UNIVERSIDADE
E D U A R D O
MONDLANE

FACULDADE DE CIÊNCIAS
Departamento de Matemática e Informática

Trabalho de Licenciatura em
Estatística

Aplicação da Aprendizagem Supervisionada na Determinação do Factores Associados a Perdas Gestacionais

Autora: Esperança António Munlela

Maputo, Julho de 2025

UNIVERSIDADE
E D U A R D O
MONDLANE

FACULDADE DE CIÊNCIAS
Departamento de Matemática e Informática

Trabalho de Licenciatura em
Estatística

Aplicação da Aprendizagem Supervisionada na Determinação do Factores Associados a Perdas Gestacionais

Autora: Esperança António Munlela

Supervisor: Prof Doutor Cachimo Assane

Maputo, Julho de 2025

Declaração de Honra

Declaro, por minha honra, que o presente Trabalho de Licenciatura é resultado da minha própria investigação, e que o processo foi concebido para ser submetido unicamente para a obtenção do grau de Licenciada em Estatística, na Faculdade de Ciências da Universidade Eduardo Mondlane.

Maputo, Junho de 2025

_____

(**Esperança António Munlela**)

Dedicatória

Dedico este trabalho à minha mãe Lúcia Ernesto Come e à memória do meu pai António Cássimo Munlela, cuja ausência nunca apagará o legado que deixou em mim.

A busca pelo conhecimento não é um caminho linear, mas um diálogo constante entre o que sabemos e o que ainda temos a descobrir.- Adaptação de Carl Sagan (Cosmos, 1980)

Agradecimentos

Agradeço, primeiramente, a Deus pela dádiva da vida e pela força que me concedeu para chegar até aqui, conservando-me ao longo de toda a minha vida e trajetória académica.

À minha mãe, Lúcia Ernesto Come, pelo amor incansável, pelas orações silenciosas e pela força que me sustentou. À você, minha gratidão eterna, porque este trabalho carrega também os sonhos que você depositou em mim.

Aos professores do curso de Estatística, pela dedicação, compromisso e pelo conhecimento partilhado ao longo de toda a formação. Agradeço também ao meu supervisor Prof. Doutor Cachimo Assane, pela pronta disposição em orientar este trabalho, pela paciência e pelas respostas rápidas e esclarecedoras que fizeram toda a diferença ao longo do processo.

Agradeço aos meus irmãos e irmãs, Sameira, Rossana, Bento, Arminda, António e Lúcia Munlela pelo apoio moral dado durante o processo académico, à minha Filha Evelyn Cassongo pela companhia durante a minha conclusão das disciplinas e motivação, e em especial à minha irmã Benilde Munlela, que foi mais que família: foi empurrão, inspiração e motivação. Teu exemplo e tua cobrança foram fundamentais para eu pudesse continuar mesmo quando quis desanimar. Ao meu noivo Bento João Cassongo, por me motivar com sua exigência e apoio firme, sempre acreditando em mim e lembrando do meu potencial, mesmo nos momentos em que eu mesma duvidava.

Um agradecimento muito especial à colega e amiga Ivone Pedro Ussivane que viveu este trabalho junto comigo. Obrigada pelas madrugadas em branco partilhadas, pelas ideias e dúvidas, motivações e gargalhadas. A tua parceria foi essencial para que este trabalho fosse concluido.

Às colegas Maida Tajú, Yura Matsinhe e Márcia Guambe, por cada partilha ao longo do curso, amigas que, mesmo à distância, torceram e vibraram comigo. Vocês fizeram os dias longos parecerem mais leves. A todos que, de alguma forma, contribuíram para esta caminhada: o meu mais profundo e sincero obrigado.

Lista de abreviaturas

AIC	Akaike Information Criterion
AUC	Área Sob a Curva (ROC)
CART	Classification and Regression Trees (Árvores de classificação e regressão)
IDS	Inquérito Demográfico de Saúde
INE	Instituto Nacional de Estatística
kNN	*k*-Nearest Neighbors
MISAU	Ministério da Saúde
OMS/WHO	Organização Mundial da Saúde/World Health Organization
OR	Odds Ratio (Razão de Chances)
ROC	Receiver Operating Characteristic
ROSE	Random Over Sampling Examples
SVM	Máquinas de vetores de suporte
UNICEF	Fundo das Nações Unidas para a Infância
VIF	Variance Inflation Factor (Fator de Inflação da Variância)
XGBoost	Extreme Gradient Boosting (Aumento de gradiente extremo)

Resumo

A perda gestacional ocorre quando a gravidez, por diversos factores, não resulta no nascimento de um bebé vivo, se tornando uma experiência de forte impacto físico e emocional para a mulher. Este estudo teve como objectivo identificar os factores associados às perdas gestacionais em Moçambique, por meio da aplicação de técnicas de Aprendizagem Supervisionada, nomeadamente a regressão logística e a árvore de decisão. Foram analisadas 2668 observações do Inquérito Demográfico e de Saúde (IDS) de 2023. foram submetidas 11 variáveis à análise descritiva, diagnóstico de multicolinearidade, tratamento de dados ausentes, balanceamento da variável resposta utilizando o método ROSE (Random Over-Sampling Examples) e construção de modelos preditivos. A idade materna avançada (35–49 anos) destacou-se como o factor de risco mais significativo para perda gestacional, com razões de chances variando entre 11,9 e 14,4. A província de residência também apresentou forte associação: mulheres residentes na Cidade de Maputo apresentaram uma probabilidade 10,5 vezes superior de perda gestacional em comparação com as residentes na província do Niassa. A idade da primeira relação sexual também se revelou influente: mulheres que iniciaram a vida sexual mais tardiamente apresentaram menor risco de perda gestacional (OR = 0,87). No que diz respeito ao desempenho dos modelos, a árvore de decisão superou a regressão logística em capacidade discriminativa, com F1-score de 0,504 contra os 0,465 e acurácia de 69,8% frente a 64,7%. A regressão logística apresentou uma sensibilidade mais elevada (81,7%) comparada com a árvore (74,5%), enquanto a árvore obteve melhor precisão (38,1% contra 32,5%). Ambas apresentaram acurácia balanceada próximas, 74,6% para a árvore e 71,6% para a regressão logística. Estes resultados sugerem uma ligeira vantagem da árvore na identificação de padrões e interacções entre variáveis relevantes.

Palavras-chaves: Aprendizagem Supervisionada, Árvore de decisão, Perdas gestacionais, Regressão logística

Abstract

Pregnancy loss occurs when a pregnancy, for various reasons, does not result in the birth of a live baby, representing an experience of significant physical and emotional impact on the woman. This study aimed to identify the factors associated with pregnancy loss in Mozambique through the application of supervised learning techniques, namely logistic regression and Árvore de decisãos. A total of 2,668 observations from the 2023 Demographic and Health Survey (DHS) were analyzed. Eleven variables were subjected to descriptive analysis, multicollinearity diagnosis, missing data treatment, response variable balancing using the ROSE (Random Over-Sampling Examples) method, and predictive model construction. Advanced maternal age (35–49 years) emerged as the most significant risk factor for pregnancy loss, with odds ratios ranging from 11.9 to 14.4. The province of residence also showed a strong association: women living in Maputo City had a 10.5 times higher probability of pregnancy loss compared to those living in Niassa Province. Age at first sexual intercourse also proved influential: women who initiated sexual activity later had a lower risk of pregnancy loss (OR = 0.87). Regarding model performance, the Árvore de decisão outperformed logistic regression in discriminative ability, with an F1-score of 0.504 versus 0.465 and an accuracy of 69.8% versus 64.7%. Logistic regression showed higher sensitivity (81.7%) compared to the tree (74.5%), while the tree obtained better accuracy (38.1% versus 32.5%). Both models showed similar balanced accuracy: 74.6% for the tree and 71.6% for logistic regression. These results suggest a slight advantage of the tree in identifying patterns and interactions between relevant variables.

Keywords: Decision tree, Logistic regression, Pregnancy losses, Supervised learning

Índice

Lista de Figuras

Lista de Tabelas

Capítulo 1

INTRODUÇÃO

1.1 Contextualização

O conceito de perda gestacional, segundo a Health Canada (2000), envolve um conjunto de situações de perda que podem ocorrer ao longo da gestação ou após o parto, incluindo aborto espontâneo, morte fetal (nado-morto), morte neonatal, interrupção médica da gravidez e interrupção voluntária da gravidez. As perdas gestacionais incluem também partos prematuros, natimortos e baixo peso ao nascer, que são a principal causa de morbidade neonatal, mortalidade e problemas físicos e psicológicos de longo prazo, segundo a Organização Mundial da Saúde (OMS, 2023).

De acordo com relatório da Fundo das Nações Unidas para a Infância (UNICEF) (2023), as crianças nascidas em África estão sujeitas ao maior risco de mortalidade infantil do mundo, num valor 15 vezes superior ao risco para as crianças na Europa e América. O risco de uma mulher ter um natimorto na África Subsaariana é sete vezes maior do que na Europa e América. Segundo Regassa *et al.* (2022), as probabilidades de um bebé nascer vivo em países de baixo rendimento são menores quando comparadas com as de um país desenvolvido.

Diversos estudos têm sido conduzidos com o objectivo de identificar os factores associados às perdas gestacionais, utilizando diferentes técnicas de Aprendizagem Supervisionada. Um exemplo é o estudo realizado por Regassa *et al.* (2022), que aplicou séries temporais e regressão logística para analisar as tendências e os determinantes das perdas gestacionais na Etiópia. Este estudo concluiu que múltiplos factores podem influenciar a ocorrência das perdas, incluindo factores biológicos, como a idade materna e condições de saúde pré-existentes como factores socioeconómicos, como nível de escolaridade e rendimento familiar; e factores ambientais, como acesso a serviços de saúde e condições de vida.

Novaes *et al.* (2018) utilizaram também a regressão logística para investigar o risco gestacional e os factores associados em mulheres atendidas pela rede pública de saúde no Brasil. Foi verificado

que os factores biológicos tendem a ser os mais influentes para gestações de risco, gestações estas que terminam em natimortos ou abortos. No contexto moçambicano, Alberto (2023) identificou a gravidez na adolescência, a baixa escolaridade, a falta de assistência pré-natal e o baixo nível socioeconómico como factores fortemente relacionados com as elevadas taxas de mortalidade neonatal, utilizando a análise de sobrevivência e modelos de regressão de Cox para identificar os factores associados aos óbitos neonatais e pós-neonatais em Moçambique.

Este estudo tem como objectivo identificar os factores associados às perdas gestacionais em Moçambique, aplicando técnicas da Aprendizagem Supervisionada para determinar os principais factores associados a esse problema.

1.2 Definição do problema

Moçambique, como país em desenvolvimento, enfrenta desafios na redução da mortalidade fetal e neonatal, ambos factores importantes para as estatísticas demográficas e para o desenvolvimento do país. A gravidez na adolescência, por exemplo, é um dos determinantes associados ao aumento do risco de complicações gestacionais e perinatais, incluindo a mortalidade neonatal.

De acordo com o IDS (2023), entre 2011 e 2022, houve uma taxa de abortos espontâneos de aproximadamente 5%, enquanto as taxas de mortalidade neonatal foram acima de 2%. Esses números estão acima das expectativas globais e das metas estipuladas pelas nações unidas na agenda 2030 para o desenvolvimento sustentável, que busca reduzir as taxas das perdas gestacionais e melhorar a saúde materna e infantil. Estas estatísticas podem estar reflectidas em problemas estruturais no sistema de saúde moçambicano, particularmente no que diz respeito ao acesso aos cuidados pré-natais de qualidade e à assistência comprometida durante o parto.

As elevadas taxas de perdas gestacionais no país podem ser atribuídas a diversos factores, desde condições biológicas e complicações durante a gestação, até questões socioeconómicas e ambientais que afectam o acesso aos cuidados de saúde. No entanto, há uma carência de estudos que analisem de maneira abrangente esses determinantes no contexto específico de Moçambique. A compreensão dos factores que influenciam directamente essas perdas é essencial não apenas para o desenvolvimento de soluções voltadas para a redução desses índices, mas também para a formação de estratégias preventivas eficazes, que possam ser aplicadas de maneira prática no país.

Diante desse cenário, surge a questão deste estudo: **quais são os principais factores associados às perdas gestacionais em Moçambique?**

1.3 Objectivos

1.3.1 Objectivo Geral

Analisar os factores associados às perdas gestacionais em Moçambique utilizando técnicas de Aprendizagem Supervisionada.

1.3.2 Objectivos Especifícos

- Descrever o perfil socio-demográfico das mulheres que tiveram perdas gestacionais em Moçambique;
- Identificar os factores socioeconômicos, biológicos e comportamentais que influenciam a ocorrência de perdas gestacionais;
- Aplicar modelos estatísticos, nomeadamente a regressão logística e a árvore de decisão, para identificar padrões e prever a probabilidade de perda gestacional;
- Comparar os desempenhos dos modelos de regressão logística e de árvore de decisão na classificação dos factores associados à perda gestacional.

1.4 Justificação

As perdas gestacionais, como abortos espontâneos e natimortos, são uma realidade dolorosa enfrentada por muitas mulheres e famílias no mundo e especificamente em Moçambique. A questão das perdas gestacionais representa um desafio significativo para a saúde pública, afectando muitas mulheres e suas famílias. O impacto das perdas gestacionais vai além das complicações físicas, estendendo-se à saúde emocional e psicológica das mulheres e de suas famílias. Embora existam dados sobre a prevalência desses eventos, é importante entender com mais profundidade os factores associados a essas perdas para que sejam criadas soluções para a sua resolução.

Este estudo é relevante porque aborda uma área de saúde materna que necessita de atenção contínua. Este estudo permitirá identificar que condições podem estar ligadas a esse problema em Moçambique, considerando o contexto social e económico do país.

1.5 Estrutura do Trabalho

Este trabalho está estruturado em 5 capítulos. O primeiro capítulo apresenta a introdução ao tema, contextualizando a perda gestacional em Moçambique, o problema de estudo, os objectivos e a relevância do estudo. O segundo capítulo apresena a revisão de literatura, abordando os fundamentos do aprendizado estatístico, os tipos de aprendizado, bem como os principais algoritmos

aplicados à previsão de perdas gestacionais e estudos anteriormente feitos relacionados ao tema. O terceiro capítulo descreve os materiais e métodos utilizados, as variáveis analisadas, os procedimentos de tratamento dos dados, escolhas dos modelos de regressão logística e árvore de decisão, e as Medidas de avaliação. No quarto capítulo são apresentados os resultados obtidos, com análise descritiva das variáveis, desempenho dos modelos, comparação entre os métodos e a discussão dos resultados relacionando-os à literatura existente . O quinto capítulo apresenta as conclusões da pesquisa, recomendações para futuras investigações e apresentação das limitações do estudo.

Capítulo 2

REVISÃO DA LITERATURA

2.1 Contextualização das Perdas Gestacionais

As perdas gestacionais compreendem eventos adversos que podem ocorrer durante a gravidez ou no período perinatal, incluindo aborto espontâneo, natimorto e morte neonatal precoce (Goldenberg *et al.*, 2011). A Organização Mundial da Saúde (OMS, 2023) define perda gestacional como a interrupção da gravidez ou a morte do recém-nascido até ao 28.º dia de vida.

Alberto (2023) evidencia que a adesão insuficiente às consultas pré-natais em Moçambique, especialmente em regiões como Nampula, contribui significativamente para o aumento das perdas gestacionais, devido à falta de detecção precoce de riscos e às limitações no maneio das complicações.

2.2 Perdas Gestacionais em Moçambique e Factores Associados

Alberto (2023) aponta a gravidez na adolescência como um factor determinante significativo para os desfechos gestacionais. Mulheres adolescentes, por conta da imaturidade biológica e psicológica, enfrentam maiores riscos de complicações obstétricas como parto prematuro, pré-eclâmpsia e anemia.

O número elevado de gestações anteriores, segundo Conde-Agudelo *et al.* (2006), é também um dos factores determinantes, pois mulheres com histórico de cinco ou mais partos apresentam risco aumentado de complicações obstétricas, incluindo parto prematuro, descolamento prematuro da placenta e hemorragia pós-parto, o que pode resultar em perdas gestacionais.

A baixa escolaridade também é fortemente correlacionada com o risco de perda gestacional. Mulheres com níveis de instrução reduzidos tendem a apresentar menor conhecimento sobre práticas

de autocuidado, sinais de risco na gravidez e direitos relacionados ao acesso à saúde. Essa limitação afecta directamente a procura por serviços de saúde reprodutiva e o cumprimento adequado do calendário de consultas pré-natais. Segundo Alberto(2023), a escolaridade influencia significativamente o comportamento de saúde e a capacidade de interpretar e seguir orientações médicas.

O nível socioeconómico reduzido é também um factor determinante das perdas gestacionais. A pobreza está associada à insegurança alimentar, habitação precária, baixa mobilidade e dificuldades de acesso a serviços médicos. Essas condições comprometem a nutrição materna e aumentam a exposição a ambientes não favoráveis, factores estes que elevam o risco de infecções e de outras complicações durante a gestação (Alberto, 2023).

Histórico de perdas gestacionais indica que mulheres com um ou mais abortos espontâneos prévios possuem maior probabilidade de vivenciar novas perdas gestacionais, especialmente se não houver investigação clínica das causas anteriores (Sundermann *et al.*, 2017).

A ausência ou precariedade da assistência pré-natal é também apontada como um dos factores mais directos e relevantes na determinação de perdas gestacionais. Quando o acompanhamento da gravidez é inexistente ou incompleto, há falhas graves na identificação de condições clínicas de risco, como hipertensão gestacional, diabetes ou infecções sexualmente transmissíveis. Alberto (2023) alerta que a ausência de um pré-natal adequado impossibilita a intervenção atempada e agrava os riscos tanto para a mãe como para o feto.

