

A abordagem de natureza quantitativa justifica-se pela utilização de métodos estatísticos e de técnicas de Machine Learning na análise das relações entre variáveis e na estimação da probabilidade de ocorrência de uniões prematuras, fundamentando-se em dados numéricos e em procedimentos formais de inferência estatística e avaliação do desempenho preditivo dos modelos.

Trata-se de um estudo observacional, uma vez que não envolve qualquer intervenção directa sobre as unidades de análise, limitando-se à análise de dados previamente recolhidos, sem manipulação experimental das variáveis em estudo.

Do ponto de vista dos objectivos, a pesquisa assume um carácter analítico e explicativo, procurando identificar, descrever e quantificar os factores demográficos, socioeconómicos, culturais e contextuais associados às uniões prematuras em Moçambique, contribuindo para a compreensão dos seus determinantes.

Quanto ao desenho temporal, o estudo apresenta carácter transversal, dado que os dados analisados correspondem a um único momento de observação, referente ao período de realização do IDS 2022–2023.

A investigação assume natureza aplicada, na medida em que os resultados obtidos visam produzir evidência empírica susceptível de subsidiar a formulação de políticas públicas e o desenvolvimento de estratégias de intervenção orientadas para a prevenção e mitigação das uniões prematuras no contexto moçambicano, recorrendo à regressão logística e a algoritmos de Machine Learning, nomeadamente o XGBoost, com potencial aplicação prática para decisores e instituições relevantes.

3.4 Métodos

A análise assentou na combinação de duas abordagens complementares de modelação: um modelo estatístico clássico e um algoritmo de aprendizagem automática. Esta articulação permitiu conciliar a interpretabilidade dos resultados com o poder preditivo, respondendo aos diferentes objectivos do estudo.

A **Regressão Logística** foi empregue numa perspectiva explicativa. Através dela, estimaram-se as associações ajustadas entre as variáveis explicativas e a ocorrência de uniões prematuras, cujo sentido e magnitude foram interpretados com recurso às razões de chances (*Odds Ratios*).

O algoritmo **Extreme Gradient Boosting (XGBoost)**, por sua vez, foi mobilizado numa óptica preditiva. Enquanto método de *Machine Learning*, destaca-se pela capacidade de captar relações não lineares e interacções complexas entre as variáveis, características que dificilmente seriam apreendidas por modelos puramente paramétricos.

3.4.1 Etapas do Processo de Análise

Com o objectivo de garantir rigor metodológico, o processo analítico foi conduzido segundo os princípios da **aprendizagem supervisionada**, seguindo uma sequência estruturada de etapas que inclui preparação dos dados, análise exploratória, treino dos modelos, validação e optimização. As etapas do processo analítico estão representadas no fluxograma apresentado na Figura 3.1.

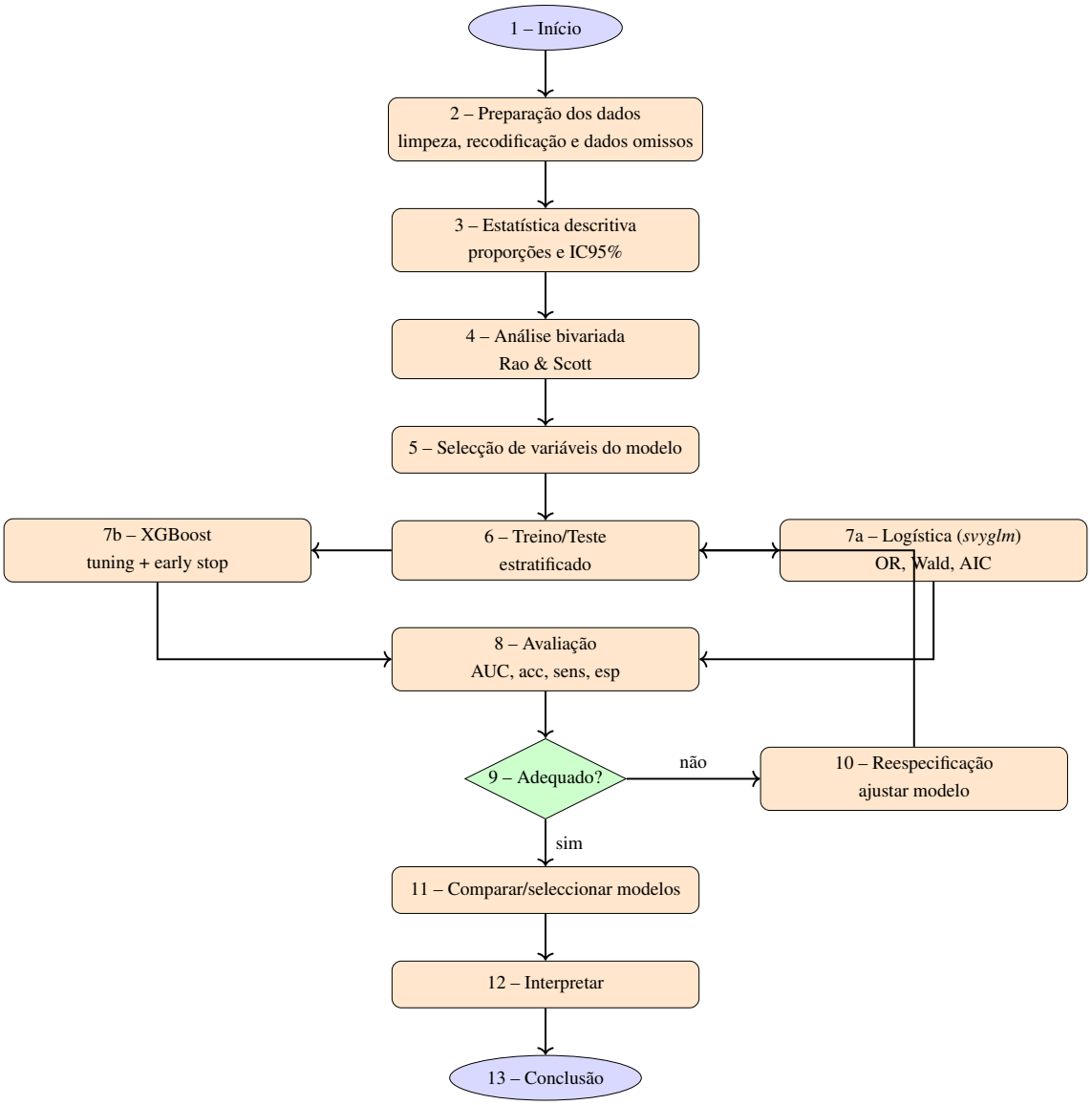


Figura 3.1: Fluxograma do processo de análise estatística e modelação preditiva

3.4.2 Análise Exploratória de Dados

A fase inicial do estudo envolveu a realização de uma exploração preliminar dos dados, com o objectivo de compreender a estrutura da base de dados, identificar possíveis inconsistências e explorar padrões iniciais entre as variáveis consideradas no estudo.

Nesta fase foram realizados os seguintes procedimentos:

- **Verificação e limpeza dos dados**, incluindo a identificação de dados ausentes, observações inconsistentes e a recodificação das categorias das variáveis explicativas;
- **Análise descritiva**, por meio do cálculo de frequências absolutas, relativas, médias e

intervalos de confiança a 95%, permitindo caracterizar o perfil sociodemográfico das mulheres incluídas na amostra;

- **Visualização gráfica dos dados**, utilizando gráficos de barras e histogramas para examinar a distribuição das variáveis e identificar possíveis padrões entre grupos;
- **Análise bivariada**, realizada através do teste de associação de Rao & Scott, apropriado para dados provenientes de inquéritos com desenho amostral complexo, com o objectivo de identificar variáveis potencialmente associadas à ocorrência de uniões prematuras.

Distância de Cook

A Distância de Cook (Cook, 1977) é uma medida de diagnóstico utilizada para identificar observações influentes, ou seja, pontos de dados cuja remoção altera de forma substancial as estimativas do modelo ajustado. Este diagnóstico é importante porque uma única observação atípica pode distorcer os resultados sem que tal seja evidente apenas pela análise dos resíduos.

A Distância de Cook para a observação i é dada por:

$$D_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(i)})^2}{p \cdot MSE} \tag{3.1}$$

em que $\hat{y}_j$ é o valor previsto para a observação j com todas as observações incluídas, $\hat{y}_{j(i)}$ é o valor previsto para a observação j quando a observação i é removida do modelo, p é o número de parâmetros estimados, e MSE é o erro quadrático médio do modelo.

Valores de D_i elevados indicam observações com forte influência sobre o modelo, sendo frequentemente utilizado como referência o limiar $D_i > 4/n$, em que n é o número total de observações. As observações que ultrapassam este limiar são identificadas como potencialmente influentes e merecem uma análise mais cuidada antes da modelação final.

Teste de Associação de Rao & Scott

O teste de associação de Rao e Scott (Rao & Scott, 1981) é uma versão corrigida do teste do Qui-quadrado, usada para testar a associação entre duas variáveis categóricas quando os dados provêm de uma amostragem complexa (por exemplo, com estratificação ou agrupamento por distrito ou unidade sanitária), situação em que o Qui-quadrado tradicional pode não ser fiável.

As hipóteses testadas são:

- H_0: não existe associação entre a variável e o desfecho;
- H_1: existe associação entre a variável e o desfecho.

A estatística de teste é calculada ajustando o Qui-quadrado de Pearson ao efeito do desenho amostral:

$$X^2_{RS} = \frac{X^2_P}{\hat{d}} \tag{3.2}$$

em que X^2_P é o valor do Qui-quadrado de Pearson e $\hat{d}$ é o factor de correcção associado ao desenho amostral. O resultado é depois comparado com uma distribuição F, em vez da distribuição Qui-quadrado habitual, tornando o teste mais adequado à estrutura dos dados (Rao & Scott, 1981).

Esta etapa permitiu identificar padrões iniciais nos dados e seleccionar as variáveis candidatas para inclusão nos modelos de regressão.

3.4.3 Divisão da Base de Dados

Com vista à avaliação da capacidade preditiva dos modelos, os dados foram particionados em dois subconjuntos distintos, de acordo com o método *hold-out*.

Foi adoptada uma divisão estratificada dos dados, com:

- **70% das observações** destinadas ao conjunto de treino;
- **30% das observações** reservadas para o conjunto de teste.

A proporção de 70% para treino e 30% para teste constitui uma das repartições mais consensuais na literatura de *Machine Learning*, por estabelecer um equilíbrio adequado entre a quantidade de dados disponível para a aprendizagem do modelo e a que é reservada para a avaliação independente do seu desempenho (Kohavi, 1995). Uma fracção de treino demasiado reduzida comprometeria a capacidade de o modelo captar os padrões subjacentes aos dados, ao passo que um conjunto de teste demasiado pequeno tornaria as estimativas de desempenho pouco fiáveis e sensíveis à variação amostral (Hastie, Tibshirani & Friedman, 2009). A opção pela divisão estratificada assegura, adicionalmente, que a distribuição da variável resposta a ocorrência de uniões prematuras se mantém proporcional em ambos os conjuntos, evitando enviesamentos decorrentes de eventuais desequilíbrios entre as classes (Kohavi, 1995).

A divisão estratificada permitiu preservar a proporção das categorias da variável resposta em ambos os subconjuntos, evitando distorções na distribuição das classes.

Esta estratégia é recomendada na literatura de aprendizagem automática (Kuhn & Johnson, 2013; James et al., 2013), pois permite avaliar o desempenho dos modelos em dados não utilizados no processo de treino, fornecendo uma estimativa mais realista da capacidade de generalização.

3.4.4 Treinamento dos Modelos

Modelo de Regressão Logística

Antes da estimação do modelo de regressão logística, foram realizados procedimentos de diagnóstico com o objectivo de verificar a adequação dos dados à modelação estatística. Nesta etapa avaliou-se a presença de multicolinearidade nas variáveis explicativas, bem como a existência de observações potencialmente influentes que pudessem afectar a estabilidade das estimativas do modelo.

Após estes procedimentos de diagnóstico, o modelo de regressão logística foi estimado considerando o desenho amostral complexo do inquérito, utilizando a função `svyglm` do pacote *survey* em `R`.

A especificação do modelo incluiu variáveis sociodemográficas, socioeconómicas e contextuais identificadas na análise exploratória. A avaliação do modelo considerou os seguintes indicadores:

- coeficientes estimados (β);
- razões de chances (*Odds Ratios*);
- intervalos de confiança de 95%;
- teste de Wald para avaliação da significância estatística;
- pseudo-R^2 de McFadden para avaliar o ajustamento global do modelo.

Modelo XGBoost

O modelo XGBoost foi treinado com base no algoritmo de *gradient boosting* aplicado a árvores de decisão. Os principais hiperparâmetros considerados foram:

- `eta` (taxa de aprendizagem);
- `max_depth` (profundidade máxima das árvores);
- `min_child_weight` (peso mínimo das observações nos nós);
- `subsample` (proporção de observações utilizadas em cada árvore);
- `colsample_bytree` (fracção de variáveis utilizadas na construção de cada árvore);
- `gamma` (parâmetro de regularização que controla a complexidade do modelo).

