UNIVERSIDADE
EDUARDO
MONDLANE

**FACULDADE DE CIÊNCIAS**
**Departamento de Matemática e Informática**

Trabalho de Licenciatura em
Estatística

# Análise dos Factores de Risco Associados à Infecção Pelo Papiloma vírus Humano no Rastreio do Cancro do Colo do Útero no Primeiro Trimestre de 2024. Caso de estudo: Comunidade Sant'Egídio, Cidade de Maputo

**Autor:** Jorge Sidumo Patrício

Maputo, Março de 2025

UNIVERSIDADE
EDUARDO
MONDLANE

**FACULDADE DE CIÊNCIAS**
**Departamento de Matemática e Informática**

Trabalho de Licenciatura em
Estatística

# Análise dos Factores de Risco Associados à Infecção Pelo Papiloma vírus Humano no Rastreio do Cancro do Colo do Útero no Primeiro Trimestre de 2024. Caso de estudo: Comunidade Sant'Egídio, Cidade de Maputo

**Autor:** Jorge Sidumo Patrício

**Supervisor:** Miranda Albino Martins Muaualo, PhD, UFRJ
**Co-Supervisor:** Elton Uamusse, MSc, UEM

Maputo, Março de 2025

# Declaração de Honra

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

Maputo, 19 de Março de 2025

---

(Jorge Sidumo Patrício)

# Dedicatória

Aos meus pais Ramiro Ofinar Patrício e Ana Lúcia Jorge Sidumo Patrício.
Em especial aos meus irmãos Renan Patrício e Patrick Patrício.
Em memória do meu tio Dario Filipe Jorge Sidumo.

*Não são as coisas que ignoramos que nos atrapalham, mas as que conhecemos. Em Deus confiamos, todos os outros devem trazer evidências.*
**Artemus Ward**

# Agradecimentos

Primeiramente, quero agradecer a Deus, à minha família, em particular aos meus pais **Ramiro Ofinar Patrício e Ana Lúcia Jorge Sidumo Patrício**, aos meus irmãos **Renan Patrício e Patrick Patrício**, aos meus tios **Amélia Sidumo, Zita Sidumo, Quito Patrício** e aos meus primos **Elias, Joyce, Léria e Káren**. Pessoas estas que acompanharam o decorrer desta jornada com zelo e perseverança.

Ao meu avô **Jorge Lázaro Sidumo**, pessoa com a qual partilho do mesmo nome por servir de fonte de referência como ser humano e ter desempenhado o papel de uma biblioteca viva para mim na busca do saber.

Quero expressar o meu enorme apreço à minha companheira **Nayara Rodrigues** com quem tenho compartilhado desafios e conquistas de vida.

Desde já ao corpo docente do DMI, em particular os docentes da área de Estatística, o meu enorme agradecimento, em especial ao meu supervisor **Miranda Muaualo** pela paciência e apoio demonstrados ao longo deste trabalho. Ao Professor **Osvaldo André Loquiha** pela pronta resposta quando a ele apresentava alguma inquietação referente ao processamento dos dados e também por servir de inspiração como profissional da área.

Agradeço ao médico **Elton Uamusse** pelo acompanhamento durante este processo, pela atenção e dedicação prestada dia após dia, graças a ele este momento de aprendizagem na área de biologia e saúde tornou-se mais eficiente.

À minha família agradeço pelo acolhimento na capital do país durante estes anos de formação superior. Aos meus amigos **Aximenes Joaquim Marcos, António Matsimbe e Gerson Macuácua** por terem fortalecido e tornado esta jornada de aprendizado muito mais agradável.

Aos meus colegas **Eduardo Sicuaio, Benigna Novela, Jaime Nhampule, Cleyton Cossa, Anísio Osias**, o meu muito obrigado pelo empenho demonstrado para um sucesso individual porém sem nunca deixar de lado o colectivo abraçando a causa de companheirismo e trabalho em equipa.

O meu muito obrigado a todos que directa ou indirectamente terão contribuído para a minha edificação como homem e ser humano.