O inadequado de métodos contraceptivos está relacionado com gravidezes não planeadas, muitas vezes em condições desfavoráveis do ponto de vista médico ou socioeconómico, o que aumenta o risco de complicações (Bearak *et al.*, 2018).

A falta de planeamento familiar, segundo Cleland *et al.* (2012), está fortemente associada a gestações em momentos inoportunos ou em intervalos muito curtos entre partos, comprometendo a recuperação do organismo materno e aumentando o risco de resultados adversos para a gravidez.

Durante a pandemia da Covid-19, cerca de 60% das gestantes não conseguiram completar as quatro consultas pré-natais recomendadas pelo Ministério da Saúde (MISAU, 2021). Esta redução foi atribuída à reorganização dos serviços, ao medo de contágio e às restrições de circulação.

2.3 Aprendizado Estatístico

O aprendizado estatístico é uma área que envolve estatística e ciência da computação, cujo objectivo principal é construir modelos capazes de inferir padrões e relações a partir de dados observa-

dos, para que se possa realizar previsões ou compreender a estrutura dos dados (Hastie *et al.*, 2009).

O aprendizado estatístico pode ser representado pela seguinte equação:

$$Y = f(X) + \varepsilon \tag{2.1}$$

onde Y é a variável resposta, $X = (X_1, X_2, \ldots, X_p)$ representa o vetor de preditores, f é a função desconhecida que relaciona os preditores à resposta, e ε é um termo de erro, geralmente assumido com média zero e variância constante.

Modelos com viés elevado podem subajustar os dados (*underfitting*), enquanto modelos com variância elevada podem ajustar-se demasiado aos dados de treino e apresentar desempenho fraco em novos dados (*overfitting*) (Hastie *et al.*, 2009).

O aprendizado estatístico compreende 3 tipos nomeadamente: Aprendizagem Não supervisionada, Aprendizagem Supervisionada e a Aprendizagem Semi-supervisionada.

2.3.1 Aprendizagem Não Supervisionada

A aprendizagem não supervisionada refere-se a um conjunto de técnicas utilizadas quando os dados disponíveis contêm apenas variáveis explicativas X, sem uma variável-resposta associada. Nesses casos, o objectivo é identificar padrões, estruturas ou agrupamentos latentes nos dados, sem o auxílio de rótulos durante o processo de treino (Hastie *et al.*, 2009).
Buscando-se dividir os dados em subgrupos de observações semelhantes que não são previamente conhecidos, com foco na descoberta de estrutura de grupo nos dados, esta abordagem é designada análise de agrupamentos (*clustering*).
A Análise de Componentes Principais (PCA) é empregue para **reduzir a dimensionalidade** dos dados ao projectá-los em direcções de maior variância, mantendo o máximo possível da informação original (Jolliffe e Cadima, 2016).

2.3.2 Aprendizagem supervisionada

A Aprendizagem Supervisionada é uma abordagem do aprendizado estatístico, como explicada pela figura 2.1, na qual um modelo é treinado a partir de um conjunto de exemplos rotulados, ou seja, pares ordenados (X_i, y_i), onde X_i representa um vector de variáveis preditoras (também chamadas de atributos) e y_i é a saída ou rótulo correspondente. O objectivo é estimar uma função $f : \mathcal{X} \to \mathcal{Y}$ que seja capaz de mapear correctamente uma nova entrada X para a sua saída y, mesmo quando esta não foi anteriormente observada (Hastie *et al.*, 2009).

A capacidade de generalizar refere-se à habilidade do modelo de manter bom desempenho fora do conjunto de treino, ou seja, em dados novos (Vapnik, 1995; Hastie *et al.*, 2009).

A Aprendizagem Supervisionada pode ser espressa como um problema de otimização, no qual se busca encontrar os parâmetros θ de um modelo f_θ que minimizem uma função de perda L:

$$\theta^* = \arg\min_{\theta} \sum_{i=1}^{n} L(f_\theta(X_i), y_i) + \lambda\, \Omega(\theta) \tag{2.2}$$

Onde:

- $L(f_\theta(X_i), y_i)$ - é a função de perda, que mede o erro entre a predição do modelo e o valor verdadeiro;
- $\Omega(\theta)$ - é um termo de regularização que penaliza modelos excessivamente complexos;
- $\lambda \geq 0$ é um hiperparâmetro que controla o grau de regularização, promovendo um equilíbrio entre a fidelidade ao conjunto de dados e a simplicidade do modelo.

Os termos de regularização têm um papel bastante importante na prevenção do sobreajuste (*overfitting*), fenómeno em que o modelo aprende excessivamente padrões específicos do conjunto de treino e falha quando aplicado a novos dados (Tibshirani, 1996).

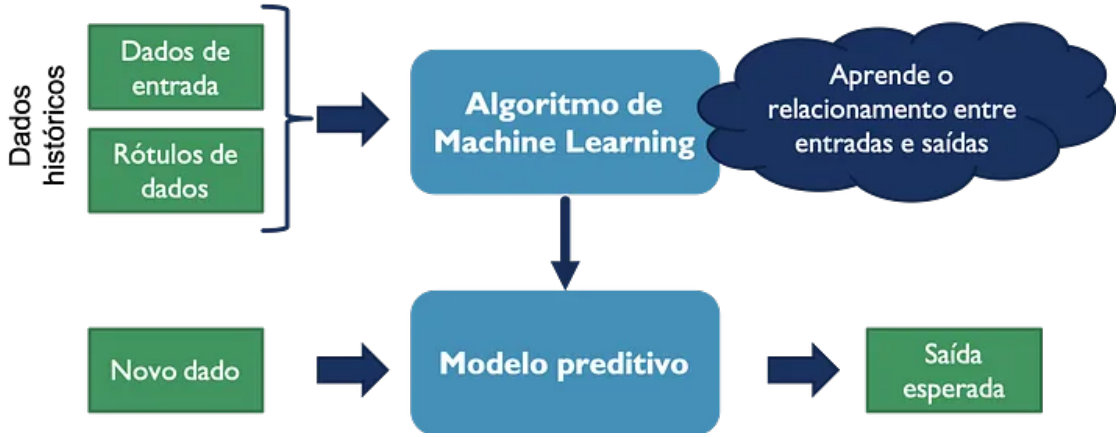

Figura 2.1: Algoritmo de aprendizagem supervisionada.

A Aprendizagem Supervisionada visa construir modelos que, a partir de dados observados, consigam inferir resultados para predição, mantendo-se interpretáveis.

2.3.3 Aprendizagem Semi-supervisionada

Aprendizagem Semi-supervisionada

A aprendizagem semi-supervisionada refere-se a um conjunto de técnicas utilizadas quando o conjunto de dados disponível contém tanto exemplos rotulados quanto não rotulados. Nesta abor-

dagem, assume-se que apenas uma pequena fração das observações possui rótulos, enquanto a maioria carece dessa informação. O objectivo é aproveitar a estrutura implícita nos dados não rotulados para melhorar a performance do modelo, explorando relações latentes entre os exemplos e generalizando melhor para novos dados (Chapelle *et al.* 2009).

Dessa forma, o modelo aprende a partir de um conjunto misto (X_i, y_i) para os exemplos rotulados e (X_j) para os exemplos não rotulados, combinando características da aprendizagem supervisionada e da não supervisionada. Esta abordagem é particularmente útil quando o custo de rotular grandes volumes de dados é elevado, permitindo extrair informações adicionais da distribuição dos dados para estimar a função $f : X \to Y$ com maior precisão.

2.3.4 Problemas em Aprendizagem Supervisionada: Regressão e Classificação

Os problemas na aprendizagem supervisionada podem ser resolvidos de dois modos, de acordo com a natureza da variável de resposta y, sendo essa diferenciação importante tanto para a formulação dos modelos como para a escolha das Medidas de avaliação e das técnicas de inferência (James *et al.*, 2021 e Hastie *et al.*, 2009):

- **Regressão** - Na regressão, o objectivo é de determinar a relação entre a(s) variável(is) explicativa(s) quando y é contínua e
- **Classificação** - Na regressão, o objectivo é de determinar a relação entre a(s) variável(is) explicativa(s) quando y é categórica.

(A) Algoritmos de Regressão

Os principais métodos para resolver problemas de regressão incluem:

Regressão Linear

O método mais utilizado, que modela uma relação linear entre as variáveis preditoras e a variável resposta. Expressa como:

$$y = \beta_0 + \sum_{j=1}^{p} \beta_j x_i + \varepsilon,$$

onde:

- β_0 é o intercepto do modelo;
- $\beta_1, \beta_2, \ldots, \beta_p$ são os coeficientes que representam a contribuição de cada preditor;
- ε é o termo de erro, geralmente ruído branco com média zero e variância constante.

Segundo James *et al.* (2021), a regressão linear é um método estatístico fundamental para modelar a relação entre uma variável dependente contínua y e um conjunto de variáveis independentes $X = (X_1, X_2, \ldots, X_p)$. Assume-se que essa relação é aproximadamente linear, descrita pela equação acima.

Os coeficientes são estimados pelo método dos mínimos quadrados ordinários, que minimiza a soma dos quadrados dos resíduos:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2.$$

Segundo Bishop (2006), essa técnica é valorizada pela sua simplicidade, interpretabilidade e eficiência computacional, sendo a base para métodos mais complexos.

Outros métodos

- **Regressão Polinimial**
- **Regressão por Árvore de Decisão**
- **Métodos de Regularização (Ridge, Lasso)**

(B) Algoritmos de Classificação

Segundo James *et al.* (2021), a classificação supervisionada dispõe de uma variedade de algoritmos que procuram aprender uma função de decisão capaz de atribuir correctamente rótulos a novas observações.

- **Árvores de Decisão**
  De acordo com Géron (2022), árvores de decisão particionam recursivamente o espaço de atributos com base em regras hierárquicas.
- **Regressão Logística**
  Segundo Géron (2022), a regressão logística é um modelo probabilístico para problemas de classificação.
- **Máquinas de Vectores de Suporte (SVM)**
  Segundo Hastie *et al.*(2009), elas funcionam tentando encontrar a melhor forma de separar os grupos com base nas suas características. O objectivo é traçar uma divisão que deixe uma margem grande entre os grupos, o que ajuda o modelo a ser mais confiável com novos

dados usando a técnica chamada função *kernel*. Essa técnica transforma os dados, permitindo encontrar uma separação mesmo quando os grupos estão organizados de forma mais complicada..

- **k-Vizinhos Mais Próximos (kNN)**
  James *et al.*(2021), define o kNN como um classificador baseado em instâncias, que atribui a classe de uma nova amostra com base na maioria das classes entre os k vizinhos mais próximos no conjunto de treinamento. É simples e eficaz para dados bem distribuídos, mas pode ser sensível ao ruído e à escala dos atributos.

- **Floresta Aleatória**
  Breiman (2001) define o método Floresta Aleatória ,como um modelo de *conjunto*,um processo em que múltiplos modelos básicos diversos são usados para prever um resultado, baseado em múltiplas árvores de decisão construídas a partir de subconjuntos aleatórios dos dados e dos atributos. Segundo James *et al.*(2021), esse método melhora a acurácia e reduz o sobreajuste.

- **Gradient Boosting Machines (GBM, XGBoost, LightGBM)**
  Segundo Friedman (2001), os métodos de *boosting* constroem modelos de forma sequencial, corrigindo os erros das etapas anteriores. Variantes modernas como XGBoost e LightGBM são altamente eficientes e costumam alcançar um desempenho excelente em competições de aprendizagem automática (Chen & Guestrin, 2016).

- **Redes Neurais Artificiais (RNA)**
  Conforme Géron (2022), redes neurais são modelos compostos por camadas de unidades interconectadas, capazes de aprender representações complexas dos dados.

2.4 Regressão Logística na Previsão de Eventos Binários

2.4.1 Fundamentos e Interpretação dos Coeficientes

A **regressão logística** é uma técnica estatística utilizada para modelar fenómenos cuja variável dependente assume apenas duas categorias exclusivas, como ”sim”ou ”não”, ”presença”ou ”ausência”. Este tipo de variável é binária. A principal finalidade da regressão logística é estimar a probabilidade de ocorrência de um determinado evento, atribuindo valores entre 0 e 1 à predição. Ao contrário da regressão linear, que estabelece uma relação linear entre as variáveis e assume que os resultados podem variar continuamente, a regressão logística trabalha com probabilidades e permite prever se um evento irá ocorrer, com base em variáveis explicativas categóricas ou numéricas (Hosmer, Lemeshow e Sturdivant, 2013).

Segundo Gujarati e Porter (2009), a forma funcional da regressão logística baseia-se na equação da função logística (**sigmoide**), que transforma uma combinação linear de variáveis num valor de probabilidade, expressa da seguinte forma:

$$P(Y=1) = \frac{1}{1+e^{\eta}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\ldots+\beta_k x_k)}} \tag{2.3}$$

Onde:

- P(Y=1) representa a probabilidade do evento ocorrer;
- β_0 é o intercepto do modelo;
- $\beta_1, \beta_2, \ldots, \beta_k$ são os **coeficientes** associados às variáveis independentes;
- $x_1, \ldots, x_k$ são as variáveis preditoras.

A Função Logit e a Razão de Chances (Odds)

Segundo Hosmer *et al.* (2013), a regressão logística utiliza uma função de ligação denominada logit, que transforma a probabilidade de ocorrência de um evento numa escala contínua ilimitada, permitindo assim modelar relações lineares entre os preditores e a variável dependente binária e para possibilitar uma modelação adequada, a ligação entre a variável resposta (probabilidade) e os preditores é feita por meio da função logit, definida como:

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} \tag{2.4}$$

O π representa a probabilidade de ocorrência do evento, o termo $\frac{\pi_i}{1-\pi_i}$ é odds ratio ou razão de chances, e representa a razão entre a probabilidade de ocorrência e a de não ocorrência do evento (Menard, 2010). Esta transformação permite que a regressão logística possa ser utilizada de forma apropriada para modelar variáveis categóricas binárias.

De acordo com Menard (2010), a função logit torna mais simples a análise de variáveis categóricas, pois permite compreender com facilidade o impacto de cada variável no modelo. Cada coeficiente indica como a razão de chances se altera quando há uma mudança numa variável, mantendo-se as outras constantes.

Segundo Hosmer *et al.* (2013), o valor de e^{β_j} (onde β_j representa um coeficiente específico do modelo, sendo j = 0,1,...) indica o fator pelo qual os odds ratio do evento se alteram a cada incremento unitário na variável preditora X_j, mantendo constantes todas as demais variáveis no modelo.

- Se $e^{\beta_j} > 1$:um aumento em X_k está associado a um aumento nos odds do evento de interesse;
- Se $e^{\beta_j} < 1$: um aumento em X_j está associado a uma diminuição nos odds;
- Se $e^{\beta_j} = 0$: X_k não tem efeito sobre os odds do evento.

Estimação dos Parâmetros

Método da Máxima Verosimilhança A verosimilhança é uma função que mede a probabilidade de observar o conjunto de dados tal como ele ocorreu, dada uma combinação específica de parâmetros. O método da máxima verosimilhança procura identificar os valores de β que maximizam essa função, ou seja, que tornam os dados observados os mais prováveis possíveis (Hosmer *et al.* 2013).

A função de verosimilhança é dada por:

$$L(\beta) = \prod_{i=1}^{n} P\left(Y_i = 1|X_i\right) \tag{2.5}$$

Mas por serem variáveis binárias ,a probabilidade condicional é ser representada como:

$$L(\beta) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i} \tag{2.6}$$

Onde π_i Representa a probabilidade predita de que Y_i=1, e y_i é o valor observado da variável dependente.

Pelo produto de várias probabilidades muito pequenas causar instabilidade numérica, é utilizado logaritmo da verosimilhança, o que transforma o produto em soma:

$$l(\beta) = \sum_{i=1}^{n} \left[y_i \log\left(\pi_i\right) + \left(1-y_i\right) \log\left(1-\pi_i\right)\right] \tag{2.7}$$

Esse procedimento não possui uma solução algébrica directa, sendo necessário derrivar em relação a π e de seguida igualado a 0. (Menard, 2010).

2.4.2 Pressupostos e Critérios de Avaliação da Regressão Logística

A aplicação correcta da regressão logística requer a verificação de pressupostos fundamentais que asseguram a validade dos resultados e a fiabilidade das inferências estatísticas.

Independência das Observações

Segundo Menard (2010), a independência das observações é um pressuposto essencial para garantir a validade estatística do modelo. Este pressuposto implica que cada unidade de análise (por exemplo, cada indivíduo da amostra) deve ser considerada como independente das restantes, sem que haja correlação entre os erros das observações. A violação dessa suposição pode ser identificada pela presença de autocorrelação nos resíduos do modelo.

Ausência de Multicolinearidade

Gujarati e Porter (2009) salientam que a multicolinearidade entre as variáveis explicativas pode comprometer a precisão das estimativas dos coeficientes. Quando duas ou mais variáveis estão altamente correlacionadas, torna-se difícil distinguir o efeito individual de cada uma sobre a variável dependente. O diagnóstico da multicolinearidade é geralmente feito através do Fator de Inflacção da Variância (VIF), cujo valor ideal deve ser inferior a 5. A fórmula do VIF é:

$$\text{VIF}_j = \frac{1}{1 - R_j^2}, \tag{2.8}$$

onde R_j^2 é o coeficiente de determinação da regressão da variável x_i sobre as outras variáveis explicativas. Um VIF elevado (geralmente maior que 5 ou 10) indica que a variável x_i está altamente correlacionada com outras variáveis no modelo, sugerindo a necessidade de eliminar ou combinar variáveis para resolver a multicolinearidade (Gujarati e Porter, 2009).

