UNIVERSIDADE
E D U A R D O
MONDLANE

FACULDADE DE CIÊNCIAS
Departamento de Matemática e Informática

Trabalho de Licenciatura em
Estatística

Análise dos Factores Associados à Não Adesão às Consultas Pré-natais: Um Estudo de Caso no Hospital Geral José Macamo

Autora: Ivone Pedro Ussivane

Maputo, Junho de 2025

UNIVERSIDADE
E D U A R D O
MONDLANE

FACULDADE DE CIÊNCIAS
Departamento de Matemática e Informática

Trabalho de Licenciatura em
Estatística

Análise dos Factores Associados a Não Adesão às Consultas Pré-natais: Um Estudo de Caso no Hospital Geral José Macamo

Autora: Ivone Pedro Ussivane

Supervisor: Prof. Doutor Cachimo Assane

Maputo, Junho 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 Licenciatura em Estatística, na Faculdade de Ciências da Universidade Eduardo Mondlane.

Maputo, Junho de 2025

_____

(**Ivone Pedro Ussivane**)

Dedicatória

Dedico este trabalho à minha mãe, Maria Macamo, mulher de força admirável, exemplo de coragem e resiliência. Pelo amor incansável, apoio incondicional e pelos inúmeros sacrifícios feitos ao longo da minha vida, esta conquista é, antes de tudo, tua.

Agradecimentos

Agradeço, em primeiro lugar, a Deus, fonte de toda força, sabedoria e esperança, por me ter sustentado ao longo deste percurso.

Ao meu supervisor, Prof. Doutor Cachimo Assane, pela paciência, dedicação e disponibilidade em estar sempre pronto a orientar-me, e pelas valiosas contribuições para a realização deste trabalho.

À minha família, em especial aos meus pais, Pedro Francisco Ussivane e Maria das Dores Macamo, e aos meus irmãos Francisco Pedro Ussivane, Xavier Pedro Ussivane e Arlência Pedro Ussivane, pelo amor incondicional, apoio constante e incentivo diário para a concretização dos meus sonhos.

Aos docentes do Departamento de Matemática e Informática, em particular aos docentes do curso de Estatística, pelo conhecimento transmitido, pela dedicação, rigor e disponibilidade demonstrados ao longo da formação.

Um agradecimento muito especial à Esperança António Munlela, que mais do que uma colega, foi um verdadeiro pilar de suporte e incentivo em todos os momentos desta jornada. Entre partilhas, dúvidas, noites longas e superações, construímos um laço de companheirismo e força que levarei para sempre.

Aos meus colegas do curso, Arlindo Massingue, Maida Tajú, Yura Francisco e Hérica Dinóvia, que caminharam comigo nesta jornada, o meu muito obrigada pelo apoio e pela amizade que construímos.

A todas as pessoas que, directa ou indirectamente, contribuíram para a realização deste trabalho, o meu sincero agradecimento.

Resumo

A não adesão às consultas pré-natais representa um relevante desafio de saúde pública, evidenciando dificuldades persistentes na garantia de um acompanhamento gestacional adequado e contínuo. Este estudo teve como objectivo identificar os factores associados à não adesão às consultas pré-natais no Hospital Geral José Macamo e desenvolver um modelo preditivo para estimar a probabilidade de não adesão. Trata-se de um estudo quantitativo, de natureza transversal, baseado numa amostra de 1026 gestantes. Inicialmente, foi realizada uma análise descritiva e aplicou-se o teste do qui-quadrado de Pearson para seleccionar as variáveis independentes com associação estatisticamente significativa à variável dependente. O modelo preditivo foi construído com recurso à regressão logística binária, utilizando o método de selecção stepwise com base no critério de informação de Akaike (AIC). Para lidar com o desbalanceamento entre as classes, aplicou-se a técnica SMOTE na base de treino (80% da amostra), e a validação do desempenho foi efectuada através de validação cruzada estratificada (10-fold).Os principais factores associados à não adesão foram o baixo nível de escolaridade, a multiparidade, a ausência de parceiro, a idade materna jovem e o início tardio das consultas. O modelo apresentou desempenho satisfatório em validação cruzada, com sensibilidade de 88.8%, especificidade de 43.1%, VPP de 61.6% e VPN de 79.2%. A acurácia global foi de 66.1% e a área sob a curva ROC (AUC) foi de 76.6%, demonstrando boa capacidade discriminativa para identificar gestantes com maior risco de não adesão ao acompanhamento pré-natal.

Palavras-chave: Não adesão; Consultas pré-natais; Regressão logística; Modelo preditivo; Validação cruzada.

Abstract

Non-adherence to antenatal care (ANC) appointments represents a significant public health challenge, highlighting persistent difficulties in ensuring adequate and continuous pregnancy monitoring. This study aimed to identify the factors associated with non-adherence to ANC at the José Macamo General Hospital and to develop a predictive model to estimate the likelihood of non-adherence. A quantitative, cross-sectional study was conducted with a sample of 1,026 pregnant women. Initially, a descriptive analysis was performed, followed by the chi-square test of independence to identify the independent variables significantly associated with the outcome. The predictive model was developed using binary logistic regression with stepwise selection based on the Akaike Information Criterion (AIC). To address class imbalance, the SMOTE technique was applied to the training set (80% of the sample), and model performance was evaluated using stratified 10-fold cross-validation. The main predictors of non-adherence were low educational level, multiparity, absence of a partner, younger maternal age, and late initiation of ANC. The model demonstrated satisfactory performance, with a sensitivity of 88.8%, specificity of 43.1%, positive predictive value (PPV) of 61.6%, and negative predictive value (NPV) of 79.2%. Overall accuracy reached 66.1%, and the area under the ROC curve (AUC) was 76.6%, reflecting good discriminatory ability to identify pregnant women at higher risk of non-adherence to ANC.

Keywords: Non-adherence; Antenatal care; Logistic regression; Predictive model; Cross-validation.

Lista de Abreviaturas

AIC	Critério de Informação de Akaike (Akaike Information Criterion)
AUC	Área sob a Curva ROC (Area Under the ROC Curve)
ANC	Cuidados Pré-Natais (Antenatal Care)
BIC	Critério de Informação Bayesiano (Bayesian Information Criterion)
HIV	Vírus da Imunodeficiência Humana (Human Immunodeficiency Virus)
IC	Intervalo de Confiança
INE	Instituto Nacional de Estatística
MISAU	Ministério da Saúde
MLG	Modelos Lineares Generalizados
OMS	Organização Mundial da Saúde
OR	Razão de Chances (Odds Ratio)
ROC	Receiver Operating Characteristic
SIDA	Síndrome da Imunodeficiência Adquirida (AIDS)
SMOTE	Synthetic Minority Over-sampling Technique
UEM	Universidade Eduardo Mondlane
VIF	Fator de Inflação da Variância (Variance Inflation Factor)
VPP	Valor Preditivo Positivo
VPN	Valor Preditivo Negativo

Lista de Figuras

Lista de Tabelas

Capítulo 1

INTRODUÇÃO

1.1 Contextualização

Segundo a Organização Mundial da Saúde (OMS, 2018), a consulta pré-natal consiste num conjunto de cuidados prestados por profissional qualificado à mulher durante o percurso de sua gravidez, com a finalidade de promover a saúde e o bem-estar da mãe e do bebê, além de rastrear, diagnosticar e prevenir doenças.

A realização do pré-natal representa papel fundamental em termos de prevenção e detecção precoce de patologias tanto maternas como fetais, permitindo um desenvolvimento saudável do bebé e reduzindo riscos para a gestante. Além disso, a consulta pré-natal auxilia a mulher a reconhecer os sinais do parto e outros sinais gerais de perigo, de modo que ela procure os cuidados de saúde em tempo oportuno, o que pode ter efeito importante na prevenção de morbimortalidade materna e infantil (OMS, 2018).

O Ministério da Saúde (MISAU, 2018), recomenda um mínimo de 4 consultas durante a gestação, sendo a idade de início de pré-natal considerada apropriada se ocorrer até a 12ª semana de gestação. Segundo Silveira et al (2020), uma das formas de minimizar a mortalidade materno-infantil pode ser uma adequação da assistência pré-natal, através da realização de um número adequado de consultas, exames e procedimentos técnicos conforme estabelecidos pelo Ministério da Saúde.

Diversos estudos têm sido realizados utilizando técnicas estatísticas com o objetivo de identificar os factores associados a adesão as consultas pré-natais. Rosa (2013) usou a regressão logística condicional para investigar os factores associados a não realização do pré-natal das mulheres residentes no município de Pelotas no Brasil, em que factores como menor escolaridade, ser mulher solteira e multípara foram apontados como principais para não realização do pré-natal. Outro exemplo é o estudo feito por Hass (2013) em que usaram a regressão logística multivariada para avaliar a adequabilidade da assistência pré-natal, em que esta foi considerada adequada para apenas 2.1% da amostra e mulheres com companheiro e maior número de filhos apresentaram maior número de consultas realizadas.

1.2 Definição do problema

O cuidado pré-natal abrange um conjunto de acções que buscam promover a saúde materna e fetal, rastrear situações de risco e tratar intercorrências o mais precocemente possível. Este cuidado proporciona efeitos positivos na gestação, tanto clínicos quanto psicológicos com impacto importante sobre a morbimortalidade materno-infantil.

Em Moçambique, apesar de uma redução significativa nos últimos anos, a mortalidade materna, a mortalidade neonatal e as taxas de nados mortos continuam elevados. No país, em cada 100 000 nascimentos morrem cerca de 451 mulheres devido a causas maternas e factores como a falta de uso de serviços pré-natais, de partos institucionais e serviços pós-natais (Instituto Nacional de Saúde, 2022).

De acordo com o Ministério da Saúde e Instituto Nacional de Estatística (MISAU & INE, 2023), verificou-se uma cobertura de cuidados pré-natais de 87% de mulheres grávidas que realizaram pelo menos uma consulta pré-natal em Moçambique. Porém, em relação ao número de consultas realizadas, apenas 49% das mulheres realizaram, pelo menos, quatro consultas durante a gravidez mais recente, incluindo 2% das mulheres que tiveram oito ou mais consultas pré-natais.

Apesar das melhorias na cobertura e da implementação de diversos programas por parte do Ministério da Saúde, promovendo o seguimento da gravidez e a consciencialização sobre a sua importância, uma parte da população-alvo continua a não aderir a estes serviços. Diante das considerações expostas, notou-se a preocupação em procurar entender quais factores estão associados à não adesão às consultas pré-natais?

1.3 Objectivos

1.3.1 Objectivo Geral

Analisar os factores associados à não adesão às consultas pré-natais.

1.3.2 Objectivos Especifícos

- Descrever o perfil das gestantes do Hospital Geral José Macamo;
- Construir um modelo de regressão logística para avaliar a relação entre a não adesão às consultas pré-natais e o perfil das gestantes;
- Utilizar o modelo estimado prever a probabilidade de uma gestante não aderir as consultas pré-natais.

1.4 Justificação

A adesão às consultas pré-natais constitui um dos pilares fundamentais para assegurar uma gestação saudável, promovendo o bem-estar materno e fetal, além de contribuir para a prevenção e o manejo precoce de complicações. Em Moçambique, apesar dos progressos alcançados na ampliação da cobertura dos cuidados pré-natais, continuam a verificar-se lacunas significativas quanto à regularidade e à continuidade da sua utilização por parte das gestantes.

A literatura nacional tende a concentrar-se na taxa de cobertura do pré-natal, com menos ênfase na análise aprofundada dos fatores determinantes da não adesão, sobretudo em contextos locais específicos, como hospitais distritais e urbanos. Esta escassez de evidências limita a formulação de intervenções direcionadas e culturalmente sensíveis.

Assim, torna-se pertinente desenvolver estudos que não apenas quantifiquem a não adesão, mas que também investiguem os seus determinantes sob uma abordagem analítica, integrando variáveis sociodemográficas, obstétricas e contextuais. Com base nisso, o presente trabalho visa contribuir para a síntese do conhecimento disponível sobre o tema, permitindo um melhor esclarecimento sobre os padrões de adesão às consultas pré-natais e os fatores que os condicionam, fornecendo subsídios valiosos para a elaboração de políticas públicas mais eficazes e adaptadas à realidade local.

1.5 Estrutura do trabalho

Este trabalho está organizado em cinco capítulos. O primeiro capítulo apresenta a introdução, na qual são descritos o problema de pesquisa, os objectivos do estudo e a sua relevância. O segundo capítulo aborda a revisão da literatura, explorando os principais conceitos relacionados à adesão às consultas pré-natais, bem como estudos anteriores sobre o tema. No terceiro capítulo, é descrita a metodologia utilizada para a realização do estudo, incluindo o tipo de pesquisa, a população-alvo, a amostra e os procedimentos de coleta e análise dos dados. O quarto capítulo apresenta e discute os resultados obtidos com base na análise dos dados. Por fim, o quinto capítulo traz as conclusões do estudo, destacando as principais descobertas, limitações e sugestões para futuras pesquisas.

Capítulo 2

REVISÃO DA LITERATURA

2.1 Cuidados pré-natais em Moçambique

De acordo com o Ministério da saúde e Instituto Nacional de Saúde (MISAU & INS, 2021), a saúde materno-infantil constitui uma prioridade na agenda governamental, dada a elevada taxa de mortalidade materna e infantil no país. Moçambique figura entre os países com maiores índices de morbimortalidade em mulheres grávidas e crianças, o que reforça a necessidade de intervenções eficazes e sustentáveis. Nos últimos anos, registaram-se progressos significativos nas ações voltadas para a saúde da mulher e da criança, refletidos em indicadores como o aumento do número de mulheres que realizam quatro ou mais consultas pré-natais, a maior proporção de partos assistidos em unidades de saúde, a ampliação da cobertura vacinal e a intensificação das estratégias de prevenção e tratamento do HIV/SIDA em contexto pré-natal. No entanto, apesar desses avanços, muitas mulheres que têm acesso aos serviços de saúde ainda não os utilizam conforme as recomendações do Ministério da Saúde, o que compromete a detecção precoce e o manejo de possíveis complicações gestacionais.

Segundo as Normas de Atendimento Pré-Natal do MISAU e INS (2021), uma mulher é considerada assistida quando realiza pelo menos uma consulta durante a gestação, e completamente assistida quando realiza quatro ou mais consultas. Essa padronização está alinhada com as diretrizes da Organização Mundial da Saúde que, em 2016, passou a recomendar um mínimo de oito contatos com os serviços de saúde ao longo da gravidez, com o objetivo de melhorar os desfechos maternos e perinatais.

Apesar dessas recomendações, dados do MISAU e INE (2023) de revelam que uma parcela significativa das gestantes inicia o acompanhamento tardiamente e não atinge o número mínimo de consultas recomendado. Esse cenário está frequentemente associado a fatores como a baixa escolaridade, pobreza, distância até as unidades de saúde, normas socioculturais, qualidade percebida dos serviços e experiências prévias negativas com o sistema de saúde.

2.2 Definição e vantagens dos cuidados pré-natais

De acordo com Benova (2018), a assistência pré-natal é definida como um conjunto de cuidados sistemáticos prestados à gestante com o objetivo de promover a saúde da mãe e do feto, prevenir complicações e identificar precocemente condições que possam comprometer o curso da gestação. Esse acompanhamento inclui não apenas a vigilância clínica, mas também ações educativas, suporte emocional e orientações sobre o parto e o cuidado neonatal.