3.4.5 Ajuste Hiperparamétrico

Para optimizar o desempenho do modelo XGBoost, foi realizado um processo de ajuste hiperaparamétrico através de *grid search* combinado com validação cruzada.

Foram avaliadas **64 combinações de hiperparâmetros**. A melhor configuração identificada foi:

- Taxa de aprendizagem $\eta = 0.05$
- Limite máximo de Profundidade da árvore `max_depth` = 3
- Peso mínimo das observações nos nós `min_child_weight` = 1
- Proporção de amostras `subsample` = 1.0
- Proporção de variáveis por árvore `colsample_bytree` = 0.8
- Parâmetro de regularização `gamma` = 0
- Número óptimo de iterações `best_nrounds` = 581

Esta configuração apresentou uma AUC média de **0.832** na validação cruzada, com desvio-padrão de **0.008**, evidenciando bom desempenho discriminatório e elevada estabilidade do modelo.

3.4.6 Técnicas de Validação

Para garantir maior robustez na avaliação do desempenho dos modelos, foram utilizadas duas estratégias de validação.

Validação Hold-out

A avaliação final do desempenho dos modelos foi realizada no conjunto de teste (30% dos dados), permitindo aferir a capacidade de generalização em dados não incluídos no conjunto de treino.

Validação Cruzada *k*-fold

Adicionalmente, aplicou-se validação cruzada *k*-fold com $k = 5$, permitindo obter estimativas mais estáveis das medidas de desempenho e reduzir a dependência de uma única partição dos dados.

3.4.7 Avaliação do Desempenho dos Modelos

A avaliação dos modelos baseou-se em diferentes indicadores de desempenho derivados da matriz de confusão (Sokolova & Lapalme, 2009; Fawcett, 2006):

- Área sob a curva ROC (AUC);
- Acurácia;
- Sensibilidade;
- Especificidade;
- Coeficiente Kappa;
- Acurácia balanceada.

A avaliação dos modelos fundamentou-se em diferentes medidas de desempenho obtidas a partir da matriz de confusão:

A AUC foi utilizada como principal medida de desempenho para a comparação dos modelos, por avaliar a capacidade discriminatória do modelo, independentemente do ponto de corte adoptado.

Adicionalmente, para avaliar se as diferenças entre as curvas ROC eram estatisticamente significativas, aplicou-se o teste de DeLong, apropriado para a comparação de curvas ROC dependentes.

O modelo considerado mais adequado foi aquele que apresentou maior AUC, melhor equilíbrio entre sensibilidade e especificidade e desempenho consistente após a etapa de optimização.

Capítulo 4

RESULTADOS E DISCUSSÃO

Nesta secção apresentam-se os principais resultados da análise estatística realizada, iniciando-se pela análise estatística descritiva dos dados, com o objectivo de caracterizar as variáveis em estudo e fornecer uma visão geral do seu comportamento na amostra analisada. Esta etapa permite compreender a distribuição da variável resposta e as variáveis explicativas mais relevantes consideradas na análise.

Posteriormente, procede-se à estimação de modelos de regressão logística bem como à aplicação de técnicas de aprendizagem automática, em particular o algoritmo XGBoost, com vista à identificação e quantificação dos factores determinantes da ocorrência de uniões prematuras. A utilização conjunta destas abordagens metodológicas permite, por um lado, avaliar os efeitos individuais das variáveis explicativas sobre o desfecho e, por outro, explorar relações mais complexas entre os diferentes factores.

4.1 Estatística Descritiva

4.1.1 Variável resposta

A Figura 4.1 ilustra a distribuição da variável resposta, correspondente à situação da união prematura entre mulheres em idade reprodutiva. Verifica-se que a maioria das mulheres não entrou em união prematura, totalizando 5 460 casos, o que corresponde a aproximadamente 54,9% da amostra. Em contrapartida, 4 488 mulheres (45,1%) entraram em união prematura, evidenciando uma proporção elevada de ocorrência do fenómeno na população em estudo. Estes resultados demonstram que a união prematura constitui uma realidade social relevante no contexto moçambicano, justificando a sua adopção como variável resposta nos modelos analíticos subsequentes, dada a sua relevância para a análise dos factores associados à sua ocorrência.

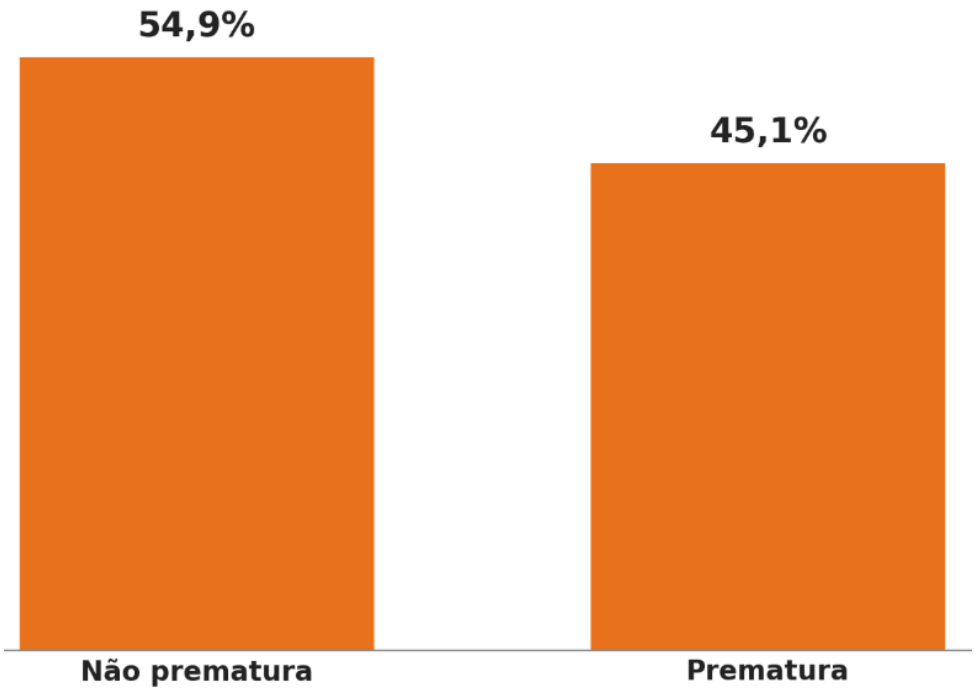


Figura 4.1: Distribuição das mulheres em idade reprodutiva segundo a União Prematura.

4.1.2 Distribuição das mulheres em idade reprodutiva por quintil de riqueza

A Figura 4.2 ilustra a distribuição das mulheres em idade fértil de acordo com o quintil de riqueza. Observa-se que a maior proporção das mulheres pertence ao quintil mais elevado de riqueza (30,06%), seguido do quarto quintil (23,63%). O quintil intermédio representa 18,36% da amostra, enquanto os quintis mais baixos apresentam proporções relativamente menores, nomeadamente o segundo quintil (14,12%) e o quintil mais baixo (13,84%).

O quintil de riqueza constitui um indicador relevante das condições socioeconómicas dos agregados familiares e pode estar associado à ocorrência de uniões prematuras. Em contextos de menor nível económico, as raparigas tendem a apresentar maior vulnerabilidade à ocorrencia de uniões prematuras, frequentemente associadas a desigualdades sociais e limitações no acesso a oportunidades educacionais e sociais. Assim, esta variável é relevante para compreender os factores associados às uniões prematuras no contexto moçambicano.

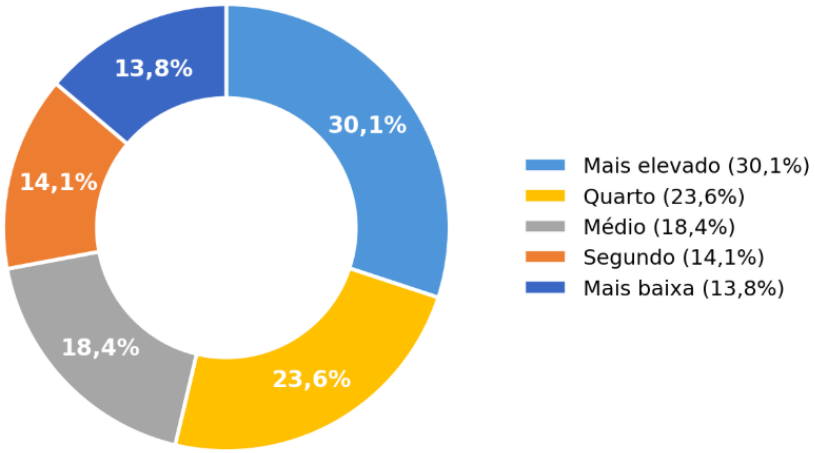


Figura 4.2: Distribuição das mulheres em idade reprodutiva por quintil de riqueza

4.1.3 Distribuição das mulheres em idade reprodutiva por área de residência

A Figura 4.3 ilustra a distribuição das mulheres em idade fértil segundo a área de residência. Observa-se uma maior proporção de mulheres residentes em meio rural (56.8%), em comparação com o meio urbano (43.2%).

Esta distribuição é relevante para a análise das uniões prematuras, uma vez que, em muitos contextos, as zonas rurais apresentam maiores níveis de vulnerabilidade social e económica, factores que podem contribuir para a ocorrência mais frequente de uniões prematuras. Assim, a área de residência constitui uma variável relevante na compreensão dos padrões e dos factores associados às uniões prematuras em Moçambique.

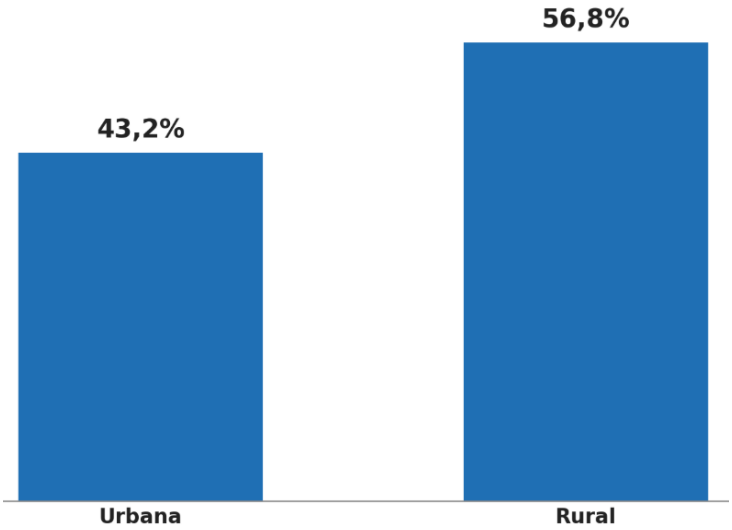


Figura 4.3: Distribuição das mulheres em idade reprodutiva por área de residência

4.1.4 Distribuição das mulheres em idade reprodutiva por nível de escolaridade

A Figura 4.4 ilustra a distribuição das mulheres em idade fértil de acordo com o nível de escolaridade. Observa-se que a maior proporção das mulheres possui ensino primário (41.2%), seguida do ensino secundário (32.3%). Por outro lado, uma parcela considerável não possui qualquer nível de escolaridade (23.0%), enquanto apenas uma pequena proporção atingiu o ensino superior (3.5%).

O nível de escolaridade constitui um factor fundamental na compreensão das uniões prematuras, uma vez que menores níveis de escolaridade encontram-se frequentemente relacionados à maior vulnerabilidade social e a uma maior probabilidade de ocorrência de uniões prematuras. Assim, esta variável revela-se relevante para a análise dos factores associados às uniões prematuras no contexto moçambicano.

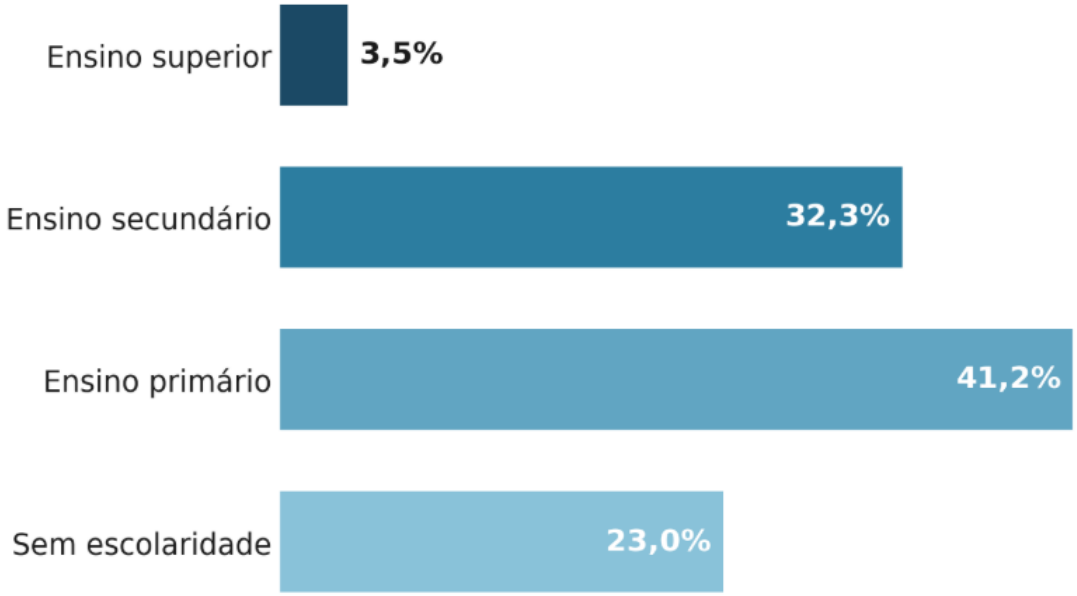


Figura 4.4: Distribuição das mulheres em idade reprodutiva por nível de escolaridade

4.2 Análise bivariada da associação entre as variáveis explicativas e as uniões prematuras

A Tabela 4.1 mostra os resultados do teste de qui-quadrado com ajustamento de Rao & Scott, aplicado na avaliação da associação entre as variáveis explicativas e a ocorrência de uniões prematuras, considerando o desenho amostral complexo do inquérito.