Qualidade do Ajuste do Modelo

A adequação global do modelo pode ser verificada pelo **teste de Hosmer-Lemeshow**, que compara os valores observados e previstos em grupos de risco. Um valor de $p > 0,05$ indica que o modelo se ajusta bem aos dados (Hosmer *et al.*, 2013). A estatística de teste de Hosmer-Lemeshow envolve a comparação dos resíduos de Pearson entre os grupos de risco:

$$\chi^2_{HL} = \sum_{g=1}^{G} \frac{(O_g - E_g)^2}{E_g(1 - \widehat{\pi}_g)}, \tag{2.9}$$

onde O_g são as observações observadas no grupo g, E_g são as expectativas para as observações no grupo g, e G é o número de grupos de risco.

Capacidade Discriminativa: Curva ROC e AUC

A capacidade discriminativa do modelo é avaliada por meio da curva ROC (Receiver Operating Characteristic), a qual relaciona a sensibilidade e a especificidade do modelo para diferentes pontos de corte. A área sob a curva (AUC) quantifica essa capacidade:

- **AUC = 0.5**: discriminação aleatória;
- **AUC** $>$ **0.7**: discriminação aceitável;
- **AUC** $>$ **0.9**: excelente discriminação (Kleinbaum & Klein, 2010).

Pseudo-R² de McFadden

O pseudo-R² de McFadden é uma medida semelhante ao R² da regressão linear, mas adaptada para modelos de máxima verosimilhança. É calculado como:

$$R^2_{\text{McFadden}} = 1 - \frac{\ln(L_{\text{modelo}})}{\ln(L_{\text{nulo}})}, \tag{2.10}$$

onde L_{modelo} é a verosimilhança do modelo com todas as variáveis explicativas e L_{nulo} é a verosimilhança do modelo sem preditores (apenas com o intercepto). Valores entre 0.2 e 0.4 são considerados satisfatórios (Menard, 2010).

2.5 Árvores de Decisão como Ferramenta de Previsão

As árvores de decisão são uma ferramenta de aprendizagem supervisionada utilizada tanto para classificação (quando a variável resposta é categórica) como para regressão (quando a variável resposta é contínua). As árvores de decisão são amplamente utilizadas em Aprendizagem Supervisionada devido à sua simplicidade, interpretabilidade e capacidade de lidar com diferentes tipos de variáveis (categóricas e contínuas). A sua estrutura hierárquica facilita a visualização do processo de decisão e é especialmente útil em cenários onde a interpretabilidade é crucial, como em áreas da saúde, finanças e negócios (Breiman *et al.*, 1986).

2.5.1 Estrutura e Funcionamento das Árvores de Decisão

Uma árvore de decisão é composta por **nós**, onde cada nó interno (não terminal) representa um teste ou decisão sobre uma das variáveis de entrada. As arestas que saem desses nós representam os valores possíveis da variável testada no nó. A **folha** (ou nó terminal) representa o valor da variável resposta obtido com base nos valores das variáveis de entrada ao longo do caminho do nó raiz até à folha. A raiz é o nó inicial da árvore.

Ela pode ser vista como uma representação gráfica de um processo de divisão de dados em subgrupos baseados nas variáveis explicativas, com o objectivo de prever a variável dependente (Quinlan, 1993).

1. **Raiz (Root):** O nó inicial da árvore, que representa o conjunto completo de dados. A partir desse ponto, a árvore começa a se ramificar com base nas variáveis que melhor dividem os dados.

2. **Nós internos (Internal Nodes):** Representam as condições ou decisões a serem feitas com base nos atributos das variáveis. Cada nó interno divide os dados de forma a maximizar a homogeneidade dos subgrupos resultantes.

3. **Folhas (Leaves):** Representam a previsão final da árvore, ou seja, a classe ou valor predito com base nas decisões feitas nos nós anteriores.

O algoritmo de Árvores Classificação e Regressão (CART), proposto por Breiman *et al.* (1984), é um dos algoritmos mais conhecidos para a construção de árvores de decisão. Ele utiliza uma abordagem de *divisão e conquista*, onde o processo de construção da árvore é feito de forma recursiva, partindo da raiz e dividindo os dados repetidamente em dois ramos (divisões binárias). Este processo continua até que um critério de paragem seja alcançado, como a homogeneidade dos dados nas folhas ou um número mínimo de observações por nó.
Para classificar uma nova observação, percorre-se a árvore a partir do nó raiz, seguindo os ramos correspondentes aos valores dos atributos da observação, até se atingir um nó folha que indica a classe prevista. No caso de árvores de regressão, o valor na folha representa a previsão para essa observação.

2.5.2 Critérios de Divisão (Índice de Gini e Entropia)

Em cada nó interno da árvore, o algoritmo CART (e outros algoritmos para árvores de decisão) deve decidir qual variável usar para dividir os dados e qual o ponto de divisão. O objectivo é escolher a divisão que resulte na maior redução da impureza dos nós filhos, ou seja, que crie nós mais homogéneos em relação à variável resposta.

Dois dos critérios de divisão mais comuns para árvores de classificação são o índice de Gini e a entropia:

- **Índice de Gini**: Utilizado no algoritmo CART. Mede a impureza de um nó, ou seja, a probabilidade de um elemento ser classificado incorretamente se for escolhido aleatoriamente de acordo com a distribuição de classe do nó (Breiman *et al.*, 1986). A fórmula do índice de Gini é dada por:

$$Gini(t) = 1 - \sum_{i=1}^{m} p_i^2 \tag{2.11}$$

  onde p_i representa a proporção de observações da classe i dentro do nó t, e m é o número total de classes.

- **Entropia**: Outra medida de impureza, baseada na teoria da informação. A entropia de um nó é calculada como:

$$Entropia(t) = -\sum_{i=1}^{m} p_i \log_2 p_i \tag{2.12}$$

Uma entropia alta indica um nó impuro (mistura de classes), enquanto uma entropia baixa (ou zero) indica um nó puro (contém apenas uma classe). O critério de divisão baseado na entropia procura maximizar o ganho de informação (a redução na entropia após a divisão).

2.5.3 Poda para Evitar sobreajuste

Uma árvore muito complexa, com muitos nós, pode ajustar-se demasiado bem aos dados de treino, incluindo o ruído presente nesses dados, e consequentemente ter um desempenho pobre em dados novos ou não vistos (James *et al.*, 2021; Rokach & Maimon, 2008).

Para evitar o sobreajuste, é usada a técnica de poda (pruning) da árvore. A poda envolve a remoção de ramos (nós) da árvore que não contribuem significativamente para a precisão preditiva em novos dados. Existem duas abordagens principais para a poda:

- **Pré-poda**: Envolve a definição de critérios de paragem durante o processo de crescimento da árvore para evitar que ela se torne demasiado complexa. Por exemplo, pode-se definir um número máximo de níveis na árvore ou um número mínimo de amostras em cada nó (Han *et al.*, 2011).
- **Pós-poda**: Primeiro, a árvore cresce até ao seu tamanho máximo, e depois os ramos que não melhoram a precisão em dados de validação (ou usando técnicas como a validação cruzada) são removidos. O algoritmo CART utiliza um método de poda de custo-complexidade, onde uma sequência de sub-árvores é gerada, e a melhor sub-árvore é escolhida com base num critério que equilibra a complexidade da árvore (número de folhas) e o seu erro de classificação (ou erro quadrático, no caso de regressão) estimado por validação cruzada. (James *et al.*, 2021).

A poda resulta numa árvore mais simples, mais fácil de interpretar e com melhor capacidade de generalização para novos dados.

2.6 Desbalanceamento de Classes

De acordo com Japkowicz e Stephen (2002), o desbalanceamento de classes ocorre quando uma ou mais classes estão significativamente sub-representadas no conjunto de dados. Essa desproporção pode comprometer o desempenho de algoritmos supervisionados, que, ao priorizarem a minimização do erro global, tendem a favorecer a classe majoritária. Como consequência, observa-se uma redução na sensibilidade e na capacidade preditiva para a classe minoritária.
Chawla *et al.*(2002) destacam que esse cenário é muito comum em aplicações como na medicina ou na segurança, onde eventos positivos (como a ocorrência de uma doença ou uma fraude) são raros,onde estes representam apenas uma fração pequena do total de observações.

Para diminuir os efeitos do desbalanceamento, estratégias de balanceamento de dados podem ser aplicadas, dentre elas:

- **Reamostragem do conjunto de dados**
  Essa abordagem visa alterar a distribuição original das classes, tornando o conjunto mais balanceado:
  - **Subamostragem:** consiste na redução do número de exemplos da classe majoritária.
  - **Sobreamostragem:** consiste em aumentar artificialmente o número de exemplos da classe minoritária.
    Proposto por Lunardon, Menardi e Torelli (2014), o Exemplos de sobreamostragem aleatória (ROSE) cria novas observações sintéticas para a classe minoritária utilizando uma abordagem baseada em estimadores de densidade com suavização (kernel). O ROSE gera amostras de forma probabilística, o que pode resultar numa representação mais variada e realista da distribuição da classe minoritária.
- **Ponderação de classes**
  Ponderação de classes é uma técnica onde diferentes pesos são atribuídos às classes durante o processo de aprendizagem, de modo que os erros cometidos em classes minoritárias tenham maior impacto na função de custo. Segundo Kuhn e Johnson (2013), a ponderação de classes ajusta a influência relativa de cada classe na função de custo, permitindo que algoritmos supervisionados aprendam com mais equilíbrio mesmo quando há predominância de uma classe sobre a outra.

2.7 Validação Cruzada

A validação cruzada é uma técnica estatística amplamente utilizada para avaliar o desempenho preditivo de modelos de aprendizagem, especialmente em contextos com dados limitados. Segundo Kohavi (1995), o seu principal objectivo é estimar como um modelo se comportará ao ser aplicado a dados não vistos, reduzindo o risco de sobreajuste e garantindo maior generalização dos resultados. A ideia central consiste em particionar o conjunto de dados em múltiplas divisões, utilizando algumas para o treino e outras para a validação, repetindo esse processo diversas vezes. De acordo com Hastie *et al.* (2009), esse procedimento permite obter Medidas mais fiáveis e estáveis em comparação com abordagens simples como a divisão *holdout*. James *et al.* (2021) reforçam que a validação cruzada é fundamental para a escolha de hiperparâmetros e para avaliar a performance real de modelos de maneira robusta.

- **Holdout:** Divide-se o conjunto de dados em dois subconjuntos fixos: um para treino e outro para teste (por exemplo, 70%–30%). Embora simples, esta abordagem pode gerar estimativas instáveis, dependendo da divisão escolhida (Han *et al.* 2011).

- **Validação Cruzada K-Fold:** O conjunto de dados é dividido em k subconjuntos de tamanho aproximadamente igual. O modelo é treinado k vezes, cada vez usando $k-1$ partes para treino e a parte restante para validação. A performance final é a média das iterações (James *et al.*, 2021). Uma escolha comum é $k = 10$, por fornecer um bom equilíbrio entre viés e variância.

- **Leave-One-Out Cross-Validation (LOOCV):** É um caso extremo do K-Fold, onde $k = n$, ou seja, cada instância do conjunto é usada uma única vez como validação. Apesar de fornecer estimativas não enviesadas, é computacionalmente caro e apresenta alta variância (Hastie, Tibshirani e Friedman, 2009).

- **Validação Cruzada de repetição K-Fold:** Consiste em repetir a validação K-Fold várias vezes com diferentes divisões aleatórias. Essa abordagem reduz a variância da estimativa e fornece resultados mais estáveis (Kuhn e Johnson, 2013).

- **Validação Cruzada Estratificada K-Fold:** Variante do K-Fold que preserva a proporção das classes em cada subdivisão. É particularmente importante em problemas de classificação com classes desbalanceadas (Fernández *et al.*, 2018).

2.8 Métodos Estatísticos e Aprendizagem Supervisionada na Saúde

2.8.1 Diferença entre Estatística Tradicional e Aprendizagem Supervisionada

De acordo com James *et al.*(2021), a aplicação de métodos estatísticos e computacionais na saúde materna tem-se mostrado importantes com os avanços tecnológicos e no controle do surgimento de varias epidemiologias. Entre as abordagens disponíveis, destacam-se a estatística tradicional, com uma base mais inferencial e a Aprendizagem Supervisionada, centrado na construção de modelos preditivos. Embora ambas usem dados para tomada de decisão, diferem-se pelos seus objectivos principais, estrutura metodológica e ênfase analítica.

Estatistica tradicional

Segundo Hosmer *et al.* (2013), a estatística tradicional, também descrita como inferencial, tem como objectivo fundamental explicar relações causais entre variáveis. Baseia-se na formulação de hipóteses e na construção de modelos paramétricos, os quais pressupõem algumas distribuições probabilísticas. A regressão linear, regressão logística, análise de variância (ANOVA) e modelos de sobrevivência são técnicas muito utilizadas, especialmente para identificação de factores de risco e para a avaliação de intervenções em saúde pública.

Aprendizagem supervisionada

Como explicam Kuhn e Johnson (2013), enquanto a estatística tradicional busca compreender os factores por detrás aos fenômenos observados, a Aprendizagem Supervisionada concentra-se em prever com elevada precisão os desfechos, mesmo que isso implique baixa interpretabilidade. Modelos supervisionados podem ser paramétricos ou não paramétricos, incluindo algoritmos como árvores de decisão, florestas aleatórias, máquinas de vetores de suporte (SVM) e redes neurais artificiais.

Muitos dos modelos da Aprendizagem Supervisionada,embora altamente precisos, carecem de interpretabilidade, o que pode comprometer sua aceitação em áreas sensíveis como a saúde pública, onde a compreensão dos factores que levam a determinado valor da previsão é tão importante quanto a própria previsão. (Lipton, 2018),

Segundo Rajkomar, Dean e Kohane (2019), apesar das suas diferenças, as duas abordagens podem ser vistas como complementares. Em contextos de pesquisa voltada à identificação de determinantes sociais ou clinicos, a estatística tradicional oferece robustez na interpretação causal. Enquanto que a Aprendizagem Supervisionada se mostra melhor em aplicações operacionais, como triagem de pacientes de risco ou previsão de perdas gestacionais.

Para Hastie *et al.*(2009), a aprendizagem estatística é uma união das duas abordanens, permitindo a construção de modelos que oferecem bom desempenho preditivo sem esquecer totalmente da interpretabilidade.

2.8.2 Exemplos de Aplicação da Aprendizagem Supervisionada na Predição de Riscos Gestacionais

Segundo Novaes, Melo e Oliveira *et al.* (2018), a aplicação de regressão logística permitiu identificar factores associados ao alto risco gestacional em mulheres atendidas pelo Sistema Único de Saúde no Brasil. O modelo utilizado destacou variáveis como a reação paterna negativa à gestação, o índice de massa corporal (IMC) pré-gestacional elevado e o número de gestações anteriores (duas ou mais) como estatisticamente significativos.

De acordo com Oliveira *et al.*(2019), outro estudo que abordou os factores associados ao nascimento pré-termo usou a regressão logística hierarquizada como principal técnica de modelação. Na pesquisa foi identificado que a ausência ou mau acompanhamento pré-natal teve papel importante na relação entre factores socioeconômicos e o risco de prematuridade. A etapa subsequente do estudo usou a modelação por equações estruturais (MEE), o que permitiu uma análise mais abrangente das relações entre os determinantes clínicos e sociais dos desfechos gestacionais.

Em um estudo de Liu, Fridline e Miller (2021), modelos de Aprendizagem Supervisionada foram aplicados a dados clínicos de unidades de terapia intensiva neonatal (UTIN) com o objetivo de prever a mortalidade de recém-nascidos. O estudo comparou diversos algoritmos, incluindo regressão logística, Floresta aleatória e redes neurais, destacando o desempenho superior de métodos baseados em árvores, especialmente o Floresta aleatória, no contexto da classificação de risco.

2.9 Estudos Relacionados

O estudo de Regassa *et al.*(2022), com o tema "Trends and determinants of pregnancy loss in eastern Ethiopia from 2008 to 2019: analysis of health and demographic surveillance data", teve como objectivo avaliar a magnitude e os determinantes da perda gestacional na Etiópia Oriental usando a Regressão de Poisson multivariada robusta para estimar razões de prevalência, identificando factores associados às perdas gestacionais.
O estudo teve como principais conclusões que, entre 2008 e 2019, foram registadas 810 perdas gestacionais em 39153 gestações, resultando numa taxa de 20.7 por 1000 nascimentos. As mortes fetais (natimortos) foram mais prevalentes do que os abortos espontâneos, com taxas de 11.14 e 9.55 por 1000 nascimentos, respectivamente. Foram identificados vários factores significativamente associados ao aumento do risco de perda gestacional: ausência de renda própria, trabalho agrícola, histórico de perda gestacional anterior, gravidez não planeada, ausência de cuidados pré-natais, não recebimento da vacina antitetânica durante a gravidez, gravidez Tardia. Além de identificar factores associados às perdas gestacionais, o estudo concluiu que a taxa geral dessas perdas se manteve estável ao longo dos 11 anos analisados.

O estudo de Novaes *et al.*(2018),com o tema"Risco gestacional e factores associados em mulheres atendidas pela rede pública de saúde", teve como objectivo classificar e estimar os factores associados ao risco gestacional em mulheres atendidas para o parto pelo Sistema Único de Saúde (SUS) no Brasil usando a regressão logística múltipla para identificar factores independentemente associados ao risco gestacional elevado.
O estudo teve como principais conclusões que, entre as 607 mulheres avaliadas, 50.9% tiveram a gravidez classificada como de risco habitual, 5.8% como risco intermédio e 43.3% como alto risco. Factores como: tabagismo, raça, distúrbios hipertensivos, índice de Massa Corporal (IMC) elevado, infecções urinárias foram identificados como determinantes para a maioria desses desfechos. Destacou a importância de considerar todos factores associados na prevenção de complicações durante a gravidez.