De acordo com o MISAU e INE (2023), o acompanhamento pré-natal tem como objetivo principal garantir o seguimento contínuo da gestante ao longo do período gestacional, com o intuito de prevenir e minimizar os riscos de complicações que possam levar à morbilidade e mortalidade tanto materna quanto infantil. Além disso, esse cuidado sistemático contribui significativamente para a diminuição de nascimentos prematuros e para a redução da mortalidade perinatal.

Silveira et al. (2020), complementam que a realização das consultas de pré-natal, quando associada a ações voltadas à qualificação da assistência prestada à mulher e à criança, representa um passo essencial para o fortalecimento das políticas públicas de saúde. Essa integração tem se mostrado eficaz na promoção de melhorias significativas nos indicadores de saúde, especialmente na redução das taxas de mortalidade materna e infantil.

Vantagens dos cuidados pré-natais

De acordo com o Ministério da Saúde (2018), os cuidados pré-natais de qualidade apresentam inúmeras vantagens, dentre elas:

- **Redução da mortalidade materno-infantil** – diminui a ocorrência de natimortos, mortes neonatais, complicações como pré-eclâmpsia e hemorragias graves.
- **Identificação precoce de doenças** – Permite o diagnóstico e tratamento de condições pré-existentes ou gestacionais, como hipertensão, diabetes, sífilis e anemias.
- **Monitoramento da gestação e do feto** – Avalia a saúde da mulher, a vitalidade fetal, a posição da placenta e detecta possíveis malformações, inclusive com possibilidade de intervenção intrauterina.
- **Promoção da saúde e educação** – Oferece orientações sobre alimentação, cuidados com o recém-nascido, e planejamento reprodutivo.
- **Prevenção de complicações** – Contribui para a diminuição de casos de prematuridade, baixo peso ao nascer e hipertensão gestacional.

2.3 Factores associados a não adesão aos cuidados pré-natais

De acordo com Barbieri et al. (2020), a não adesão aos cuidados pré-natais está associada a uma série de impactos negativos para a saúde materno-infantil. Entre as consequências relatadas, destacam-se o aumento de partos prematuros e o surgimento de recém-nascidos com baixo peso. Esse cenário é agravado por fatores como baixa escolaridade, gestações na adolescência e gravidez indesejada, que frequentemente ocorrem paralelamente à escassez de orientações precisas e à dificuldade de acesso aos serviços de saúde.

Muleva et al. (2021) ressaltaram que a ausência de um acompanhamento pré-natal adequado pode favorecer o desenvolvimento silencioso de condições crônicas, como anemia e hipertensão, as quais muitas vezes só são detectadas no momento do parto. Além disso, a falta de apoio familiar e de informações sobre os riscos decorrentes da não adesão dificulta o preparo emocional das gestantes, especialmente das mais jovens, levando ao abandono escolar e a uma preparação insuficiente para os desafios da maternidade. Essa combinação de fatores evidencia a complexidade do problema e a necessidade de intervenções integradas para melhorar a adesão aos cuidados pré-natais.

2.3.1 Factores Demográficos

Os factores demográficos relacionados a não adesão as consultas pré-natais englobam aspectos como a idade da gestante, estado civil e área de residência.

De acordo com Muleva et al. (2021), gestantes adolescentes tendem a iniciar o pré-natal mais tardiamente e a realizar um número menor de consultas. Isso pode ser atribuído à falta de conhecimento sobre a importância do pré-natal, medo de estigmatização e ausência de apoio familiar. Por outro lado, mulheres com idade avançada podem subestimar a necessidade do acompanhamento pré-natal, especialmente se já tiverem passado por gestações anteriores, o que pode levar à negligência das consultas recomendadas.

O estado civil também influencia a adesão ao pré-natal. Boene et al. (2019) constataram que mulheres solteiras, separadas ou viúvas podem enfrentar desafios adicionais, como menor apoio emocional e financeiro, o que dificulta a continuidade das consultas. Em contrapartida, gestantes casadas ou em união estável geralmente apresentam maior probabilidade de cumprir o calendário recomendado de consultas pré-natais, possivelmente devido ao suporte do parceiro. A ausência de um companheiro pode afetar negativamente a motivação e a capacidade da gestante de buscar os cuidados necessários.

Segundo Barbieri et al. (2020), mulheres que vivem em áreas rurais enfrentam obstáculos significativos, como a distância até as unidades de saúde, falta de transporte adequado e escassez de profissionais qualificados. Essas barreiras logísticas contribuem para o início tardio do pré-natal e para a realização de um número insuficiente de consultas. Em Moçambique, por exemplo, estudos

mostram que a cobertura do pré-natal é menor nas zonas rurais em comparação com as urbanas, refletindo desigualdades no acesso aos serviços de saúde.

2.3.2 Factores Socioeconómicos

Segundo Benova et al. (2018), os fatores socioeconômicos constituem uma barreira significativa para o acesso e a continuidade dos cuidados pré-natais. Mulheres em situação de vulnerabilidade econômica, especialmente aquelas com baixos níveis de escolaridade, apresentam menor adesão aos serviços de saúde durante a gestação.

O MISAU e INE (2023), relataram que mulheres com ensino secundário ou superior tendem a ter uma cobertura pré-natal substancialmente maior do que aquelas que não completaram o ensino básico, sugerindo que a educação desempenha um papel crucial na tomada de decisão e na compreensão da importância desse acompanhamento.

De acordo com Viellas et al. (2014), mulheres que vivem em situações de pobreza enfrentam desafios adicionais relacionados ao acesso aos serviços de saúde, mesmo quando estes são oferecidos de forma gratuita. Esses desafios incluem custos indiretos, como transporte, alimentação, tempo de espera nas unidades e a necessidade de se ausentar do trabalho ou de suas responsabilidades familiares.

A ocupação da gestante também influencia sua adesão ao pré-natal. Segundo Silveira et al (2020), mulheres desempregadas ou com empregos informais e precários tendem a enfrentar maiores obstáculos, como jornadas exaustivas e falta de flexibilidade de horários. Já aquelas que trabalham em contextos formais, com direitos garantidos, têm mais facilidade para comparecer às consultas. Muleva et al. (2021), apontam que o desemprego feminino, que atinge altos índices em várias regiões de Moçambique, tem sido apontado como um fator que contribui para a dependência econômica dos parceiros, o que pode limitar a autonomia da mulher na tomada de decisão sobre sua saúde reprodutiva.

2.3.3 Factores obstétricos e de saúde reprodutiva

A idade gestacional no início do pré-natal é um indicador-chave de adesão adequada. De acordo com as diretrizes do Ministério da Saúde de Moçambique (2018), o ideal é que a primeira consulta ocorra até a 12ª semana de gestação. No entanto, Segundo Silveira et al. (2020), muitas mulheres iniciam tardiamente (após o segundo trimestre) perdendo oportunidades cruciais de diagnóstico precoce, imunizações e orientações fundamentais. Esse atraso está geralmente relacionado à falta de informação, dificuldade de acesso, ou mesmo ao desconhecimento da própria gestação nas fases iniciais.

A paridade influencia a percepção da necessidade do pré-natal. Segundo Viellas et al. (2014),

mulheres primíparas, por estarem vivenciando a gravidez pela primeira vez, tendem a buscar mais informações e a aderir melhor às consultas pré-natais. Em contrapartida, multíparas podem considerar desnecessário o acompanhamento rigoroso, baseando-se em experiências anteriores, o que pode resultar em menor frequência às consultas e, consequentemente, em riscos não identificados para a saúde materna e fetal.

De acordo com Domingues et al. (2012), gestações não planeadas ou indesejadas estão fortemente associadas à baixa procura por cuidados, especialmente no primeiro trimestre, que é o período mais crítico para o início adequado do pré-natal. Nessas situações, a mulher pode negar ou ocultar a gravidez, adiar o início do acompanhamento e, em alguns casos, não buscar assistência nenhuma até o parto. Pesquisa de Silva et al. (2021) mostrou que mulheres com gravidez não planejada apresentaram até 40% menos chances de comparecer às consultas recomendadas.

2.4 Modelos Lineares Generalizados

Os Modelos Lineares Generalizados (MLG), introduzidos por Nelder e Wedderburn em 1972, representam uma extensão dos modelos lineares clássicos que permite a modelação de variáveis dependentes que não seguem necessariamente uma distribuição normal. Essa abordagem proporciona uma estrutura estatística mais flexível, adequada à análise de diferentes tipos de dados, como variáveis binárias, de contagem ou contínuas com distribuição assimétrica.

Segundo Nelder e Wedderburn (1972), os MLGs permitem que a variável resposta siga qualquer distribuição pertencente à família exponencial, desde que expressa na forma canônica. Além disso, oferecem maior flexibilidade na especificação da função que relaciona a média da variável resposta ao preditor linear, permitindo ajustar melhor os modelos às características dos dados empíricos. Do ponto de vista teórico e conceitual, os MLGs representam uma síntese abrangente das principais abordagens de modelação estatística desenvolvidas até então.

São considerados casos particulares dos MLGs os seguintes modelos:

- Modelo de regressão linear clássico;
- Modelos de análise de variância e covariância;
- Modelo de regressão logística;
- Modelo de regressão de Poisson;
- Modelos log-lineares para tabelas de contingência multidimensionais, etc.

Família exponencial

Segundo Alvarenga (2015), os Modelos Lineares Generalizados pressupõem que a variável resposta segue uma distribuição pertencente à família exponencial, ou seja, assume-se que a função

densidade de probabilidade (ou a função massa de probabilidade) da variável resposta pode ser expressa na chamada forma canônica da família exponencial, dada por:

$$f(y \mid \theta; \varphi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\varphi)} + c(y, \varphi)\right\} \tag{2.1}$$

Onde:

- Y é a variável de interesse;
- θ é o parâmetro de localização;
- φ é o parâmetro de dispersão;
- $a(\cdot)$, $b(\cdot)$ e $c(\cdot)$ são funções reais conhecidas ou também denominadas funções específicas.

Características do Modelo Linear Generalizado

Segundo Figueira (2006), um Modelo Linear Generalizado (MLG) é estruturado com base em três componentes fundamentais:

Componente Aleatória

Esta parte do modelo descreve a distribuição de probabilidade da variável dependente. Assume-se que, dado o vetor de covariáveis x_i, as variáveis aleatórias Y_i são condicionalmente independentes e seguem uma distribuição pertencente à família exponencial. Consequentemente, a média da variável resposta é expressa por:

$$f(Y_i; x_i) = \mu_i = b(\theta), \quad i = 1, 2, \ldots, n \tag{2.2}$$

Componente Sistemática

A componente sistemática corresponde à combinação linear dos preditores. Essa estrutura define como as variáveis explicativas influenciam a resposta, mesmo quando a relação entre elas e a média da resposta não é diretamente linear. Define-se o preditor linear η_i como:

$$\eta_i = \sum_{q=1}^{r} X_{iq}\beta_q = X_i^T\beta \tag{2.3}$$

Onde:

- $X = (x_1, \ldots, x_n)$ é a matriz de covariáveis;
- $\beta = (\beta_1, \ldots, \beta_r)$ é o vetor de parâmetros desconhecidos;
- $\eta = (\eta_1, \ldots, \eta_n)$ representa o vetor de preditores lineares.

Função de Ligação

A função de ligação estabelece uma relação matemática entre o valor esperado da variável resposta μ_i e o preditor linear η_i. Nesta equação, $g(\cdot)$ é uma função real, estritamente monótona e diferenciável, que transforma a média da resposta em uma escala compatível com o preditor linear. Formalmente, define-se:

$$g(\mu_i) = \eta_i \tag{2.4}$$

A interação dessas três componentes, determina o tipo de modelo ajustado dentro da filosofia dos MLGs. Por exemplo, quando a distribuição da resposta é normal, a função de ligação é a identidade e as covariáveis são contínuas, tem-se o modelo clássico de regressão linear. Dependendo da especificação de cada componente, diferentes tipos de MLGs podem ser utilizados, como mostra a tabela a seguir (Fávero & Belfiore, 2017).

Tabela 2.1: Variações dos modelos lineares generalizados

Tipo de Modelo	Distribuição	Função de Ligação	Tipo de Variável Resposta
Regressão Linear	Normal	Identidade	Contínua (simétrica)
Regressão Logística	Binomial	Logit	Binária (0 ou 1)
Regressão de Poisson	Poisson	Logarítmica	Contagem (inteiros ≥ 0)
Regressão Binomial Neg.	Binomial Negativa	Logarítmica	Contagem com sobredispersão
Regressão Gama	Gama	Inversa, Log, Identidade	Contínua (assimétrica, positiva)
Regressão Inversa Gauss	Inversa Gaussiana	Inversa quadrática	Contínua (positiva)

2.5 Modelo de Regressão Logística

Segundo Agresti e Finlay (2009), o modelo de regressão logística representa um dos casos mais relevantes dos Modelos Lineares Generalizados (MLG), especialmente adequado quando o objetivo é modelar uma variável resposta categórica dicotômica, ou seja, com dois resultados possíveis

De acordo com Hosmer et al. (2013), a regressão logística se fundamenta no uso da função logística para assegurar que as probabilidades estimadas permaneçam dentro do intervalo [0,1], sendo apropriada para situações em que a variável resposta assume dois possíveis valores, como presença/ausência, sucesso/fracasso ou sim/não. A equação básica do modelo é:

$$\pi(\mathbf{x}) = \frac{e^{\eta}}{1+e^{\eta}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\ldots+\beta_k x_k)}} \tag{2.5}$$

onde:

- $\pi(\mathbf{x})$: probabilidade de ocorrência do evento de interesse
- $\eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$: preditor linear
- β_j: coeficientes do modelo ($j = 0, \ldots, k$)

De acordo com Baptista (2015), em qualquer problema, a quantidade a ser modelada é o valor médio da variável resposta, dados os valores das variáveis independentes. A ocorrência dos eventos sucesso e fracasso dá-se com probabilidades $\pi(\mathbf{x}) = P(Y = 1 \mid X = \mathbf{x})$ e $1 - \pi(\mathbf{x}) = P(Y = 0 \mid X = \mathbf{x})$, respectivamente. Como Y só pode assumir os valores 0 e 1, a probabilidade será igual a $\mathbb{E}(Y|X = \mathbf{x})$, que é a média condicional de Y dado $\mathbf{x}$.

Segundo Belfiore (2015), na regressão logística, a variável dependente Y é geralmente binária, ou seja, assume apenas dois valores possíveis, representando sucesso ou fracasso. Nessa configuração, considera-se que Y segue uma distribuição de Bernoulli, caracterizada por uma única tentativa e por uma probabilidade de sucesso desconhecida p, tal que $0 \leq p \leq 1$. Lembrando que a distribuição de Bernoulli é um caso particular da distribuição binomial, correspondente à situação em que o número de ensaios n é igual a 1, ou seja, trata-se da realização de um único experimento.

Vantagens da regressão logística

De acordo com Hosmer et al. (2013), o modelo de regressão logística oferece diversas vantagens práticas e teóricas. Entre as principais, destacam-se:

1. **Interpretação direta das probabilidades e odds**: ao estimar diretamente a probabilidade de ocorrência do evento e sua razão de chances (*odds ratio*), facilita-se a comunicação dos resultados em termos intuitivos para pesquisadores e tomadores de decisão.