# Resumo

O cancro do colo do útero é uma das doenças mais comuns entre mulheres no mundo e a sua prevalência tem uma forte ligação com a infecção pelo papilomavírus humano (HPV). É crucial identificar factores de risco específicos à infecção pelo HPV na população da cidade de Maputo, ajudando a orientar políticas de saúde pública e intervenções preventivas. Com o objectivo de analisar os factores de risco associados à infecção pelo HPV no rastreio do cancro do colo do útero em mulheres atendidas nos centros DREAM da comunidade Sant'Egídio na cidade de Maputo no primeiro trimestre de 2024. Para o alcance do objectivo adopta-se a imputação de valores perdidos para lidar com dados incompletos e a regressão logística para a estimação de parâmetros que indiquem uma relação entre variáveis socioeconómicas, comportamentais e biológicas com a infecção por HPV. Durante o estudo, foram identificados factores de risco como idade da primeira relação sexual, ciclo menstrual, estado HIV positivo, e uso de métodos de planeamento familiar como significativos para a infecção por HPV. Os factores aqui obtidos condizem com os encontrados na literatura sugerindo uma necessidade de implementar-se assim programas educacionais sobre práticas sexuais seguras, melhoria no acesso ao rastreamento e tratamento, suporte contínuo para mulheres HIV positivas e monitoramento contínuo de factores de risco.



**Palavras-chave**: *Papilomavírus humano, cancro do colo do útero, HIV, factores de risco, regressão logística*.

# Abstract

Cervical cancer is one of the most common diseases among women in the world and its prevalence is strongly linked to human papillomavirus (HPV) infection. It is crucial to identify specific risk factors for HPV infection in the population of Maputo city, helping to guide public health policies and preventive interventions. The aim of this study was to analyse the risk factors associated with HPV infection in cervical cancer screening in women attending DREAM centres in the Sant'Egídio community in Maputo city in the first quarter of 2024. To achieve the objective, lost value imputation was adopted to deal with incomplete data and logistic regression was used to estimate parameters that indicate a relationship between socioeconomic, behavioural and biological variables with HPV infection. During the study, risk factors such as age at first sexual intercourse, menstrual cycle, HIV-positive status, and use of family planning methods were identified as significant for HPV infection. The factors obtained here are consistent with those found in the literature, suggesting a need to implement educational programmes on safe sex practices, improved access to screening and treatment, ongoing support for HIV-positive women and continuous monitoring of risk factors.



**Keywords**: *Human papillomavirus, cervical cancer, HIV, risk factors, logistic regression.*

# Lista de Abreviaturas

**AIC** - Critério de Informação Akaike
**AUC** - Área Sob a Curva (Area Under the Curve)
**BIC** - Critério de Informação Bayesiana.
**Cm** - centímetros
**DC** - Doenças Crónicas
**DCNT** - Doenças Crónicas não Transmissíveis
**DMI** - Departamento de Matemática e Informática
**DREAM** - Disease Relief through Excellent and Advanced Means ( Alívio de Doenças Através de Meios de Excelência e Avançados)
**DNA/ADN** - Ácido Desoxirribonucleico
**FN** - Falso negativo
**FP** - Falso positivo
**GL** - Grau de Liberdade
**H0** - Hipótese Nula
**H1** - Hipótese Alternativa
**HCM** - Hospital Central de Maputo
**HIV/VIH** - Vírus de imunodeficiência humana
**HPV** - Papilomavírus Humano
**IC** - Intervalo de Confiança
**Kg** - Quilograma
**LL** - Log Likelihood (Log de probabilidade)
**MISAU** - Ministério da Saúde
**MLG** - Modelos Lineares Generalizados
**OMS** - Organização Mundial da Saúde
**ORFs** - Open Reading Frame (Quadro de Leitura Aberto)
**RC** - Razão de Chance
**RL** - Regressão Logística
**ROC** - Receiver Operating Characteristic
**SPSS** - Statistical Package for the Social Sciences
**VIF** - Factor de Inflação da Variância
**VN** - Verdadeiro negativo
**VP** - Verdadeiro positivo

# Índice

# Lista de Figuras

# Lista de Tabelas

# Capítulo 1

# INTRODUÇÃO

## 1.1 Contextualização

As doenças crónicas (DC) são um problema de saúde pública que têm surgido com maior frequência a partir do estilo de vida. Neste cenário, as DC fazem parte de um grupo de doenças resultantes de diversas causas, representadas maioritariamente por factores comportamentais que levam um certo tempo até expressarem os seus efeitos (Figueiredo *et al.* 2021).

Entre as Doenças Crónicas Não Transmissíveis (DCNT) mais comuns destacam-se as doenças cardiovasculares, cancros, diabetes mellitus (DM) e doenças respiratórias com aproximadamente 33,2 milhões de mortes anualmente no mundo (Estrela *et al*., 2020). Este problema tornou-se mais frequente durante a pandemia da COVID-19, um período em que as pessoas com doenças crónicas foram responsáveis por 70% das mortes na fase grave da doença. Assim, as condições crónicas foram estabelecidas devido às formas graves da infecção por SARS-CoV-2, resultando em sequelas substanciais.