A hipótese nula assume a inexistência de associação entre as variáveis analisadas e as uniões prematuras, sendo rejeitada quando $p - valor < 0{,}05$, o que indica evidência estatisticamente significativa de associação.

Tabela 4.1: Teste de Associação de Rao & Scott entre as Variáveis Explicativas e a Variável Resposta

Variável	**Estatística F**	**P-valor**
Idade na primeira relação	396.244	3.664×10^{-250}
Estado civil	382.002	8.207×10^{-128}
Língua materna	389.457	4.544×10^{-67}
Nível de escolaridade	109.655	3.614×10^{-63}
Quintil de riqueza	39.090	6.715×10^{-30}
Idade actual	38.484	1.051×10^{-43}
Residência	77.597	1.380×10^{-17}
Província	21.902	7.088×10^{-33}
Exposição às mídias	19.724	4.253×10^{-24}
Ocupação (agrupada)	19.704	8.346×10^{-16}
Religião	9.017	2.178×10^{-9}
Tamanho do agregado familiar	6.958	1.480×10^{-4}
Início da gravidez	3.104	4.805×10^{-2}

Os resultados do teste de associação de Rao & Scott indicam que as variáveis analisadas apresentam associação estatisticamente significativa com as uniões prematuras (p-valor $< 0{,}05$). As associações mais fortes avaliadas pela magnitude da estatística F observam-se para a idade na primeira relação, o estado civil e a língua materna, o que evidencia que estas variáveis são as mais relevantes na diferenciação da ocorrência de uniões prematuras.

O nível de escolaridade, assim como o quintil de riqueza também apresentam associação estatisticamente significativa, sugerindo que factores socioeconómicos desempenham um papel relevante na ocorrência das uniões prematuras. De igual modo, a idade actual e a área de residência revelam associação estatisticamente significativa, indicando diferenças relevantes entre grupos demográficos e entre o meio urbano e o meio rural.

Adicionalmente, variáveis contextuais e culturais, como a província, a exposição aos meios de comunicação social, a ocupação, a religião e o tamanho do agregado familiar, apresentam associação estatisticamente significativa, evidenciando a relevância do contexto sociocultural e regional na ocorrência das uniões prematuras. Por fim, o início da gravidez revela igualmente associação estatisticamente significativa, embora com menor magnitude, sugerindo uma relação menos pronunciada com a ocorrência de uniões prematuras.

4.3 Modelação Estatística

Nesta secção descrevem-se os métodos de modelação estatística utilizados para analisar os factores associados à ocorrência de uniões prematuras em Moçambique. A modelação tem como objectivo avaliar a associação entre a variável resposta e um conjunto de covariáveis explicativas de natureza demográfica, socioeconómica e contextual.

4.3.1 Diagnóstico de Multicolinearidade

Antes do ajustamento do modelo de regressão logística, procedeu-se à avaliação da multicolinearidade entre as variáveis explicativas, uma vez que a existência de forte correlação entre os preditores pode afectar a estabilidade das estimativas e dificultar a interpretação dos coeficientes do modelo.

Para tal, utilizou-se o Factor de Inflação da Variância Generalizado (*Generalized Variance Inflation Factor* – GVIF), apropriado para estruturas de modelação que incluem variáveis categóricas com múltiplos graus de liberdade. Com vista a possibilitar a comparação entre variáveis com distintos números de categorias, considerou-se a medida ajustada $GVIF^{1/(2Df)}$, onde Df representa os graus de liberdade associados a cada variável.

A Tabela 4.2 apresenta os resultados observados. De acordo com a literatura, valores de $GVIF^{1/(2Df)}$ abaixo de 5 sugerem ausência de multicolinearidade significativa entre os preditores. Os resultados observados mostram que todas as variáveis apresentam valores abaixo deste limiar, sugerindo que não existe multicolinearidade relevante entre as variáveis explicativas incorporadas no modelo.

Assim, conclui-se que as variáveis consideradas fornecem informação complementar e podem ser incluídas simultaneamente na modelação estatística sem afectar a estabilidade das estimativas da modelação por regressão logística.

Tabela 4.2: Diagnóstico de Multicolinearidade das Variáveis Explicativas

Variável	**GVIF**	**Df**	$GVIF^{1/(2Df)}$
Diferença de idade entre cônjuges	2.937	1	1.714
Estado civil	4.516	2	1.458
Área de residência	2.052	1	1.432
Quintil de riqueza	4.218	4	1.197
Nível de escolaridade	2.085	3	1.130
Província de residência	9.875	10	1.121
Religião	3.523	6	1.111
Ocupação	2.035	4	1.093
Exposição aos meios de comunicação	2.273	7	1.060
Idade na primeira relação sexual	1.555	4	1.057
Grupo etário da mulher	1.805	6	1.050
Tamanho do agregado familiar	1.163	3	1.026
Número de filhos no agregado	1.035	2	1.009

Adicionalmente, procedeu-se à análise de observações com potencial influência, recorrendo à medida da distância de Cook, aos resíduos deviance e às medidas de alavancagem. A distância de Cook foi utilizada para identificar observações com impacto excessivo nas estimativas do modelo. A Figura 4.5 apresenta a distribuição desta medida para todas as observações da amostra.

Observa-se que a maioria dos valores se encontra abaixo do limiar de referência, indicado pela linha horizontal, sugerindo que não existem observações com influência excessiva sobre os coeficientes estimados. De forma consistente, a análise dos resíduos deviance e das medidas de alavancagem não revelou a presença de casos extremos que pudessem comprometer a estabilidade do modelo.

Assim, conclui-se que não existem observações influentes com impacto significativo nas estimativas da regressão logística ajustada, permitindo prosseguir com o processo de modelação sem necessidade de remoção de casos.

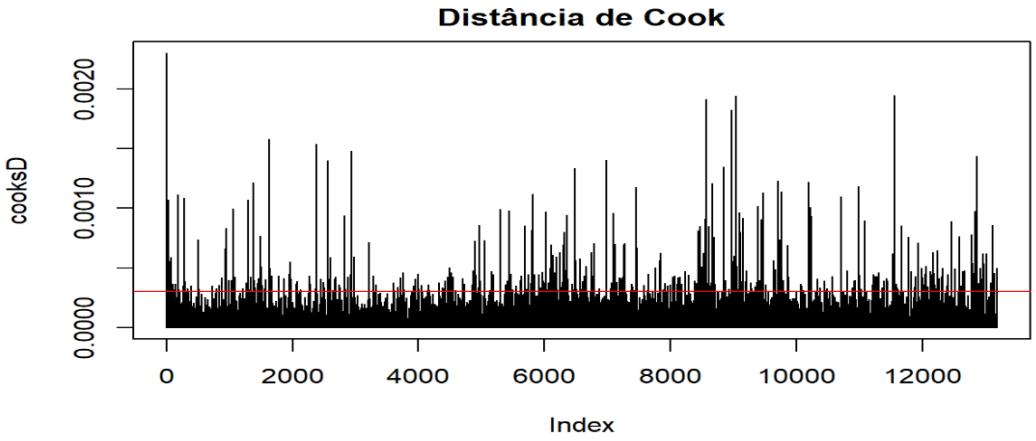


Figura 4.5: Diagnóstico de Observações Influentes através da Distância de Cook

4.3.2 Estimação do Modelo de Regressão Logística

A definição do modelo de regressão logística fundamentou-se em diferentes etapas de diagnóstico e selecção de variáveis. Inicialmente, realizou-se uma análise bivariada entre as variáveis explicativas e a ocorrência de uniões prematuras através do teste de associação de Rao & Scott, adequado para dados provenientes de inquéritos com desenho amostral complexo. Esta análise permitiu identificar as variáveis potencialmente associadas ao desfecho e consideradas na modelação multivariada.

Posteriormente, avaliou-se a existência de multicolinearidade entre as variáveis explicativas através do factor de Inflação da variância (GVIF). Os resultados evidenciaram valores inferiores aos limites recomendados na literatura, sugerindo ausência de multicolinearidade relevante entre os preditores.

Adicionalmente, foram analisadas observações potencialmente influentes com base na distância de Cook, bem como nos resíduos deviance e nas medidas de alavancagem. Não foram identificados casos com influência excessiva sobre as estimativas do modelo.

Após estes procedimentos de diagnóstico, procedeu-se à estimação do modelo de regressão logística, integrando variáveis de natureza demográfica, socioeconómica e contextual associadas à ocorrência de uniões prematuras.

Valores de *Re* acima de 1 sugerem aumento da probabilidade de ocorrência de união prematura, enquanto valores inferiores a 1 indicam efeito protector em relação à categoria de referência.

A Tabela 4.3 apresenta os principais resultados do modelo, incluindo os coeficientes estimados, as razões de chances (*RC*), os intervalos de confiança de 95% e os respectivos níveis de significância estatística.

Tabela 4.3: Modelo de Regressão Logística para os Factores Associados à União Prematura

Variável	β	**RC**	**IC 95% (RC)**	**p-valor**
Idade 15–19 (ref.)	–	1	–	–
Idade 20–24	0.254	1.290	[1.011 ; 1.645]	0.040
Idade 25–29	-0.017	0.983	[0.744 ; 1.299]	0.905
Idade 30–34	-0.406	0.666	[0.519 ; 0.855]	0.002
Idade 35–39	-0.384	0.681	[0.516 ; 0.899]	0.007
Idade 40–44	-0.541	0.582	[0.425 ; 0.798]	0.001
Idade 45–49	-0.898	0.408	[0.290 ; 0.572]	<0.001
Primeira relação <15 (ref.)	–	1	–	–
Primeira relação 15–17	-1.306	0.271	[0.231 ; 0.319]	<0.001
Primeira relação 18–24	-19.134	0.000	–	<0.001
Primeira relação 25+	-18.446	0.000	–	<0.001
Nunca teve relação / missing	-1.453	0.234	[0.167 ; 0.327]	<0.001
Sem escolaridade (ref.)	–	1	–	–
Ensino primário	0.407	1.503	[1.252 ; 1.804]	<0.001
Ensino secundário	-0.189	0.828	[0.661 ; 1.038]	0.102
Ensino superior	-0.781	0.458	[0.266 ; 0.790]	0.005
Residência urbana (ref.)	–	1	–	–
Residência rural	0.160	1.173	[0.985 ; 1.396]	0.073
Nunca casou/uniu (ref.)	–	1	–	–
Em união	1.900	6.684	[4.635 ; 9.638]	<0.001
Fora de união	2.108	8.227	[5.942 ; 11.393]	<0.001
Quintil mais baixo (ref.)	–	1	–	–
Segundo quintil	-0.148	0.862	[0.668 ; 1.113]	0.255
Quintil médio	0.066	1.068	[0.831 ; 1.372]	0.608
Quarto quintil	0.358	1.430	[1.081 ; 1.891]	0.012
Quintil mais elevado	0.404	1.497	[1.057 ; 2.121]	0.023
Niassa (ref.)	–	1	–	–
Cabo Delgado	-0.327	0.721	[0.551 ; 0.943]	0.017
Nampula	0.245	1.278	[0.974 ; 1.678]	0.078
Zambézia	0.108	1.114	[0.806 ; 1.540]	0.514
Tete	-0.292	0.747	[0.540 ; 1.032]	0.077
Manica	0.053	1.055	[0.739 ; 1.506]	0.768

Continua na página seguinte

Tabela 4.3: Modelo de Regressão Logística para os Factores Associados à União Prematura (continuação)

Variável	β	**RC**	**IC 95% (RC)**	**P-valor**
Sofala	0.044	1.045	[0.748 ; 1.459]	0.797
Inhambane	-0.083	0.920	[0.639 ; 1.325]	0.654
Gaza	-0.470	0.625	[0.432 ; 0.904]	0.013
Maputo Província	-0.502	0.606	[0.422 ; 0.868]	0.007
Cidade de Maputo	-0.356	0.700	[0.476 ; 1.030]	0.071
Católica (ref.)	–	1	–	–
Islâmica	0.110	1.116	[0.866 ; 1.439]	0.395
Zion	0.347	1.415	[1.115 ; 1.795]	0.004
Evangélica/Pentecostal	0.193	1.212	[1.001 ; 1.468]	0.049
Anglicana	-0.199	0.819	[0.524 ; 1.282]	0.382
Sem religião	0.112	1.119	[0.788 ; 1.589]	0.529
Outra religião	0.760	2.138	[1.072 ; 4.264]	0.031
Sem exposição (ref.)	–	1	–	–
Só rádio	0.257	1.293	[1.065 ; 1.569]	0.010
Rádio e TV	0.251	1.286	[1.047 ; 1.578]	0.016
Só jornal	0.474	1.607	[0.705 ; 3.661]	0.258
Todos os meios	0.202	1.224	[0.904 ; 1.658]	0.191
Pequeno (1–3) (ref.)	–	1	–	–
Médio (4–6)	0.253	1.288	[1.076 ; 1.541]	0.006
Grande (7–9)	0.279	1.321	[1.077 ; 1.621]	0.008
Muito grande (10+)	0.351	1.420	[1.094 ; 1.843]	0.009

Os resultados indicam que vários factores permanecem estatisticamente associados às uniões prematuras, após o ajustamento pelas restantes variáveis do modelo.