2. **Robustez a diferentes distribuições dos preditores**: não exige que as variáveis independentes sigam distribuição normal, o que amplia o leque de situações em que o modelo pode ser aplicado sem fortes suposições.

3. **Flexibilidade na especificação funcional**: permite incluir termos de interação, polinômios e variáveis categóricas sem grandes dificuldades, adaptando-se a relações não-lineares entre preditores e a resposta.

4. **Estimação eficiente em amostras moderadas**: por meio da máxima verossimilhança, apresenta boas propriedades estatísticas (consistência e eficiência) mesmo quando o tamanho da amostra não é muito grande, desde que o número de eventos seja adequado.

2.5.1 Transformação Logit

De acordo com Agresti (2015), a regressão logística lida com probabilidades, as quais devem estar contidas no intervalo $[0, 1]$. No entanto, modelar diretamente a probabilidade como uma função linear pode resultar em valores fora desse intervalo, violando a lógica probabilística. Para resolver esse problema, utiliza-se a transformação logit, que converte a probabilidade de ocorrência do evento em uma escala contínua e ilimitada.

Segundo Hosmer et al (2013), a função logit é definida como o logaritmo da razão entre a probabilidade de sucesso π e a probabilidade de insucesso $(1-\pi)$. Essa razão é conhecida como *odds* ou razão de chances, que representa quantas vezes é mais provável que o evento de interesse ocorra em relação a ele não ocorrer. Assim, a transformação logit é dada por:

$$\text{logit}(\pi) = \ln\left(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{2.6}$$

Essa função transforma o intervalo $[0,1]$, onde estão as probabilidades, no conjunto dos números reais $\mathbb{R}$. A principal vantagem dessa transformação, conforme argumenta Agresti (2013), é que ela permite utilizar uma estrutura linear para modelar a relação entre a variável explicativa e a resposta, mantendo as propriedades desejadas do modelo probabilístico.

2.5.2 Regressão logística simples ou univariada

Segundo Figueira (2006) a regressão logística univariada é uma forma simples do modelo logístico, na qual se considera apenas uma variável explicativa para modelar a variável resposta binária.

De acordo com Hair et al. (2005), a regressão logística simples é uma técnica estatística utilizada para descrever a relação entre uma variável dependente dicotómica e uma única variável independente. A variável resposta Y assume dois valores possíveis, geralmente codificados como 1 para representar o evento de interesse e 0 para indicar o evento complementar. Esses dois resultados são mutuamente exclusivos, o que significa que a ocorrência de um impede a ocorrência do outro, e sua soma compreende todas as possibilidades de desfecho da variável resposta analisada.

A probabilidade de sucesso do modelo logístico simples é dada por:

$$\pi_i = \pi(x_i) = P(Y_i = 1 \mid X_i = x_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \tag{2.7}$$

E a probabilidade de fracasso é dada por:

$$1-\pi_i = 1-\pi(x_i) = P(Y_i = 0 \mid X_i = x_i) = \frac{1}{1+e^{\beta_0+\beta_1 x_i}} \tag{2.8}$$

Onde β_0 e β_1 sao os parametros do modelo, $P(Y_i = 1 \mid X_i = x_i)$ representa a probabilidade de ocorrência ou não do evento de interessse em razão dos valores assumidos pela variável explicativa X.

Quando $\beta_1 < 0$ a função $\pi(x)$, que representa a probabilidade do evento de interesse, é decrescente em relação à variável explicativa x. Por outro lado, quando $\beta_1 > 0$, essa função torna-se crescente. À medida que x tende ao infinito, $\pi(x)$ tende a zero no caso de $\beta_1 < 0$ e tende a 1 no caso $\beta_1 > 0$.

Já quando $\beta_1 = 0$, a probabilidade $\pi(x)$ permanece constante, indicando que a variável explicativa x não exerce influência sobre a variável resposta Y, ou seja, são independentes.

Estimação dos parâmetros do modelo de regressão logística univariada

Segundo Hosmer et al. (2013), os parâmetros β_0 e β_1 são estimados através do método da máxima verossimilhança (MLE - Maximum Likelihood Estimation). Este método consiste em encontrar os valores dos parâmetros que maximizam a probabilidade de observar os dados disponíveis.

A função de verossimilhança para uma amostra de n observações independentes é dada por:

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} \pi_i^{Y_i}(1 - \pi_i)^{1-Y_i} \tag{2.9}$$

Tomando o logaritmo da função de verossimilhança (log-verossimilhança), obtêm-se:

$$l(\beta_0, \beta_1) = \sum_{i=1}^{n} [Y_i \ln(\pi_i) + (1 - Y_i) \ln(1 - \pi_i)] \tag{2.10}$$

Substituindo π_i:

$$l(\beta_0, \beta_1) = \sum_{i=1}^{n} \left[Y_i(\beta_0 + \beta_1 x_i) - \ln(1 + e^{\beta_0 + \beta_1 x_i})\right] \tag{2.11}$$

Para maximizar $l(\beta_0, \beta_1)$ deriva-se em relação a β_0 e β_1, igualando-as a zero:

$$\frac{\partial l}{\partial \beta_0} = \sum_{i=1}^{n} (Y_i - \pi_i) = 0 \tag{2.12}$$

$$\frac{\partial l}{\partial \beta_1} = \sum_{i=1}^{n} x_i(Y_i - \pi_i) = 0 \tag{2.13}$$

Estas equações não têm solução analítica fechada, e por isso é necessário utilizar métodos numéricos, como o método de Newton-Raphson ou o método iterativo de Fisher scoring (Agresti, 2013).

Segundo Menard (2000), o método de Newton-Raphson é uma técnica iterativa que atualiza os parâmetros usando a seguinte fórmula:

$$\beta^{(t+1)} = \beta^{(t)} - \left[H\left(\beta^{(t)}\right)\right]^{-1} S\left(\beta^{(t)}\right) \tag{2.14}$$

Onde:

- $S(\beta)$ é o vetor *score* (gradiente da log-verossimilhança)
- $H(\beta)$ é a matriz Hessiana (matriz das segundas derivadas)

Este processo continua até que a diferença entre os valores dos parâmetros em iterações consecutivas seja suficientemente pequena (convergência).

2.5.3 Regressão logística múltipla

De acordo com Hosmer e Lemeshow (1989), a regressão logística múltipla representa uma extensão do modelo de regressão logística simples, que considera apenas uma variável preditora, permitindo agora a inclusão de múltiplos preditores, contínuos ou categóricos, para explicar simultaneamente a probabilidade de ocorrência de um evento binário.

Para formalizar essa generalização, considera-se um vetor de p variáveis explicativas, representado por $\mathbf{x}_i^T = \{x_{i1}, x_{i2}, \ldots, x_{ip}\}$ em que $\mathbf{x}_i$ é o vetor de covariáveis associado ao i-ésimo indivíduo da amostra. O modelo assume que a probabilidade π_i de ocorrência do evento de interesse está relacionada a uma combinação linear dos preditores, através da função logit, conforme:

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \tag{2.15}$$

Neste contexto, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ representa o vetor de parâmetros desconhecidos a serem estimados, onde cada β_j está associado à variável explicativa x_{ij}, com $j = 1, 2, \ldots, p$. Assim, a função logit permanece como elo entre a componente sistemática e a probabilidade de sucesso, mantendo a estrutura do modelo dentro do arcabouço dos Modelos Lineares Generalizados (McCullagh e Nelder, 1989).

Estimação dos parâmetros no modelo de regressão logística múltipla

Segundo Moura (2018), a estimação dos parâmetros do modelo é feita pelo método da máxima verossimilhança de modo que os estimadores maximizem o logaritmo da função de verossimilhança.

Segundo Hosmer e Lemeshow (2000), a função de verossimilhança de um modelo de regressão logística é construída a partir da distribuição binomial, dado que a variável resposta Y_i assume valores 0 ou 1. Assim, para uma amostra de n observações, a função de verossimilhança é dada por:

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

Onde:

$$\pi_i = P(Y_i = 1|x_i) = \frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}}$$

Tomando o logaritmo natural da função de verossimilhança, obtém-se a log-verossimilhança:

$$l(\beta) = \sum_{i=1}^{n} [Y_i \ln(\pi_i) + (1-Y_i)\ln(1-\pi_i)]$$

Como não existe uma solução analítica direta para maximizar essa função, utiliza-se um método numérico iterativo, como o método de Newton-Raphson ou o Fisher Scoring, baseados nas derivadas da log-verossimilhança.

A função score é o vetor gradiente da log-verossimilhança, ou seja, é formado pelas derivadas

parciais da função em relação a cada parâmetro β_j:

$$\frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} (Y_i - \pi_i) x_{ij}$$

A forma matricial do vetor score é:

$$U(\beta) = X^T (Y - \pi)$$

Onde:

- X é a matriz de design (com cada linha X_i^T)
- Y é o vetor das observações y_i
- π é o vetor das probabilidades π_i

A matriz de covariância dos coeficientes estimados é obtida a partir das derivadas parciais de segunda ordem do logaritmo da função de verossimilhança:

$$\frac{\partial^2 l(\beta)}{\partial \beta_j \partial \beta_k} = -\sum_{i=1}^{n} \pi_i (1 - \pi_i) x_{ij} x_{ik}$$

Em notação matricial, a Hessiana assume a forma:

$$H(\beta) = -X^T W X$$

Em que W é uma matriz diagonal com elementos $w_{ii} = \pi_i(1 - \pi_i)$, $i = 1, 2, 3, \ldots, n$ e X a matriz dos dados, e sua inversa $l^{-1}(\beta)$ é a matriz de variância e covariância dos estimadores de máxima verossimilhança dos parâmetros.

2.6 Aprendizagem Supervisionada em Modelos de Classificação

A aprendizagem supervisionada é uma das áreas centrais da ciência de dados e da inteligência artificial, baseada na ideia de construir modelos preditivos a partir de um conjunto de dados rotulados, isto é, cujos resultados (ou classes) já são conhecidos. Conforme referem James et al. (2021), o objetivo desta abordagem é aprender uma função que, a partir de variáveis explicativas, seja capaz de prever com precisão a classe ou valor de uma variável resposta.

Tipos de algoritmos supervisionados

Os algoritmos de aprendizagem supervisionada dividem-se geralmente em dois grupos: regressão (para variáveis resposta contínuas) e classificação (para variáveis categóricas). Dentre os algoritmos de classificação mais utilizados destacam-se:

- **Regressão logística** - Apesar da sua fundamentação estatística no âmbito dos Modelos Lineares Generalizados, a regressão logística é amplamente utilizada no domínio da aprendizagem supervisionada, destacando-se pela sua interpretabilidade e eficácia em problemas de classificação binária (Hastie et al., 2009).

- **Árvores de decisão** - Este método constrói uma estrutura hierárquica em forma de árvore, onde os nós representam decisões baseadas nos valores das variáveis explicativas. A árvore é gerada a partir da divisão sucessiva do conjunto de dados em subconjuntos homogéneos, maximizando a pureza de cada nó final (ou folha). A sua principal vantagem é a simplicidade e interpretabilidade, embora possa sofrer de sobreajustamento se não for devidamente controlada (James et al., 2021).

- **Floresta aleatória**- É um modelo de ensemble (combinação de múltiplos modelos), baseado em múltiplas árvores de decisão construídas a partir de subconjuntos aleatórios dos dados e das variáveis. O resultado final é obtido por votação (para classificação) ou média (para regressão). Segundo Breiman (2001), esta técnica oferece grande robustez a outliers e colinearidade, e tende a apresentar elevado desempenho preditivo, embora com perda de interpretabilidade face a uma única árvore.

- **Máquinas de Boosting por Gradiente (Gradient Boosting Machines — GBM)** constituem uma família de algoritmos de aprendizagem supervisionada baseados no princípio do boosting, em que modelos fracos são combinados de forma sequencial para formar um modelo forte. Entre as implementações mais conhecidas encontram-se o XGBoost e o LightGBM, amplamente utilizados pela sua eficiência e elevado desempenho em problemas de classificação e regressão (Friedman, 2001).

- **Máquinas de Vectores de Suporte (SVM)** - são algoritmos de aprendizagem supervisionada utilizados em problemas de classificação e regressão. O princípio fundamental consiste em encontrar o hiperplano que melhor separa as classes no espaço das variáveis explicativas, maximizando a margem entre os vectores de suporte. Trata-se de uma técnica robusta, especialmente eficaz em problemas de elevada dimensionalidade (Hastie et al., 2009).

Apesar de muitos destes modelos mais avançados oferecerem um poder preditivo elevado, apresentam frequentemente desvantagens em termos de interpretabilidade, o que pode limitar a sua aplicação em contextos clínicos, onde a explicação de resultados é essencial para a tomada de decisão (Kuhn & Johnson, 2013).

2.7 Técnicas de pré-processamento de dados

O pré-processamento de dados constitui uma etapa fundamental na aprendizagem supervisionada, na medida em que garante que os dados se encontram em condições adequadas para o treino dos modelos. Esta fase tem como principais objectivos: tratar inconsistências e dados ausentes,

codificar variáveis, normalizar escalas e, especialmente em contextos de classificação binária, lidar com a desproporção entre as classes (desbalanceamento).

2.7.1 O Problema do Desbalanceamento

Em muitos problemas de classificação no domínio da saúde, verifica-se que as classes não se encontram representadas de forma equitativa, o que pode comprometer a aprendizagem do modelo e conduzir a predições enviesadas. Este fenómeno, conhecido como desbalanceamento de classes, ocorre quando uma das categorias da variável resposta tem uma frequência substancialmente inferior à outra. Japkowicz e Stephen (2002) referem que os algoritmos de classificação tendem a favorecer a classe majoritária, o que pode reduzir a sensibilidade para detectar eventos menos comuns, mas frequentemente mais importantes do ponto de vista clínico.

Para amenizar os efeitos do desbalanceamento, podem ser adoptadas várias estratégias de balanceamento de classes:

- **Subamostragem da classe majoritária**: consiste em reduzir aleatoriamente o número de observações da classe mais representada, o que pode, no entanto, implicar perda de informação relevante.
- **Sobreamostragem da classe minoritária**: aumenta artificialmente o número de observações da classe menos representada, replicando casos existentes, oque pode levar ao sobreajustamento.
- **Técnica de Sobreamostragem da Classe Minoritária (SMOTE)**: proposto por Chawla et al. (2002), é um método de sobreamostragem mais avançado que gera exemplos sintéticos, combinando instâncias reais da classe minoritária com os seus vizinhos mais próximos. O SMOTE é amplamente utilizado em problemas de saúde por preservar a diversidade da classe minoritária e reduzir a variância do modelo (Fernández et al., 2018).

2.7.2 Outras técnicas de pré-processamento

Além do balanceamento de classes, existem outras operações de pré-processamento que são essenciais para garantir a qualidade dos dados e a eficácia dos modelos preditivos:

- **Codificação de variáveis categóricas**: Variáveis qualitativas devem ser convertidas em formato numérico antes de serem utilizadas nos modelos. A codificação mais comum é a codificação dummy, na qual cada categoria é representada por uma variável binária (0 ou 1). Alternativamente, pode ser usada a codificação one-hot. Segundo Kuhn e Johnson (2013), uma codificação adequada evita problemas de interpretação e garante que o modelo trata correctamente as variáveis categóricas.
- **Tratamento de valores ausentes**: Dados ausentes podem introduzir viés ou reduzir a eficácia dos modelos. As abordagens incluem a exclusão dos casos incompletos, imputação com a

média, moda ou mediana, e métodos mais avançados como imputação múltipla ou k-NN imputation. Segundo Little e Rubin (2019), a escolha do método depende da proporção e do padrão de ausência.