O cancro do colo do útero é o quarto cancro mais comum nas mulheres em todo o mundo, com cerca de 660 000 novos casos e cerca de 350 000 mortes em 2022 (OMS, 2024a). As taxas de incidência e mortalidade mais elevadas registaram-se em países de baixa e média renda.

Em Moçambique, de acordo com o Ministério da Saúde (2020), o cancro do colo do útero representou cerca de 32 em cada 100 novos casos de todos os cancros em mulheres, representando 3690 casos com 2356 mortes por ano. Isto significa que 64 em cada 100 mulheres que contraem cancro do colo do útero morrem por terem sido diagnosticadas numa fase avançada da doença.

Na Cidade de Maputo, segundo o Ministério da Saúde (2013), dados dos Serviços de Anatomia Patológica (SAP) do HCM de 1991 a 2008 e 2009 a 2010 na cidade de Maputo, mostram

que nas mulheres os cancros mais frequentes são o cancro do colo do útero 31%, seguido do cancro da mama 10% e do sarcoma de Kaposi 7%.

## 1.2 Definição do problema

Segundo os dados da OMS, (2024b), à infecção pelo papilomavírus humano (HPV) é causadora de 95% dos casos do cancro do colo do útero. A prevalência do HPV é maior entre mulheres que vivem com HIV, indivíduos imunocomprometidos, pessoas com co-infecção com outras infecções sexualmente transmissíveis (IST), pessoas que recebem medicamentos imunossupressores e crianças que foram abusadas sexualmente.

A maioria das mortes por cancro do colo do útero associadas a presença do HPV, pode estar associada ao consumo de tabaco, uma alimentação inadequada, inatividade física e ao consumo excessivo de bebidas alcoólicas, entre outros factores. Os factores de riscos importantes relacionados com doenças crónicas estão ligados ao estilo de vida e podem ser sensíveis a intervenções de prevenção e promoção da saúde (OMS, 2024a). Isto reflecte a padrões de saúde, associados a factores de risco, incluindo a prevalência do vírus da imunodeficiência humana (HIV), determinantes sociais e económicos.

Sendo o HPV uma infecção sexualmente transmissível comum que pode afectar a pele, a zona genital e a garganta, é provável que quase todas as pessoas sexualmente activas sejam infectadas pelo vírus em algum momento da vida, mas podem permanecer assintomáticas. Na maioria dos casos, o sistema imunológico elimina o HPV do corpo, mas a infecção persistente com HPV de alto risco pode causar o desenvolvimento de células anormais, que se transformam em cancro.

As estratégias para reduzir o aparecimento e o agravamento de algumas destas condições do cancro do colo do útero incluem a detecção precoce, aumento da actividade física, redução do consumo de tabaco e álcool, bem como incentivar uma alimentação saudável.

Relacionado com a incerteza na determinação dos possíveis factores que podem causar mortes por cancro do colo do útero, associados a presença do HPV nas mulheres atendidas nos centros *Disease Relief through Excellent and Advanced Means* (DREAM) na comunidade de Sant'Egidio na cidade de Maputo, surgiu a necessidade de se realizar pesquisas que possam reduzir esta incerteza, o que motivou o presente estudo com a seguinte questão de pesquisa: Quais são os factores de risco associados à infecção pelo papilomavírus humano nas mulheres atendidas no rastreio do cancro do colo do útero nos centros DREAM da comunidade Sant'Egídio na cidade de Maputo?

## 1.3 Objectivos

### 1.3.1 Objectivo Geral

Analisar os factores de risco associados à infecção pelo papilomavírus humano nas mulheres atendidas no rastreio do cancro do colo do útero nos centros DREAM da comunidade Sant'Egídio na cidade de Maputo

### 1.3.2 Objectivos Específicos

- Descrever o perfil das mulheres atendidas nos centros DREAM da comunidade Sant'Egídio na cidade de Maputo;
- Identificar os factores de risco associados à infecção pelo papilomavírus humano nas mulheres atendidas no rastreio do cancro do colo do útero nos centros DREAM da comunidade Sant'Egídio na cidade de Maputo;
- Estimar os coeficientes do modelo de regressão logística para descrever a probabilidade de uma mulher ter como resultado positivo ou resultado negativo para o papilomavírus humano no rastreio do cancro do colo do útero nas mulheres atendidas nos centros DREAM da comunidade Sant'Egídio na cidade de Maputo;
- Verificar se a imputação de valores perdidos nos dados tem algum impacto na explicação dos factores de risco à infecção pelo papilomavírus humano no rastreio do cancro do colo do útero nas mulheres atendidas nos centros DREAM da comunidade Sant'Egídio na cidade de Maputo.