No que respeita às características demográficas, verifica-se que os grupos etários mais avançados apresentam menor probabilidade de ocorrência de uniões prematuras relativamente ao grupo de referência (15–19 anos). A chance diminui progressivamente com a idade, passando de uma redução de cerca de 33% no grupo dos 30–34 anos (RC = 0,67; IC 95%: 0,52–0,86) para uma redução de aproximadamente 59% no grupo dos 45–49 anos (RC = 0,41; IC 95%: 0,29–0,57). A idade na primeira relação sexual revela igualmente uma associação forte: as mulheres cuja primeira relação ocorreu entre os 15 e os 17 anos apresentam uma chance cerca de 73% inferior (RC = 0,27; IC 95%: 0,23–0,32) à das que a iniciaram antes dos 15 anos, o que

evidencia a relação entre o início precoce da vida sexual e a entrada em uniões prematuras.

No domínio socioeconómico, o nível de escolaridade surge como um factor relevante, embora com efeitos distintos consoante o grau. As mulheres com ensino superior apresentam uma chance cerca de 54% inferior de entrar em união prematura (RC = 0,46; IC 95%: 0,27–0,79) face às sem escolaridade, ao passo que as mulheres com ensino primário evidenciam, pelo contrário, uma chance cerca de 50% superior (RC = 1,50; IC 95%: 1,25–1,80). O quintil de riqueza apresenta associações significativas nos estratos mais elevados: as mulheres do quarto quintil e do quintil mais elevado têm chances, respectivamente, cerca de 43% e 50% superiores às do quintil mais baixo (RC = 1,43; IC 95%: 1,08–1,89 e RC = 1,50; IC 95%: 1,06–2,12).

O estado civil constitui o factor de maior magnitude no modelo. As mulheres em união apresentam uma chance cerca de 6,7 vezes superior (RC = 6,68; IC 95%: 4,64–9,64) e as que se encontram fora de união uma chance cerca de 8,2 vezes superior (RC = 8,23; IC 95%: 5,94–11,39) de ocorrência de união prematura, comparativamente à categoria de referência.

Relativamente às variáveis contextuais, a área de residência apresenta uma associação apenas marginal (RC = 1,17; IC 95%: 0,99–1,40; $p = 0{,}073$), sugerindo uma tendência para maior ocorrência no meio rural, ainda que não estatisticamente significativa ao nível de 5%. Observam-se ainda diferenças regionais significativas: em Gaza e na Província de Maputo, as chances são cerca de 38% e 39% inferiores às da província de referência (RC = 0,63; IC 95%: 0,43–0,90 e Re = 0,61; IC 95%: 0,42–0,87, respectivamente).

Os factores socioculturais apresentam também associações significativas em determinadas categorias. Quanto à religião, as mulheres de confissão Zion e Evangélica/Pentecostal apresentam chances superiores às da categoria de referência (RC = 1,42; IC 95%: 1,12–1,80 e RC = 1,21; IC 95%: 1,00–1,47, respectivamente). No que respeita à exposição aos meios de comunicação social, a exposição apenas ao rádio e ao conjunto de rádio e televisão associa-se a chances cerca de 29% superiores (RC = 1,29; IC 95%: 1,07–1,57 e RC = 1,29; IC 95%: 1,05–1,58), sugerindo que as normas culturais e o acesso à informação influenciam o padrão de ocorrência das uniões prematuras.

Por fim, o tamanho do agregado familiar apresenta uma associação positiva e gradual: face aos agregados de menor dimensão, a chance aumenta cerca de 29% nos agregados médios (RC = 1,29; IC 95%: 1,08–1,54), 32% nos grandes (RC = 1,32; IC 95%: 1,08–1,62) e 42% nos muito grandes (RC = 1,42; IC 95%: 1,09–1,84), indicando que as mulheres de agregados mais numerosos têm maior probabilidade de entrar em uniões prematuras.

De forma geral, os resultados evidenciam que as uniões prematuras resultam da interacção de factores demográficos, socioeconómicos e contextuais, reforçando a necessidade de abordagens integradas na formulação de políticas públicas destinadas à prevenção deste fenómeno em Moçambique.

4.3.3 Avaliação do Desempenho do Modelo de Regressão Logística

No presente estudo, as medidas de desempenho foram calculadas com base na matriz de confusão, a partir da qual se avalia a capacidade preditiva do modelo. Nesta matriz, os Verdadeiros Positivos (TP) correspondem aos casos de união prematura correctamente classificados pelo modelo, enquanto os Verdadeiros Negativos (TN) representam os casos sem ocorrência de união prematura correctamente identificados. Os Falsos Positivos (FP) referem-se aos casos em que o modelo indicou a ocorrência de união prematura quando esta não se verificou, ao passo que os Falsos Negativos (FN) dizem respeito aos casos de união prematura que não foram detectados pelo modelo. A acurácia representa a proporção global de classificações correctas, enquanto as medidas de sensibilidade e a especificidade avaliam, respectivamente, a capacidade do modelo para identificar correctamente os casos classificados como positivos e negativos.

O Modelo 1, estimado por regressão logística ajustada ao desenho amostral complexo (*svyglm*), apresentou um Pseudo-R^2 de McFadden igual a 0,269, indicando um nível moderado de ajustamento. Este resultado sugere que o conjunto de variáveis sociodemográficas, económicas e familiares incluídas no modelo explica uma parcela relevante da variabilidade associada à ocorrência de união prematura, embora permaneça espaço para a inclusão de outros factores explicativos não observados.

No que respeita ao desempenho preditivo, o modelo alcançou uma acurácia de 74,8%, evidenciando uma capacidade global satisfatória de classificação dos casos. Todavia, observa-se um desempenho assimétrico na identificação dos casos positivos e negativos. A sensibilidade, correspondente à proporção de casos de união prematura correctamente identificados, foi de 56,2%, indicando que o modelo apresenta maior dificuldade em detectar todos os casos positivos. Em contrapartida, a especificidade atingiu 84,3%, revelando uma elevada capacidade do modelo para identificar correctamente os casos nos quais não ocorreu união prematura.

A avaliação detalhada da matriz de confusão mostra que o modelo classificou correctamente 1 766 casos de união prematura (verdadeiros positivos) e 5 151 casos sem união prematura (verdadeiros negativos). No entanto, registaram-se 1 378 falsos negativos, correspondentes a casos de união prematura não identificados pelo modelo, e 958 falsos positivos, em que o modelo previu a ocorrência de união prematura quando esta não se verificou. Estes resultados confirmam a tendência do modelo para privilegiar a correcta identificação dos casos negativos, em detrimento da detecção completa dos casos positivos.

Tabela 4.4: Desempenho do Modelo de Regressão Logística considerando o desenho amostral

Modelo	**Pseudo-R^2**	**Acurácia**	**Sensibil**	**Especif**	**TP**	**TN**	**FP**	**FN**
Modelo 1 (svyglm)	0.2688	0.7475	0.5617	0.8432	1766	5151	958	1378

4.3.4 Curva ROC e Área sob a Curva

A Figura 4.6 ilustra a curva ROC referente ao Modelo 1, estimado por modelo de regressão logística ajustada ao desenho amostral complexo. A curva ROC representa a relação entre a sensibilidade e a especificidade do modelo para diferentes pontos de corte da probabilidade prevista, permitindo avaliar a sua capacidade discriminatória de forma independente do limiar de classificação adoptado.

Observa-se que a curva se encontra consistentemente acima da linha diagonal de referência, correspondente a um classificador aleatório, evidenciando que o modelo apresenta capacidade discriminatória claramente superior ao acaso. A Área sob a Curva (AUC) apresentou um valor de **AUC = 0,821**, evidenciando uma boa capacidade de discriminação do modelo na distinção entre indivíduos com e sem ocorrência de união prematura, Em termos práticos, este resultado sugere que o modelo apresenta forte capacidade para atribuir valores de probabilidade mais elevados para os casos positivos do que os casos negativos.

Este desempenho é consistente com as medidas obtidas a partir da matriz de confusão, reforçando a adequação do modelo para a análise das uniões prematuras, ainda que se observe margem para melhorias na identificação completa dos casos classificados como positivos.

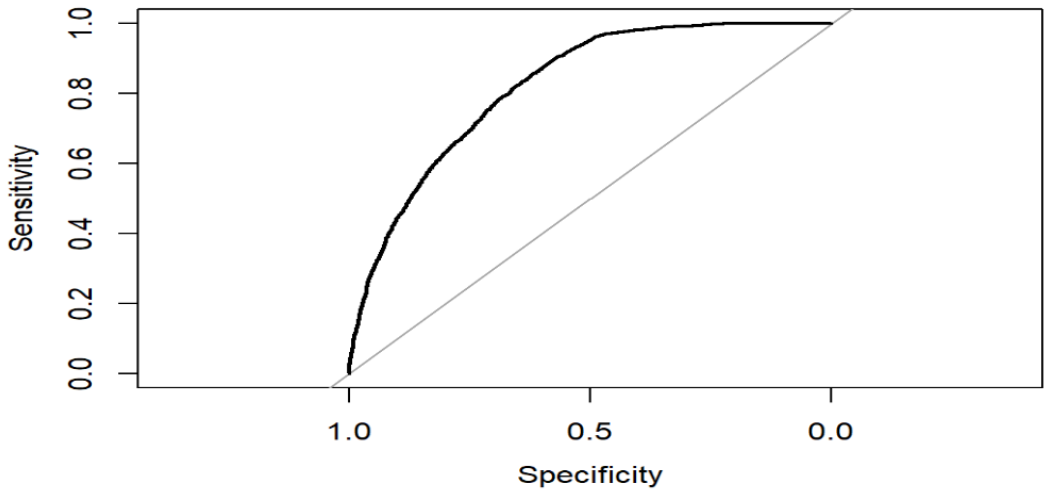


Figura 4.6: Curva ROC do Modelo

4.3.5 Validação Cruzada do Modelo de Regressão Logística

O desempenho do modelo de regressão logística foi avaliado através de um procedimento de validação cruzada do tipo *k*-fold, com $k = 5$, com vista a assegurar maior robustez e rigor na estimação do desempenho preditivo. Este método baseia-se na divisão do conjunto de dados em cinco subconjuntos aproximadamente iguais, sendo o modelo ajustado em quatro deles e validado no subconjunto remanescente, repetindo-se o processo até que todos os subconjuntos sejam utilizados como conjunto de validação.

Esta abordagem permite reduzir a variabilidade associada a uma única partição dos dados e fornece estimativas mais estáveis e fiáveis das medidas de desempenho. Na sequência, apresentam-se os principais resultados observados no processo de validação cruzada.

A Tabela 4.5 apresenta as medidas de desempenho associadas ao modelo de Regressão Logística obtidas por procedimento de validação cruzada do tipo *k*-fold, com $k = 5$. O modelo revelou uma acurácia média de 74,7%, com intervalo de confiança a 95% $(0,7364;\ 0,7577)$, indicando estabilidade do desempenho ao longo das partições.

O coeficiente Kappa foi de 0,4201, evidenciando concordância moderada entre as classificações previstas e observadas, para além do acaso. A sensibilidade estimada foi de 56,9%, enquanto a especificidade atingiu 83,9%, reflectindo um desempenho assimétrico, com maior capacidade de identificação correcta dos casos negativos.

Os valores preditivos apresentaram-se consistentes com este padrão, com PPV de 64,6% e NPV de 79,0%. A prevalência observada da união prematura foi de 34,0%, sendo a prevalência detectada pelo modelo de 30,0%, o que indica uma subestimação moderada dos casos positivos e a taxa de detecção foi de 19,4%.

A acurácia balanceada foi de 70,4%, confirmando desempenho global consistente mesmo na presença de desbalanceamento entre classes. Adicionalmente, a Área sob a curva ROC (AUC) apresentou um valor de 0,8194, evidenciando boa capacidade discriminatória do modelo.

Em termos globais, os resultados confirmam a adequação da regressão logística para fins preditivos, apresentando desempenho robusto e estabilidade estatística.