- **Identificação e tratamento de outliers**: Valores extremos podem distorcer os parâmetros estimados e reduzir o desempenho dos modelos. Métodos como o uso de gráficos boxplot, scores z ou distância de Mahalanobis ajudam a detectar outliers. Dependendo do contexto, estes podem ser removidos ou transformados (Batista et al., 2004).

- **Análise de colinearidade**: A presença de elevada correlação entre variáveis explicativas (multicolinearidade) pode comprometer a estabilidade dos coeficientes no modelo. É prática comum verificar o Índice de Inflação da Variância (VIF) antes de treinar o modelo, e remover ou combinar variáveis com VIF elevado (Gujarati, 2011).

2.8 Métodos de seleção de variáveis

De acordo com Menard (2000), a escolha do conjunto de variáveis explicativas que deve compor o modelo de regressão logística múltipla é uma etapa fundamental para garantir a parcimônia, a precisão dos coeficientes estimados e a capacidade preditiva do modelo.

Segundo Hosmer et al. (2013), destacam-se três abordagens amplamente utilizadas: o método forward, o método backward e o método stepwise.

- No **método forward** o processo inicia-se com um modelo vazio, ou seja, sem nenhuma variável explicativa. As variáveis são então adicionadas uma a uma, com base num critério estatístico, como o valor do teste de Wald ou a significância do p-valor. A cada passo, adiciona-se a variável que mais melhora o modelo, até que nenhuma outra variável satisfaça o critério de inclusão pré-estabelecido.

- O **método backward** funciona de forma inversa. Começa com um modelo completo, contendo todas as variáveis candidatas, e procede-se com a eliminação sequencial daquelas que não contribuem significativamente para o modelo, de acordo com um critério de exclusão. Este processo continua até que apenas permaneçam no modelo as variáveis estatisticamente relevantes.

- O **método stepwise** combina as abordagens anteriores. As variáveis são adicionadas como no método forward, mas a cada nova inclusão, verifica-se também se alguma das variáveis já incluídas deve ser excluída, permitindo assim um modelo mais ajustado e parcimonioso.

2.9 Critérios para seleção do modelo

Após a aplicação de um método de seleção de variáveis seja forward, backward ou stepwise, é fundamental dispor de um critério que permita comparar os modelos candidatos de forma objectiva.

2.9.1 Critério de Informação de Akaike (AIC)

Segundo Emiliano et al. (2009), o critério de Informação de Akaike, procura identificar o modelo que melhor equilibra a qualidade de ajuste e simplicidade estrutural. O AIC é definido da seguinte forma:

$$AIC = -2\ln(L(\hat{\beta})) + 2q \tag{2.16}$$

Onde:

- $\ln(L(\hat{\beta}))$ é o logaritmo natural da função de verossimilhança do modelo com os parâmetros estimados
- q representa o número de variáveis explicativas incluídas no modelo

Este critério penaliza a inclusão de variáveis desnecessárias, evitando que um modelo excessivamente complexo seja preferido apenas por apresentar melhor ajuste. Assim, o modelo com o menor valor de AIC é considerado o mais parcimonioso, ou seja, o que apresenta o melhor compromisso entre bom ajuste e simplicidade.
Segundo Emiliano et al. (2009), não é apropriado escolher modelos unicamente com base na qualidade do ajuste, pois isso tenderia a favorecer modelos com muitas variáveis, inclusive irrelevantes. Ao penalizar a complexidade, o AIC ajuda a evitar esse problema.

2.9.2 Critério de informação Bayesiano

O Critério de Informação Bayesiano (BIC), proposto por Schwarz (1978), é uma alternativa ao AIC com um enfoque mais conservador, particularmente útil quando o tamanho da amostra é elevado. A fórmula do BIC é:

$$BIC = -2\ln(L(\hat{\beta})) + q\ln(n) \tag{2.17}$$

Onde:

- n é o número de observações

O BIC impõe uma penalização mais severa do que o AIC à inclusão de novos parâmetros, dado que o termo $\ln(n)$ tende a crescer com o tamanho da amostra. Assim, o BIC tende a favorecer modelos mais simples, o que pode ser vantajoso em situações onde se pretende evitar o sobreajuste.
Segundo Burnham e Anderson (2002), o BIC assume uma fundamentação bayesiana mais rigorosa e tem como objectivo seleccionar o modelo mais plausível, assumindo que o verdadeiro modelo está entre os candidatos. Em contrapartida, o AIC é mais focado na minimização da perda de informação.

2.10 Interpretação dos coeficientes

A interpretação dos coeficientes estimados num modelo de regressão logística é essencial para compreender de que forma cada variável explicativa influencia a probabilidade do evento de interesse ocorrer. De acordo com Hosmer et al. (2013), na regressão logística, o foco não está na variação média da resposta, como na regressão linear, mas sim na alteração da log-odds (logaritmo da razão de probabilidades) associada a mudanças nas variáveis preditoras.

Sinal do coeficiente

O sinal de cada coeficiente β_j indica a direção da associação entre a variável explicativa x_j e a variável resposta:

- **Coeficiente positivo** ($\beta_j > 0$): Sugere que o aumento em x_j está associado a um aumento da *log-odds* do evento, o que implica maior probabilidade de ocorrência do desfecho.
- **Coeficiente negativo** ($\beta_j < 0$): Indica que o aumento em x_j reduz a *log-odds* do evento, resultando numa menor probabilidade de ocorrência.

Razão de chances (*odds ratio*-OR)

De acordo com Agresti (2013), para facilitar a interpretação prática, os coeficientes podem ser exponenciados, ou seja, e^{β_j}, de modo a fornecer a razão de chances (*odds ratio*). Esta transformação indica a multiplicação esperada nas razões de chances do evento para cada incremento unitário na variável preditora.

- Se $e^{\beta_j} > 1$: há um aumento nas chances do evento ocorrer.
- Se $e^{\beta_j} < 1$: há uma diminuição nas chances do evento ocorrer.
- Se $e^{\beta_j} = 1$: não há associação entre o preditor e o evento.

Intervalos de Confiança

Para além da estimativa pontual, é essencial considerar o intervalo de confiança (IC) dos coeficientes ou das odds ratio. Este intervalo fornece uma faixa de valores plausíveis para o parâmetro verdadeiro e permite avaliar a precisão da estimativa. Se o intervalo de confiança para a OR não incluir o valor 1, assume-se geralmente que o efeito é estatisticamente significativo ao nível de 5% (Hosmer et al., 2013).

2.11 Validação Cruzada

A validação cruzada é uma técnica estatística amplamente utilizada para avaliar o desempenho preditivo de modelos em contextos de aprendizagem supervisionada. Segundo Arlot e Celisse (2010),

a validação cruzada visa estimar o erro de generalização de um modelo, isto é, a sua capacidade de realizar previsões precisas em dados não observados durante o treino. Esta abordagem fornece uma estimativa mais fiável do desempenho do modelo do que a simples avaliação sobre os dados de treino, ajudando a evitar o sobreajustamento

De acordo com James et al. (2021), a principal finalidade da validação cruzada é fornecer uma estimativa realista do erro de previsão, permitindo assim uma comparação mais robusta entre diferentes modelos ou configurações. Kuhn e Johnson (2013) complementam que a validação cruzada actua como um mecanismo de controlo para garantir que o modelo não apenas se ajusta bem aos dados amostrais, mas também mantém bom desempenho quando aplicado a novos dados.

Existem diversos métodos de validação cruzada, sendo os mais usados os seguintes:

- **Validação Cruzada k-fold:** a amostra de dados é dividida em k subconjuntos de tamanho aproximadamente igual (chamados de *folds*). O modelo é treinado em $k-1$ destes subconjuntos e testado no subconjunto restante. Este processo é repetido k vezes, cada vez com um *fold* diferente reservado para teste. Por fim, os erros de previsão de cada iteração são agregados, geralmente por média, para estimar o desempenho global do modelo (James et al., 2021).

- **Validação Cruzada com Exclusão de Uma Observação (Leave-One-Out Cross-Validation — LOOCV):** Este é um caso particular do método anterior em que $k = n$, sendo n o número total de observações. Ou seja, cada observação é deixada de fora uma única vez como conjunto de teste, enquanto as restantes são usadas para treino. O LOOCV fornece uma estimativa quase sem viés, mas pode ser computacionalmente intensivo para grandes conjuntos de dados(Kuhn & Johnson, 2013).

- **Validação Cruzada Estratificada:** Em problemas de classificação com classes desbalanceadas, como frequentemente acontece na regressão logística binária, é recomendada a validação cruzada estratificada. Neste método, garante-se que a proporção das classes da variável dependente é aproximadamente a mesma em cada *fold*. Segundo Kuhn e Johnson (2013), este procedimento melhora a estabilidade da validação e reduz a variância da estimativa do erro.

Conforme afirmam James et al. (2021), no contexto da regressão logística, a validação cruzada permite avaliar métricas como a acurácia, sensibilidade, especificidade, área sob a curva ROC (AUC) e erro quadrático médio do logaritmo (Log Loss). Estas medidas ajudam a determinar se o modelo tem capacidade discriminativa suficiente e se generaliza bem.

2.12 Estudos Relacionados

Diversas técnicas estatísticas e de aprendizagem de máquina têm sido usadas na resolução de muitos problemas de diferentes áreas. Silveira et al. (2020), aplicaram regressão logística binária para

identificar fatores associados à realização de menos consultas pré-natais. Os resultados indicaram que mulheres solteiras, separadas ou viúvas tinham até três vezes mais chances de realizar um número reduzido de consultas. Além disso, adolescentes e mulheres com baixa escolaridade também apresentaram maior risco de não adesão ao pré-natal adequado.

Um estudo transversal realizado por Muleva et al. (2021), que investigou o número de consultas e a idade gestacional no início do pré-natal entre 393 puérperas em Nampula, revelou que, embora todas as mulheres tenham realizado pelo menos uma consulta de pré-natal, apenas 39.9% iniciaram o acompanhamento até a 16ª semana de gestação e 49.1% completaram quatro ou mais consultas, conforme recomendado pelo MISAU. A escolaridade foi um fator determinante, sendo o ensino secundário ou superior fortemente associado à maior probabilidade de início precoce e de número adequado de consultas. Outros fatores associados incluíram trabalho remunerado e coabitação com o parceiro. Além disso, entre os principais factores relatados para o início tardio e o número reduzido de consultas destacaram-se barreiras culturais como "considerar a barriga ainda pequena", desconhecimento da gravidez, ausência de acompanhante e dificuldade de acesso aos serviços de saúde.

Outro estudo conduzido por Nascimento e Barbosa (2021) realizou uma revisão sistemática com o objetivo de identificar os principais fatores associados ao início tardio do pré-natal. A partir da análise de dez artigos publicados entre 2016 e 2021, os autores classificaram os fatores em três grandes categorias: sociodemográficos, obstétricos e psíquicos. Entre os fatores sociodemográficos, destacaram-se a baixa escolaridade, residência em zona rural, idade materna jovem e baixa condição económica como preditores significativos do tardiamento. No domínio obstétrico, a multiparidade e a gravidez não planeada estiveram associadas a maior atraso no início das consultas. Já entre os fatores psíquicos, o uso de substâncias como álcool e drogas, bem como a ocorrência de violência por parceiro íntimo, foram apontados como elementos que contribuem negativamente para a adesão precoce ao pré-natal.

Costa et al (2014), aplicaram a regressão logística para analisar os fatores associados a ausência de realização de pré-natal em município de grande porte, em que no modelo final, as variáveis que mostraram associação com a não realização de pré-natal foram: ter menor escolaridade, especialmente menos de quatro anos de estudo, e ser multípara.

Reza e Salma (2024) utilizaram algoritmos de aprendizagem supervisionada para prever o risco de baixo peso ao nascer, com base em dados demográficos e obstétricos. Inicialmente, aplicaram o algoritmo Boruta para selecção de variáveis, identificando onze preditores relevantes, incluindo idade materna, nível de escolaridade, índice de massa corporal, idade à primeira relação sexual, ordem de nascimento e se a criança era gémea. Em seguida, compararam o desempenho da regressão logística com vários modelos de aprendizagem supervisionada, como Árvore de Decisão, Naive Bayes, Random Forest, XGBoost e AdaBoost. O modelo Random Forest destacou-se com

especificidade de 99%, sensibilidade de 58%, acurácia de 85,86%, F1-score de 92% e AUC de 55%. Mesmo com a selecção de apenas cinco variáveis significativas, a regressão logística manteve elevado desempenho, com F1-score de 93% e acurácia de 87,12%. Os autores sublinham a utilidade dos modelos de aprendizagem supervisionada, na identificação precoce de gestantes com maior risco de desfechos adversos, como o baixo peso à nascença.

Zhang et al. (2023) conduziram um estudo prospectivo com 2.000 gestantes para desenvolver modelos preditivos de pré-eclâmpsia de início tardio (após 34 semanas de gestação). Utilizando regressão logística, máquinas de vectores de suporte (SVM) e XGBoost, os autores incorporaram dados clínicos e laboratoriais coletados entre 6 e 13 semanas de gestação. A regressão logística identificou fatores de risco significativos, como hipertensão crônica, diabetes e níveis elevados de proteína urinária. O modelo XGBoost apresentou a melhor performance preditiva, com uma área sob a curva (AUC) superior a 0.90.

Bosschieter et al. (2023) desenvolveram modelos preditivos interpretáveis utilizando o método Explainable Boosting Machine (EBM) para identificar fatores de risco associados a complicações na gravidez, como morbidade materna grave, distocia de ombro, pré-eclâmpsia pré-termo e natimorto anteparto. Comparando o desempenho do EBM com modelos de regressão logística e outras técnicas de aprendizado de máquina, os autores observaram que o EBM ofereceu alta acurácia preditiva e melhor interpretabilidade. O estudo destacou fatores como altura materna e idade gestacional como preditores significativos, sugerindo que modelos interpretáveis podem auxiliar na compreensão e prevenção de complicações na gravidez.

Capítulo 3

MATERIAL E MÉTODOS

Neste capítulo é apresentada a classificação do presente estudo, o material e os procedimentos aplicados para a obtenção dos resultados.

3.1 Classificação da Pesquisa

3.1.1 Quanto à Natureza

Trata-se de uma pesquisa aplicada, uma vez que tem como finalidade gerar conhecimento voltado para a resolução de um problema concreto, neste caso, a identificação dos factores associados à não adesão às consultas pré-natais e a construção de um modelo preditivo com potencial de aplicação prática em contextos hospitalares. Segundo Prodanov e Freitas (2013), a pesquisa aplicada visa gerar conhecimentos para aplicação imediata, dirigidos à solução de problemas específicos da realidade.