## 1.4 Justificação

Dada a prevalência e ocorrência regular do cancro do colo do útero aliado a presença do HPV nas mulheres da cidade de Maputo, há uma necessidade de verificar se atitudes comportamentais podem estar por detrás da ocorrência regular desta infecção e consequentemente deste tipo de cancro resultante. A necessidade de verificar de forma rotineira os padrões de saúde das mulheres numa determinada sociedade, especificamente as mulheres da cidade de Maputo, está na base da escolha deste tema. Esta necessidade está associada a factores como os socioeconômicos, socioambientais e sociodemográficos.

Com a realização deste estudo, será possível a posteriori possibilitar ao governo e entidades competentes a criação de diretrizes para a definição de novas políticas de gestão e controlo de doentes com cancro do colo do útero e doentes crónicos como um todo, ou cidadãos vulneráveis

à ocorrência destas doenças, melhorando assim este serviço público e a qualidade de vida dos munícipes da cidade de Maputo, entre outros benefícios.

## 1.5 Estrutura do trabalho

Estruturalmente, o seguinte trabalho está dividido em 5 capítulos, que respectivamente apresentam a parte introdutória sobre o cancro do colo do útero e o HPV, objectivos da pesquisa, relevância, revisão de literatura que apresenta a fundamentação teórica que embalsa os conceitos a serem abordados no decorrer do trabalho com base na literatura ou artigos científicos publicados relacionados ao tema, material e métodos utilizados para o desenvolvimento do trabalho, desde a definição do instrumento de recolha de dados, e métodos de estudo a aplicar, a parte dos resultados que engloba os principais pontos obtidos relativamente ao tratamento da informação recolhida com base nos métodos estabelecidos, e por fim as conclusões retiradas sobre o trabalho, conclusões essas que remetem para os objectivos definidos.

# Capítulo 2

# REVISÃO DA LITERATURA

## 2.1 Doenças Crónicas

De acordo com Figueiredo *et al*. (2021), as doenças crónicas são doenças causadas por vários factores, incluindo factores comportamentais que levam um certo tempo para manifestar suas consequências.

Para Jamison (2006), as doenças crónicas são sequelas resultantes de patologias cardiovasculares, respiratórias, genito-urinárias, reumatológicas, endocrinológicas, digestivas, neurológicas e psiquiátricas, bem como de outras situações que provocam incapacidade precoce ou uma redução significativa da esperança média de vida.

Segundo Victorino (2012), as doenças crónicas são aquelas que habitualmente se desenvolvem lentamente, duram longos períodos de tempo, normalmente mais de 2 meses, e têm efeitos a longo prazo difíceis de prever. Segundo Victorino, seis atributos estão presentes na condição de doença crónica, que permitem reconhecer os efeitos sobre a pessoa, a família e a comunidade:

- Natureza de longa duração;
- Às vezes causam incapacidade;
- Requer maior esforço paliativo ou seja para que o doente atinja o nível desejado de autocuidado, ele necessita do apoio da família e da comunidade;
- Favorece o aparecimento de múltiplas doenças;
- Necessitam de monitoria contínua;
- Requerem muito acompanhamento de serviços (saúde/social) além de ser onerosa.

**Classificação das Doenças Crónicas**

De acordo com a OMS (2011), as doenças crónicas dividem-se em dois grupos principais, as **doenças crónicas não Transmissíveis** como a hipertensão, diabetes, cancro, epilepsia, doenças respiratórias crónicas e outras doenças cardiovasculares; **doenças crónicas Transmissíveis** especialmente a tuberculose, lepra, HIV e a SIDA.

Segundo Goulart (2011), as doenças crónicas não transmissíveis (DCNT) são hoje responsáveis pela maioria das doenças e mortes em muitos países, sejam eles de alto, médio ou baixo nível socioeconómico.