O estudo de Alberto (2023), com o tema "Condições de nascimento e factores gestacionais associados, antes e durante a pandemia da COVID-19, no distrito de Nampula–Moçambique", teve

como objectivo avaliar as condições de nascimento e os factores gestacionais associados antes e durante a pandemia da COVID-19 no distrito de Nampula usando a regressão logística binária para identificar factores preditores do baixo peso ao nascer.
Este estudo concluiu que ouve uma associação significativa entre o baixo peso ao nascer e a pandemia da COVID-19, com uma maior prevalência de baixo peso durante a pandemia. As variáveis preditoras para o baixo peso ao nascer foram o não cumprimento do número mínimo de consultas pré-natais recomendado pelo Ministério da Saúde de Moçambique e a idade gestacional no parto inferior a 37 semanas. Evidenciou que a pandemia da COVID-19 impactou negativamente o funcionamento do sistema de saúde, influenciando no aumento da prevalência de baixo peso ao nascer durante esse período.

O estudo de Silva (2021), com o tema ”Avaliação de modelos de aprendizado de máquina para classificação de gestantes e predição de gravidez de risco usando o histórico de consultas médicas”teve como objectivo avaliar diferentes modelos de aprendizado de máquina (Regressão logística, Árvore de decisão, Floresta aleatória,k- Vizinhos mais próximos, Máquinas de vetores de suporte) para classificar gestantes e prever o risco gestacional com base no histórico de consultas médicas, comparando a performance dos modelos preditivos para identificar o método mais eficiente na classificação de gravidez de risco.
O estudo identificou que algumas variáveis presentes nas primeiras semanas de gestação foram altamente preditivas para a classificação do risco gestacional. Dentre elas: Idade materna, Pressão arterial, Diabetes e hipertensão, Número de gestações anteriores, Número de abortos prévios, Resultados de exames laboratoriais iniciais, Tempo de início do pré-natal. O modelo Floresta aleatória foi o que apresentou os melhores resultados. Modelos supervisionados, especialmente árvores de decisão e floresta aleatória, apresentaram melhor desempenho na classificação das gestantes em risco e não risco. O estudo demonstrou que a precisão dos modelos depende fortemente da qualidade e completude dos dados.

O estudo de Liu *et al.*(2024), com o tema ”Aprendizado de Máquina model-based preterm birth prediction and clinical nomogram”, teve como objectivo desenvolver um modelo preditivo usando quatro algoritmos de aprendizagem supervisionada: regressão logística, regressão lasso adaptativa, florestas bootstrap e árvores impulsionadas para o risco de parto prematuro, visando fornecer aos profissionais clínicos uma ferramenta para prevenção precoce.
As conclusões foram que todos os quatro modelos de aprendizado de máquina demonstraram alta acurácia na predição de parto prematuro. Este destacou nove variáveis como importantes para a predição do modelo: idade materna, índice de massa corporal (IMC), número de gestações anteriores, complicações obstétricas (como pré-eclâmpsia), intervalo entre partos, factores comportamentais (como tabagismo), condições médicas pré-existentes, tipo de assistência pré-natal, peso fetal estimado.

O estudo de Qi *et al.* (2024), com o tema ”Building a Aprendizado de Máquina-based risk prediction model for second-trimester miscarriage”, teve como principal objectivo desenvolver modelos preditivos baseados em Aprendizagem Supervisionada para estimar o risco de aborto espontâneo no segundo trimestre da gestação, comparando sete modelos baseados nos seguintes algoritmos de Aprendizagem Automática: Regressão Logística (LR),k- Vizinhos mais próximos (KNN), Máquinas de vetores de suporte (SVM), Árvore de Decisão (DT), Floresta aleatória (RF), Aumento de gradiente extremo (XGBoost) e Rede Neural Artificial (ANN).
Neste estudo, o modelo XGBoost foi identificado como o algoritmo com melhor desempenho na previsão de aborto espontâneo no segundo trimestre, superando os demais modelos, com uma acurácia de 85.8%, precisão de 51.9%, recall de 77.8%, F1-score de 62.2% e AUC (Área sob a Curva ROC) de 88.3%. Dez características clínicas foram identificadas como preditoras relevantes para os desfechos de perda gestacional, sendo elas: dor abdominal, hemorragia vaginal, corrimento vaginal, comprimento cervical, hematoma subcoriónico, miomas uterinos, contagem de glóbulos brancos, percentagem de neutrófilos, nível de proteína C reactiva (CRP) e anomalias placentárias. Sintomas clínicos como hemorragia vaginal e corrimento vaginal patológico também demonstraram forte associação com esses desfechos.

O estudo de Yehuala, Mengesha e Baykemagn (2025),”Predicting pregnancy loss and its determinants among reproductive-aged women using supervised Aprendizado de Máquina algorithms in Sub-Saharan Africa”, teve como principal objectivo identificar os preditores da perda gestacional e construir modelos preditivos para a ocorrência de perdas gestacionais na região da África Subsaariana. para a obtencão dos resultados foram aplicaram algoritmos de aprendizagem supervisionada, como Floresta aleatória, Árvore de Decisão, Regressão Logística, XGBoost e Gaussian Naive Bayes.
Os autores mostram que modelo Floresta aleatória apresentou o melhor desempenho entre os algoritmos avaliados, com uma precisão de 98%, recall de 77%, F-measure de 83% e AUC de 94%. Os principais preditores identificados para a perda gestacional foram: estado civil, paridade, uso de contraceptivos, nível de riqueza, local de residência, idade materna, peso da gestante e grau de independência. O estudo concluiu que mulheres solteiras apresentaram maior risco de perda gestacional, enquanto mulheres com mais de um filho demonstraram menor probabilidade de sofrer perdas, em comparação com gestantes pela primeira vez. Foi verificado também que a aderência aos cuidados pré-natais e serviços de planeamento familiar esteve associada à redução do risco de perda gestacional. Também foi evidenciado que níveis socioeconômicos mais baixos e menor escolaridade dificultaram o acesso e a utilização dos serviços de saúde.

O estudo de Islam *et al.*(2022), intitulado ”Aprendizado de Máquina to predict pregnancy outcomes: a systematic review, synthesizing framework and future research agenda”, teve como ob-

jectivo principal explorar as abordagens existentes na utilização de algoritmos de Aprendizado de Máquina (ML) para prever desfechos gestacionais, identificar complicações durante a gravidez e determinar o modo de parto mais adequado. A revisão sistemática sintetizou os resultados de 26 estudos publicados entre 2000 e 2020, analisando criticamente seus métodos, dados utilizados, algoritmos aplicados e contextos geográficos.
A revisão evidenciou que os algoritmos de Aprendizagem Supervisionada foram os mais amplamente utilizados, com destaque para Árvore de decisão (DT), Máquinas de vetores de suporte, Floresta aleatória (RF), Naive Bayes (NB) e J48. Esses algoritmos demonstraram desempenhos promissores em diferentes aplicações, como:

- Previsão do tipo de parto (vaginal ou cesariana);
- Previsão de recém-nascidos com baixo peso;
- Identificação de complicações gestacionais.

Além disso, observou-se que as variáveis preditoras mais recorrentes incluíam factores demográficos, obstétricos e clínicos.

Capítulo 3

MATERIAL E MÉTODOS

3.1 Classificação do estudo

Este estudo é de natureza aplicada, pois busca solucionar um problema prático: a identificação de factores associados à perda gestacional (Gil, 2019).
A abordagem é quantitativa, pois está fundamentada em dados numéricos e técnicas estatísticas para analisar relações entre variáveis (Creswell, 2010).
Quanto aos objectivos, trata-se de um estudo exploratório e descritivo: exploratório por investigar um fenómeno ainda pouco discutido sob determinadas variáveis sociodemográficas, e descritivo por caracterizar a população estudada (mulheres que já tiveram pelo menos uma gestação) (Gil, 2019).
Quanto aos procedimentos, classifica-se como documental e levantamento, por utilizar dados secundários provenientes de registos já consolidados e por se basear em informações colectadas por questionário (Marconi & Lakatos, 2005).

3.2 Materiais

A pesquisa foi realizada com recurso a uma base de dados secundária fornecida pelo Instituto Nacional de Estatística (INE), referente ao Inquérito Demográfico de Saúde (IDS), realizado em Moçambique entre o período 2022 a 2023. A base de dados está disponível no site da Demographic Health Survey (DHS), `https://www.dhsprogram.com`, mediante um pedido de autorização para a realização do relatório.
Para o estudo foi usada uma base de dados composta por 2668 observações, que representam mulheres com idades compreendidas entre os 15 e 49 anos, as quais estiveram grávidas nos três anos anteriores à recolha das informações. O conjunto de dados inclui 11 variáveis que abrangem aspectos sociodemográficos, clínicos e relacionados à assistência ao pré-natal. A variável dependente, denominada **perda gestacional**, foi criada para identificar mulheres que tiveram perdas durante o

período de análise. Esta variável foi criada com base na ocorrência de perdas gestacionais, sendo categorizada em dois valores conforme mostra a tabela 3.1 **sim**, para mulheres que relatam uma ou mais perdas gestacionais, e **não**, para aquelas que não apresentaram nenhuma perda gestacional. Para processamento foi usado o software Microsoft Office Excel 2016 e R versão 4.1.2.

Tabela 3.1: Descrição das variáveis

Variável	**Descrição**	**Categorias**	**Classificação**
Província	Província de residência da inquirida	Niassa, Cabo Delgado, Nampula, Zambézia, Tete, Manica, Sofala, Inhambane, Gaza, Maputo Província, Maputo cidade	Categórica nominal
Idade	Idade actual do respondente	14-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59	Categórica ordinal
Estado Civil	Estado civil do respondente	Solteira, casada, união estável, viúva, divorciada, separada	Categórica nominal
Nível de riqueza	Nível de riqueza do respondente	Muito pobre, pobre, média, rica, muito rica	Categórica ordinal
Idade da primeira relação sexual	Idade em que a respondente teve a sua primeira relação sexual	-	Numérica contínua
Número de filhos já nascidos	Número de filhos que a respondente teve até a data do inquérito	-	Numérica discreta
Idade da primeira menstruação	Idade em que a respondente teve a sua primeira menstruação	-	Numérica contínua
Estado de fumante	Se a respondente fuma ou não	Fumante, não fumante	Categórica nominal
Uso de contraceptivo	Se a respondente usa algum método contraceptivo	Moderno, tradicional, não usa mas tenciona usar, não usa nem tenciona usar	Categórica nominal
Nível de escolaridade	Nivel de escolaridade actual da respondente	Sem escolaridade, primária, secundária, superior	Categórica ordinal
Perda gestacional	Se a respondente já teve alguma perda gestacional	Não, Sim	Categórica nominal (Dependente)

3.3 Métodos

A modelação estatística deste trabalho foi realizada por meio da Aprendizagem Supervisionada, uma abordagem que, segundo James *et al.* (2021), combina métodos estatísticos e computacionais para construir modelos capazes de prever respostas para novas observações ou explicar relações entre variáveis explicativas e uma variável. Hastie *et al.* (2009) explicam que a Aprendizagem Supervisionada utiliza conjuntos de dados rotulados para treinar algoritmos preditivos, que depois são avaliados em dados não utilizados no treino, para garantir a capacidade de generalização dos modelos.

No presente estudo, a Aprendizagem Supervisionada foi empregue para identificar os factores

associados à ocorrência de perdas gestacionais. Para isso, foram comparados dois modelos preditivos: a regressão logística, que é amplamente utilizada para modelar variáveis dependentes binárias e interpretar efeitos das covariáveis (Hosmer & Lemeshow, 2013), e a árvore de decisão, que oferece uma representação gráfica e interpretável dos critérios de decisão para classificação (Breiman *et al.*, 1984).

3.4 Análise Exploratória de Dados

A análise exploratória foi feita para a visualização da distribuição inicial das variáveis, identificar valores extremos e orientar as escolhas de modelação. Segundo Tukey (1977), a representação gráfica e numérica dos dados é importante pois revela padrões, tendências e efeitos que podem comprometer a validade das inferências estatísticas caso não sejam identificadas.

Estatística Descritiva das Variáveis Numéricas

As estatísticas descritivas foram feitas com base na média, mediana, desvio-padrão, mínimo e máximo, para as variáveis numéricas. Segundo Montgomery *et al.*(2021), a descrição de tendência central e dispersão permite avaliar assimetria, curtose e possível presença de outliers. Esses indicadores são importantes, pois a regressão logística pressupõe que a relação entre as variáveis preditoras numéricas e o logit da variável resposta seja aproximadamente linear, desvios extremos podem requerer transformações (Kleinbaum & Klein, 2010).

Comparação Bivariada

Foi construído gráfico de barras emparelhadas da distribuição de perda gestacional por província e idade. Para o restante das variáveis foi construída uma tabela de frequências cruzada para visualizar a proporção das variáveis em relação à variável dependente.

Foi também construído boxplots para as variáveis numéricas, permitindo inspecção simultânea de mediana, dispersão e outliers nos dois grupos.

A análise exploratória permitiu uma caracterização completa do conjunto de dados, fornecendo uma base empírica sólida para as etapas de modelação supervisionada que virão nos resultados.

Divisão da Base de Dados

No contexto do aprendizado de máquina, a divisão dos dados em conjuntos de treino e teste é fundamental para garantir a validade do modelo preditivo. Esse processo permite avaliar o desempenho do modelo em dados não vistos durante o processo de ajuste, simulando sua capacidade de generalização a novos cenários.

Segundo Géron (2022), esse procedimento é essencial para evitar o sobreajuste e garantir que o modelo aprenda padrões reais, e não apenas se ajuste aos dados do conjunto de treino. A divisão foi realizada de forma a preservar a proporção original das classes (*"sim"* e *"não"*) em ambos os subconjuntos.

Para este trabalho, a base foi separada em dois subconjuntos:

- **Conjunto de Treinamento (80%)** – utilizado para o ajuste dos modelos de Aprendizagem Supervisionada;
- **Conjunto de Teste (20%)** – utilizado para avaliação final do desempenho do modelo.

3.5 Balanceamento dos dados

Em problemas de classificação binária com variável resposta desequilibrada, como é o caso da variável aleatória, onde há número desproporcional de casos com perda e sem perda, o desempenho dos algoritmos preditivos pode ser prejudicado, especialmente em relação à classe minoritária (Batista *et al.*,2004). Este desequilíbrio tende a induzir o modelo a favorecer a classe maioritária, resultando em baixa sensibilidade e falhas na detecção dos eventos menos frequentes (Santos *et al.*,2021).

c

3.6 Construção do Modelo com a Regressão Logística

A construção do modelo de Regressão Logística foi realizada com base nos dados do conjunto de treino. A regressão logística binária modela a probabilidade $Pr\ (Y = 1|x)$ de ocorrência do evento $(Y = 1)$ em função do vector de preditores $x = (x_{1,\ \ldots},x_p)$. O modelo é especificado por:

$$\log\left[\frac{Pr\ (Y = 1|x)}{1 - Pr\ (Y = 1|x)}\right] = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \tag{3.1}$$

A função de verosimilhança é maximizada para obter os estimadores $\widehat{\beta}_j$. O *odds ratio* de cada covariável é dado por $exp(\widehat{\beta}_j)$, interpretado como a multiplicação da oportunidade de perda gestacional para cada unidade de aumento em x_j, mantendo-se as demais variáveis constantes (Hosmer *et al.*, 2013).

3.6.1 Avaliação da Multicolinearidade

Para cada preditor, efectuou-se o cálculo do Factor de Inflação da Variância (VIF):

$$VIF_j \frac{1}{1 - R_j^2} \tag{3.2}$$

Em que R_j^2 é o coeficiente de determinação obtido ao regredir x_j contra todos os demais preditores.

3.6.2 Selecção de Variáveis: Critério de Informação de Akaike

A selecção de variáveis utilizou o procedimento *stepwise* bidireccional, baseado no Critério de Informação de Akaike (AIC):

$$AIC = -2ln\left(\widehat{L}\right) + 2k \tag{3.3}$$

Onde $\widehat{L}$ é a verosimilhança do modelo ajustado e k é o número de parâmetros livres (Burnham & Anderson, 2004).

Segundo Hastie, Tibshirani e Friedman (2009), a seleção de variáveis em modelos estatísticos pode ser realizada por métodos stepwise, que alternam entre a inclusão (*forward*) e a exclusão (*backward*) de preditores. Esses métodos buscam otimizar critérios como o AIC, que balanceia a qualidade do ajuste do modelo com sua parcimônia, evitando o sobreajuste e garantindo modelos mais interpretáveis.

3.6.3 Avaliação do Modelo

A etapa de avaliação teve como objectivo verificar se o modelo de regressão logística se ajusta adequadamente aos dados e se possui uma boa capacidade explicativa. Foram aplicados testes de ajuste global, medidas de qualidade do modelo e diagnósticos de resíduos e influência, segundo recomendações de Hosmer *et al.* (2013) e Kleinbaum e Klein (2010)..

Teste de Hosmer-Lemeshow

O teste de Hosmer-Lemeshow foi usado para avaliar a concordância entre as probabilidades previstas e as frequências observadas dos dados. Neste teste as observações são classificados em *g* grupos (decis = 10) com base nas probabilidades estimadas $\widehat{\pi}_i$. A estatistica de teste é dado por:

$$X_{HL}^2 = \sum_{k=1}^{g} \frac{(O_j - E_j)^2}{E_j(1 - \widehat{\pi}_i)} \tag{3.4}$$

em que O_j é o número observado de eventos no grupo *k* e E_j é o número esperado. Valores de *p* $>$ 0,05 indicam bom ajuste do modelo (Hosmer *et al.*,2013).