3.1.2 Quanto aos objectivos

Do ponto de vista dos seus objectivos, este estudo enquadra-se nas pesquisas de natureza descritiva e explicativa. Segundo Gil (2008), as pesquisas descritivas têm como propósito observar, registrar, analisar e correlacionar factos ou fenómenos sem manipulá-los, enquanto as explicativas procuram identificar os factores que determinam ou contribuem para a ocorrência dos fenómenos observados, estabelecendo relações de causa e efeito. Neste trabalho, além de descrever o perfil das gestantes, pretendeu-se compreender os factores que explicam a não adesão ao acompanhamento pré-natal.

3.1.3 Quanto à abordagem

Quanto à abordagem metodológica, trata-se de um estudo de natureza quantitativa. Esta abordagem permite quantificar relações entre variáveis e testar hipóteses de forma objectiva, conferindo maior rigor científico aos resultados. Richardson (2017) afirma que a abordagem quantitativa busca quantificar variáveis e aplicar instrumentos estatísticos para análise dos dados, sendo indicada quando o objectivo é mensurar, comparar e generalizar conclusões.

3.1.4 Quanto ao Delineamento

Trata-se de um estudo de delineamento transversal, pois os dados foram recolhidos num único momento no tempo, permitindo avaliar a associação entre variáveis num determinado ponto temporal. Segundo Gil (2008), estudos transversais são apropriados para estimar a prevalência de fenómenos e examinar relações entre variáveis numa população específica, sendo amplamente utilizados em pesquisas em saúde pública e ciências sociais.

3.2 Material

A amostra do presente estudo é constituída por todas as mulheres que foram atendidas nos serviços de cuidados pré-natais do Hospital Geral José Macamo, no período de Maio de 2023 a Maio de 2024. Para a recolha desta informação, proveniente de fonte secundária, foi solicitado à direcção da referida instituição o devido consentimento para utilização dos dados com fins académicos.
A recolha e posterior organização da informação, no sentido de constituir uma base de dados apta à análise estatística, foram efectuadas a partir dos registos clínicos das consultas pré-natais realizadas pelas gestantes no hospital em questão.
A base de dados resultante integra informações relativas a 1026 mulheres atendidas nos serviços de cuidados pré-natais do Hospital Geral José Macamo durante o referido período.
A variável dependente, designada por "não adesão", foi derivada da variável "número de consultas realizadas"e definida como o não cumprimento do mínimo de quatro consultas pré-natais, conforme as recomendações do Ministério da Saúde. A seguir, apresenta-se a descrição detalhada das variáveis incluídas na análise:

Tabela 3.1: Descrição das variáveis do estudo

Variável	Descrição	Categorias	Classificação
Nível de escolaridade	Nível de escolaridade da mulher	Sem escolaridade, Primário, Secundário, Superior	Categórica ordinal
Idade	Faixa etária da mulher	15–19 anos, 20–24 anos, >25 anos	Categórica ordinal
Idade gestacional	Idade gestacional na primeira consulta	1º trimestre (<12 semanas), 2º trimestre (12–27 semanas), 3º trimestre (28+ semanas)	Categórica ordinal
Estado civil	Estado civil da mulher	Solteira, Casada	Categórica nominal
Parceiro presente	Se o parceiro se fez presente às consultas	Sim, Não	Categórica nominal
Gravidez desejada	Se desejava a gravidez ou não	Sim, Não	Categórica nominal
Paridade	Gestações que a mulher já teve	Primípara, Multípara	Categórica nominal
Histórico de abortos	Se já sofreu abortos anteriormente	Sim, Não	Categórica nominal
Teste HIV	Resultado do teste de HIV da mulher	Positivo, Negativo	Categórica nominal
Ocupação	Se a mulher possui alguma fonte de renda	Sim, Não	Categórica nominal
Não adesão às consultas	Se aderiu ou não às consultas pré-natais	Não adesão, Adesão	Categórica nominal

Os dados foram processados com o auxílio do Microsoft Office Excel 2016, software R versão 4.1.2, as hipóteses foram testadas a um nível de significância de 5% e para efeitos de avaliação da regra de decisão o p-valor associado a estatística do teste.

3.3 Métodos

O processo metodológico deste estudo seguiu as etapas padrão de aprendizagem supervisionada desde o pré-processamento dos dados até a validação do modelo final. O principal objetivo foi identificar os factores associados à não adesão às consultas pré-natais e desenvolver um modelo preditivo capaz de classificar gestantes quanto ao risco de não comparecimento às consultas. Para esse fim, foi utilizada a regressão logística binária como algoritmo de classificação, dada a sua elevada capacidade explicativa e a vantagem de permitir interpretar o impacto individual de cada variável preditora sobre a probabilidade de ocorrência do evento de interesse (Agresti, 2013). A avaliação da robustez do desempenho do modelo foi realizada através de validação cruzada estratificada do tipo k-fold, uma técnica que permite estimar, de forma mais estável e realista, a capacidade de generalização do modelo a novos dados (James et al., 2021).

3.3.1 Análise descritiva

A análise descritiva, também conhecida como análise exploratória de dados, é uma etapa essencial no processo de modelação estatística e aprendizagem de máquina, pois permite compreender o comportamento das variáveis antes da aplicação de técnicas preditivas ou inferenciais. Segundo Tukey (1977), a análise exploratória visa identificar padrões, outliers, possíveis inconsistências e relações preliminares entre as variáveis, além de auxiliar na formulação de hipóteses e na escolha de estratégias de modelação.

Considerando que todas as variáveis deste estudo são de natureza categórica, a análise descritiva foi conduzida por meio da distribuição de frequências absolutas e relativas. De acordo com Pestana e Gageiro (2014), esse tipo de análise facilita a compreensão da proporção de indivíduos em cada categoria, permitindo não apenas a caracterização da amostra, mas também comparações entre subgrupos e a identificação de possíveis assimetrias na distribuição dos dados.

3.3.2 Divisão da amostra

A divisão da amostra em subconjuntos de treino e teste constitui uma etapa crucial no desenvolvimento de modelos preditivos. Segundo Kuhn e Johnson (2013), essa separação visa evitar o sobreajustamento, permitindo que o modelo seja avaliado com dados que não foram utilizados durante o processo de treino, o que garante uma estimativa mais realista da sua capacidade de generalização.

Neste estudo, a base de dados foi dividida aleatoriamente em:

- **Conjunto de Treino (80%):** utilizado para ajustar o modelo, ou seja, para estimar os coeficientes dos parâmetros com base nos dados disponíveis.
- **Conjunto de Teste (20%):** serviu exclusivamente para avaliar a performance preditiva do modelo em dados não vistos.

A divisão foi feita de forma estratificada, assegurando que a proporção entre os casos de adesão e não adesão fosse mantida em ambos os subconjuntos, o que é particularmente importante em situações de classes desbalanceadas (James et al. 2021).

3.3.3 Balanceamento das classes

Dado que a variável de interesse neste estudo, a não adesão às consultas pré-natais, apresentou distribuição assimétrica entre suas categorias, foi necessário aplicar técnicas de balanceamento para evitar o viés do modelo em favor da classe majoritária. Em problemas de classificação binária, a presença de classes desbalanceadas pode comprometer seriamente o desempenho do modelo, resultando em alta acurácia aparente, mas baixa capacidade de detecção da classe minoritária (Fernández et al., 2018).

Neste estudo, para amenizar este problema, foi utilizado o método de superamostragem denominado SMOTE (Synthetic Minority Over-sampling Technique), o qual gera novos exemplos sintéticos da classe minoritária a partir da interpolação entre observações reais e seus vizinhos mais próximos (Chawla et al., 2002). O objetivo é criar um conjunto de treino com proporção aproximadamente equilibrada entre as classes (cerca de 50% para cada uma), aumentando a capacidade do modelo de aprender os padrões associados à classe minoritária, neste caso, as gestantes que não aderiram às consultas. A aplicação do SMOTE foi feita apenas no conjunto de treino, respeitando a integridade do conjunto de teste, de forma a garantir que a avaliação do modelo fosse feita em um cenário realista e imparcial

3.3.4 Teste de independência do qui-quadrado

A selecção inicial das variáveis com potencial associação à não adesão às consultas pré-natais foi realizada através do teste de independência do qui-quadrado. Este teste permite verificar a existência de associação estatisticamente significativa entre variáveis qualitativas, comparando as frequências observadas com as esperadas sob a hipótese de independência (Pestana & Gageiro, 2014).
Para cada variável explicativa, foi testada a hipótese nula de ausência de associação com a variável dependente. As variáveis que apresentaram valor de p inferior a 0.05 foram consideradas estatisticamente significativas e, consequentemente, seleccionadas para inclusão no modelo de regressão logística. O valor do qui-quadrado pode ser calculado por meio da equação:

$$\chi^2 = \sum_{i=1}^{l}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \sim \chi^2_{(l-1)(c-1)}$$

Onde:
O_{ij} – São as frequências observadas
E_{ij} – São as frequências esperadas
l – É o número de linhas

c – É o número de colunas

$\chi^2_{(l-1)(c-1)}$ – é a distribuição qui-quadrado com $(l-1)(c-1)$ graus de liberdade

3.3.5 Construção do Modelo de Regressão Logística

O modelo de regressão logística foi construído usando a base de treino balanceada. A Regressão Logística estima a probabilidade de ocorrência de um determinado evento (neste caso, a não adesão), em função de um conjunto de variáveis explicativas. O modelo assume a seguinte forma matemática:

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

Onde $P(Y=1)$ representa a probabilidade de não adesão, β_0 é o intercepto, β_1, β_2,..., β_k são os coeficientes associados às variáveis preditoras x_1, x_2,..., x_k. Estes coeficientes podem ser interpretados em termos de razões de chances (*odds ratios*), permitindo avaliar o impacto de cada preditor sobre a probabilidade do desfecho.

Verificação da Multicolinearidade

A avaliação da multicolinearidade foi realizada por meio do cálculo do Fator de Inflação da Variância (VIF) e da tolerância.
O VIF mede quanto a variância de um coeficiente estimado aumenta devido à multicolinearidade. Para uma dada variável X_j, o VIF é definido por:

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

Onde R_j^2 é o coeficiente de determinação obtido ao regredir X_j em todas as outras variáveis independentes do modelo.

- $\text{VIF}_j = 1$: não há correlação entre X_j e as outras variáveis;
- $1 < \text{VIF}_j < 5$: indica baixa multicolinearidade;
- $\text{VIF}_j > 5$: alerta para possível multicolinearidade;
- $\text{VIF}_j > 10$: considerado forte sinal de multicolinearidade.

A tolerância é o inverso do VIF:

$$\text{Tol}_j = \frac{1}{\text{VIF}_j} = 1 - R_j^2$$

Valores baixos de tolerância (menores que 0,10) indicam elevada multicolinearidade.

Seleção das variáveis do modelo

A optimização do modelo foi feita através da selecção automática de variáveis pelo método *stepwise*, orientado pelo Critério de Informação de Akaike (AIC). O AIC é definido pela fórmula:

$$\text{AIC} = -2\ln\left(L(\widehat{\beta})\right) + 2q$$

Onde $\ln\left(L(\widehat{\beta})\right)$ é o logaritmo natural da função de verosimilhança do modelo com os parâmetros estimados e q representa o número de variáveis explicativas incluídas no modelo. O procedimento *stepwise* foi aplicado em ambas as direcções (*forward* e *backward*), avaliando iterativamente os modelos e seleccionando aquele com o menor valor de AIC.

3.3.6 Avaliação do ajuste do modelo

Concluída a seleção das variáveis, procedeu-se à avaliação da qualidade do ajustamento do modelo de regressão logística, com o intuito de verificar a sua adequação aos dados observados. O ajuste do modelo foi avaliado pelos testes de Hosmer-Lemeshow, teste de razão de verossimilhanças e pelas medidas pseudo-R^2 (Cox e Snell, Nagelkerke e McFadden).

Teste de Hosmer-Lemeshow

Segundo Hosmer e Lemeshow (2000), este teste é especificamente desenhado para avaliar o ajuste global do modelo de regressão logística. Divide a amostra em grupos (tipicamente 10) com base nos valores preditos da probabilidade ($\widehat{\pi}$) e compara as frequências observadas e esperadas de ocorrências dentro de cada grupo. A estatística é:

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

Onde:

- O_g é o número de eventos observados no grupo g;
- E_g é o número esperado de eventos no grupo g;
- G é o número de grupos.

Hipóteses:

- H_0: O modelo ajusta-se bem aos dados
- H_1: O modelo não se ajusta bem aos dados

Interpretação: Um valor-p superior a 0.05 indica que não se rejeita H_0 sugerindo que o modelo apresenta um bom ajuste.

Teste de razão de verossimilhanças

Segundo Hosmer e Lemeshow (2000), o teste da razão de verosimilhança compara o modelo completo com um modelo reduzido (nulo), que não inclui preditores. Este teste avalia se a introdução das variáveis explicativas melhora significativamente o ajuste do modelo. A estatística de teste é dada por:

$$G = -2\ln\left(\frac{L_0}{L_1}\right) = -2[\ln(L_0) - \ln(L_1)]$$

Onde:

- L_0 é a verosimilhança do modelo nulo (com apenas o intercepto);
- L_1 é a verosimilhança do modelo ajustado com os preditores.

Esta estatística segue aproximadamente uma distribuição qui-quadrado com graus de liberdade iguais ao número de variáveis adicionadas. Segundo Hosmer e Lemeshow (2000), este teste é uma das ferramentas mais robustas para avaliar a significância global do modelo.

Hipóteses:

- H_0: Todos os coeficientes das variáveis independentes são iguais a zero ($\beta_1 = \beta_2 = \cdots = \beta_k = 0$).
- H_1: Pelo menos um coeficiente é diferente de zero

Interpretação: Se o valor-p do teste for inferior a 0.05, rejeita-se H_0, indica que o modelo com as variáveis explicativas se ajusta significativamente melhor do que o modelo sem preditores.

Medidas Pseudo-R^2

Segundo Menard (2000), embora o verdadeiro R^2 não se aplique à regressão logística, existem várias medidas alternativas conhecidas como pseudo-R^2, que procuram quantificar a proporção da variabilidade explicada pelo modelo. Neste trabalho foram usadas as seguintes medidas:

- R^2 **de McFadden**: baseia-se na comparação entre o logaritmo da verossimilhança do modelo ajustado e o de um modelo nulo (sem preditores), sendo interpretado de forma análoga ao R^2 tradicional, embora com valores geralmente mais baixos. Tem a forma:

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

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

- R^2 **de Cox e Snell**: baseia-se na verossimilhança e estima a proporção da variabilidade explicada, embora seu valor máximo seja inferior a 1. É dada por:

$$R^2_{\text{CS}} = 1 - \left(\frac{L_0}{L_M}\right)^{\frac{2}{n}}$$

onde L_0 é a verossimilhança do modelo nulo, L_M é a verossimilhança do modelo ajustado e n o número de observações.

- R^2 **de Nagelkerke**: ao contrário do R^2 de Cox e Snell, cujo valor máximo é inferior a 1, o R^2 de Nagelkerke corrige essa limitação, possibilitando alcançar o extremo superior do intervalo. Em termos práticos, valores superiores a 0,30 são geralmente considerados indicativos de um bom ajustamento do modelo. É dada por:

$$R^2_N = \frac{R^2_{\text{CS}}}{1 - L_0^{\frac{2}{n}}}$$

3.3.7 Avaliação do desempenho preditivo

Após o ajuste do modelo final de regressão logística, procedeu-se à avaliação do seu desempenho preditivo, utilizando-se o conjunto de teste previamente separado. Esta etapa tem como finalidade avaliar a capacidade de generalização do modelo, ou seja, a sua aptidão para realizar previsões fiáveis quando aplicado a novos dados não utilizados durante o processo de treino. Para esse fim, foram analisadas diversas métricas de desempenho obtidas a partir da matriz de classificação.