Tabela 4.5: Medidas de desempenho do modelo de Regressão Logística

Medida	**Valor**
Acurácia	0,7470
Intervalo de Confiança (95%)	(0,7364 ; 0,7577)
Kappa	0,4201
Sensibilidade (Recall)	0,5691
Especificidade	0,8389
Valor Preditivo Positivo (PPV)	0,6460
Valor Preditivo Negativo (NPV)	0,7904
Prevalência	0,3404
Taxa de Detecção	0,1937
Prevalência Detectada	0,3000
Acurácia Balanceada	0,7040
Área sob a Curva (AUC)	0,8194

Nota: O intervalo de confiança a 95% refere-se à acurácia global, calculado pelo método binomial exacto (Clopper-Pearson). As restantes medidas são apresentadas como estimativas pontuais.

4.3.6 Curva ROC do Modelo de Regressão Logística

A Figura 4.7 ilustra a curva ROC associada ao modelo de regressão logística obtida através de validação cruzada do tipo *k*-fold ($k = 5$). A curva ROC permite avaliar a capacidade discriminatória do modelo entre indivíduos com e sem ocorrência de união prematura, considerando diferentes pontos de corte das probabilidades previstas.

Observa-se que a curva se encontra consistentemente acima da linha diagonal de referência, correspondente a um classificador aleatório, indicando que o modelo apresenta um desempenho superior ao acaso. A AUC média foi de **0,8196**, resultado que evidencia boa capacidade discriminatória do modelo.

Este desempenho é consistente com as medidas obtidas a partir da matriz de confusão, reforçando a adequação do modelo para a análise das uniões prematuras, ainda que se observe margem para melhorias na identificação dos casos positivos.

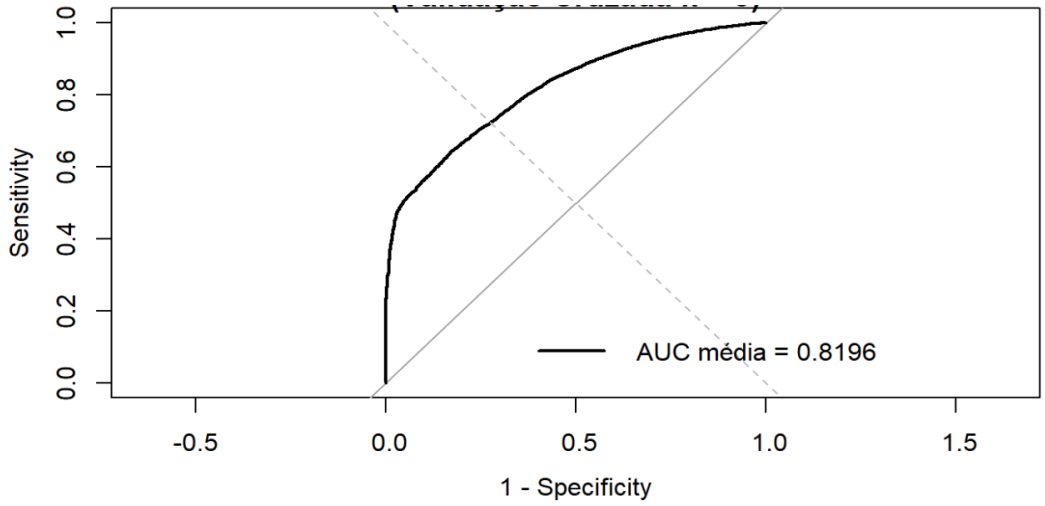


Figura 4.7: Curva ROC de regressão logística com Validação Cruzada

4.3.7 Comparação do Desempenho entre o Modelo Inicial e o Modelo com Validação Cruzada

Com o objectivo de avaliar a robustez do desempenho preditivo, procedeu-se à comparação entre o Modelo Inicial de Regressão Logística ajustado ao desenho amostral complexo (*svyglm*) e o modelo avaliado através de validação cruzada *k*-fold ($k = 5$). Enquanto o modelo inicial reflecte o desempenho obtido num único ajuste, a validação cruzada permite obter estimativas mais estáveis e menos dependentes de uma partição específica dos dados.

No Modelo Inicial, observou-se um Pseudo-R^2 de McFadden = 0,269, indicando um nível moderado de ajustamento e sugerindo que o conjunto de variáveis incluídas explica uma parcela relevante da variabilidade associada à ocorrência de uniões prematuras. Em termos preditivos, o modelo apresentou Acurácia = 74,8%, Sensibilidade = 56,2% e Especificidade = 84,3%, evidenciando maior capacidade de identificação correcta dos casos negativos. A capacidade discriminatória global, avaliada pela curva ROC foi elevada com AUC = 0,821, confirmando desempenho claramente superior ao acaso.

Por sua vez, o modelo avaliado por validação cruzada (k = 5) apresentou medida consistentes: Acurácia média = 74,7% (IC95%: $(0,7364;\ 0,7577)$), Sensibilidade = 56,9%, Especificidade = 83,9% e AUC média = 0,8194. A proximidade entre as medida do modelo inicial e as estimativas médias obtidas na validação cruzada indica que o desempenho observado no ajuste único é estável e generalizável, não sendo resultado de uma partição particular dos dados.

Em síntese, a validação cruzada confirma a robustez do modelo inicial, mantendo níveis semelhantes de acurácia e capacidade discriminatória. No entanto, o padrão assimétrico (especificidade superior à sensibilidade) mantém-se, sugerindo margem para melhorias na detecção

completa dos casos positivos.

Tendo em conta as limitações observadas no desempenho preditivo da Regressão Logística, nomeadamente ao nível da sensibilidade, procede-se, na secção seguinte, à aplicação de métodos de aprendizagem automática, em particular o algoritmo de Gradient Boosting, com o objectivo de avaliar eventuais ganhos na capacidade de previsão das uniões prematuras.

4.4 Modelo XGBoost

Nesta secção são apresentados os resultados obtidos com a aplicação do modelo XGBoost, com destaque para o seu desempenho preditivo, avaliado por meio da curva ROC e da respectiva área sob a curva (AUC), bem como para a identificação das variáveis com maior importância relativa na predição das uniões prematuras entre mulheres moçambicanas.

4.4.1 Avaliação do Desempenho do Modelo XGBoost

O desempenho global do modelo XGBoost foi avaliado com base na sua capacidade de discriminar entre mulheres em união prematura e não prematura, utilizando como principal medidas a área sob a curva ROC. Os resultados demonstram que o modelo alcançou uma AUC de 0,839 no conjunto de teste, evidenciando uma boa capacidade discriminatória.

A Figura 4.8 ilustra a curva ROC do modelo, observando-se um afastamento consistente da linha de referência associada a uma classificação aleatória, o que demonstra a aptidão do modelo para captar padrões relevantes nos dados. Este desempenho sugere que o XGBoost consegue equilibrar de forma adequada a sensibilidade e a especificidade na predição das uniões prematuras.

Adicionalmente, a diferença entre a AUC do conjunto de treino (0,888) e do conjunto de teste (0,839) revela-se moderada, sugerindo adequada capacidade de generalização e inexistência de sobreajustamento relevante. O recurso ao critério de early stopping permitiu determinar o número óptimo de iterações, contribuindo para a robustez do modelo.

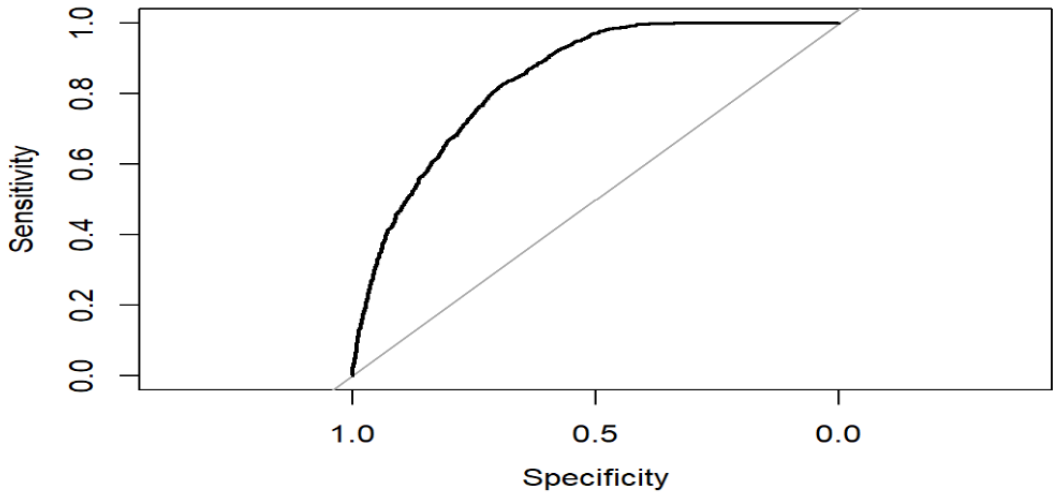


Figura 4.8: Curva ROC do modelo XGBoost

A Tabela 4.6 apresenta, de forma sintética, as principais medidas de desempenho do modelo XGBoost. Para além da AUC, observa-se uma acurácia de 0,756 e um coeficiente Kappa de 0,439, indicando um nível moderado de concordância entre as classificações previstas pelo modelo e os valores observados.

Relativamente ao desempenho por classe, o modelo registou sensibilidade de 0,578, evidenciando capacidade moderada para identificar correctamente os casos de união prematura, e uma especificidade de 0,848, indicando elevada capacidade na identificação dos casos sem ocorrência do fenómeno. A acurácia balanceada de 0,713 reforça o bom equilíbrio do modelo na classificação entre as duas classes.

Tabela 4.6: Medidas de desempenho preditivo do modelo XGBoost

Medida	**Valor**
Desempenho global do modelo	
AUC	0.839
IC 95% da AUC	[0.827; 0.851]
AUC (Conjunto de Teste)	0.839
Acurácia	0.756
Kappa	0.439
Desempenho por Classe	
Sensibilidade	0.578
Especificidade	0.848
Valor preditivo positivo (PPV)	0.663
Valor preditivo negativo (NPV)	0.795
Acurácia balanceada	0.713

4.4.2 Avaliação do Desempenho do Modelo XGBoost por Validação Cruzada

A Tabela 4.7 apresenta as medidas de desempenho preditivo do modelo XGBoost obtidas através do procedimento de validação cruzada *k*-fold (k=5), reportando a média e o desvio-padrão das medidas avaliadas. Os resultados evidenciam um desempenho consistente do modelo ao longo dos diferentes subconjuntos de validação.

Em termos globais, o modelo revelou uma AUC média de 0,831 (DP = 0,009), indicando boa capacidade discriminatória na distinção entre mulheres em união prematura e não prematura. O reduzido desvio-padrão observado sugere estabilidade do desempenho entre os diferentes *folds*, reforçando a robustez do modelo.

A acurácia média de 0,751 (DP = 0,008) reflecte uma elevada proporção de classificações correctas, enquanto o coeficiente Kappa médio de 0,424 (DP = 0,016) indica um nível moderado de concordância entre as classificações estimadas e os valores observados, para além do esperado por acaso.

No que respeita ao desempenho por classe, o modelo apresentou uma sensibilidade média de 0,559 (DP = 0,013), evidenciando uma capacidade moderada de identificar correctamente os casos de união prematura. Em contrapartida, a especificidade média de 0.850 (DP = 0.012)

indica elevada capacidade do modelo para identificar correctamente os casos sem ocorrência de união prematura.

Os valores médios do valor preditivo positivo (PPV) e o valor preditivo negativo (NPV) foram de 0.658 (DP = 0.017) e 0.789 (DP = 0.005), respectivamente, indicando níveis aceitáveis de fiabilidade das previsões realizadas pelo modelo.

Por fim, a acurácia balanceada média de 0.704 (DP = 0.007) confirma um desempenho equilibrado do modelo na presença de possíveis desequilíbrios entre as classes, reforçando a adequação do XGBoost para a análise da união prematura.

Tabela 4.7: Medidas de desempenho do modelo XGBoost obtidas por validação cruzada

Medida	**Média**	**DP**
AUC	0,831	0,009
Acurácia	0,751	0,008
Kappa	0,424	0,016
Sensibilidade	0,559	0,013
Especificidade	0,850	0,012
Valor Preditivo Positivo (PPV)	0,658	0,017
Valor Preditivo Negativo (NPV)	0,789	0,005
Acurácia Balanceada	0,704	0,007

4.4.3 Avaliação do Desempenho do Modelo XGBoost por meio da Curva ROC

A Figura 4.9 ilustra a curva ROC associada ao modelo XGBoost ajustado para a predição das uniões prematuras. Observa-se que a curva se afasta de forma consistente da linha diagonal de referência, correspondente a uma classificação aleatória, evidenciando que o modelo apresenta capacidade discriminatória relevante.