Pseudo-R^2 de McFadden

Conforme aponta Menard (2022), para quantificar a capacidade explicativa global foi calculado o pseudo-R^2 de McFadden:

$$R^2_{McF} = 1 - \frac{ln(L_{modelo})}{ln(L_{nulo})} \tag{3.5}$$

Onde L_{modelo} é a verosimilhança do modelo ajustado e L_{nulo} é a verosimilhança do modelo sem preditores. Valores entre 0.20 e 0.40 são considerados indicativos de bom ajuste em estudos aplicados (Domencich & McFadden, 1975).

Diagnóstico de Resíduos e Valores Influentes

Para verificar a adequação local do modelo foram analisados:

- ***Leverage*** h_{ii}: identifica observações com influência potencial sobre os coeficientes; pontos com h_{ii} muito acima da média ($\frac{p+1}{n}$) merecem atenção.
- **Estatística de Cook** (D_i):

$$D_i = \frac{r^2_{D_i} h_{ii}}{(p+1)(1-h_{ii})^2} \tag{3.6}$$

  Onde *p* é o número de preditores. Valores $D_i > 1$ indicam alto impacto da observação sobre os parâmetros estimados (Belsley, Kuh & Welsch, 2005).

Teste de Wald para Significância dos Coeficientes

Para avaliar a significância estatística dos coeficientes estimados no modelo de regressão logística, aplicou-se o teste de Wald. Este teste tem como objectivo verificar se o parâmetro estimado (β) é estatisticamente diferente de zero, ou seja, se a variável explicativa associada possui um efeito significativo sobre a variável dependente (Field, 2013; Hosmer, Lemeshow & Sturdivant, 2013).

O teste de Wald é calculado com a seguinte fórmula:

$$W = \left(\frac{\widehat{\beta}}{SE(\widehat{\beta})} \right)^2 \tag{3.7}$$

Como descrito por Menard (2022), esta estatística segue uma distribuição qui-quadrado com um grau de liberdade, sob a hipótese nula de que o coeficiente é igual a zero (H_0: $\beta = 0$). Valores de *p* inferiores ao nível de significância $\alpha = 0.05$ indicam que o preditor tem um efeito estatisticamente significativo sobre a variável resposta.

Teste da Razão de Verosimilhança

A qualidade global do modelo de regressão logística foi examinada por meio do teste da razão de verosimilhança (*Likelihood-Ratio Test*, LR). Este teste compara o modelo completo (com todos os preditores seleccionados) ao modelo nulo (contendo apenas o intercepto), verificando se a inclusão das variáveis explicativas produz ganho estatisticamente significativo em termos de verosimilhança (Agresti, 2019).

A estatística do teste é definida por:

$$LR = -2[\ln(L_0) - \ln(L_1)], \tag{3.8}$$

onde:

- L_0 = verosimilhança do modelo nulo
- L_1 = verosimilhança do modelo completo

Sob a hipótese nula ($H_0 : \beta_1 = \ldots = \beta_p = 0$) segue uma distribuição qui-quadrado com *p* graus de liberdade, em que *p* é o número de parâmetros adicionais no modelo completo. Valores de *p* inferiores a 0.05 indicam que o modelo com preditores apresenta ajuste significativamente melhor do que o modelo nulo, evidenciando a relevância conjunta das variáveis explicativas (Hosmer, Lemeshow & Sturdivant, 2013).

3.6.4 Árvore de Decisão

Para este trabalho, foi usado o algoritmo CART para a construção do modelo de árvore de decisão usando a mesma base de treino usada para a construção do Modelo logístico. Segundo Breiman *et al.*(1984), essa técnica baseia-se em divisões sucessivas dos dados, com o objectivo de formar grupos internamente mais homogéneos em relação à variável perda gestacional. Para melhor visualização e interpretação fez-se também a renomeação de algumas variáveis.

Critério de Divisão

O critério usado para as divisões foi o índice de Gini, que quantifica o grau de impureza de um nó. O índice é calculado conforme a fórmula:

$$Gini = 1 - \sum_{i=1}^{C} p_i^2 \tag{3.9}$$

onde p_i representa a proporção de observações da classe i dentro do nó, e C é o número total de classes. Quanto menor o valor do Gini, maior a pureza do nó - ou seja, mais homogéneo em relação às categorias da variável dependente. As divisões são realizadas sempre buscando a maior redução possível nessa impureza (Loh, 2011).

Ajuste de Hiperparâmetros

Com o intuito de evitar o sobreajuste, foram aplicadas técnicas de pré-poda, limitando a profundidade máxima da árvore a 6 níveis e ajustando um número mínimo de 50 observações por nó para permitir novas divisões. A restrição da profundidade da árvore é uma medida recomendada para manter a interpretabilidade do modelo e prevenir o crescimento excessivo da árvore, o qual pode levar à captura de ruídos nos dados (Kuhn & Johnson, 2013). Segundo Hastie *et al.* (2009), árvores mais profundas tendem a ter menor viés, porém maior variância, o que pode comprometer a sua capacidade de generalização. Ajustou-se o parâmetro de complexidade (*cp*), que regula crescimentos a mais da árvore, sendo essencial na escolha de um modelo que equilibre acurácia e simplicidade. O parâmetro *cp* define o decréscimo mínimo na impureza necessário para uma divisão ser considerada. Modelos com valores muito baixos de *cp* tendem a ser mais complexos, enquanto valores mais altos favorecem árvores mais simples, com menos divisões (Breiman *et al.*, 1984).

Treinamento do Modelo Final

Após a selecção do valor óptimo do parâmetro de complexidade (*cp*) e a definição dos limites de pré-poda (*maxdepth* e *minsplit*), a árvore de decisão foi treinada utilizando todo o conjunto de treino balanceado. A selecção destes hiperparâmetros foi realizada com o auxílio de validação cruzada estratificada, permitindo identificar a configuração com melhor desempenho médio.

Importância das variáveis

A importância das variáveis no modelo final foi extraída da árvore ajustada. As variáveis mais relevantes foram apresentadas em um gráfico de barras ordenado, indicando sua contribuição relativa na classificação da perda gestacional (Breiman *et al.*, 1984).

3.6.5 Validação Cruzada Estratificada

Neste trabalho, será utilizada a técnica de validação cruzada estratificada do tipo k-fold, com $k = 10$, como estratégia para avaliação dos modelos de classificação. A validação cruzada consiste em particionar o conjunto de dados em k subconjuntos aproximadamente iguais, utilizando $k - 1$ para treino e o restante para validação, em ciclos sucessivos. Segundo James *et al.* (2021), essa técnica permite obter estimativas mais estáveis da performance do modelo, reduzindo a variabilidade decorrente de partições específicas dos dados.

A versão estratificada da validação *k*-fold será empregada com o intuito de preservar, em cada fold, a proporção original das classes da variável resposta. De acordo com Fernández *et al.* (2018), a estratificação é fundamental em cenários de desbalanceamento, pois evita que alguns subconjuntos contenham apenas exemplos da classe majoritária, o que comprometeria a avaliação real do

modelo.

Além disso, será incorporado o método método ROSE como estratégia de balanceamento. Conforme descrito por Lunardon, Menardi e Torelli (2014), o ROSE gera exemplos sintéticos de forma estocástica a partir da distribuição local das observações, promovendo equilíbrio entre as classes. No presente estudo, o ROSE será aplicado exclusivamente aos conjuntos de treino gerados em cada iteração da validação cruzada, mantendo o conjunto de validação inalterado. Essa prática evita a introdução de viés artificial na avaliação, garantindo que o desempenho dos modelos reflita a capacidade de generalização diante de dados com distribuição real.

3.7 Avaliação dos modelos preditivos

Com o modelo final de regressão logística e Árvore de decisão construído, foi conduzida a etapa de avaliação preditiva e de desempenho, utilizando-se o conjunto de teste. O objectivo é avaliar a capacidade do modelo em generalizar para novos dados, ou seja, medir sua eficácia preditiva fora da amostra de treinamento (James *et al.*, 2021).

3.7.1 Matriz de classificação

Segundo Han *et al.*(2011), a **matriz de classificação** é uma ferramenta fundamental para avaliação de classificadores binários. Ela resume o desempenho do modelo ao comparar as classes previstas com as reais.

A matriz é estruturada em quatro elementos principais:

Tabela 3.2: Matriz de classificação

	Classe Real: Não	**Classe Real: Sim**
Previsto: Não	Verdadeiro Negativo (VN)	Falso Negativo (FN)
Previsto: Sim	Falso Positivo (FP)	Verdadeiro Positivo (VP)

Estes quatro componentes permitem o cálculo das Medidas de desempenho:

Sensibilidade (Recall)

A sensibilidade representa a proporção de casos positivos correctamente identificados pelo modelo. No contexto deste estudo, indica a capacidade de o modelo detectar correctamente gestantes que tiveram perda gestacional.

$$\mathbf{Sensibilidade} = \frac{\mathbf{VP}}{\mathbf{VP} + \mathbf{FN}} \tag{3.10}$$

Uma alta sensibilidade é desejável quando o objectivo é minimizar falsos negativos.

Especificidade

A especificidade mede a proporção de negativos correctamente classificados. Representa a capacidade do modelo de reconhecer correctamente os casos em que não houve perda gestacional.

$$\mathbf{Especificidade} = \frac{\mathbf{VN}}{\mathbf{VN} + \mathbf{FP}} \tag{3.11}$$

É uma Medida importante para evitar alarmes falsos (falsos positivos).

Precisão (Positive Predictive Value)

A precisão refere-se à proporção de casos classificados como positivos que realmente são positivos. Avalia a confiança que se pode ter nas previsões positivas feitas pelo modelo.

$$\mathbf{Precisão} = \frac{\mathbf{VP}}{\mathbf{VP} + \mathbf{FP}} \tag{3.12}$$

F1-Score

O F1-score é a média harmónica entre sensibilidade e precisão. Para Saito e Rehmsmeier (2015), esta é uma Medida útil especialmente em situações de classes desbalanceadas, por ser mais informativa que a acurácia, pois captura melhor o desempenho do modelo ao lidar com classes minoritárias.

$$\mathbf{F1 - Score} = 2 \times \frac{\mathbf{Precisão} \times \mathbf{Sensibilidade}}{\mathbf{Precisão} + \mathbf{Sensibilidade}} \tag{3.13}$$

Acurácia

A acurácia representa a proporção total de classificações correctas (positivas e negativas) feitas pelo modelo.

$$\mathbf{Acurácia} = \frac{\mathbf{VP} + \mathbf{VN}}{\mathbf{VP} + \mathbf{FP} + \mathbf{FN} + \mathbf{VN}} \tag{3.14}$$

Porém, Saito e Rehmsmeier (2015) referem que a acurácia não é informativa quando a distribuição de classes é desbalanceada. Eles recomendam o uso de curvas de Precisão, Recall e F1-score como alternativas.

3.7.2 Curva ROC e AUC

A curva ROC (*Receiver Operating Characteristic*) é um gráfico que mostra a relação entre sensibilidade e a taxa de falsos positivos para diferentes pontos de corte. A área sob a curva (AUC) indica a capacidade geral de discriminação do modelo:

- AUC = 0.5: desempenho equivalente ao acaso;

- AUC entre 0.7 e 0.8: aceitável;
- AUC entre 0.8 e 0.9: excelente;
- AUC $>$ 0.9: excepcional (Hosmer *et al.*2013).

Para este trabalho, o ponto de corte óptimo foi definido com base no índice de Youden, que maximiza a soma da sensibilidade e especificidade.(Youden, 1950)

Capítulo 4

RESULTADOS

4.1 Análise exploratória

A amostra analisada no presente estudo é composta por 2668 respondentes submetidas ao um inquérito sobre perda gestacional, com idades compreendidas entre os 15 aos 49 anos, com uma média de aproximadamente 28 anos, tendo estas visto a sua primeira menstruação entre os 10 aos 21 anos, apresentando assim uma média de aproximadamente 14 anos, a idade da primeira relação sexual mostrou-se entre 8 a 27 anos de idade e em média, estas mulheres tiveram a sua primeira relação sexual aos 15 anos. Em relação ao número de filhos, no mínimo tiveram 1 filho e no máximo 12 filhos.

Tabela 4.1: Estatísticas descritivas das variáveis numéricas

Variáveis	Mínimo	Máximo	Média	Desvio padrão
Número de filhos já nascidos	1	12	3.382	2.105
Idade da primeira menstruação	10	21	13.8	2.039
Idade da primeira relação sexual	8	27	15.18	2.047

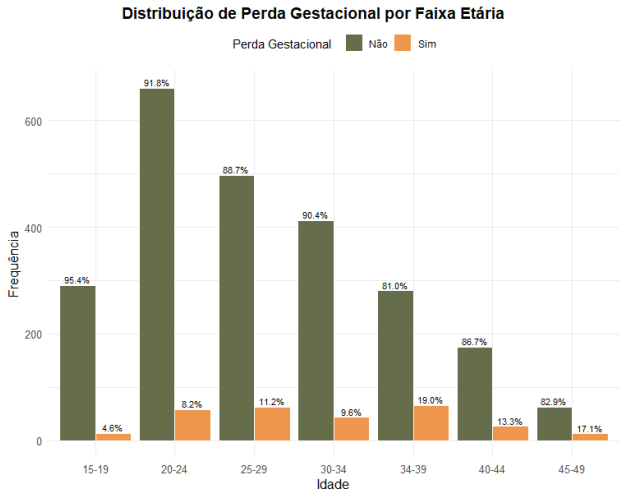

Figura 4.1: Distribuição das perdas gestacionais por idade

De acordo com a Figura 4.1, a maior parte das perdas gestacionais ocorreu em mulheres com idade entre 35 e 39 anos (19% dos casos). Em seguida, verificou-se que mulheres entre 45 e 49 anos apresentaram 17% das perdas. O gráfico mostra que a menor ocorrência de perdas gestacionais foi observada em mulheres mais jovens (15 a 19 anos), com apenas 5% dos casos.

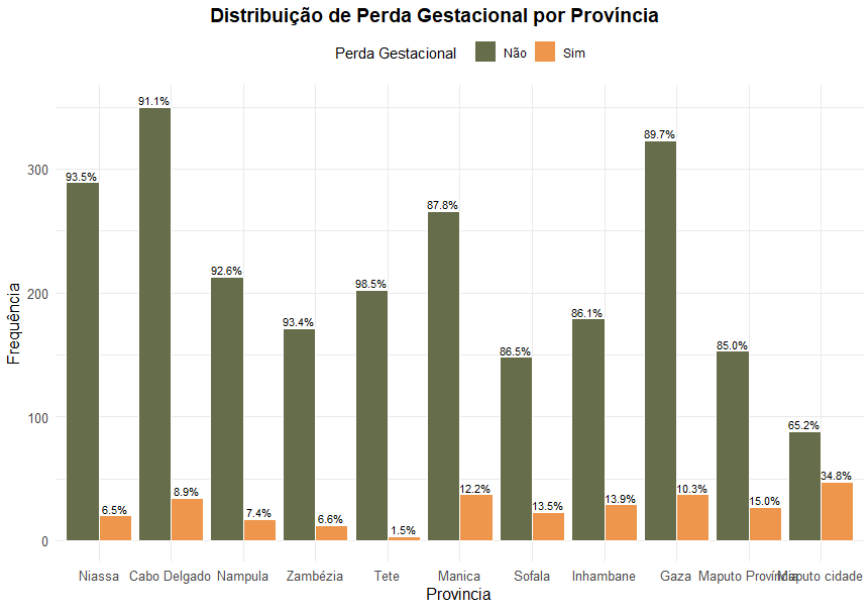

Figura 4.2: Distribuição das perdas gestacionais por província

Indo de acordo com a Figura 4.2, aproximadamente 35% das perdas gestacionais ocorreram em mulheres residentes na cidade de Maputo. A região Sul, composta pelas províncias de Inhambane, Gaza, Maputo Província e Maputo Cidade, concentrou o maior percentual de perdas gestacionais (74% dos casos).

A província de Tete apresentou o menor percentual de perdas, com apenas 1% das mulheres tendo vivenciado perdas gestacionais.

A Tabela 4.2 apresenta a relação entre diferentes factores sociais, de saúde e reprodutivos, com os percentuais de ”Perdas de gravidez”e ”Sem perdas de gravidez”em diferentes categorias.

Tabela 4.2: Tabela de Contingencia da distribuição das Perdas Gestacionais

Co-variáveis	Não (n=2382)	Sim (n=286)	Total (N=2668)
Índice de riqueza			
Muito Pobre	525 (93%)	40 (7%)	565
Pobre	347 (90%)	39 (10%)	386
Média	564 (92%)	49 (8%)	613
Rica	603 (89%)	74 (11%)	677
Muito Rica	337 (79%)	90 (21%)	427
Estado de Fumante			
Fumante	54 (91%)	5 (9%)	59
Não Fumante	2322 (89%)	287 (11%)	2609
Estado Civil			
Solteira	270 (92%)	24 (8%)	294
Casada	496 (92%)	44 (8%)	540
União estável	809 (87%)	124 (13%)	933
Viúva	126 (88%)	17 (12%)	143
Divorciada	173 (92%)	16 (8%)	189
Separada	508 (89%)	61 (11%)	569
Uso de Contraceptivo			
Moderno	737 (86%)	116 (14%)	853
Tradicional	15 (79%)	4 (21%)	19
Não usa mas tenciona usar	630 (89%)	74 (11%)	704
Não usa e nem tenciona usar	1000 (92%)	92 (8%)	1092
Nível de Escolaridade			
Sem escolaridade	621 (92%)	52 (8%)	673
Primário	1117 (90%)	130 (10%)	1247
Secundário	615 (87%)	89 (13%)	704
Superior	29 (66%)	15 (34%)	44

No que diz respeito ao estado civil actual as mulheres casadas, divorciadas ou viúvas têm percentuais semelhantes de perdas de gravidez (8%), mulheres vivendo com o parceiro apresentam uma maior proporção de perdas (13%). Em relação ao uso de contraceptivos, as que usam métodos modernos correspondem a 14% das mulheres com perdas gestacionais, quem usa métodos tradicionais tem o maior percentual de perdas (21%). No que concerne ao nível de escolaridade mulheres sem escolaridade correspondem a 8% das perdas, mulheres com educação superior apresentam o maior percentual de perdas de gravidez chegando a 34%. No que diz respeito ao índice de riqueza, as mulheres mais ricas tendem a ter mais perdas gestacionais, 21% das perdas gestacionais, ocorrerem em mulheres ricas. 8% das perdas correram em mulheres fumantes, contra os 92% das mulheres que não fumam, o que sugere que a variável estado de fumante não seja um determinante para a ocorrência de uma perda gestacionais.