As características gerais das principais DCNT do ponto de vista clínico e de impacto populacional, são apresentadas a seguir:

Tabela 2.1: Características gerais das principais DCNT

| |
|---|
| DOENÇAS CARDIOVASCULARES (DCV) |
| Doenças do coração e vasos sanguíneos, incluindo variadas condições derivadas de suprimento sanguíneo diminuído a diversos órgãos do corpo. Cerca de 80% da mortalidade diz respeito três condições deste grupo, a saber: doença coronariana isquêmica (infarto do miocárdio), acidente vascular cerebral, doença hipertensiva e insuficiência cardíaca congestiva. Ao longo da última década as DCV se tornaram as principais causas de mortalidade em todo o mundo, representando cerca de 30% de todas as mortes e até 50% da mortalidade pelo conjunto das DCNT. As DCV sozinhas causam 17 milhões de mortes e 50 milhões de DALY no mundo. Factores de risco de fundo comportamental bem conhecidos e definidos, como uso de tabaco, inatividade física e alimentação pouco saudável explicam perto da 80% da carga total de DCV. |
| CANCRO |
| Multiplicação anormal de células em determinados órgãos do corpo, afectando as células normais e produzindo novos focos invasivos à distância, conhecidos como metástases. Existem mais de 100 tipos de cancro, com factores de risco igualmente múltiplos. O cancro é a segunda principal causa de morte no mundo, sendo responsável por 13% do total, ou seja cerca de oito milhões de mortes por ano. Estudos recentes mostram que quase 13 milhões de novos casos de cancro todos os anos e que este número atingirá 17 milhões até ao final da presente década |
| DOENÇAS RESPIRATÓRIAS CRÓNICAS |
| Doenças de natureza crónica que afectam as vias aéreas e também outras estruturas dos pulmões. As mais comuns são: asma, doença pulmonar obstrutiva crónica (DPOC), estados alérgicos, hipertensão pulmonar, além de algumas doenças relacionados ao processo de trabalho. Juntas, elas representam cerca de 7% da mortalidade global, causando 4,2 milhões de óbitos anuais. Somente a DPOC, associada geralmente ao hábito de fumar, além de outras causalidades, afecta mais de 200 milhões de pessoas em todo o mundo, representando de 4 a 8% das mortes nos países mais ricos e até mais do que isso nos mais pobres. |
| DIABETES |
| O diabetes é uma doença de fundo metabólico na qual existe, por parte do organismo, incapacidade total ou parcial de retirar a glicose (além de outras substâncias) do sangue e levá-las para dentro das células, provocando e mantendo níveis sanguíneos altos dessas substâncias. A não regulação da glicose no sangue dos diabéticos tem como causa a baixa sensibilidade ou a pouca produção da insulina , que é o hormônio natural dotado de tal função, no pâncreas. O tipo 2 do diabetes, que acomete pessoas mais velhas, é o mais frequente, responsabilizando-se por mais de 90% dos casos. O diabetes, em si, não tem mortalidade elevada, quando comparado a outras DCNT (1,3 milhões de mortes no mundo), mas constitui um importante factor de risco e de disfunção (comorbidade) para outras condições mais graves, tais como, as DCV, insuficiência renal e a cegueira. |
| DOENÇAS MENTAIS |
| Trata-se de um termo genérico, que designa condições variadas que afectam as atitudes, o pensamento, os sentimentos, além a capacidade de se relacionar socialmente. Elas afectam centenas de milhões de pessoas em todo mundo. No início da presente década, a depressão, por exemplo, atingia cerca de 150 milhões de pessoas em todo o mundo; 25 milhões sofriam de esquizofrenia; mais de 100 milhões apresentavam abuso de álcool e drogas. Além disso, estima-se que perto de um milhão de pessoas se suicidam a cada ano. As doenças mentais contribuem fortemente para os anos de vida perdidos por incapacidade (DALY) em todo o mundo, com cifras estimadas em 13%, no ano de 2004. Existem evidências de comorbidade entre as doenças mentais, a diabetes e as DCV |

**Fonte:** Goulart (2011).

## 2.2 Cancro do colo do útero

De acordo com De Oliveira *et al.* (2019), o cancro é um tumor que pode aparecer em qualquer parte do corpo.

Ainda na perspectiva dos autores acima, o cancro do colo do útero é uma neoplasia caracterizada pela replicação desordenada de células anormais, afectando o epitélio de revestimento do colo do útero, sendo a ectocérvice a principal área afectada. O carcinoma de células escamosas é responsável por cerca de 90% dos casos de cancro do colo do útero. Outra neoplasia importante é o adenocarcinoma, que afecta o endocérvix nas células glandulares e é responsável por cerca de 10% dos casos.

Segundo Matos (2008), o cancro do colo do útero é a segunda neoplasia maligna mais frequente nas mulheres a nível mundial. Cerca de 80% dos casos ocorrem em países em desenvolvimento, onde em muitas regiões é o cancro mais comum nas mulheres.