A **área sob a curva (AUC = 0,817)** evidencia um **bom desempenho preditivo**, reflectindo a aptidão do modelo para distinguir correctamente entre mulheres em união prematura e não prematura. De acordo com critérios amplamente utilizados na literatura, valores de AUC superiores a 0,80 são indicativos de modelos com boa capacidade de discriminação. Adicionalmente, a forma da curva sugere um equilíbrio adequado entre **sensibilidade** e **especificidade**, o que indica que o modelo consegue identificar uma proporção considerável de casos de união prematura sem comprometer excessivamente a correcta classificação dos casos não prematuros. Estes resultados confirmam a adequação do modelo XGBoost para a análise do fenómeno em estudo e reforçam a sua utilidade como ferramenta preditiva no contexto das uniões prematuras.

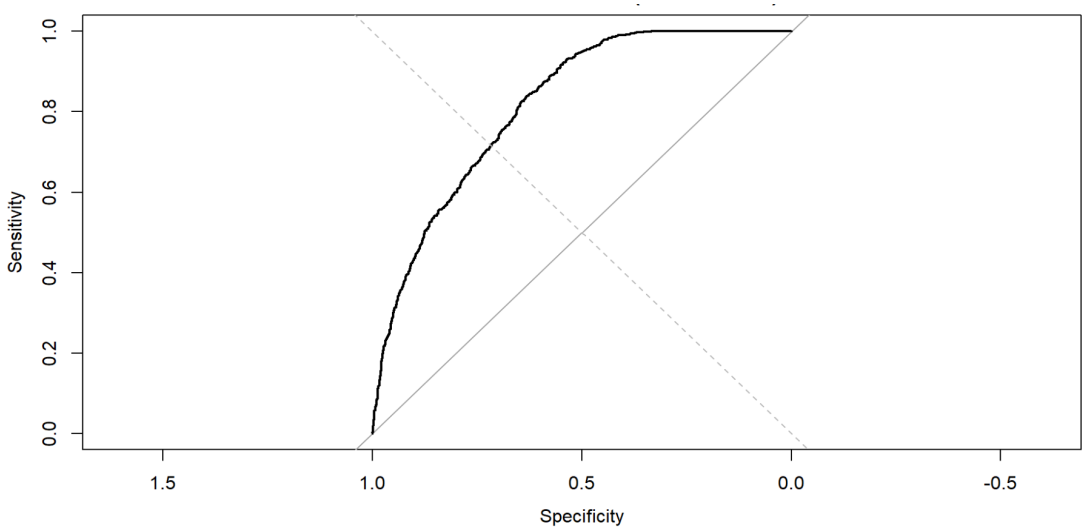


Figura 4.9: Curva ROC do modelo XGBoost para a predição da união prematura

4.4.4 Ajuste Hiperparamétrico do Modelo XGBoost

De modo a optimizar o desempenho do modelo, procedeu-se ao ajuste hiperaparamétrico utilizando o método de pesquisa em grelha (*grid search*) combinado com validação cruzada. Este procedimento permitiu avaliar sistematicamente diferentes combinações de hiperparâmetros do algoritmo *Extreme Gradient Boosting* (XGBoost), com o objectivo de identificar a configuração que maximiza a capacidade preditiva do modelo.

Foram considerados os principais hiperparâmetros associados à construção das árvores de decisão e ao processo de aprendizagem do modelo, nomeadamente:

- taxa de aprendizagem (*learning rate*), representada por η (`eta`);
- profundidade máxima das árvores (`max_depth`);
- peso mínimo das observações em cada nó terminal (`min_child_weight`);
- proporção de observações utilizadas em cada árvore (`subsample`);
- proporção de variáveis utilizadas na construção de cada árvore (`colsample_bytree`);
- parâmetro de regularização associado à redução de divisões nas árvores (`gamma`).

A selecção da melhor combinação de hiperparâmetros foi realizada com base no desempenho médio obtido durante a validação cruzada, utilizando como critério principal a área sob a curva ROC (*Area Under the Curve* – AUC). Este procedimento contribuiu para a diminuição do risco de sobreajustamento (*overfitting*) e para a melhoria da capacidade de generalização do modelo.

4.4.5 Avaliação das Combinações de Hiperparâmetros e Identificação da Melhor Configuração do Modelo XGBoost

Os resultados do ajuste hiperparamétrico apresentados na Tabela 4.8 indicam que, entre as 64 combinações avaliadas, a melhor configuração do modelo foi obtida com taxa de aprendizagem $\eta = 0.05$, profundidade máxima das árvores igual a 3 e peso mínimo das observações por nó terminal igual a 1. Adicionalmente, o modelo utilizou todas as observações em cada árvore (`subsample` = 1.0) e 80% das variáveis disponíveis em cada divisão (`colsample_bytree` = 0.8), não sendo aplicada penalização adicional através do parâmetro de regularização `gamma`.

Conforme evidenciado na Tabela 4.8, esta configuração apresentou uma AUC média de 0,832 na validação cruzada, com desvio-padrão de 0,008, indicando bom desempenho discriminatório e elevada estabilidade entre os subconjuntos de validação. O número óptimo de iterações foi estimado em 581 árvores, sugerindo um equilíbrio adequado entre a complexidade do modelo e a capacidade de generalização.

Tabela 4.8: Resumo do ajuste hiperparamétrico do modelo XGBoost

Parâmetro	**Valor**
Número de combinações avaliadas	64
Melhor combinação de hiperparâmetros	
Taxa de aprendizagem (η, `eta`)	0.05
Profundidade máxima da árvore (`max_depth`)	3
Peso mínimo das observações por nó (`min_child_weight`)	1
Proporção de amostras (`subsample`)	1.0
Proporção de variáveis por árvore (`colsample_bytree`)	0.8
Parâmetro de regularização (`gamma`)	0
Número óptimo de iterações (`best_nrounds`)	581
AUC média na validação cruzada	0.832413
Desvio-padrão da AUC	0.008247

4.4.6 Avaliação do Desempenho do Modelo XGBoost na Validação Cruzada e no Conjunto de Teste

Os resultados apresentados na Tabela 4.9 demonstram o desempenho do XGBoost quer na validação cruzada quer no conjunto de teste. Durante a validação cruzada, o modelo apresentou uma AUC média de 0,832, e desvio-padrão de 0,008, indicando bom desempenho discriminatório e elevada estabilidade entre os diferentes subconjuntos de validação. O reduzido desvio-padrão

observado sugere ainda baixa variabilidade nas estimativas ao longo dos *folds*, reforçando a robustez do modelo.

Relativamente ao conjunto de teste, o modelo ajustado apresentou uma AUC igual a 0,817 e uma acurácia de 0,741, evidenciando uma boa capacidade de discriminação entre indivíduos com e sem ocorrência de união prematura. O coeficiente Kappa foi de 0,408, evidenciando concordância moderada entre os valores observados e os previstos pelo modelo.

Em termos de capacidade de classificação, verificou-se que a especificidade (0.832) foi superior à sensibilidade (0.566), indicando que o modelo apresenta maior capacidade de identificar correctamente os casos negativos (indivíduos sem união prematura) do que os casos positivos. O valor da acurácia balanceada (0.699) sugere um desempenho global razoável do modelo, considerando o equilíbrio entre sensibilidade e especificidade.

Tabela 4.9: Desempenho do modelo XGBoost na validação cruzada e no conjunto de teste

Medidas	**Validação cruzada**	**Conjunto de teste**
AUC	0.832	0.817
Desvio-padrão da AUC	0.008	–
Acurácia	–	0.741
Kappa	–	0.408
Sensibilidade	–	0.566
Especificidade	–	0.832
Valor preditivo positivo (PPV)	–	0.634
Valor preditivo negativo (NPV)	–	0.788
Acurácia balanceada	–	0.699

4.4.7 Importância das Variáveis no Modelo

A Tabela 4.10 sintetiza a importância relativa das principais variáveis explicativas consideradas no modelo XGBoost, utilizado na predição da ocorrência de uniões prematuras. Os resultados indicam que a idade da primeira relação sexual constitui o factor mais relevante para o desempenho do modelo, destacando-se particularmente as categorias entre os 18 e os 24 anos e entre os 15 e os 17 anos, que apresentam os maiores valores da medida de importância baseada em *gain*. Este resultado sugere que o momento de iniciação da vida sexual está fortemente associado à ocorrência de uniões prematuras.

Entre as variáveis relacionadas com a situação conjugal, observa-se que o estado civil, tanto na condição de *em união* como *fora de união*, apresenta contributo significativo para a capacidade

preditiva do modelo. Estes resultados indicam que a condição conjugal da mulher está estreitamente associada a probabilidade de ocorrencia de uniões prematuras. A presença da categoria *nunca teve relação sexual / dados ausentes* entre as variáveis mais relevantes evidencia-se ainda a heterogeneidade dos perfis reprodutivos considerados na análise.

No que se refere às características socioeconómicas, destaca-se o nível de escolaridade, particularmente o ensino primário, bem como a variável que indica a ausência do parceiro no agregado familiar. Embora apresentem menor contribuição relativa em comparação com as variáveis comportamentais e conjugais, estes factores indicam que aspectos ligados à educação e à estrutura familiar também influenciam o risco de ocorrência de uniões prematuras.

Outras variáveis, como o quintil de riqueza e alguns grupos etários, apresentam menor importância relativa no modelo, embora contribuam para a explicação global do fenómeno. A inclusão destes factores reforça o carácter multifactorial das uniões prematuras, resultante da interacção entre características demográficas, comportamentais, socioeconómicas e contextuais.

De forma geral, a análise da relevância relativa das variáveis no modelo XGBoost evidencia que as uniões prematuras estão maioritariamente associadas a características relacionadas com as trajectórias sexuais e conjugais das mulheres, sendo também influenciadas por factores socioeconómicos e demográficos.

Tabela 4.10: Importância das variáveis no modelo XGBoost

Variável	**Gain**	**Cover**	**Frequência**
Idade da primeira relação sexual (18–24 anos)	0.318	0.116	0.057
Idade da primeira relação sexual (15–17 anos)	0.138	0.077	0.055
Estado civil: Em união	0.124	0.045	0.040
Nunca teve relação sexual	0.104	0.063	0.071
Estado civil: Fora de união	0.092	0.046	0.036
Nível de escolaridade: Primário	0.037	0.027	0.029
Marido não reside no agregado	0.030	0.011	0.034
Quintil de riqueza mais elevado	0.014	0.007	0.019
Grupo etário 45–49 anos	0.013	0.043	0.028
Grupo etário 20–24 anos	0.011	0.018	0.017

A Figura 4.10 apresenta a representação gráfica da importância relativa das variáveis no modelo, permitindo visualizar de forma comparativa o contributo de cada factor na melhoria do desempenho preditivo.

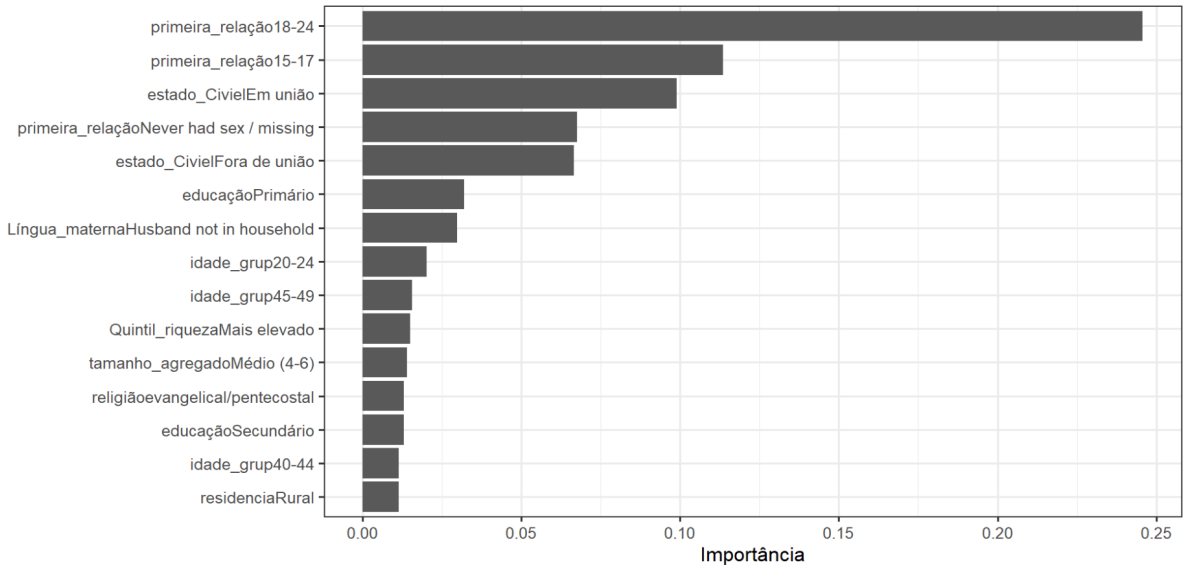


Figura 4.10: Importância das variáveis no modelo XGBoost

4.4.8 Comparação entre Regressão Logística e XGBoost

A comparação entre a Regressão Logística ajustada ao desenho amostral complexo (*svyglm*) e o modelo XGBoost permite avaliar diferenças no desempenho preditivo, na estabilidade dos resultados e na capacidade de generalização dos modelos. Ambos os métodos apresentaram desempenho satisfatório na predição da ocorrência das uniões prematuras, evidenciando uma boa capacidade discriminatória e níveis semelhantes de acurácia.