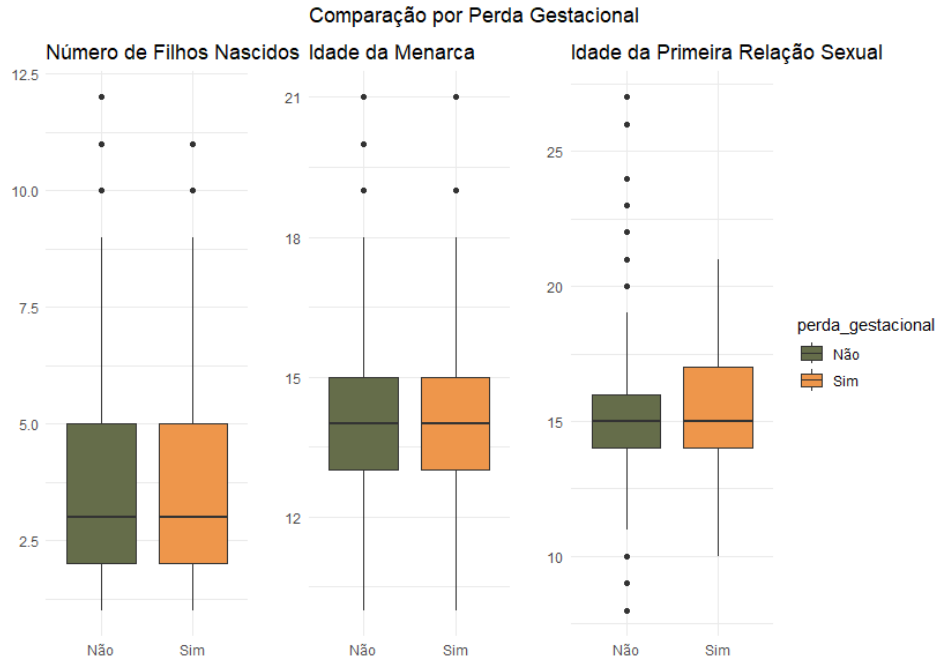

Figura 4.3: Box-plots das variáveis numéricas

A Figura 4.3 mostra que, no caso do número de filhos nascidos, a mediana é de aproximadamente 3 filhos tanto para mulheres com quanto sem perda gestacional. Entretanto, o grupo sem perdas apresentou valores mais elevados, incluindo casos de mulheres com 6 ou mais filhos e outliers acima de 10 filhos.
Quanto à idade da primeira menstruação (menarca), observaram-se distribuições semelhantes entre os grupos. Nota-se também a presença de valores acima de 18 anos, sem diferença visível entre os grupos.
Para a idade da primeira relação sexual, a distribuição foi mais dispersa, variando de 10 até mais de 25 anos, com presença de outliers tanto em idades precoces quanto tardias.

4.2 Modelação da Regressão logística

Inicialmente, a variável dependente estava presente em apenas cerca de 10% das observações da base de dados original mostrando assim um grande desbalanceamento das classes. Para a construção modelos preditivos supervisionado foi aplicado um procedimento de balanceamento apenas no conjunto de treino, e assim, a proporção de observações com perda gestacional foi ajustada para aproximadamente 49%, permitindo uma aprendizagem mais equilibrada por parte do modelo.

A partir da base de treino balanceada, estimou-se os parâmetros do modelo inicial de regressão logística, utilizando todas as variáveis disponíveis. Após o ajuste, foram calculados os factores

de inflacção da variância (VIF) para verificar a presença de multicolinearidade entre as variáveis explicativas, mostrando os seguintes resultados:

4.2.1 Avaliação da multicolinearidade

Os valores de VIF encontra-se todos abaixo de 10, conforme demonstra a tabela 4.3. A tolerância também está para todas variáveis acima de 0,1, mostrando assim que não há indícios de multicolinearidade que comprometam a estabilidade dos coeficientes da regressão logística.

Tabela 4.3: Valores de VIF (Fator de Inflação da Variância) e Tolerância das Variáveis Independentes

Variável	**VIF**	**Tolerância**
Idade	1.108	0.902
Província	1.085	0.922
Tipo de residência	1.465	0.683
Nível de escolaridade	1.171	0.854
Índice de riqueza	1.215	0.823
Nº de filhos nascidos	1.519	0.658
Idade da primeira menstruação	1.161	0.861
Uso de contraceptivo	1.077	0.928
Estado civil	1.085	0.922
Idade da 1ª relação sexual	1.223	0.818
Estado fumante	1.068	0.936

4.2.2 Estimação de parâmetros do Modelo de Regressção logística

As faixas etárias de 20-24, 25-29, 30-34, 35-39, 40-44 e 45-49 anos, segundo a tabela 4.4, apresentam razões de chances (odds ratios) significativamente maiores em comparação com o grupo de referência de 15-19 anos. As faixas de 35 a 39 anos (OR = 11.90), 40 a 44 anos (OR = 14.39) e 45 a 49 anos (OR = 12.83), demostram um crescimento na probabilidade de perda gestacional à medida que a idade materna aumenta.

Em relação à região, o local de residência mostrou-se fortemente associado à perda gestacional. Mulheres residentes na Cidade de Maputo apresentaram uma chance mais de dez vezes superior (OR = 10.55) de ter perda gestacional em comparação com as da província de Niassa (grupo de referência). Outras províncias, como Manica (OR = 4.83), Sofala (OR = 3.17), Gaza (OR = 2.07) e Inhambane (OR = 2.65) também apresentaram valores elevados.

Quanto ao nível de escolaridade, observou-se que mulheres com ensino primário, secundário e superior apresentaram uma probabilidade significativamente maior de relatar perda gestacional em relação às sem escolaridade. As razões de chances foram de 1.61 para o ensino primário, 1.85 para o secundário e 3.28 para o superior. A variável ”idade da primeira relação sexual”apresentou associação negativa com a perda: quanto mais tarde se inicia a relação sexual, reduz-se em cerca de 13% a chance de perda (OR = 0.87).

Tabela 4.4: Estimativas dos parâmetros do Modelo Logistico

Variável	B	S.E.	Wald	p-valor	Exp(B)	IC 95%
Intercepto	-1.2696	0.4603	7.6069	**0.0058**	0.2809	[0.11; 0.69]
Idade (ref: 15-19)						
20-24	1.1625	0.2205	27.7872	<**0.0001**	3.1980	[2.0927; 4.9757]
25-29	1.6476	0.2348	49.2547	<**0.0001**	5.1947	[3.3043; 8.3054]
30-34	1.4213	0.2480	32.8559	<**0.0001**	4.1423	[2.5651; 6.7888]
35-39	2.4766	0.2683	85.2074	<**0.0001**	11.9010	[7.0895; 20.3156]
40-44	2.6671	0.3177	70.4670	<**0.0001**	14.3983	[7.7822; 27.0655]
45-49	2.5521	0.3657	48.6954	<**0.0001**	12.8343	[6.3152; 26.5277]
Província (ref: Niassa)						
Cabo Delgado	0.6440	0.2164	8.8536	**0.0029**	1.9040	[1.2493; 2.9206]
Nampula	0.5049	0.2641	3.6564	0.0559	1.6569	[0.9867; 2.7816]
Zambézia	0.9707	0.2649	13.4271	**0.0002**	2.6398	[1.5728; 4.4486]
Tete	-2.0091	0.4612	18.9794	<**0.0001**	0.1341	[0.0492; 0.3087]
Manica	1.5740	0.2228	49.9076	<**0.0001**	4.8259	[3.1322; 7.5066]
Sofala	1.1523	0.2514	21.0033	<**0.0001**	3.1654	[1.9404; 5.2037]
Inhambane	0.9758	0.2382	16.7861	<**0.0001**	2.6533	[1.6684; 4.2474]
Gaza	0.7285	0.2219	10.7783	**0.0010**	2.0719	[1.3452; 3.2126]
Maputo Província	1.2610	0.2516	25.1287	<**0.0001**	3.5291	[2.1636; 5.8048]
Maputo Cidade	2.3565	0.2956	63.5508	<**0.0001**	10.5540	[5.9734; 19.0625]
Nível de escolaridade (ref: Sem escolaridade)						
Primário	0.4768	0.1437	11.0100	**0.0009**	1.6109	[1.2168; 2.1379]
Secundário	0.6157	0.1867	10.8781	**0.0010**	1.8510	[1.2852; 2.6725]
Superior	1.1895	0.4027	8.7233	**0.0031**	3.2854	[1.5313; 7.4898]
Índice de riqueza (ref: Muito pobre)						
Pobre	0.5496	0.1832	8.9983	**0.0027**	1.7326	[1.2107; 2.4840]
Média	0.2177	0.1728	1.5870	0.2077	1.2432	[0.8864; 1.7459]
Rica	0.4270	0.1778	5.7694	**0.0163**	1.5326	[1.0826; 2.1738]
Muito rica	0.5105	0.2263	5.0913	**0.0240**	1.6662	[1.0702; 2.5994]
Número de filhos nascidos	-0.0443	0.0308	2.0694	0.1503	0.9566	[0.9004; 1.0161]
Estado civil (ref: Solteira)						
Casada	0.0253	0.2043	0.0153	0.9015	1.0256	[0.6875; 1.5325]
União estável	0.5839	0.1778	10.7856	**0.0010**	1.7929	[1.2670; 2.5451]
Viúva	0.1262	0.2841	0.1973	0.6569	1.1345	[0.6506; 1.9841]
Divorciada	-0.1087	0.2636	0.1701	0.6800	0.8970	[0.5343; 1.5031]
Separada	0.0352	0.1938	0.0330	0.8559	1.0358	[0.7087; 1.5156]
Idade da primeira relação sexual	-0.1403	0.0251	31.3004	<**0.0001**	0.8691	[0.8272; 0.9126]
Estado de fumante (ref: Não fumante)						
Fumante	0.9464	0.3422	7.6497	**0.0057**	2.5763	[1.3248; 5.0999]

4.2.3 Teste de razão de Verossimilhança e pseudo-R^2

Na Tabela 4.5, mostra o teste de Verossimilhança, que representa uma medida do grau de ajustamento do modelo aos dados observados. Embora esse valor não mostre que o modelo é perfeito, ele pode ser considerado razoável, isso significa que o modelo ainda pode ajudar a entender os principais factores, mesmo sem representar todos os detalhes com precisão.

Os coeficientes de determinação pseudo-R^2, Cox & Snell (0.206) e Nagelkerke (0.275), mostram que o modelo é capaz de explicar entre 20,6% e 27,5% da variabilidade de perda gestacional.

Tabela 4.5: Teste de razão de verossimilhança e pseudo-R^2

Modelo	Deviance	Dif. Deviance	p-valor	(Cox & Snell)	(Nagelkerke)
Nulo	2959.56	-	-	-	-
Final	2467.42	492.1475	<0,001	0.20587	0.275

4.2.4 Teste de Hosmer e Lemeshow

O teste de Hosmer-Lemeshow, com o p-valor é maior que 0,05, sugere que não há evidências estatísticas para rejeitar a hipótese nula de que o modelo apresenta um bom ajuste aos dados. Isso indica que as probabilidades estimadas pelo modelo não diferem significativamente dos valores observados, sugerindo que o modelo é adequado para representar a realidade dos dados analisados.

Tabela 4.6: Teste de Hosmer-Lemeshow

Estatística	Valor
Qui-quadrado	8.9348
Graus de liberdade	8
p-valor	0.3478

4.2.5 Diagnóstico de Resíduos e Valores Influentes

Na Figura 4.4, observou-se que a grande maioria das observações apresenta valores muito baixos. As observações com maiores valores foram identificadas pelos índices 565, 1261, 1355, 1714 e 1775.

Porém, nenhuma observação ultrapassa o valor de 1, que é indicativo de influência exagerada. Isso sugere que não há observações altamente influentes no modelo, ou seja, não há necessidade de excluir ou tratar especificamente essas unidades no presente modelo de regressão logística.

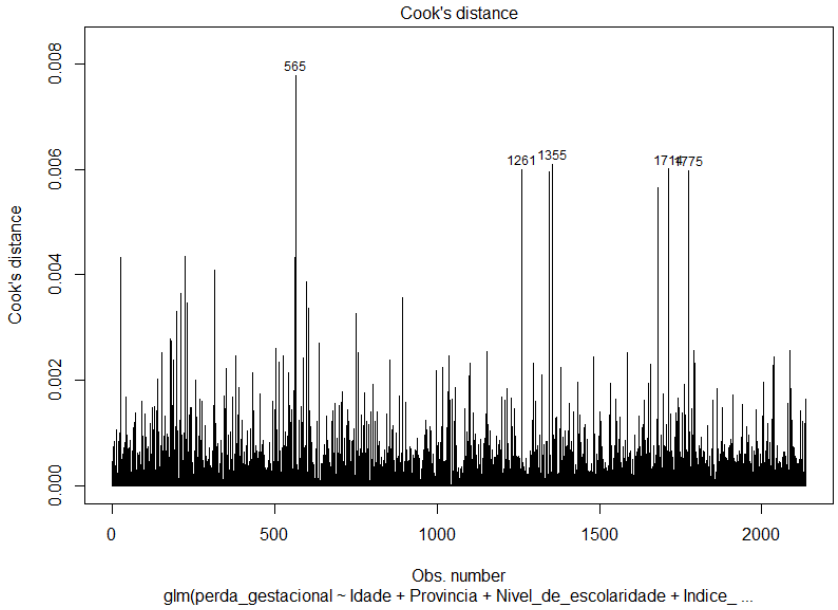

Figura 4.4: Gráfico de influência das observações

4.2.6 Avaliação do desempenho do modelo

Abaixo estão apresentados os resultados obtidos considerando Medidas como acurácia, sensibilidade, especificidade, valor preditivo positivo usados para avaliação do desempenho do modelo, foi usando dados correspondentes a 20% da base original.

Tabela 4.7: Matriz de Classificação do Modelo de Regressão logistica

	Classe observada	
Classe prevista	Não	Sim
Não	265	168
Sim	19	81

Tabela 4.8: Medidas de desempenho do Modelo de Regressão logistica

Medida	Valor
Acurácia	0.6473
No Information Rate	0.813
Precisão	0.325
Sensibilidade	0.817
F1-score	0.464
Detection Prevalence	0.467
Acurácia Balanceada	0.690

O modelo apresentou uma acurácia de 64.73%, indicando que a maior parte das observações do conjunto de teste foi classificada corretamente. A precisão do modelo foi de 32.53%, o que significa que, das vezes em que o modelo previu perda gestacional, aproximadamente um terço dessas previsões estava correta, o que ainda revela uma quantidade considerável de falsos positivos. Em contrapartida, a sensibilidade foi elevada, com 81.70%, demonstrando que o modelo conseguiu identificar corretamente a maior parte dos casos reais de perda gestacional. O modelo classificou como positivos 46.72% dos casos. A acurácia balanceada foi de 69.07%, sugerindo um desempenho relativamente estável mesmo diante do desbalanceamento das classes. No geral, os resultados indicam que o modelo tem boa capacidade de detectar casos positivos de forma sensível, mas ainda requer melhorias no controle dos falsos positivos.

4.2.7 Avaliação do Modelo de regressão logística com Validação Cruzada

Com base nos resultados iniciais obtidos na base de teste, observou-se que, embora o modelo de regressão logística tenha alcançado uma sensibilidade satisfatória, a precisão e o equilíbrio geral entre as classes ainda eram limitados, sobretudo devido ao desbalanceamento da base de dados. Para contornar essas limitações e aumentar a robustez do modelo, adotou-se uma abordagem de validação cruzada, que permite estimar o desempenho de forma mais confiável, reduzindo a variância associada à divisão única dos dados.

Tabela 4.9: Medidas de validação cruzada do Modelo de Regressão logistica

Medida	Valor
Acurácia	0.705
No Information Rate	0.591
Precisão	0.374
Sensibilidade	0.821
F1-score	0.529
Detection Prevalence	0.403
Acurácia Balanceada	0.731

O modelo de regressão logística apresentou uma acurácia de 70.5%, e um No Information Rate de 59.1%, indicando um desempenho melhor que o acaso na predição de perdas gestacionais. Possui uma alta sensibilidade de 82.1%, que mostra boa capacidade do modelo em identificar corretamente a maioria dos casos reais, mas a precisão foi baixa (37.4%), sugerindo a ocorrência de um número significativo de falsos positivos.
O F1-score de 0.529 e a acurácia balanceada de 73.1% indicam um desempenho moderado e equilibrado entre as classes.

4.3 Modelação da Árvore de Decisão

Para a construção do modelo de árvore de decisão foi usada a mesma base de treino usada para a construção do Modelo de regressão logística. Segundo Breiman *et al.*(1984) essa técnica baseia-se em divisões sucessivas dos dados, com o objectivo de formar grupos internamente mais homogêneos em relação à variável perda gestacional. Para melhor visualização e interpretação fez-se também a renomeação de algumas variáveis.