Segundo o Ministério da Saúde (2020), a nível mundial, o cancro do colo do útero afecta mais de 570 000 mulheres e é responsável por cerca de 311 000 mortes por ano. A maioria das mortes ocorre em países de baixo e médio rendimento, como é o caso de Moçambique.

De acordo com Jamison (2006), o cancro do colo do útero é o principal cancro nas mulheres da África subsariana, com uma estimativa de 70.700 novos casos em 2002 (o total para todo o continente foi de 78 900 casos), o que representou 13,3% de todos os cancros em mulheres adultas, equivalente a uma incidência anual ajustada à idade de 33 novos casos por 100 000 habitantes.

De acordo com Jamison (2006), a incidência do cancro do colo do útero difere de acordo com variáveis demográficas clássicas, como a classe social, o estado civil, a etnia e a religião. Posteriormente, estudos epidemiológicos, principalmente estudos de caso-controlo, mostraram uma associação consistente entre o risco e a idade precoce de início da atividade sexual, o aumento do número de parceiros sexuais das mulheres ou dos seus parceiros sexuais e outros indicadores de comportamento sexual. Estes resultados sugeriam fortemente um papel causal de um agente sexualmente transmissível, um agente mais tarde identificado como o vírus do papiloma humano (HPV).

## 2.3 Papilomavírus Humano

Segundo Andersson (2009), em 1974, Harald Zur Hausen, um médico alemão, publicou o seu primeiro relatório tentando encontrar ADN do HPV no cancro do colo do útero e nas verrugas genitais. A teoria de Harald Zur Hausen sobre a etiologia do cancro do colo do útero pelo HPV foi apoiada por trabalhos posteriores de Meisels e Fortin, que descreveram células coilocitóticas atípicas como uma manifestação de alteração citopática induzida pelo vírus do papiloma na displasia do colo do útero. A identificação subsequente de partículas semelhantes ao vírus do papiloma nessas células, por microscopia eletrónica, apoiou ainda mais a ideia. A equipa de Harald Zur Hausen utilizou posteriormente ADN purificado de partículas virais de diferentes verrugas plantares para gerar sondas que permitiram a deteção de padrões distintos de clivagem por enzimas de restrição em isolados de HPV de muitos doentes. Isto levou à identificação de múltiplas estirpes de HPV 1 a 3.

De acordo com Zur Hausen (2002), os primeiros tipos de HPV foram isolados directamente de biópsias de cancro do colo do útero especificamente o HPV16 e HPV18 foram clonados em 1983 e 1984, respectivamente. Os papilomavírus estavam etmologicamente envolvidos neste tipo de cancro e a infecção por mais de um tipo de HPV poderia resultar em cancro do colo do útero. Finalmente ele demonstrou que partes do ADN do HPV estavam integradas no hospedeiro genoma em linhas celulares de cancro do colo do útero, incluindo os genes virais HPV16 e 18 que foram preferencialmente retidos nos tumores.

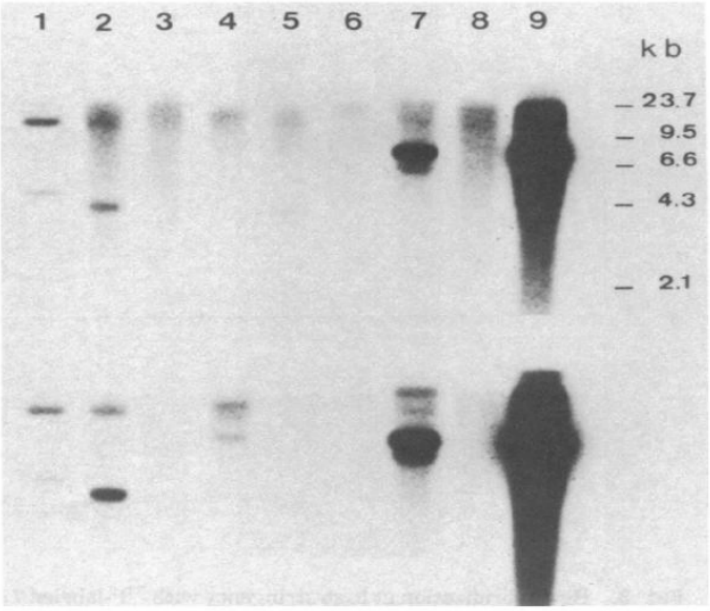


Figura 2.1: Hibridização por transferência com sonda de ADN de HPV16 marcada
**Fonte:** Andersson (2009).