Matriz de classificação

Segundo Hair et al (2005), a matriz de classificação permite avaliar a capacidade do modelo em classificar correctamente os casos como eventos (Y = 1) ou não eventos (Y = 0), com base num limiar de decisão. Ela é composta por quatro elementos principais que são usados para calcular as métricas de desempenho do modelo:

Tabela 3.2: Matriz de Classificação

	Real Positivo	**Real Negativo**
Previsto Positivo	Verdadeiro Positivo (VP)	Falso Positivo (FP)
Previsto Negativo	Falso Negativo (FN)	Verdadeiro Negativo (VN)

A partir da matriz de classificação, foram calculadas as seguintes métricas:

- **Acurácia**: mede a proporção de previsões corretas. É dada por:

$$\text{Acurácia} = \frac{VP + VN}{VP + FP + FN + VN}$$

- **Precisão (Valor Preditivo Positivo - VPP)**: corresponde à proporção de casos correctamente classificados como não aderentes entre todos os classificados como tal:

$$\text{Precisão} = \frac{VP}{VP + FP}$$

- **Valor Preditivo Negativo (VPN)**: refere-se à proporção de aderentes corretamente classificados:

$$\text{VPN} = \frac{VN}{VN + FN}$$

- **Sensibilidade (Recall)**: indica a capacidade do modelo em identificar corretamente os casos de não adesão:

$$\text{Sensibilidade} = \frac{VP}{VP + FN}$$

- **Especificidade**: mede a proporção de aderentes corretamente classificados:

$$\text{Especificidade} = \frac{VN}{VN + FP}$$

- **F1-score**: representa a média harmónica entre a precisão e a sensibilidade, oferecendo uma medida equilibrada entre as duas:

$$F1 = 2 \cdot \frac{\text{Precisão} \cdot \text{Sensibilidade}}{\text{Precisão} + \text{Sensibilidade}}$$

- **Coeficiente Kappa**: quantifica o grau de concordância entre as classificações do modelo e os valores reais, ajustando a estimativa para o acaso. Valores de Kappa entre 0.61 e 0.80 indicam concordância substancial (Landis & Koch, 1977).

Segundo Powers (2011), para métricas como acurácia, sensibilidade, especificidade, VPP e VPN, valores superiores a 70% são geralmente considerados bons indicadores de desempenho em modelos preditivos.

Curva ROC e Área sob a Curva (AUC)

A Curva ROC (Receiver Operating Characteristic) foi utilizada para avaliar o poder discriminativo do modelo. Ela analisa a relação entre a sensibilidade e a especificidade para diferentes pontos de corte. A área sob a curva (AUC) é uma medida agregada da performance do modelo:

- AUC = 0.5: modelo aleatório;
- AUC entre 0,7 e 0.8: bom;
- AUC entre 0,8 e 0.9: muito bom;
- AUC $>$0.9: excelente.

O ponto de corte ideal foi determinado com base na menor distância até o canto superior esquerdo da curva ROC, critério que maximiza simultaneamente a sensibilidade e a especificidade (Fawcett, 2006). Com base nessa análise, foi adotado um ponto de corte de 0.35 para a classificação final.

Validação Cruzada

Para avaliar a robustez e generalização do modelo construído, foi aplicada a técnica de validação cruzada estratificada do tipo k-fold, com k=10. Esta abordagem consiste em subdividir aleatoriamente a base de dados em 10 subconjuntos (folds), mantendo em cada um deles aproximadamente a mesma proporção entre as categorias da variável dependente, o que é especialmente relevante em contextos com classes desbalanceadas. Em cada iteração, o modelo é treinado em 9 subconjuntos e testado no subconjunto restante, repetindo-se o processo até que cada fold tenha servido uma vez como conjunto de validação.

As métricas de desempenho como acurácia, sensibilidade, especificidade, precisão, valor preditivo negativo (VPN), F1-score, AUC e coeficiente Kappa, foram calculadas em cada iteração e, posteriormente, as médias foram utilizadas como estimativas finais da performance do modelo. Esta estratégia permite uma avaliação mais estável e fiável, reduzindo a variabilidade decorrente de uma única divisão treino-teste (James et al., 2021).

Capítulo 4

RESULTADOS E DISCUSSÃO

Neste capítulo, são apresentados os resultados do estudo de acordo com a metodologia apresentada no capítulo anterior. Também são apresentadas estatísticas descritivas da amostra e a discussão dos principais resultados obtidos.

4.1 Análise descritiva

A amostra do estudo é composta por 1026 mulheres com idade superior a 15 anos que realizaram acompanhamento pré-natal no Hospital Geral José Macamo. Dentre elas, 24.9% foram classificadas como não aderentes por terem comparecido a menos de quatro consultas durante a gestação, enquanto 75.1% foram consideradas aderentes, conforme ilustra a Figura 4.1.

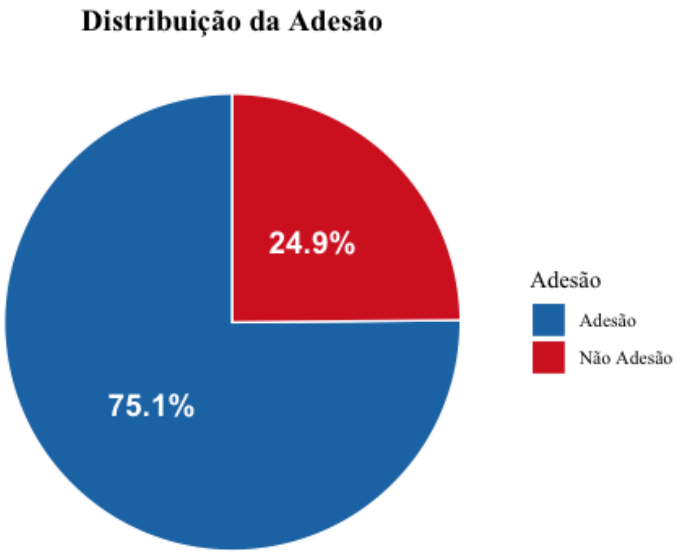

Figura 4.1: Distribuição das mulheres segundo a adesão às consultas pré-natais

Na Figura 4.2, são apresentadas as distribuições de frequências das variáveis nível de escolaridade e idade gestacional na primeira consulta. Pode-se observar que do total de mulheres, mais da metade (52.4%) frequentou o ensino secundário, 7.5% não possuíam escolaridade, 28.0% tinham

o nível primário e 12.1% possuíam nível superior de educação. Em relação à idade gestacional na primeira consulta, 17.2% das mulheres iniciaram o pré-natal com menos de 12 semanas de gestação, 75.5% entre 12 a 27 semanas e 7.3% com mais de 28 semanas de gestação.

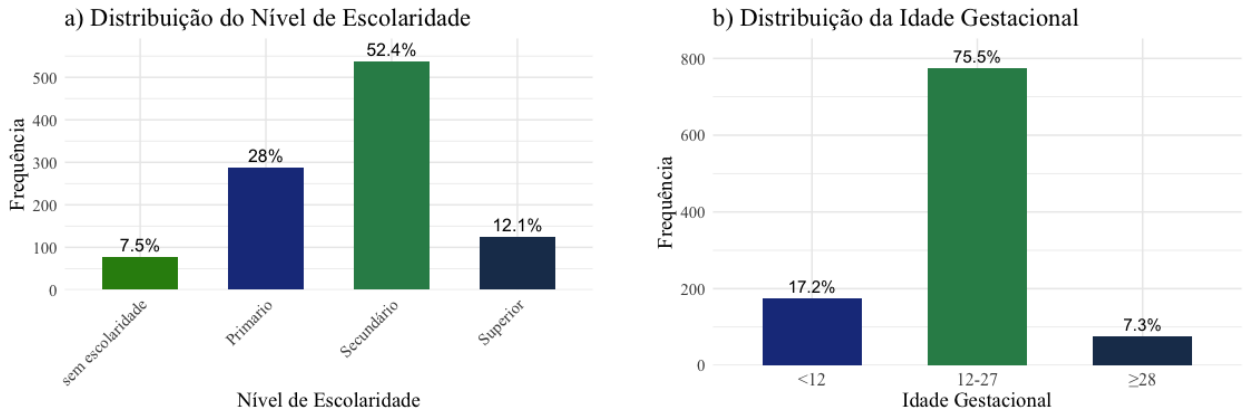

Figura 4.2: Distribuição das mulheres por nível de escolaridade e idade gestacional na primeira consulta (em semanas).

Na Figura 4.3, são apresentadas as distribuições de frequências da não adesão às consultas por faixa etária (alínea a)) e pela presença do parceiro nas consultas (alínea b)). Pode-se observar que, entre as gestantes com idade inferior a 20 anos, 45.2% não aderiram às consultas pré-natais, enquanto 54.8% aderiram. No grupo com idade entre 20 a 24 anos, 22.7% não aderiram e 77.3% aderiram; entre aquelas com mais de 25 anos, 24.6% não aderiram e 75.4% aderiram. Quanto à presença do parceiro, verifica-se que, entre as mulheres sem parceiro presente, 40.0% apresentaram não adesão às consultas, enquanto 60.0% aderiram. Já entre aquelas com parceiro presente, apenas 19.3% não aderiram, ao passo que 80.7% aderiram.

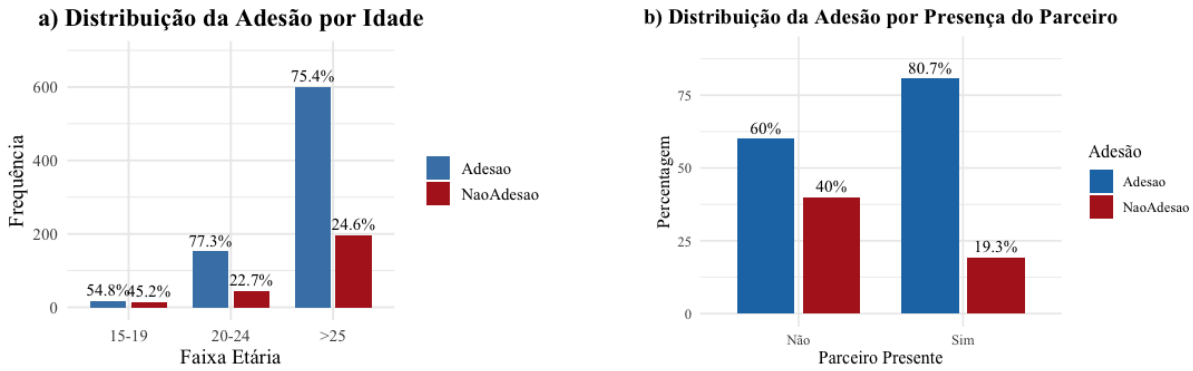

Figura 4.3: Distribuição da adesão às consultas por faixa etária e presença do parceiro.

A tabela 4.1, apresenta a distribuição de frequências absolutas e relativas das gestantes que realizaram consultas pré-natais no Hospital Geral José Macamo, de acordo com suas características demográficas, socioeconómicas e de saúde reprodutiva.

No que diz respeito ao estado civil, observa-se que, entre as 168 mulheres casadas, 17.3% não aderiram às consultas pré-natais, enquanto 82.7% aderiram. Entre as solteiras, 26.3% não aderiram, enquanto 73.7% aderiram às consultas.

Em relação ao teste de HIV, entre as mulheres com resultado negativo, 24.2% não aderiram às consultas pré-natais, enquanto 75.8% aderiram. Já entre aquelas com teste positivo, 28.6% não aderiram, enquanto 71.4% aderiram.

No que concerne ao nível de escolaridade, das mulheres sem escolaridade, 44.2% não aderiram às consultas, enquanto 55.8% aderiram. Entre aquelas com nível primário, 26.5% não aderiram, enquanto 73.5% aderiram. No nível secundário, 24.4% não aderiram, enquanto 75.6% aderiram. Por fim, entre as mulheres com ensino superior, apenas 11.3% não aderiram, enquanto 88.7% aderiram.

No que se refere ao histórico de abortos, observa-se que entre as mulheres que não sofreram abortos, 24.1% não aderiram às consultas pré-natais, enquanto 75.9% aderiram. Já entre aquelas que sofreram abortos, 26.1% não aderiram, enquanto 73.9% aderiram.

Quanto à gravidez desejada, entre as mulheres que não desejaram a gravidez, 23.0% não aderiram às consultas pré-natais, enquanto 77.0% aderiram. Já entre aquelas que desejaram a gravidez, 30.2% não aderiram, enquanto 69.8% aderiram. No que diz respeito à ocupação, observa-se que entre as mulheres sem ocupação, 24.4% não aderiram às consultas pré-natais, enquanto 75.6% aderiram. Já entre aquelas que possuem ocupação, 25.9% não aderiram, enquanto 74.1% aderiram.

Quanto à paridade, observa-se que entre as primíparas, 16.0% não aderiram às consultas pré-natais, enquanto 84.0% aderiram. Já entre as multíparas, 28.2% não aderiram, enquanto 71.8% aderiram.

Por fim, em relação à idade gestacional, observa-se que entre as mulheres com idade gestacional inferior a 12 semanas, 16.5% não aderiram às consultas pré-natais, enquanto 83.5% aderiram. Entre aquelas com idade gestacional entre 12 e 27 semanas, 22.3% não aderiram, enquanto 77.7% aderiram. Já entre as mulheres com idade gestacional igual ou superior a 28 semanas, 70.7% não aderiram, enquanto 29.3% aderiram às consultas pré-natais.

Tabela 4.1: Distribuição de frequências absolutas e relativas das gestantes segundo características sociodemográficas e de saúde reprodutiva

Variável	Não Aderiu (n = 255)	Aderiu (n = 771)	Total (N = 1026)
Estado civil			
Casada	29 (17.3%)	139 (82.7%)	168
Solteira	226 (26.3%)	632 (73.7%)	858
Teste HIV			
Negativo	209 (24.2%)	656 (75.8%)	865
Positivo	46 (28.6%)	115 (71.4%)	161
Escolaridade			
Sem escolaridade	34 (44.2%)	43 (55.8%)	77
Primário	76 (26.5%)	211 (73.5%)	287
Secundário	131 (24.4%)	407 (75.6%)	538
Superior	14 (11.3%)	110 (88.7%)	124
Histórico de abortos			
Não	150 (24.1%)	474 (75.9%)	624
Sim	105 (26.1%)	297 (73.9%)	402
Gravidez desejada			
Não	176 (23.0%)	588 (77.0%)	764
Sim	79 (30.2%)	183 (69.8%)	262
Ocupação			
Sem ocupação	176 (24.4%)	545 (75.6%)	721
Com ocupação	79 (25.9%)	226 (74.1%)	305
Paridade			
Primípara	45 (16.0%)	236 (84.0%)	281
Multípara	210 (28.2%)	535 (71.8%)	745
Idade gestacional (1ª consulta)			
$<$12 semanas	29 (16.5%)	147 (83.5%)	176
12–27 semanas	173 (22.3%)	602 (77.7%)	775
$\geq$ 28 semanas	53 (70.7%)	22 (29.3%)	75

4.2 Análise da associação entre as variáveis independentes e a variável dependente

Para análise da existência de associação entre a variável de interesse e as demais variáveis do estudo foi usado o teste Qui-quadrado de Pearson cujos resultados são apresentados na tabela 4.2. Consideraram-se como variáveis significativas aquelas cujo p-value foi menor do que o nível de significância de 5%.