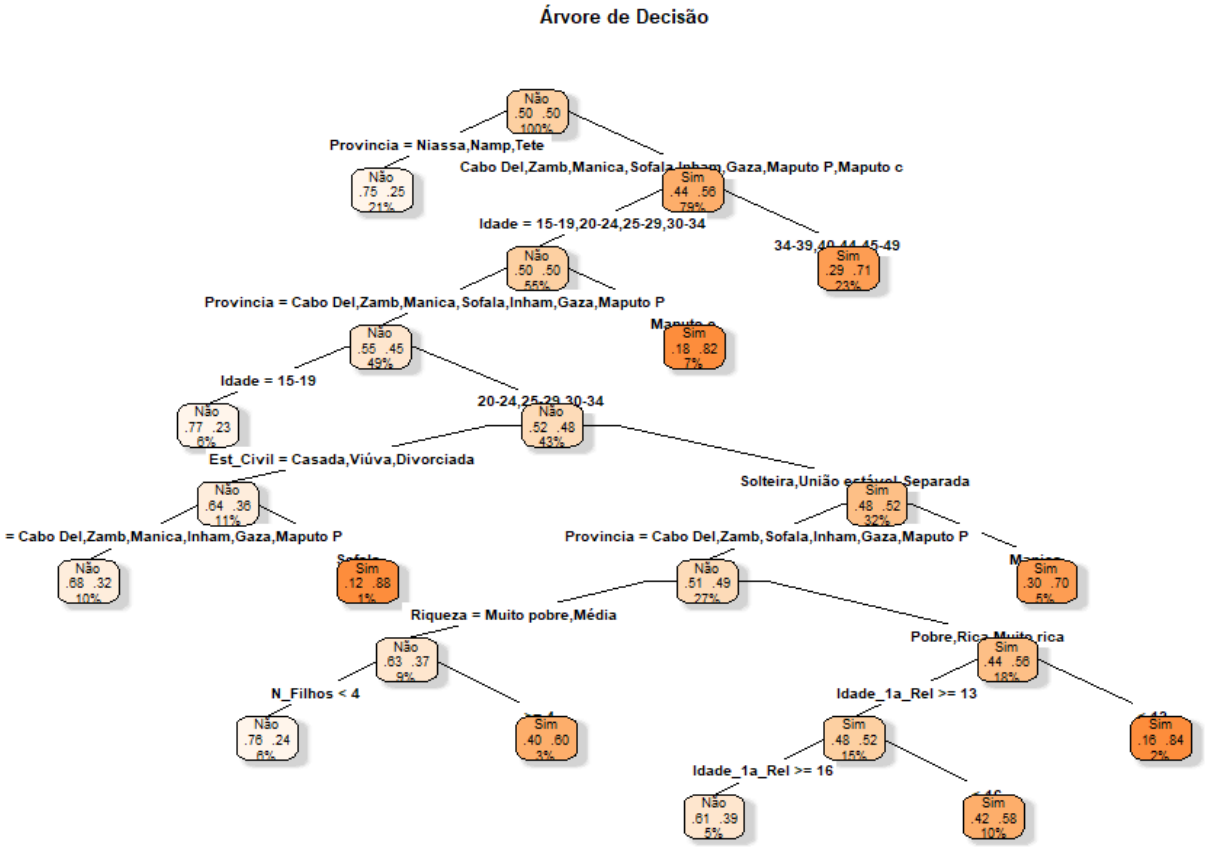

Figura 4.5: Árvore de decisão inicial

A árvore de decisão inicial acima gerada apresenta uma estrutura ramificada, com múltiplas divisões baseadas nas variáveis província, idade, estado civil, número de filhos e nível de riqueza. É verificado que o nível de complexidade dessa árvore tende a demonstrar um sobreajuste (*overfitting*), pois verifica-se que algumas ramificações possuem pequenos subconjuntos da amostra. Isso pode comprometer a capacidade de generalização do modelo para novos dados.

4.3.1 Avaliação do modelo da árvore de decisão

Tabela 4.10: Matriz de classificação do modelo da árvore de decisão

	Classe observada	
Classe prevista	Não	Sim
ão	290	166
Sim	21	34

Tabela 4.11: Medidas de desempenho do modelo da árvore de decisão

Medida	Valor
Acurácia	0.614
No Information Rate	0.893
Precisão	0.172
Sensibilidade	0.614
F1-score	0.269
Detection Prevalence	0.380
Acurácia Balanceada	0.631

A acurácia geral do modelo foi de 61.4%, mas menor que o NIR de 89.3%, o que indica que o modelo não supera de forma significativa a escolha da classe majoritária. A precisão baixa demonstra que, nos casos classificados como perdas gestacionais, a maioria foi classificada incorretamente. O F1-score de 26.9% demonstra desequilíbrio entre precisão e sensibilidade, sugerindo fraca performance na identificação efetiva dos casos positivos.

4.3.2 Ajuste de hiperparâmetros

A construção do modelo final foi baseada na curva de validação cruzada do parâmetro de complexidade (cp), esta que demonstrou que os maiores valores de AUC foram obtidos com árvores menos podadas (cp próximo de zero) segndo a Figura 4.6. Com base nisso, optou-se pela modelação uma árvore de decisão final mais extensa que a inicial baseada no cp ótimo, uma vez que essa estrutura mais complexa pode apresentar melhor desempenho preditivo.

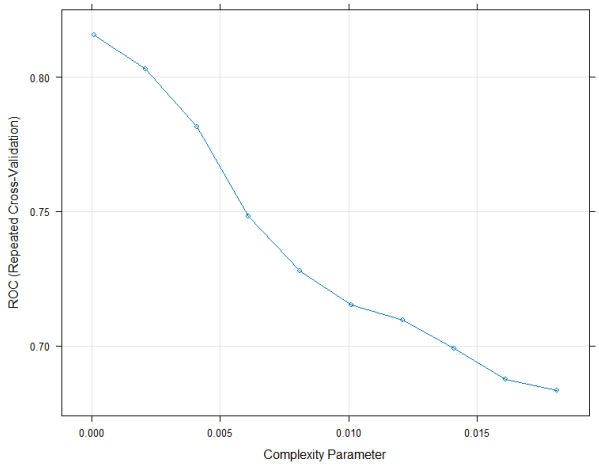

Figura 4.6: Curva de validação para seleção do parâmetro de complexidade (cp)

Após o processo de ajuste dos hiperparâmetros, como validação cruzada estratificada k= 10 folds , parâmetro de complexidade (cp), obteve-se a árvore final apresentada na Figura 4.7.

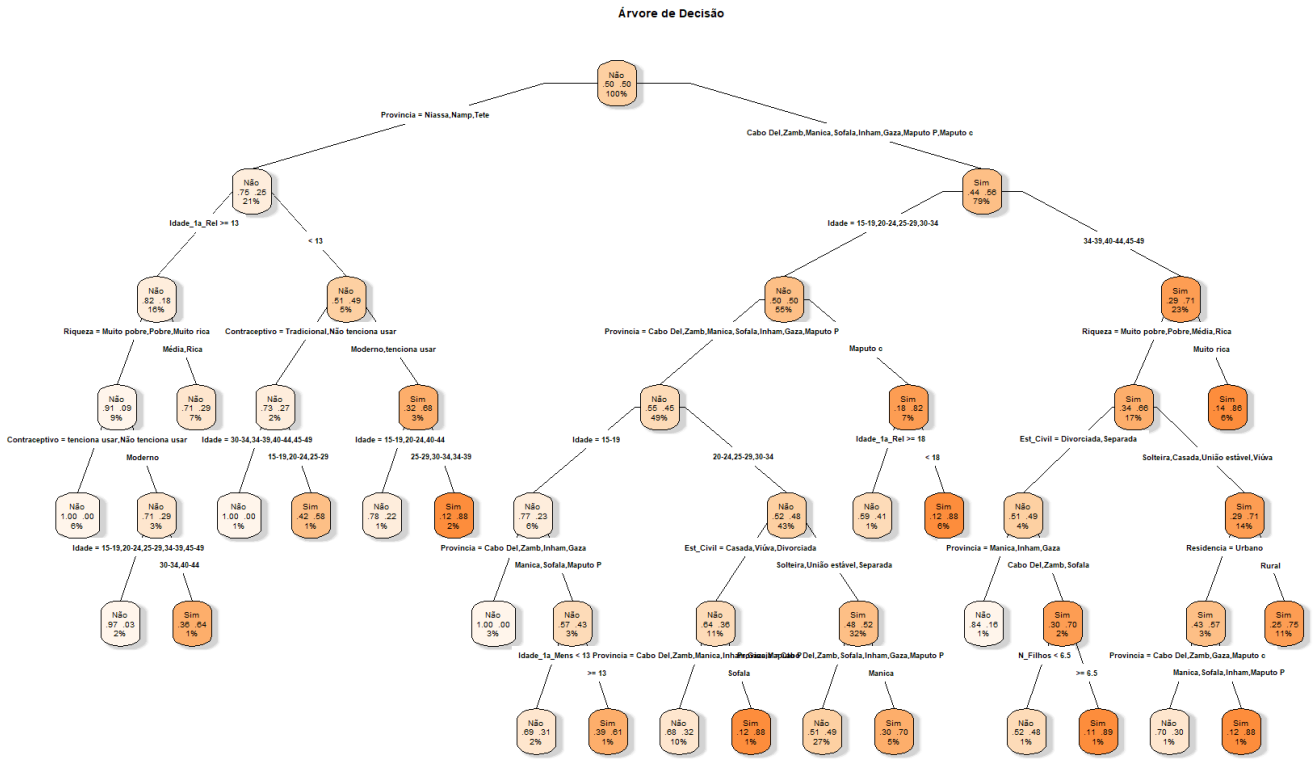

Figura 4.7: Modelo final da árvore de decisão

A árvore mostra em cada nó: a classe majoritária, a proporção de casos e as percentagens de observações em relação ao total. As cores representam a proporção de casos da classe ”Sim”,

quanto mais escura a cor laranja, maior a proporção de perda gestacional.

A província de residência foi a variável mais discriminante para prever a ocorrência de perda gestacional. Isso significa que o local onde a mulher vive está altamente associado ao risco de perda gestacional. Mulheres que residem nas províncias de Cabo Delgado, Zambézia, Manica, Sofala, Inhambane, Gaza, Maputo Província e Maputo Cidade apresentaram uma proporção de 50% de perda gestacional, enquanto nas províncias de Niassa, Nampula e Tete essa proporção foi de apenas 25%.

4.3.3 Importância das variáveis

Segundo a Figura 4.8 variável com maior influência foi **Província**, indicando que o local de residência pode estar fortemente associado à probabilidade de perda gestacional. Em seguida, a **idade da mulher** também mostrou-se importante. As variáveis **idade da primeira relação sexual**, **número de filhos**, **estado civil** e **índice de riqueza** também demonstraram contribuição relevante para a classificação das perdas. As variáveis **residência (urbana/rural)** e **status de fumante** apresentaram baixa importância no modelo, sugerindo que seu impacto preditivo foi bastante reduzido

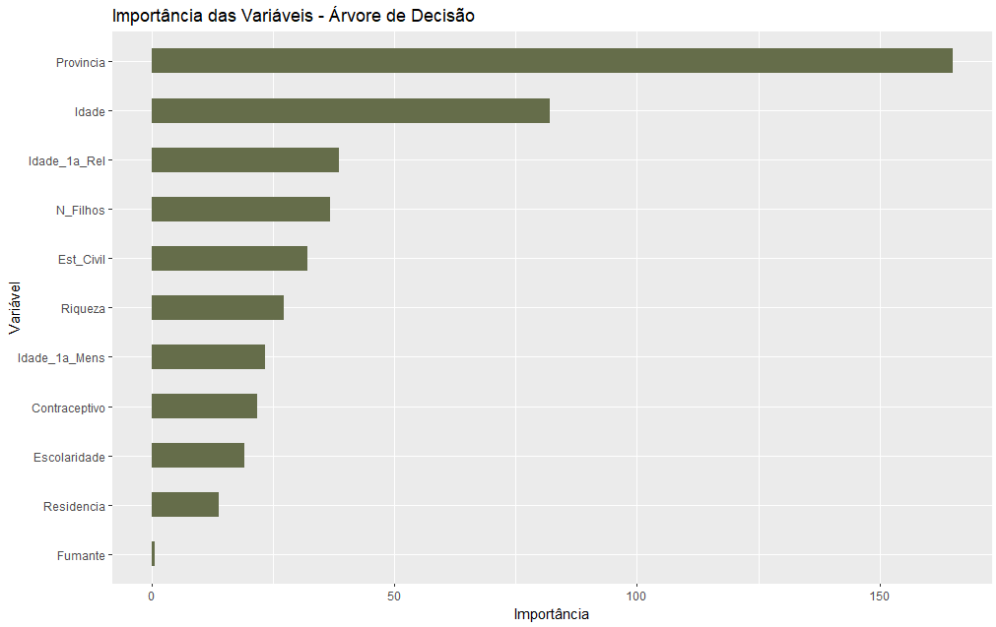

Figura 4.8: Importância das variáveis no modelo final

.

4.3.4 Avaliacão do modelo final da Árvore de decisão

Tabela 4.12: Matriz de Classificação do modelo final da Árvore de decisão

	Classe observada	
Classe prevista	Não	Sim
Não	290	133
Sim	28	82

Tabela 4.13: Medidas de desempenho do modelo final da Árvore de decisão

Medida	Valor
Acurácia	0.698
No Information Rate	0.793
Precisão	0.381
Sensibilidade	0.745
F1-score	0.504
Detection Prevalence	0.403
Acurácia Balanceada	0.716

O modelo final da árvore de decisão apresentou uma acurácia de 69.8%. A precisão foi de 38.1%, revelando que pouco mais de um terço das previsões positivas feitas pelo modelo correspondiam, de fato, a casos reais de perda gestacional. A sensibilidade foi de 74.5%, o que demonstra que o modelo foi eficaz na identificação da maioria dos casos reais da classe minoritária. O F1-score, que representa a média harmônica entre precisão e sensibilidade, foi de 50.4%, sugerindo um desempenho equilibrado nas predições positivas. O modelo classificou como positivos 40.3% dos casos. A acurácia balanceada, de 71.6%, revela um desempenho razoável considerando o desbalanceamento das classes.

4.3.5 Validação cruzada

Para obter estimativas mais fiáveis do desempenho, recorreu-se à validação cruzada para avaliar o modelo de forma mais estável e generalizável.

Tabela 4.14: Medidas de desempenho do modelo de árvore de decisão após validação cruzada

Medida	**Valor**
Acurácia	0.726
No Information Rate	0.591
Precisão	0.447
Sensibilidade	0.866
F1-score	0.587
Detection Prevalence	0.403
Acurácia Balanceada	0.754

O modelo apresentou uma acurácia geral de 72,6%, superando o No Information Rate de 59,1%, o que indica desempenho melhor que a predição da classe majoritária. A sensibilidade elevada (86,6%) demonstra boa capacidade de detectar casos positivos. O F1-score de 58,7% evidencia um equilíbrio razoável entre sensibilidade e precisão, enquanto a acurácia balanceada (75,4%) confirma que o modelo lida adequadamente com o desbalanceamento das classes.

4.4 Comparação de desempenho (Regressão Logística VS Árvore de decisão)

Tabela 4.15: Comparação entre modelo de Regressão Logística e modelo de Árvore de Decisão

Medida	Regressão Logística	Árvore de Decisão
Acurácia	0.647	0.698
No Information Rate	0.813	0.793
Precisão	0.325	0.381
Sensibilidade	0.817	0.745
F1-score	0.465	0.504
Detection Prevalence	0.468	0.403
Acurácia Balanceada	0.716	0.746

O modelo de regressão logística apresentou uma boa sensibilidade (81.7%), indicando que identificou correctamente a maioria das mulheres que tiveram perdas gestacionais. No entanto, a precisão foi relativamente baixa (32.5%), o que revela uma elevada proporção de falsos positivos.

A acurácia geral foi de 64.7%, inferior à obtida pelo modelo de árvore de decisão, e o F1-score situou-se em 0.465, reflectindo o desequilíbrio entre sensibilidade e precisão. A acurácia balanceada foi de 71.6%, evidenciando um desempenho razoável do modelo na presença de classes desbalanceadas.

O modelo de árvore de decisão demonstrou desempenho ligeiramente superior em várias Medidas: acurácia de 69.8%, F1-score de 0.504 e acurácia balanceada de 74.6%. Embora a sua sensibilidade (74.5%) tenha sido inferior à da regressão logística, a precisão foi mais elevada (38.1%), indicando que o modelo cometeu menos falsos positivos. O melhor equilíbrio entre sensibilidade e precisão resultou num F1-score superior ao da regressão, sugerindo maior eficácia global na identificação de casos positivos.