Na Figura 2.1, os ADNs celulares preparados a partir de crescimento invasivo carcinomas cervicais (pistas 2, 4, 5, 7 e 9), uma displasia (pista 6), dois carcinomas *in situ* do colo do útero

(pistas 1 e 5) e um carcinoma vulvar (pista 3) foram clivados por BamHl. Hibridização com baixo rigor (superior) e alto rigor (inferior). A hibridização positiva com ADN do HPV 16 foi evidente nas amostras 1, 2, 4, 5, 7 e 9. A reação positiva em 2, 4 e 7 foi observada mais claramente após a remoção do fundo inespecífico em alta temperatura.

Bosch FX *et al*. (2002), definem o HPV como a infecção viral mais comum do trato reprodutivo e causa uma série de condições em homens e mulheres, incluindo lesões pré-cancerosas que podem evoluir para o cancro. Embora a maioria das infecções por HPV sejam assintomáticas e se resolvam espontaneamente, a infecção persistente por HPV pode resultar em doença. Nas mulheres, a infecção persistente por tipos oncogênicos de HPV pode levar à neoplasia intraepitelial cervical (NIC) que se não tratado, pode evoluir para cancro do colo do útero invasivo.

Segundo Mondiale de la Santé (2022), o HPV pertence à família Papillomaviridae. Os viriões são não envelopados e contêm um genoma de ADN de cadeia dupla. O material genético é rodeado por um capsídeo icosaédrico composto por proteínas estruturais maiores e menores, L1 e L2, respetivamente. Os vírus são altamente específicos dos tecidos e infectam a pele e o epitélio das mucosas.

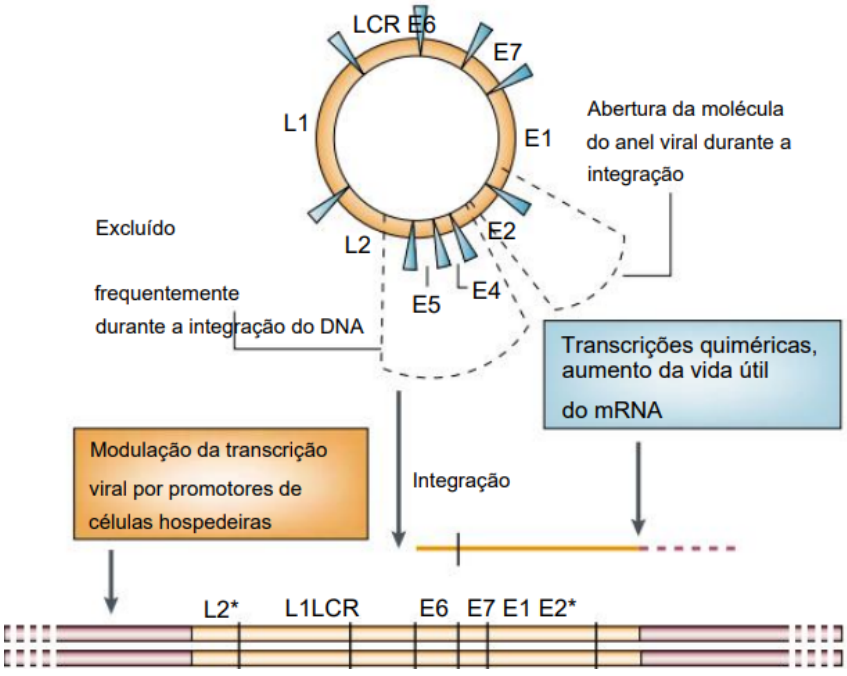


Figura 2.2: Organização do ADN circular do HPV e sua integração no ADN da célula hospedeira

Fonte: Zur Hausen (2002).

Na Figura 2.2 observamos a organização do ADN circular do HPV e sua integração no ADN da célula hospedeira O genoma do papilomavírus humano (HPV) contém entre 6.800 e 8.000 pares de bases e é dividido em oito quadros de leitura abertos - E6, E7, E1, E2, E4, E5 e L2 e L1 - codificando para 'precoce' (E) ou ' funções tardias' (L). No decurso do desenvolvimento

do cancro, a molécula viral torna-se frequentemente integrada no ADN da célula hospedeira. A molécula do anel é mais frequentemente aberta dentro do quadro de leitura aberto E2, interrompendo a continuidade desse gene. Parte de E2 e quadros de leitura abertos adjacentes a E2 - E4, E5 e parte de L2 - são regularmente excluídos após a integração (genes parciais são representados por um asterisco). Os transcritos virais, que abrangem uniformemente a região E6 e E7, e estão frequentemente ligados a sequências celulares flanqueadoras, estão presentes e a transcrição pode ser modulada (melhorada) por promotores flanqueadores da célula hospedeira. LCR, região de controle longa.