Se o p-valor for inferior a 5%, rejeita-se a hipótese nula, indicando que há evidências estatísticas de uma associação significativa entre as variáveis analisadas. Por outro lado, quando o p-valor é igual ou superior a 5%, não se encontram indícios suficientes para afirmar que exista uma relação estatística entre elas.

Com base nisso, foram encontradas associações estatisticamente significativas entre a variável Não adesão as consultas pré-natais e as seguintes co-variáveis: idade da mulher, idade gestacional na

primeira consulta, paridade, presença do parceiro, nível de escolaridade, estado civil e a variável se desejava a gravidez.

Tabela 4.2: Teste de associação entre as variáveis independentes e a variável dependente

Co-variável	χ^2	**Valor-p**
Idade gestacional na 1ª consulta	89.698	<**0.001**
Presença do parceiro	49.270	<**0.001**
Nível de escolaridade	31.548	<**0.001**
Paridade	15.082	<**0.001**
Estado civil	4.697	**0.030**
Gravidez desejada	4.234	**0.040**
Idade da mulher	6.286	**0.043**
Teste de HIV	3.400	0.065
Histórico de abortos	0.359	0.549
Ocupação	0.000	1.000

4.3 Modelação da Regressão Logística

Dado o desequilíbrio observado na variável dependente, com 75.1% das gestantes classificadas como aderentes e apenas 24.9% como não aderentes, procedeu-se ao balanceamento do conjunto de treino utilizando a técnica SMOTE (Synthetic Minority Over-sampling Technique). Esta abordagem permitiu gerar novos exemplos sintéticos da classe minoritária, ajustando a proporção entre as categorias para cerca de 55% de adesão e 45% de não adesão. Este procedimento visou reduzir o viés do modelo e melhorar a sua capacidade de identificar correctamente os casos de não adesão. A partir da base balanceada, avançou-se para a análise da multicolinearidade entre as variáveis explicativas, seguida da construção do modelo final de regressão logística.

4.3.1 Multicolinearidade

Para verificar a presença de multicolinearidade entre as variáveis independentes, foram calculados o Fator de Inflação da Variância (VIF) e a Tolerância. Todos os valores observados do VIF ficaram abaixo de 10, o que indica ausência de multicolinearidade entre os preditores e os valores da tolerância se aproximam de 1 reforçando a inexistência de multicolinearidade relevante entre os preditores, conforme apresentado na tabela 4.3.

Tabela 4.3: Multicolinearidade entre as variáveis explicativas

Variável	VIF	Tolerância
Idade	1.066	0.938
Parceiro presente	1.121	0.892
Estado civil	1.026	0.975
Teste HIV	1.099	0.910
Escolaridade	1.286	0.778
Histórico de abortos	1.079	0.926
Gravidez desejada	1.015	0.985
Ocupação	1.089	0.918
Paridade	1.088	0.919
Idade gestacional	1.376	0.727

4.3.2 Construção do modelo final

Inicialmente, foi ajustado um modelo contendo todas as variáveis independentes que apresentaram associação estatisticamente significativa ao nível de 5%, conforme identificado pelo teste do qui-quadrado de Pearson (ver Tabela 4.2), aplicado à base original (não balanceada). Em seguida, aplicou-se o procedimento de seleção stepwise baseado no Critério de Informação de Akaike (AIC), com o objetivo de obter um modelo mais parcimonioso, eliminando preditores com menor relevância estatística e retendo aqueles com maior capacidade explicativa para o desfecho de interesse.

Na tabela 4.4 são apresentados os coeficientes estimados do modelo de regressão logística final.

Tabela 4.4: Modelo de regressão logística final

Variável	Estimativa	EP	P-valor	OR	IC 95%
Intercepto	-1.2775	0.3226	**<0.001**	0.279	[0.148; 0.523]
Idade – >25 anos (ref)					
15–19 anos	1.2163	0.414	**0.003**	3.375	[1.512; 7.531]
Parceiro presente – Não (ref)					
Sim	-1.0286	0.1523	**<0.001**	0.358	[0.266; 0.482]
Estado civil – Casada (ref)					
Solteira	0.3869	0.2033	0.057	1.472	[0.987; 2.196]
Escolaridade – Primário (ref)					
Secundário	-0.3028	0.1571	0.054	0.739	[0.541; 1.010]
Sem escolaridade	0.6107	0.2727	**0.025**	1.842	[1.077; 3.150]
Superior	-1.4468	0.3018	**<0.001**	0.235	[0.130; 0.426]
Gravidez desejada – Não (ref)					
Sim	-0.3101	0.1741	0.075	0.733	[0.522; 1.030]
Paridade – Primípara (ref)					
Multípara	1.2455	0.1865	**<0.001**	3.475	[2.426; 4.976]
Idade gestacional – <12 semanas (ref)					
12–27 semanas	0.6207	0.202	**0.002**	1.860	[1.242; 2.786]
≥28 semanas	2.6252	0.3404	**<0.001**	13.808	[6.987; 27.290]

(Ref) – categoria de referência

Controlando o efeito das demais variáveis do modelo, observou-se que o momento de início do pré-natal esteve fortemente associado à não adesão. Gestantes que iniciaram as consultas entre 12 á 27 semanas, apresentaram 1.86 vezes mais chance de não adesão em comparação com aquelas que iniciaram antes das 12 semanas. A associação foi ainda mais expressiva para aquelas que iniciaram com 28 semanas ou mais, cuja chance de não adesão foi aproximadamente 14 vezes maior (OR = 13.808; p <0.001).

A presença do parceiro durante a gestação demonstrou ser um fator protetor significativo contra a não adesão. Ter o parceiro presente reduz em 64.3% a chance de não adesão, comparando com quem não tem parceiro presente. Quanto a gravidez desejada, mulheres que desejavam a gravidez tiveram 27% menos chance de não adesão, embora com significância fraca.

A escolaridade mostrou-se uma variável determinante na adesão ao pré-natal. Mulheres sem escolaridade tiveram aproximadamente 2 vezes mais chance de não adesão em relação àquelas com ensino primário (OR = 1.82; p = 0.025), enquanto ter ensino secundário reduz em cerca de 26% a chance de não adesão, em relação ao ensino primário. E gestantes com ensino superior apresentaram uma redução significativa do risco, com uma chance 76.5% menor de não adesão (OR = 0.24; p <0.001).

A multiparidade foi positivamente associada à não adesão, indicando que mulheres com dois ou mais filhos apresentaram 3.47 vezes mais probabilidade de não adesão ao pré-natal em comparação com as primíparas (OR = 3.47; p <0.001).

A idade materna também influenciou significativamente a não adesão. Gestantes adolescentes, com idades entre 15 á 19 anos, apresentaram 3.37 vezes mais chance de não adesão em comparação com mulheres de 25 anos ou mais (OR = 3.37; p = 0.003), controlando o efeito das outras variáveis.

4.3.3 Verificação do Ajuste do Modelo

Teste de razão de verossimilhanças

Para avaliar a significância global do modelo de regressão logística múltipla ajustado, foi utilizado o Teste da Razão de Verossimilhanças, o qual compara o modelo ajustado com um modelo nulo (sem variáveis explicativas). Este teste verifica se o conjunto de variáveis independentes contribui significativamente para a explicação da variabilidade na adesão às consultas pré-natais. Como mostra a tabela 4.5, o resultado do teste revelou um valor de p <0.05, indicando que o modelo ajustado oferece um melhor ajuste aos dados do que o modelo nulo, ou seja, pelo menos uma das variáveis independentes incluídas no modelo, tem efeito significativo na probabilidade de não adesão.

Tabela 4.5: Teste da razão de verossimilhança entre o modelo nulo e o modelo final

Modelo	Deviância Res.	Dif. GL	Dif. Deviância	Valor-p
Modelo Nulo	1483.8	–	–	–
Modelo Final	1238.0	10	245.81	<0.001

Medidas Pseudo- R^2

As medidas de pseudo-R^2 foram utilizadas para avaliar a qualidade do ajuste do modelo, oferecendo uma estimativa da proporção da variância explicada pela combinação de variáveis independentes. O R^2 de McFadden, com valor de 0.166, indica um ajuste razoável do modelo, considerando que valores entre 0.15 e 0.30 são geralmente interpretados como adequados para estudos em ciências sociais e saúde pública. O R^2 de Cox Snell apresentou valor de 0.204, sugerindo que aproximadamente 20.4% da variabilidade na não adesão pode ser explicada pelo modelo. Já

o R² de Nagelkerke, uma versão ajustada que pode atingir o valor máximo de 1, atingiu 0.273, indicando um ajuste moderado e reforçando a capacidade explicativa do modelo, como mostra a tabela 4.6.

Tabela 4.6: Medidas Pseudo-R^2

Medida	R^2 **de McFadden**	R^2 **de Cox & Snell**	R^2 **de Nagelkerke**
Valor obtido	0.166	0.204	0.273

Teste de Hosmer-Lemeshow

O ajuste do modelo de regressão logística foi avaliado por meio do teste de ajuste de Hosmer e Lemeshow (Tabela 4.7), considerando um nível de significância de 5%. O resultado obtido (p = 0.1752) indica que não se rejeita a hipótese nula do teste, a qual postula que não há diferença significativa entre os valores observados e os valores previstos pelo modelo. Portanto, conclui-se que o modelo apresenta bom ajuste aos dados, reforçando sua capacidade explicativa para a não adesão às consultas pré-natais com base nas variáveis selecionadas.

Tabela 4.7: Teste de Hosmer e Lemeshow

Teste	**Estatística** (χ^2)	**G.L.**	**P-valor**
Hosmer-Lemeshow	11.495	8	0.1752

4.3.4 Desempenho do Modelo

A performance do modelo ajustado foi avaliada sobre uma base de teste correspondente a 20% da amostra total, intencionalmente não balanceada, refletindo a distribuição real das classes. Essa abordagem é essencial para estimar a capacidade de generalização do modelo, isto é, a sua habilidade de realizar previsões precisas em dados novos e não utilizados no processo de ajuste.

A matriz de confusão apresentada a seguir (Tabela 4.8) ilustra como o modelo se comporta em condições operacionais reais, oferecendo uma visão realista da sua eficácia prática na identificação de gestantes não aderentes às consultas pré-natais. A partir dessa matriz, foram calculadas as principais métricas de desempenho, cujos resultados estão sintetizados na Tabela 4.9.

Tabela 4.8: Matriz de classificação

	Observado	
Previsto	Não adesão	Adesão
Não adesão	86	10
Adesão	66	44

Tabela 4.9: Métricas do desempenho do modelo

Métrica	Valor
Acurácia	63.11%
Sensibilidade	81.48%
Especificidade	56.58%
VPP (Precisão)	40.00%
VPN	89.58%
F1-Score	52.63%
Kappa	28.52%

Com base nas métricas apresentadas na Tabela 4.9, o modelo demonstrou desempenho razoável na predição da não adesão às consultas pré-natais. A sensibilidade de 81.48% indica que o modelo foi eficaz em identificar a maioria das gestantes que não aderiram ao acompanhamento. A especificidade de 56.58%, embora mais modesta, mostra que o modelo teve desempenho moderado na identificação correta das aderentes. A acurácia global de 63.11% revela um desempenho geral satisfatório.

A precisão (VPP) de 40% indica que, entre os casos classificados como "não adesão", 40% realmente não aderiram, enquanto o VPN elevado (89.58%) sugere que o modelo é bastante confiável ao identificar as gestantes aderentes. Por fim, o índice Kappa de 28.52% revela um acordo moderado entre as previsões do modelo e os dados reais, acima do esperado ao acaso.

Com uma área sob a curva ROC (AUC-ROC) de 73.1%, o modelo demonstrou boa capacidade discriminativa, ou seja, foi capaz de distinguir corretamente, em aproximadamente 73% das vezes, uma gestante que não aderiu das que aderiram às consultas pré-natais, como mostra a Figura 4.4. Esse valor enquadra-se na faixa considerada aceitável à boa e reforça a utilidade preditiva do modelo.

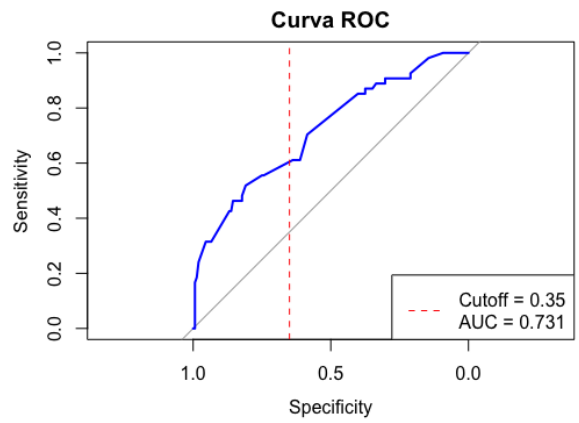

Figura 4.4: Curva ROC.

4.3.5 Validação Cruzada

Para obter uma estimativa mais robusta do desempenho do modelo, foi utilizada validação cruzada do tipo *k-fold* estratificada, com $k = 10$ subdivisões, assegurando-se que cada subconjunto de validação mantivesse aproximadamente a mesma proporção entre as categorias da variável dependente.

A sensibilidade média de 88.8% demonstra que o modelo tem excelente desempenho em identificar corretamente os casos de não adesão. Por outro lado, a especificidade de 42.6% indica limitações na identificação precisa das aderentes. A precisão (VPP) foi de 60.7%, significando que mais da metade das mulheres previstas como não aderentes de facto não aderiram, enquanto o VPN foi de 79.2%, indicando boa confiança nas previsões de adesão. A métrica *F1-score* atingiu 72.1%, revelando um equilíbrio entre sensibilidade e precisão. A acurácia geral foi de 65.7%, com um índice Kappa de 31.4%, que representa um acordo moderado além do acaso.

Tabela 4.10: Validação cruzada $k = 10$ folds

Métrica	**Valor Médio**
Acurácia	66.1%
Sensibilidade	88.8%
Especificidade	43.1%
Precisão (VPP)	61.6%
VPN	79.2%
F1-score	72.4%
Kappa	31.4%
AUC-ROC	76.6%

4.4 Discussão dos Resultados

A secção de resultados deste estudo revelou um conjunto de fatores associados de forma estatisticamente significativa à não adesão às consultas pré-natais no Hospital Geral José Macamo. Entre estes, destacaram-se o baixo nível de escolaridade, a multiparidade, a idade materna, o início tardio do pré-natal e a ausência de parceiro presente como preditores mais fortes do não comparecimento. Na presente discussão, retoma-se cada um desses achados, confrontando-os com a literatura existente.

No modelo de regressão logística múltipla, as mulheres com menor escolaridade apresentaram probabilidades superiores de não comparecerem às consultas pré-natais, resultado que se alinha com achados de Costa et al. (2014) e Baptista (2015). Estes autores identificaram que a escolaridade reduzida limita a compreensão dos benefícios das consultas pré-natais, levando a menor adesão. Da mesma forma, Muleva et al. (2021), ao estudarem puérperas na província de Nampula, observaram que mulheres com ensino secundário ou superior tinham maior probabilidade de iniciar o acompanhamento precocemente e de completar o número mínimo recomendado de consultas.