No caso da Regressão Logística, o modelo apresentou AUC de 0.821 no conjunto de teste e AUC média de 0.819 na validação cruzada ($k = 5$), indicando boa capacidade de discriminação entre mulheres em união prematura e não prematura. Em termos de classificação, o modelo apresentou acurácia de 0.748, sensibilidade de 0.562 e especificidade de 0.843 no conjunto de teste. Os resultados obtidos indicam que o modelo apresenta melhor desempenho na identificação correcta dos casos negativos em comparação com os positivos. Na validação cruzada, o desempenho manteve-se estável, com acurácia média de 0.747, sensibilidade de 0.569, especificidade de 0.839 e acurácia balanceada de 0.704.

Relativamente ao modelo XGBoost, procedeu-se inicialmente ao ajuste hiperaparamétrico utilizando pesquisa em grelha (*grid search*) combinada com validação cruzada, tendo sido avaliadas 64 combinações de hiperparâmetros. A melhor configuração foi obtida com taxa de aprendizagem $\eta = 0.05$, profundidade máxima das árvores igual a 3 (`max_depth = 3`), peso mínimo das observações por nó terminal igual a 1 (`min_child_weight = 1`), utilização de todas as observações em cada árvore (`subsample = 1.0`) e utilização de 80% das variáveis disponíveis em cada divisão (`colsample_bytree = 0.8`), não sendo aplicada penalização

adicional através do parâmetro de regularização (`gamma = 0`). O número óptimo de iterações estimado foi de 581 árvores.

Durante a validação cruzada, o modelo XGBoost apresentou AUC média de 0.832 com desvio-padrão de 0.008, indicando bom desempenho discriminatório e elevada estabilidade entre os subconjuntos de validação. Contudo, no conjunto de teste, o modelo apresentou AUC de 0.817, acurácia de 0.741, sensibilidade de 0.566, especificidade de 0.832 e acurácia balanceada de 0.699. Estes valores são ligeiramente inferiores aos observados na Regressão Logística, sugerindo que os ganhos observados durante o processo de validação cruzada não se traduziram integralmente em melhor desempenho fora da amostra.

Para além da avaliação do desempenho preditivo, o modelo XGBoost permitiu ainda analisar a importância relativa das variáveis explicativas na predição das uniões prematuras. Os resultados indicaram que a idade na primeira relação sexual constitui o factor mais relevante no desempenho do modelo, destacando-se particularmente as categorias entre 18–24 anos e 15–17 anos. Variáveis relacionadas com o estado civil, nomeadamente as categorias *em união* e *fora de união*, também apresentaram contributo significativo para a capacidade preditiva do modelo.

Entre as variáveis socioeconómicas, destacaram-se o nível de escolaridade e a presença do parceiro fora do agregado familiar, enquanto outras variáveis demográficas e contextuais apresentaram menor contribuição relativa. A possibilidade de identificar e quantificar a contribuição relativa das variáveis constitui uma vantagem adicional do modelo XGBoost na exploração de padrões complexos nos dados.

Em síntese, os dois modelos apresentaram desempenho preditivo semelhante na análise das uniões prematuras. A Regressão Logística evidenciou maior estabilidade entre os indicadores obtidos no conjunto de teste e durante a validação cruzada, particularmente no que se refere aos valores de AUC e acurácia, que se mantiveram próximos entre si. Esta proximidade sugere menor variabilidade no desempenho do modelo, indicando maior robustez na sua capacidade de generalização. Adicionalmente, a Regressão Logística apresenta a vantagem de permitir uma interpretação mais directa dos efeitos das variáveis explicativas, através das razões de chances (*odds ratios*).

Por outro lado, o modelo XGBoost mostrou-se útil sobretudo para fins preditivos e exploratórios, permitindo a optimização automática do desempenho através do ajuste de hiperparâmetros, bem como a identificação da importância relativa das variáveis no modelo. Assim, enquanto a Regressão Logística se revela mais adequada para análises explicativas e inferenciais, o XGBoost constitui uma ferramenta complementar relevante para a análise preditiva e para a identificação de relações não lineares entre as variáveis.

Tabela 4.11: Comparação do desempenho preditivo entre Regressão Logística e XGBoost

Medida	**Logística (Teste)**	**Logística (CV $k = 5$)**	**XGBoost (Teste)**	**XGBoost (CV $k = 5$)**
AUC	0,821	0,819	0,817	0,831 $\pm$ 0,009
Acurácia	0,748	0,747	0,741	0,751 $\pm$ 0,008
Sensibilidade	0,562	0,569	0,566	0,559 $\pm$ 0,013
Especificidade	0,843	0,839	0,832	0,850 $\pm$ 0,012
Kappa	–	0,420	0,408	0,424 $\pm$ 0,016
PPV	–	0,646	0,634	0,658 $\pm$ 0,017
NPV	–	0,790	0,788	0,789 $\pm$ 0,005
Acurácia balanceada	–	0,704	0,699	0,704 $\pm$ 0,007

4.5 Discussão

Os resultados obtidos indicam que tanto a Regressão Logística, ajustada ao desenho amostral complexo, como o modelo XGBoost apresentam um desempenho adequado na análise das uniões prematuras entre mulheres moçambicanas. Os valores de AUC próximos de 0,821 evidenciam a boa capacidade discriminatória da Regressão Logística na distinção entre mulheres que entraram e não entraram em união prematura, sendo este desempenho considerado satisfatório de acordo com critérios amplamente discutidos na literatura (Hosmer & Lemeshow, 2000; Kuhn & Johnson, 2013). De forma semelhante, o modelo XGBoost revelou igualmente um desempenho discriminatório elevado, com AUC média de 0,832 na validação cruzada e 0,817 no conjunto de teste, reforçando a adequação dos algoritmos de *boosting* na modelação de fenómenos sociais complexos (Friedman, 2001; Chen & Guestrin, 2016).

A optimização do modelo XGBoost foi realizada através de ajuste hiperparamétrico utilizando pesquisa em grelha (*grid search*) combinada com validação cruzada, tendo sido avaliadas 64 combinações de hiperparâmetros. A melhor configuração foi obtida com taxa de aprendizagem $\eta = 0.05$, profundidade máxima das árvores igual a 3 e peso mínimo das observações por nó terminal igual a 1, resultando num modelo com boa capacidade de generalização. A AUC média de validação cruzada (0,832) e o baixo desvio-padrão observado (0,008) indicam elevada estabilidade do desempenho do modelo entre os diferentes subconjuntos de validação.

Apesar do desempenho global satisfatório, ambos os modelos evidenciaram um comportamento assimétrico entre sensibilidade e especificidade, demonstrando maior capacidade na correcta classificação dos casos sem união prematura do que na identificação completa dos casos positivos. Este padrão está em consonância com evidências reportadas na literatura sobre modelos de classificação com desfechos binários em contextos de dados desbalanceados, nos quais se observa uma tendência para privilegiar a classe maioritária (Kuhn & Johnson, 2013; Silva, 2019).

No presente estudo, a sensibilidade moderada sugere que uma proporção das uniões prematuras não é captada pelos modelos, possivelmente devido à sobreposição de perfis individuais e à ausência de variáveis que representem de forma mais directa normas sociais, dinâmicas familiares e factores comunitários, frequentemente apontados como determinantes relevantes na literatura sobre uniões prematuras (Hodges, 2015; Alem et al., 2020).

A Regressão Logística destacou-se como uma ferramenta particularmente relevante para fins explicativos, ao permitir identificar e quantificar associações ajustadas por meio de razão de chance. Variáveis como a idade à primeira relação sexual, o nível de escolaridade, o quintil de riqueza, o estado civil e a dimensão do agregado familiar apresentaram associações estatisticamente significativas, corroborando evidências de estudos conduzidos em contextos africanos e asiáticos, que apontam a educação feminina, a condição socioeconómica e o contexto familiar como determinantes da ocorrência de uniões prematuras (Alem et al., 2020; Mathabatha & Tsawe, 2023). Conforme salientado por Figueira (2006) e Cabral (2013), a Regressão Logística apresenta como uma das suas principais vantagens a obtenção de resultados facilmente interpretáveis, essenciais para a formulação de políticas públicas.

Por sua vez, o modelo XGBoost complementou a análise ao permitir captar relações não lineares e interacções complexas entre as variáveis explicativas, características inerentes aos métodos de aprendizagem automática baseados em árvores e *boosting* (Breiman, 2001; Friedman, 2001). A análise da importância das variáveis evidenciou a centralidade da idade na primeira relação sexual e do estado civil, seguidas por factores educacionais e características do agregado familiar. Estes resultados reforçam a interpretação de que as uniões prematuras estão fortemente associadas às trajectórias sexuais e conjugais das mulheres, sendo simultaneamente influenciadas por condições socioeconómicas e contextuais.

A avaliação do desempenho através de múltiplas medidas, nomeadamente AUC, sensibilidade, especificidade, coeficiente Kappa e acurácia balanceada, revelou-se essencial para uma apreciação mais completa da qualidade dos modelos. Conforme salientado por James et al. (2013) e Kuhn & Johnson (2013), a utilização exclusiva da acurácia pode conduzir a interpretações enviesadas, sobretudo em cenários de desbalanceamento da variável resposta. A aplicação da validação cruzada k-fold permitiu ainda confirmar a estabilidade e a capacidade de generalização dos modelos, reduzindo o risco de sobreajustamento e reforçando a robustez das conclusões.

Do ponto de vista comparativo, ambos os modelos evidenciaram desempenho preditivo semelhante. A Regressão Logística destacou-se pela maior consistência dos resultados entre o conjunto de teste e a validação cruzada, para além de possibilitar a interpretação directa dos efeitos das variáveis explicativas. Por sua vez, XGBoost, revelou-se particularmente eficaz na optimização do desempenho preditivo e na identificação da importância relativa das variáveis, oferecendo uma abordagem complementar baseada em aprendizagem automática. Assim, a uti-

lização combinada de métodos estatísticos clássicos e técnicas de *Machine Learning* revela-se adequada para o estudo das uniões prematuras, permitindo simultaneamente uma interpretação substantiva dos determinantes e exploração de padrões complexos nos dados.

No conjunto, os resultados indicam que as uniões prematuras em Moçambique são associadas por um conjunto interligado de factores demográficos, socioeconómicos, comportamentais e contextuais, reflectindo a complexidade do fenómeno. Estes resultados reforçam a necessidade de abordagens multidimensionais na formulação de políticas públicas orientadas para a prevenção das uniões prematuras e para a promoção dos direitos e do bem-estar das raparigas em Moçambique.

Capítulo 5

CONCLUSÕES E RECOMENDAÇÕES

5.1 Conclusões

O presente estudo aplicou técnicas de *Machine Learning* na análise dos factores associados às uniões prematuras em Moçambique, com base nos dados do Inquérito Demográfico e de Saúde (IDS) 2022–23 e recorrendo a uma abordagem integrada que combina métodos estatísticos clássicos e algoritmos de aprendizagem automática. Os resultados monstram que a união prematura constitui um fenómeno social relevante e persistente, influenciado por múltiplos factores de natureza demográfica, socioeconómica, comportamental e contextual.

A análise descritiva e bivariada revelou associações estatisticamente significativas entre a ocorrência de uniões prematuras e variáveis como a idade na primeira relação sexual, o nível de escolaridade, o quintil de riqueza, a área de residência, a província, a religião, a exposição aos meios de comunicação social e a dimensão do agregado familiar, confirmando o carácter multifactorial do fenómeno.

A Regressão Logística, ajustada ao desenho amostral complexo, apresentou boa capacidade discriminatória, com um valor de AUC próximo de 0,82, revelando-se particularmente adequada para fins explicativos. Este modelo permitiu identificar a direcção e a magnitude das associações entre os factores analisados e o desfecho, por meio das razões de chances (*odds ratios*). Verificou-se que níveis mais elevados de escolaridade se associam a uma menor probabilidade de entrada em união prematura, ao passo que determinados factores socioeconómicos e contextuais contribuem para o aumento do risco.

O modelo XGBoost foi optimizado através de ajuste hiperparamétrico, recorrendo a pesquisa em grelha (*grid search*) combinada com validação cruzada, tendo sido avaliadas 64 combinações de hiperparâmetros. A melhor configuração alcançou uma AUC média de 0,832 na validação cruzada, com desvio-padrão de 0,008, o que traduz elevada estabilidade do desempe-

nho preditivo. No conjunto de teste, o modelo registou uma AUC de 0,817 e uma acurácia de 0,741, valores próximos dos obtidos pela Regressão Logística e indicativos de um desempenho preditivo semelhante entre as duas abordagens.

A análise da importância das variáveis no modelo XGBoost mostrou que a idade na primeira relação sexual e o estado civil constituem os factores com maior contributo para o desempenho preditivo, seguidos do nível de escolaridade, da ausência do parceiro no agregado familiar e do quintil de riqueza. Estes resultados reforçam a interpretação de que as uniões prematuras estão fortemente associadas às trajectórias sexuais e conjugais das mulheres, em articulação com factores socioeconómicos e familiares.