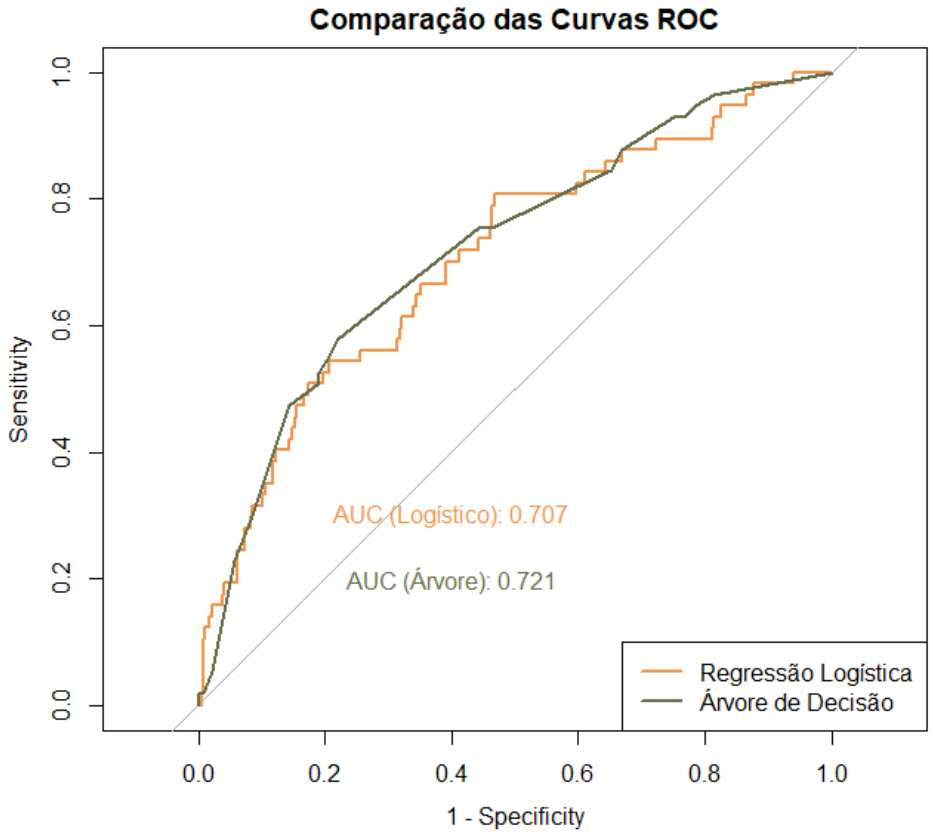

Figura 4.9: Curva ROC para comparação dos modelos

A Área Sob a Curva (AUC) para o modelo logístico foi de 0.707 , este valor indica que o modelo apresenta capacidade moderada de discriminação entre gestantes que sofreram perda gestacional e aquelas que não sofreram esse desfecho. Isso significa que, ao escolher aleatoriamente uma gestante com perda gestacional e outra sem, há aproximadamente 70.7% de probabilidade de o modelo atribuir uma probabilidade maior à gestante que de facto teve perda.

No caso do modelo de Árvore de Decisão, a AUC foi de 0.721, este valor indica que o modelo apresenta capacidade de discriminação moderada, ou seja, ao escolher aleatoriamente uma gestante com perda gestacional e outra sem, há 72.1% de probabilidade de o modelo atribuir uma probabilidade maior de perda à gestante que de facto sofreu esse desfecho. A Figura 4.9 atribui uma ligeira vantagem para a abordagem baseada na Árvore de Decisão.

4.5 Discussão dos Resultados

Os dois métodos aplicados, Regressão Logística e Árvore de Decisão, apresentaram resultados parecidos em termos de desempenho preditivo. As variáveis com maior influência na perda gestacional foram praticamente as mesmas em ambos os modelos, ainda que com diferentes intensidades, tendo-se destacado a Província como variável com maior importância em ambos os modelos.

A idade materna avançada (35–49 anos) foi mostrada como o principal factor de risco. A regressão logística revelou razões de chances entre 11 e 14 vezes maiores nessa faixa etária, resultado que vai de acordo com estudos de Regassa *et al.* (2022) e Liu *et al.* (2024), que identificaram a gravidez tardia como determinante para perdas gestacionais.

A escolaridade elevada e o maior nível de riqueza, surpreendentemente, ao contrário do que Alberto (2023), Regassa *et al.* (2022) e Yehuala *et al.* (2025) destacam, mostraram associação positiva com perdas gestacionais. Essa contradição pode ser explicada por factores como: mulheres com maior escolaridade e rendimento tendem a projectar uma gravidez já em idade mais avançada, e o maior acesso a exames e registos pode aumentar a probabilidade de identificação de perdas gestacionais (Liu *et al.*, 2024).

Outro factor importante foi a província de residência. Ambos os modelos apontaram Maputo Cidade como a cidade com maior risco, bem como outras províncias do sul. Tal como concluído por Alberto (2023), esta questão pode dever-se a uma maior concentração de hospitais, melhor diagnóstico e mais registos.

A idade da primeira relação sexual apresentou uma associação negativa com a perda gestacional, indicando que mulheres que iniciaram a vida sexual mais tarde apresentaram menor risco. A UNICEF (2023) associa a gravidez precoce à imaturidade biológica e à ausência de cuidados.

A ausência de variáveis clínicas pode ter influênciado directamente no desempenho dos modelos pois como visto por Qi *et al.*(2024) estas variaveis podem aumentar a capacidade de detenção desses desfechos.

sensibilidade registada pelo modelo logístico (81,7%) foi superior à da árvore de decisão

(74,5%). Todavia, as precisões obtidas por estes modelos foram de 32,5% para o modelo logístico e 38,1% para a árvore, o que demonstra que, embora o modelo logístico identifique a maioria dos casos de mulheres que tiveram perdas gestacionais, a sua precisão é inferior à da árvore. Em termos práticos, isto significa que o modelo logístico apresenta uma proporção mais elevada de falsos positivos.

Essas diferenças são coerentes com a literatura, pois, segundo Hosmer *et al.*(2013), modelos lineares como a regressão logística tendem a ser mais sensíveis, o que é desejável em contextos onde o custo de perder um caso verdadeiro é alto, enquanto modelos baseados em árvore frequentemente oferecem melhor acurácia geral e robustez preditiva (Silva, 2021; Liu *et al.*2024).

Capítulo 5

CONCLUSÕES E RECOMENDAÇÕES

5.1 Conclusão

Este estudo aplicou modelos de Aprendizagem Supervisionada para identificar os factores associados à perda gestacional em Moçambique, utilizando dados sociodemográficos e comportamentais. Os resultados demonstraram que variáveis como idade materna avançada, província de residência, escolaridade elevada e nível de riqueza estão significativamente associadas ao aumento do risco de perda gestacional.

Neste estudo, onde se buscou a compreensão dos factores associados e a aplicabilidade prática, os resultados indicaram que, enquanto a regressão logística apresentou maior sensibilidade, sendo mais eficaz para detectar casos positivos, a árvore de decisão teve melhor desempenho global, com maior acuidade e equilíbrio entre sensibilidade e precisão. Ambos os modelos, portanto, mostraram-se úteis e complementares: a regressão é mais adequada para fins explicativos e análise de risco, enquanto a árvore de decisão é mais prática para uso em triagens e sistemas de apoio à decisão.

De forma geral, os resultados confirmam que a idade materna, localização, escolaridade e condição económica exercem influência significativa no desfecho de gestações. Além disso, a comparação entre os modelos mostrou que, apesar de diferenças nas Medidas, ambos são eficazes e podem ser usados de forma complementar.

5.2 Recomendações

Com base na discussão e nas conclusões obtidas neste estudo, recomenda-se:

- A implementação de campanhas de conscientização e palestras que abordem o uso adequado de métodos contraceptivos, bem como os riscos associados à gravidez precoce ou tardia.

- Melhorar o acesso às consultas pré-natais e a exames de diagnóstico em todas as províncias, com atenção especial às regiões remotas ou carentes, de forma a garantir a detecção precoce de complicações gestacionais.
- Expandir a análise para incluir variáveis como histórico médico, complicações gestacionais, acesso a água potável, nutrição, que não foram aqui abordadas.
- Utilizar outras técnicas de Aprendizagem Supervisionada, visando comparar os desempenhos e melhorar a precisão preditiva dos modelos.

5.3 Limitações

- A base de dados usada não contém variáveis clínicas detalhadas, como hipertensão gestacional, diabetes ou histórico familiar de complicações obstétricas, as quais poderiam enriquecer a análise dos factores de risco.
- Apesar da aplicação do método ROSE para balanceamento da base de dados, a classe minoritária (perdas gestacionais) permaneceu pouco representada, o que pode ter comprometio a generalização dos resultados obtidos pelos modelos preditivos.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Agresti, A. (2019). *An introduction to categorical data analysis* (3rd ed.). Wiley.

[2] Alberto. J. (2023). *Condições de nascimento e fatores gestacionais associados, antes e durante a pandemia da COVID-19, no distrito de Nampula–Moçambique* (Dissertação de mestrado). Universidade Federal de Viçosa. https://locus.ufv.br/handle/123456789/31784

[3] Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004).A study of the behavior of several methods for balancing Aprendizado de Máquina training data. *SIGKDD Explorations*, *6*(1), 20–29. https://doi.org/10.1145/1007730.1007735

[4] Bearak, J., Popinchalk, A., Alkema, L., & Sedgh, G. (2018). Global, regional, and subregional trends in unintended pregnancy and its outcomes from 1990 to 2014: Estimates from a Bayesian hierarchical model. *The Lancet Global Health, 6*(4), e380–e389. https://doi.org/10.1016/S2214-109X(18)30029-9

[5] Belsley, D. A., Kuh, E., & Welsch, R. E. (2005). *Regression diagnostics: Identifying influential data and sources of collinearity*. John Wiley & Sons.

[6] Bishop, C. M. (2006). *Pattern recognition and Aprendizado de Máquina*. Springer.

[7] Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. Chapman & Hall/CRC.

[8] Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics, 24*(6), 2350–2383.

[9] Breiman, L. (2001). *Floresta aleatórias*. Aprendizado de Máquina, 45(1), 5–32.

[10] Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research, 33*(2), 261–304. https://doi.org/10.1177/0049124104268644

[11] Chapelle, O., Schölkopf, B., & Zien, A. (2009). *Semi-supervised learning*. IEEE Transactions on Neural Networks, 20(3), 542–542. https://doi.org/10.1109/TNN.2009.2015974

[12] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321–357. https://doi.org/10.1613/jair.953

[13] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Em *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

[14] Cleland, J., Conde-Agudelo, A., Peterson, H., Ross, J., & Tsui, A. (2012). *Contraception and health*.The Lancet, 368(9549), 149–156. https://doi.org/10.1016/S0140-6736(06)69481-0

[15] Conde-Agudelo, A., Rosas-Bermúdez, A., & Kafury-Goeta, A. C. (2006). Birth spacing and risk of adverse perinatal outcomes: A meta-analysis. *JAMA, 295*(15), 1809–1823. https://doi.org/10.1001/jama.295.15.1809

[16] Creswell, J. W. (2010). *Projeto de pesquisa: métodos qualitativo, quantitativo e misto* (3. ed.). Artmed.

[17] Domencich, T. A., & McFadden, D. (1975). *Urban travel demand: A behavioral analysis*. North-Holland.

[18] Escovedo,T. (2020).*Aprendizado de Máquina: conceitos e modelos*. Medium.https://tatianaesc.medium.com/machine-learning-conceitos-e-modelos-f0373bf4f445

[19] Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from imbalanced data sets*. Springer.

[20] Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. Sage Publications.

[21] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

[22] Géron, A. (2022). *Hands-on machine learning with scikit-learn, keras, and tensorflow: Concepts, tools, and techniques to build intelligent systems* (3rd ed.). O'Reilly Media.

[23] Gil, A. C. (2019). *Métodos e técnicas de pesquisa social* (7. ed.). Atlas. .

[24] Goldenberg, R. L., McClure, E. M., Bhutta, Z. A., et al. (2011). Stillbirths: The vision for 2020. *The Lancet, 377*(9777), 1798–1805. https://doi.org/10.1016/S0140-6736(10)62235-0

[25] Gujarati, D. N., & Porter, D. C. (2009). *Basic econometrics* (5th ed.). New York: McGraw-Hill.

[26] Han, J., Kamber, M., & Pei, J. (2011). *Data mining: concepts and techniques* (3rd ed.). Elsevier.

[27] Hastie, T., Tibshirani, R., and Friedman, J. (2009).*The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.

[28] Health Canada. (2000). *Canadian Perinatal Health Report 2000*. Minister of Public Works and Government Services Canada.

[29] Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

[30] IDS. (2023). *Inquérito Demográfico e de Saúde de Moçambique*. Instituto Nacional de Estatística.

[31] Islam, M. N., Mustafina, S. N., Mahmud, T., & Khan, N. I. (2022). Aprendizado de Máquina to predict pregnancy outcomes: a systematic review, synthesizing framework and future research agenda. *BMC Pregnancy and Childbirth, 22*, 348. https://doi.org/10.1186/s12884-022-04594-2

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

[33] Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis, 6*(5), 429–449. https://doi.org/10.3233/IDA-2002-6504

[34] Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

[35] Kleinbaum, D. G., & Klein, M. (2010). *Logistic regression: A self-learning text* (3rd ed.). Springer.

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

[37] Kuhn, M., and Johnson, K. (2013). *Applied predictive modeling*. Springer.

[38] Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, *61*(10), 36–43. https://doi.org/10.1145/3233231

[39] Liu, P.-Y., Fridline, M., & Miller, B. (2021). Prediction of neonatal deaths in NICUs: development and validation of Aprendizado de Máquina models. *BMC Medical Informatics and Decision Making*, *21*(1), 1–10. https://doi.org/10.1186/s12911-021-01497-8

[40] Liu, Y., Liu, J., & Shen, H. (2024). Aprendizado de Máquina model-based preterm birth prediction and clinical nomogram: A big retrospective cohort study. *International Journal of Gynecology & Obstetrics*, 169(1), 332–340. https://doi.org/10.1002/ijgo.16036

[41] Loh, W.-Y. (2011). Classification and regression trees. *Data Mining and Knowledge Discovery*, *1*(1), 14–23. https://doi.org/10.1002/widm.8

[42] Lunardon, N., Menardi, G., & Torelli, N. (2014). ROSE: a Package for Binary Imbalanced Learning. *The R Journal*, *6*(1), 79–89. https://doi.org/10.32614/RJ-2014-008

[43] Marconi, M. A., & Lakatos, E. M. (2005). *Fundamentos de metodologia científica* (6ª ed.). São Paulo, SP: Atlas.

[44] Menard, S. (2010). *Logistic regression: From introductory to advanced concepts and applications*. Sage Publications.

[45] Menard, S. (2022). *Applied logistic regression analysis*. Sage.

[46] Ministério da Saúde (MISAU). (2021).Impacto da pandemia da COVID-19 nos serviços de saúde materna e infantil em Moçambique: Relatório técnico. Maputo: Direcção Nacional de Saúde Pública.

[47] Montgomery, D. C., Peck, E. A., and Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.

[48] Novaes, E. S., Melo, E. C., Ferracioli, P. L. R. V., Oliveira, R. R. de, & Mathias, T. A. de F. (2018). Risco gestacional e fatores associados em mulheres atendidas pela rede pública de saúde. *Ciência, Cuidado e Saúde*, *17*(3), e45232. https://doi.org/10.4025/ciencuidsaude.v17i3.45232

[49] Oliveira, A. A. de, Almeida, M. F. de, Silva, Z. P. da, Assunção, P. L. de, Silva, A. M. R., Santos, H. G. dos, Alencar, G. P. (2019). Fatores associados ao nascimento pré-termo: da regressão logística à modelação com equações estruturais. *Cadernos de Saúde Pública*, *35*(1), e00211917. https://doi.org/10.1590/0102-311X00211917

[50] Organização das Nações Unidas. (2015). *Transformando nosso mundo: a Agenda 2030 para o Desenvolvimento Sustentável*.

[51] Qi, S., Zheng, J., Lu, M., Chen, A., Chen, Y., & Fu, X. (2024). Building a Aprendizado de Máquina-based risk prediction model for second-trimester miscarriage. *BMC Pregnancy and Childbirth*, *24*, 738. https://doi.org/10.1186/s12884-024-06942-w

[52] Quinlan, J. R. (1993). *C4.5: Programs for Aprendizado de Máquina*. Morgan Kaufmann.

[53] Rajkomar, A., Dean, J., & Kohane, I. (2019). Aprendizado de Máquina in medicine. *New England Journal of Medicine*, *380*(14), 1347–1358. https://doi.org/10.1056/NEJMra1814259

[54] Regassa, L. D., Tola, A., Daraje, G., & Dheresa, M. (2022). Trends and determinants of pregnancy loss in eastern Ethiopia from 2008 to 2019: Analysis of health and demographic surveillance data. *BMC Pregnancy and Childbirth*, *22*, 538. https://doi.org/10.1186/s12884-022-04994-4

[55] Rokach, L., & Maimon, O. (2008). *Data mining with Árvore de decisãos: Theory and applications*. World Scientific.

[56] Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, *10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

[57] Santos, L. I., Camargos, M. O., D'Angelo, M. F. S. V., & Mendes, J. B. (2021). Árvore de decisão and artificial immune systems for stroke prediction in imbalanced data. *Expert Systems with Applications*, *176*, 116221. https://doi.org/10.1016/j.eswa.2021.116221

[58] Silva, M. A. da. (2021). Avaliação de modelos de aprendizado de máquina para classificação de gestantes e predição de gravidez de risco usando o histórico de consultas médicas (Dissertação de Mestrado). Universidade Federal de Juiz de Fora.

[59] Sundermann, A. C., Velez Edwards, D. R., Bray, M. J., Jones, S. H., Latham, S. M., & Hartmann, K. E. (2017). Leiomyomas in pregnancy and spontaneous abortion: A systematic review and meta-analysis. *Obstetrics & Gynecology*, *130*(5), 1065–1072. https://doi.org/10.1097/AOG.0000000000002313

[60] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288.

[61] Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.

[62] UNICEF. (2023). *Situação da infância no mundo*. Fundo das Nações Unidas para a Infância. https://www.unicef.org/reports/state-of-worlds-children-2023

[63] Vapnik, V. N. (1995).*The nature of statistical learning theory*.Springer

[64] World Health Organization. (2023). Stillbirths and neonatal deaths. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/stillbirths

[65] Yehuala, T. Z., Mengesha, S. B., & Baykemagn, N. D. (2025). Predicting pregnancy loss and its determinants among reproductive-aged women using supervised Aprendizado de Máquina algorithms in Sub-Saharan Africa. *Frontiers in Global Women's Health*, *6*, 1456238. https://doi.org/10.3389/fgwh.2025.1456238

[66] Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*(1), 32–35. https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3