A multiparidade surgiu como outro fator de risco estatisticamente significativo para a não adesão. Mulheres que já tinham experimentado partos anteriores mostraram-se menos propensas a comparecer regularmente às consultas. Esta tendência foi previamente relatada por Rosa (2013) e Costa et al. (2021), que indicam que a confiança acumulada em experiências anteriores, associada à sobrecarga de cuidados domésticos e familiares, pode dificultar a assiduidade às consultas. Esta conclusão é reforçada na revisão sistemática de Nascimento e Barbosa (2021), que identificaram a multiparidade como factor de risco recorrente em diversos contextos.

Relativamente a idade gestacional no início do acompanhamento, este revelou-se o factor com maior razão de chance no modelo final. Iniciar o pré-natal apenas no segundo ou terceiro trimestre compromete significativamente as possibilidades de intervenção precoce e está associado a realização de menor número de consultas. Estes resultados estão alinhados aos estudos de Felício (2013), Bosschieter et al. (2023), bem como a revisão sistemática de Nascimento e Barbosa (2021), que destacam o impacto negativo do início tardio no seguimento completo do pré-natal.

A ausência de parceiro também surgiu como um factor fortemente associado à não adesão. Este achado está alinhado com os resultados de Hass (2013), Boene et al. (2019), bem como os de Muleva et al. (2021), os quais evidenciam que a coabitação com o parceiro representa um factor protector, ao facilitar o suporte emocional e logístico necessário para a realização periódica das consultas.

A idade maternal também se mostrou estatisticamente significativa no modelo final, indicando que mulheres mais jovens apresentaram maior probabilidade de não adesão as consultas pré-natais.

Este resultado é consistente com estudos prévios, como o de Barbieri et al. (2020), que demonstraram que adolescentes e mulheres com menos de 20 anos têm maior risco de negligenciar o acompanhamento gestacional, muitas vezes por falta de conhecimento, medo do julgamento social ou ausência de suporte familiar. De igual forma, Assis et al. (2022), ao analisarem a reincidência da gravidez na adolescência, concluíram que a juventude esta associada a piores indicadores de cuidado pré-natal.

A variável gravidez desejada não apresentou significância estatística no modelo final. No entanto, literatura relevante, como a de Silva et al. (2021) e Costa et al. (2014), indica que gestações não planeadas tendem a estar associadas a atrasos ou ausências no acompanhamento, particularmente quando acompanhadas de sentimentos de rejeição, insegurança emocional ou negação inicial da gravidez.

Capítulo 5

CONCLUSÕES E RECOMENDAÇÕES

5.1 Conclusões

Este estudo teve como objectivo principal identificar os factores associados à não adesão às consultas pré-natais entre gestantes atendidas no Hospital Geral José Macamo, em Maputo, com base em dados clínicos recolhidos no local. Para tal, foi utilizada a regressão logística binária como modelo de classificação, enquadrada no contexto da aprendizagem supervisionada, permitindo não só identificar variáveis associadas ao fenómeno estudado, como também construir um modelo preditivo com capacidade discriminativa satisfatória.

Os resultados obtidos mostraram que as variáveis idade materna, idade gestacional na primeira consulta, presença de parceiro, paridade e nível de escolaridade estiveram significativamente associadas à não adesão às consultas pré-natais. De modo particular, gestantes mais jovens, multíparas, com escolaridade mais baixa e sem parceiro presente apresentaram maior probabilidade de não realizar o número mínimo de consultas recomendadas. Além disso, iniciar o acompanhamento em fases mais avançadas da gestação revelou-se um factor determinante para o não cumprimento do pré-natal adequado.
A validação cruzada k-fold estratificada assegurou uma avaliação mais robusta da performance do modelo, garantindo maior fiabilidade na generalização dos resultados.

5.2 Recomendações

Com base nos resultados obtidos, recomenda-se:

1. Reforçar estratégias de sensibilização para o início precoce do pré-natal, sobretudo entre adolescentes e jovens adultos, através de campanhas de educação em saúde específicas para esta faixa etária.

2. Implementar mecanismos de rastreio e acompanhamento personalizado para gestantes que iniciam o pré-natal após o primeiro trimestre, garantindo maior suporte para completar o número mínimo de consultas.

3. Aperfeiçoar os sistemas de registo clínico e informação hospitalar, permitindo capturar um maior número de variáveis socioeconómicas e contextuais relevantes, facilitando análises mais abrangentes e precisas.

4. Que sejam desenvolvidos estudos em cada província ou região do pais para analisar detalhadamente quais variáveis estão associadas a não adesão ás consultas em cada região, de modo que sejam tomadas medidas adequadas para cada região.

5. Além disso, recomenda-se que estudos futuros explorem métodos mais avançados de aprendizagem supervisionada, os quais poderão capturar relações não lineares e interacções mais complexas entre as variáveis.

5.3 Limitações

Os dados utilizados foram recolhidos a partir dos registos clínicos físicos disponíveis nos livros do Hospital Geral José Macamo, o que implica restrições quanto à completude, padronização e qualidade da informação.
A ausência de registos electrónicos dificultou a automatização da recolha e limitou o acesso a variáveis complementares que poderiam enriquecer a análise, como rendimento familiar, distância até à unidade de saúde, meio de transporte utilizado ou apoio familiar.

REFERÊNCIAS

[1] Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Wiley.

[2] Agresti, A. (2015). *Foundations of linear and generalized linear models*. John Wiley & Sons.

[3] Agresti, A., & Finlay, B. (2009). *Statistical methods for the social sciences* (4th ed.). Prentice Hall.

[4] Alvarenga, A. (2015). *Modelos lineares generalizados: Aplicação a dados de acidentes rodoviários* [Dissertação de Mestrado]. Lisboa, Portugal.

[5] Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys, 4*, 40–79.

[6] Assis, G. M. P., Vasconcelos, I. C. F., & Lima, L. S. (2022). Reincidência da gravidez na adolescência: uma análise sobre o pré-natal e seus desfechos. *Revista Enfermagem Atual In Derme, 96*(40), e022041.

[7] Barbieri, A. R., Couto, M. T., & Azevedo, R. C. S. (2020). Fatores associados à inadequação do pré-natal em uma coorte de gestantes de Minas Gerais. *Revista Mineira de Enfermagem, 24*, e-1304.

[8] Baptista, A. M. S. (2015). *Regressão logística: Uma introdução ao modelo estatístico – Exemplo de aplicação ao revolving credit*. Vida Económica Editorial.

[9] Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter, 6*(1), 20–29.

[10] Belfiore, P. (2015). *Estatística aplicada à administração, contabilidade e economia com Excel e SPSS* (1ª ed.). Rio de Janeiro.

[11] Benova, L., Tunçalp, Ö., Moran, A. C., & Campbell, O. M. R. (2018). Not just a number: Examining coverage and content of antenatal care in low-income and middle-income countries. *BMJ Global Health, 3*(2), e000779.

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

[13] Bosschieter, T. M., Xu, Z., Lan, H., Lengerich, B. J., Nori, H., Painter, I., Souter, V., & Caruana, R. (2023). Interpretable predictive models to understand risk factors for maternal and fetal outcomes. *Journal of Healthcare Informatics Research, 8*(1), 65–87.

[14] Boene, H., Vidler, M., Sacoor, C., Nhama, A., Nhacolo, A., Bique, C., & Sevene, E. (2019). Determinants of antenatal care utilization in rural Mozambique: A cross-sectional study. *PLOS ONE, 14*(4), e0214847.

[15] Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. Springer.

[16] 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.

[17] Costa, M. C. N., Leal, M. C., Esteves-Pereira, A. P., Ayres, B. V., & Sánchez, A. R. A. (2014). Fatores associados à não realização de pré-natal em município de grande porte. *Revista de Saúde Pública, 48*(6), 977–984.

[18] Cox, D. R., & Snell, E. J. (1989). *The analysis of binary data* (2nd ed.). Chapman and Hall.

[19] Domingues, R. M. S. M., Leal, M. C., Hartz, Z. M. A., & Dias, M. A. B. (2012). Avaliação da adequação da assistência pré-natal na rede SUS do Município do Rio de Janeiro, Brasil. *Cadernos de Saúde Pública, 20*(S1), S39–S49.

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

[21] Emiliano, P. C., Vivanco, M. J. F., Menezes, F. S. M., & Avelar, F. G. (2009). Fundamentos e comparação de critérios de informação: Akaike and Bayesian. *Revista Brasileira de Biomatemática, 27*(3), 394–411.

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

[23] Fávero, L. P., & Belfiore, P. (2017). *Análise de dados: Estatística e modelação multivariada com Excel, SPSS e Stata*. Rio de Janeiro: Campus Elsevier.

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

[25] Felício, L. S. (2013). *Fatores associados ao absenteísmo às consultas pré-natais do SUS em Aracruz – ES* [Dissertação de Mestrado]. Universidade Federal do Rio Grande do Sul.

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

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

[28] Gil, A. C. (2008). *Métodos e técnicas de pesquisa social* (6.ª ed.). São Paulo: Atlas.

[29] Gujarati, D. N. (2011). *Econometria Básica* (5.ª ed.). AMGH Editora.

[30] Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (2005). *Análise multivariada de dados* (5.ª ed.). Bookman.

[31] Hass, C. (2013). Adequabilidade da assistência pré-natal em uma estratégia de saúde da família de Porto Alegre-RS. Brasil.

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

[33] Hosmer, D. W., & Lemeshow, S. (1989). *Applied Logistic Regression*. John Wiley & Sons.

[34] Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). Wiley.

[35] Instituto Nacional de Saúde. (2022). *Relatório do Inquérito Demográfico e de Saúde (IDS) 2022–2023*. Maputo.

[36] Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5), 429–449.

[37] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R*. Springer.

[38] Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer.

[39] Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174.

[40] Little, R. J. A., & Rubin, D. B. (2019). *Statistical Analysis with Missing Data* (3rd ed.). Wiley.

[41] Martins, J. M. (2019). Factores associados a não utilização da consulta pós-natal por crianças lactantes em Moçambique. Universidade Nova de Lisboa, Instituto de Higiene e Medicina Tropical.

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

[43] Menard, S. (2000). *Applied Logistic Regression Analysis* (2nd ed.). Sage Publications.

[44] Ministério da Saúde & Instituto Nacional de Saúde. (2021). *Normas para atenção pré-natal e cuidados pós-natal para mulheres e recém-nascidos*. Maputo.

[45] Ministério da Saúde. (2018). *Relatório anual do balanço do sector de saúde*.

[46] Ministério da Saúde & Instituto Nacional de Estatística. (2023). *Inquérito Demográfico e de Saúde 2022–2023: Relatório preliminar*. Maputo, Moçambique: INE.

[47] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis* (5th ed.). Wiley.

[48] Moura, M. R. de. (2018). *Modelos lineares generalizados: Aplicações em R* [Dissertação de mestrado, Universidade Estadual Paulista]. Repositório Institucional UNESP.

[49] Muleva, B. R., Duarte, L. S., Silva, C. M., Gouveia, L. M. R., & Borges, A. L. V. (2021). Assistência ao pré-natal em Moçambique: Número de consultas e idade gestacional no início do pré-natal. *Revista Latino-Americana de Enfermagem*, 29, e3481.

[50] Nascimento, J. W. A., & Da Silva Barbosa, L. M. (2021). Principais fatores associados ao tardiamento do pré-natal: uma revisão sistemática. *Research, Society and Development*, 10(4).

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

[52] Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, 135, 370–384.

[53] Organização Mundial da Saúde. (2018). *Recomendações sobre cuidados pré-natais para uma experiência positiva na gravidez.*

[54] Pestana, M. H., & Gageiro, J. N. (2014). *Análise de Dados para Ciências Sociais – A Complementaridade do SPSS* (6ª ed.). Edições Sílabo.

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

[56] Prodanov, C. C., & Freitas, E. C. (2013). *Metodologia do trabalho científico: Métodos e Técnicas da Pesquisa e do Trabalho Académico* (2.ª ed.). Feevale.

[57] Reza, M. M., & Salma, A. (2024). Predicting and feature selection of low birth weight using machine learning algorithms. *J Health Popul Nutr*.

[58] Richardson, R. J. (2017). *Pesquisa social: métodos e técnicas* (4.ª ed.). Atlas.

[59] Rosa, C. (2013). Fatores associados à não realização do pré-natal no município de Pelotas. Rio de Janeiro.

[60] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464.

[61] Silva, F. G. R., Souza, K. V., & Rocha, M. S. (2021). Gravidez não planejada: Implicações para a saúde da mulher e do recém-nascido. *Revista Brasileira de Saúde Materno Infantil*, 21(1), 145–153.

[62] Silveira, M. F., Victora, C. G., Barros, A. J. D., & Santos, I. S. (2020). Adequação da assistência pré-natal no Brasil: Revisão sistemática. *Revista de Saúde Pública*, 54, 8.

[63] Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.

[64] Viellas, E. F., Domingues, R. M. S. M., Dias, M. A. B., Gama, S. G. N., Theme Filha, M. M., Costa, J. V., & Leal, M. C. (2014). *Assistência pré-natal no Brasil*. Cadernos de Saúde Pública.

[65] Zhang, Y., Gu, X., Yang, N., Xue, Y., Ma, L., Wang, Y., Zhang, H., & Jia, K. (2025). Modelos de predição para pré-eclâmpsia de início tardio: Um estudo baseado em regressão logística, máquina de vetores de suporte e modelos de gradiente extremo. *Biomedicines*, 13(2), 347.

REPÚBLICA DE MOÇAMBIQUE

CIDADE DE MAPUTO

CONSELHO DOS SERVIÇOS DE REPRESENTAÇÃO DO ESTADO

SERVIÇO DE SAÚDE DA CIDADE

Á

Universidade Eduardo Mondlane

Maputo

N/Ref. n.2108/SSCM/050°/2024 Data:13 de Maio de 2024

ASSUNTO: Resposta ao pedido de autorização para desencadear Estudo " Análise dos Factores associados a não adesão as consultas pré-natais no Hospital Geral José Macamo"

O Serviço de Saúde da Cidade de Maputo acusa a recepção do pedido Sra. Ivone Pedro Ussivane, estudante do curso de Licenciatura em Estatística na Universidade Eduardo Mondlane, com o teor retro -mencionado.

Sobre a matéria, comunica-se que o Serviço de Saúde da Cidade de Maputo (SSCM) autoriza a realização da actividade, devendo apresentar os resultados no SSCM.

Sem mais de momento, queiram aceitar as nossas calorosas saudações.

Maputo, 13 de Maio de 2024

A Directora

Dra. Sheila Márcia Tajú Lobo de Castro

(Médica de Clínica Geral Principal)

Cc. Sra. Ivone Pedro Ussivane
Hospital Geral José Macamo

REPÚBLICA DE MOÇAMBIQUE
MAPUTO
Entrada 1044
28 / 05 de 24
Rubrica

Endereço: Serviço de Saúde da Cidade de Maputo
C.P. 2217
Av. Maguiguana nº 1240
E-mail: dscm.gabdirector@gmail.com

Telefone: 21-360276/7
Telefax: 21-048658/ 21-430212
MAPUTO - República de Moçambique