De forma geral, o estudo evidencia que a utilização combinada de modelos explicativos e preditivos constitui uma abordagem metodológica robusta para o estudo das uniões prematuras. Enquanto a Regressão Logística se distingue pela sua capacidade interpretativa e inferencial, o XGBoost permite captar relações não lineares e quantificar a importância relativa das variáveis explicativas. No seu conjunto, os resultados confirmam que as uniões prematuras em Moçambique resultam da interacção de factores demográficos, socioeconómicos, comportamentais e contextuais, fornecendo evidência empírica relevante para apoiar a formulação de políticas públicas e o desenvolvimento de estratégias de intervenção orientadas para a prevenção e a redução deste fenómeno, bem como para a promoção do bem-estar das raparigas no país.

5.2 Recomendações

Com base nos resultados obtidos, apresentam-se as seguintes recomendações, orientadas para a prevenção e redução da ocorrência das uniões prematuras em Moçambique:

- **Reforço do acesso à educação feminina:** Promover políticas que incentivem a permanência das raparigas no sistema de ensino, sobretudo ao nível do ensino secundário e superior, tendo em conta o efeito protector da escolaridade na redução das uniões prematuras.
- **Promoção da educação sexual e reprodutiva:** Desenvolver programas de educação sexual abrangentes e adaptados ao contexto sociocultural, com foco no adiamento da iniciação sexual e na consciencialização sobre os riscos associados às uniões prematuras.
- **Intervenções em contextos socioeconómicos vulneráveis:** Implementar estratégias direccionadas para famílias em situação de maior vulnerabilidade económica, incluindo programas de apoio social e económico, de modo a reduzir factores de risco associados às uniões prematuras.

- **Fortalecimento das campanhas de comunicação social:** Expandir o acesso à informação através dos meios de comunicação social, promovendo campanhas de sensibilização sobre os impactos negativos das uniões prematuras e os benefícios da educação e autonomia feminina.

- **Integração de abordagens comunitárias e culturais:** Envolver líderes comunitários e religiosos na promoção de mudanças de normas sociais e culturais que perpetuam as uniões prematuras, incentivando práticas mais favoráveis ao desenvolvimento das raparigas.

- **Desenvolvimento de políticas públicas baseadas em evidência:** Utilizar os resultados de estudos empíricos, como o presente, para orientar a formulação de políticas públicas eficazes e contextualizadas, com foco na redução das desigualdades regionais e sociais.

- **Aperfeiçoamento de modelos preditivos:** Recomenda-se a incorporação de variáveis adicionais relacionadas com normas sociais, dinâmicas familiares e contextos comunitários em estudos futuros, de modo a melhorar a capacidade preditiva dos modelos e a identificação de grupos de risco.

- **Continuidade da investigação científica:** Incentivar a realização de estudos longitudinais e análises espaciais que permitam compreender melhor a evolução temporal e a distribuição geográfica das uniões prematuras em Moçambique.

REFERÊNCIAS BIBLIOGRÁFICAS

Agresti, A. (2013). *Categorical Data Analysis* (3rd ed.). Hoboken, NJ: John Wiley & Sons.

Aldrich, J. (1997). R. A. Fisher and the making of maximum likelihood 1912–1922. *Statistical Science*, 12(3), 162–176.

Alem, A. Z., & Sahu, S. (2020). Determinants and spatial distribution of child marriage in Bangladesh: A machine learning approach. *PLoS ONE*, 15(7), e0235597.

Batista, A. M. S. (2015). *Regressão logística: uma introdução ao modelo estatístico – exemplo de aplicação ao revolving credit*. In *Regressão Logística* (pp. 31–33). Lisboa: Vida Económica.

Belachew, T., Worku, M. G., Tesema, G. A., Teshale, A. B., Alem, A. Z., Yeshaw, Y., & Liyew, A. M. (2022). Determinants of early marriage among women in sub-Saharan Africa: A multilevel analysis. *BMC Women's Health*, 22(1), 1–12.

Berman, D. S., Buczak, A. L., Chavis, J. S., & Corbett, C. L. (2019). A survey of deep learning methods for cyber security. *Information*, 10(4), 122.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

Brownlee, J. (2018). *Machine Learning Mastery with XGBoost*. Machine Learning Mastery.

Bruce, P., & Bruce, A. (2019). *Practical Statistics for Data Scientists*. Sebastopol, CA: O'Reilly Media.

Cabral, C. I. S. (2013). *Aplicação do Modelo de Regressão Logística num Estudo de Mercado* (Projecto de Mestrado em Matemática Aplicada à Economia e à Gestão). Universidade de Lisboa, Lisboa, Portugal.

Cavalcante, L. C., Meneguette, R. I., Torres, R. H., Mano, L. Y., Gonçalves, V. P., Ueyama, J., Pessin, G., Nze, G. D. A., & Filho, G. P. R. (2022). Federated system for transport mode detection. *Energies*, 15(23), 9256.

Chagas, D. A. M. (2024). *Detecção de ataques de negação de serviço em SGBDs a partir de logs internos usando abordagens supervisionada e não supervisionada* (Dissertação de Mestrado em Engenharia Eléctrica). Universidade de Brasília, Brasília, DF.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, 19(1), 15–18.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, 34(2), 187–220.

Cox, D. R., & Snell, E. J. (1989). *Analysis of Binary Data* (2nd ed.). London: Chapman and Hall.

ENAP – Escola Nacional de Administração Pública. (2020). *Análise de dados em linguagem R*. Disponível em: `https://www.escolavirtual.gov.br/curso/325`.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874.

Figueira, C. V. (2006). *Modelos de regressão logística* (Dissertação de Mestrado). Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232.

Gama, J., Carvalho, A. C. P. L. F., Faceli, K., Lorena, A. C., & Oliveira, M. (2012). *Extração de conhecimento de dados: Data mining*. Lisboa: Edições Sílabo.

Gashaw, B. T., Schelling, E., Worku, A., & Melkamu, Y. A. (2019). Determinants of early marriage among women in Ethiopia. *BMC Women's Health*, 19(1), 1–6.

Géron, A. (2017). *Hands-on Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. Sebastopol, CA: O'Reilly Media.

Goldstein, H. (2011). *Multilevel Statistical Models* (4th ed.). Chichester: Wiley.

Gormez, Y., Aydin, Z., Karademir, R., & Gungor, V. C. (2020). A deep learning approach with Bayesian optimization and ensemble classifiers for detecting denial of service attacks. *International Journal of Communication Systems*, 33(11), e4401.

Governo de Moçambique. (2016). *Estratégia Nacional para a Prevenção e Combate às Uniões Prematuras*. Maputo: Governo de Moçambique.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2009). *Análise Multivariada de Dados* (6. ed., pp. 32–55). Porto Alegre: Bookman.

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.).

Morgan Kaufmann.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. New York: Springer.

Hauck, W. W., Jr., & Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association*, 72(360), 851–853.

Heeringa, S., West, B. T., & Berglund, P. A. (2017). *Applied Survey Data Analysis* (2nd ed.). Boca Raton: CRC Press.

Hodges, J. (2015). *Marriage, Childbearing and School Attendance Among Girls in Mozambique*. Maputo: UNICEF Moçambique.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression* (2nd ed.). New York: Wiley.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). New York: John Wiley & Sons.

Instituto Nacional de Estatística, Ministério da Saúde, & ICF. (2023). *Inquérito Demográfico e de Saúde 2022–2023*. Maputo, Moçambique e Rockville, Maryland, EUA: INE, MISAU & ICF.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R*. New York: Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning* (2nd ed.). New York: Springer.

Kaur, H., Pannu, H. S., & Malhi, A. K. (2019). A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Computing Surveys*, 52(4), 1–36.

Kleinbaum, D. G., & Klein, M. (2012). *Survival Analysis: A Self-Learning Text* (3rd ed.). New York: Springer.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 2, 1137–1143.

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. New York: Springer.

Malerba, D. (2023). *Predição de evasão universitária através de algoritmos de aprendizado supervisionado* (Dissertação de Mestrado). Universidade Federal de Alagoas, Maceió.

Maroco, J. (2007). *Análise Estatística com Utilização do SPSS*. Lisboa: Edições Sílabo.

Mathabatha, S., & Tsawe, M. (2023). Factors associated with early marriage among women in

South Africa: A machine learning approach. *BMC Women's Health*, 23(1), 1–12.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models* (2nd ed.). London: Chapman and Hall.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics* (pp. 105–142). New York: Academic Press.

Menard, S. (2002). *Applied Logistic Regression Analysis* (2nd ed.). Thousand Oaks: Sage Publications.

Mendes, F. R. (2023). *Modelos de Machine Learning aplicados à previsão de infeções hospitalares* (Dissertação de Mestrado). Universidade de Lisboa, Lisboa, Portugal.

Mesquita, P. S. B. (2014). *Um modelo de regressão logística para avaliação dos programas de pós-graduação no Brasil* (Dissertação de Mestrado). Universidade Estadual do Norte Fluminense – UENF, Campos dos Goytacazes, RJ.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78(3), 691–692.

Nelder, J., & Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, 135(3), 370–384. Obtido em `https://docs.ufpr.br/~taconeli/CE225/Artigo.pdf`

Organização das Nações Unidas (ONU). (2011). *Objetivos de Desenvolvimento Sustentável*. Nova Iorque: ONU.

Paixão, R., Silva, M., & Andrade, L. (2022). Aplicação de técnicas de aprendizado de máquina na análise de dados de saúde. *Revista Brasileira de Informática em Saúde*, 18(2), 45–58.

Pereira, A. L. (2023). *Aprendizado semi e auto-supervisionado aplicado à classificação multi-label de imagens de inspecções submarinas* (Dissertação de Mestrado em Engenharia Eléctrica). Pontifícia Universidade Católica do Rio de Janeiro, Brasil.

Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M. P., Shyu, M.-L., Chen, S.-C., & Iyengar, S. S. (2019). A survey on deep learning: Algorithms, techniques, and applications. *ACM Computing Surveys*, 51(5).

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC. *Journal of Machine Learning Technologies*, 2(1), 37–63.

Rao, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proceedings of the Cambridge Philosophical Society*, 44, 50–57.

Rao, J. N. K., & Scott, A. J. (1981). The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *Journal*

of the American Statistical Association, 76(374), 221–230.

Santos, R., Pereira, A., & Nogueira, F. (2019). Análise preditiva em saúde utilizando técnicas de machine learning. *Revista de Saúde Pública*, 53, 1–10.

Silva, A. F. (2019). *Ensaios sobre Economia Aplicada: Doações eleitorais, compras públicas, análise de políticas afirmativas e reprovação no Ensino Superior* (Tese de Doutoramento em Economia). Universidade Federal da Paraíba. Disponível em: `https://repositorio.ufpb.br/jspui/handle/123456789/20028`

Silva, C. M. (2023). *Uma ferramenta para análise e previsão dos dados da MIMIC-III* (Dissertação de Mestrado em Engenharia de Redes e Sistemas Informáticos). Universidade do Porto.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427–437.

Souza, A. M. de. (2020). *Machine Learning e a evasão escolar: Análise preditiva no suporte à tomada de decisão* (Tese de Doutoramento em Sistemas de Informação e Gestão do Conhecimento).

Steyerberg, E. W., et al. (2010). Assessing the performance of prediction models. *Epidemiology*, 21(1), 128–138.

Turkman, A., & Silva, G. (2000). *Modelos Lineares Generalizados – da teoria à prática*. Lisboa: Universidade Técnica de Lisboa. Obtido em `https://docs.ufpr.br/~taconeli/CE225/tp.pdf`

United Nations Children's Fund (UNICEF). (2015). *Child Marriage and Adolescent Pregnancy in Mozambique*. Maputo: UNICEF.

United Nations Children's Fund (UNICEF). (2016). *Child Marriage: A Statistical Overview*. New York: UNICEF.

United Nations Children's Fund (UNICEF). (2021). *Global Progress and Prospects on Child Marriage*. New York: UNICEF.

United Nations Children's Fund (UNICEF). (2022). *Child Marriage*. New York: UNICEF.

United Nations Department of Economic and Social Affairs (UNDESA). (2022). *World family planning: Meeting the changing needs for family planning – contraceptive use by age and method*. New York: United Nations.

United Nations Population Fund (UNFPA). (2014). *Adolescents, youth and the transformation of the future: The state of world population 2014*. New York: UNFPA.

United Nations Population Fund (UNFPA). (2016). *Child Marriage in Africa*. New York: UNFPA.

Universidade de Coimbra. (2004). *Introdução à inteligência artificial*. Departamento de Engenharia Informática. Disponível em: http://www.ppgia.pucpr.br/~picolotto/Weka/1020283132.pdf. Acesso em: 02 jan. 2025.

Vaz, F. E. C. R. (2020). *Modelação linear e extensões: Aplicação da regressão logística no estudo do cancro da mama* (Dissertação de Mestrado). Universidade Aberta, Lisboa, Portugal.

XGBoost Developers. (2022). *XGBoost Documentation*. Disponível em: https://xgboost.readthedocs.io/en/stable/.