Segundo Andersson (2009), os vírus do papiloma humano são vírus de ADN icosaédricos pequenos, sem envelope, que têm um diâmetro de 52–55 nm. As partículas virais contêm uma única molécula de ADN de fita dupla com cerca de 8.000 pares de bases contida em uma capsídeo composto por 72 proteínas de capsômero pentaméricas. O capsídeo contém duas proteínas estruturais – a L1 tardia e L2 – que são codificados viralmente e expressos no final do ciclo de replicação. Os genomas de todos os HPV tipos contêm aproximadamente oito quadros de leitura abertos (ORFs) que são transcritos de apenas uma fita de ADN.
As ORFs são classificadas em três partes funcionais: a região inicial (E) que codifica proteínas (E1 – E7) necessária para a replicação viral, a região tardia (L) que codifica as proteínas estruturais (L1-L2) necessárias para a montagem do vírion, e uma parte em grande parte não codificante que é referida como a região de controle longa (LCR) que contém elementos cis necessários para replicação e transcrição viral.

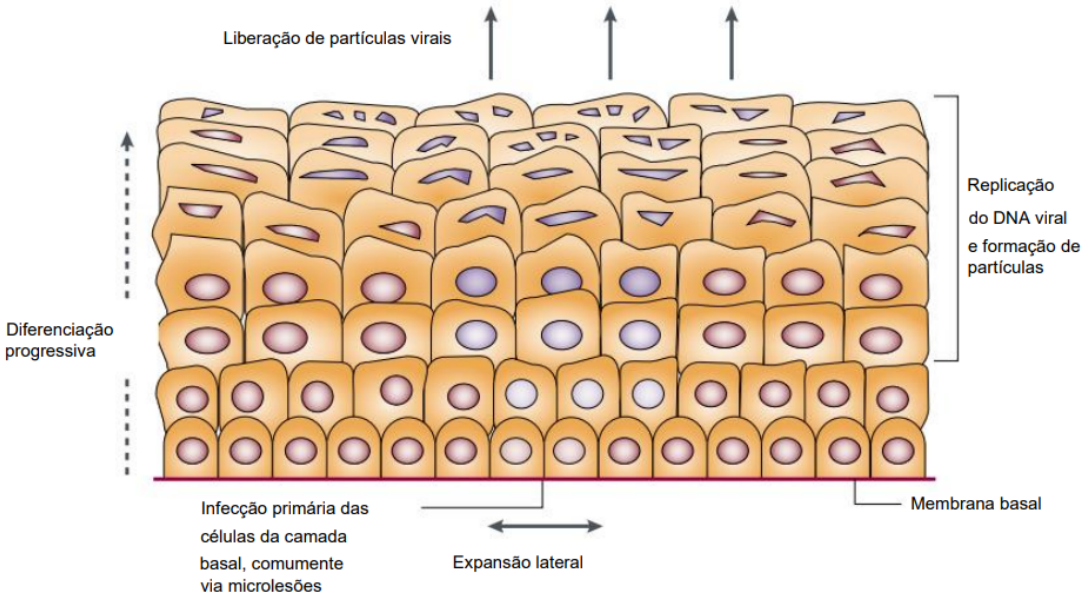


Figura 2.3: Ciclo de vida do papilomavírus humano

Fonte: Zur Hausen (2002).

A infecção pelo HPV requer a disponibilidade de uma célula da camada basal. Isso geralmente ocorre em micro lesões de pele ou mucosa. A célula infectada se divide e a população

se espalha lateralmente. Parte da progênie migra para as camadas celulares de diferenciação supra-basais, onde os genes virais são activados, o ADN viral é replicado e as proteínas do capsídeo são formadas.

### 2.3.1 Tipos de Papilomavírus Humano

Segundo Mondiale de la Santé (2022), com base na sequência genómica de L1, o gene que codifica a principal proteína da cápside, foram identificados e caracterizados mais de 200 tipos de HPV. Os tipos de HPV são classificados de várias maneiras, incluindo pelo seu potencial para induzir o cancro, ou seja, tipos de alto risco versus tipos de baixo risco. Atualmente 12 tipos de HPV são definidos como de alto risco (oncogênicos) e causam cancro em humanos (tipos 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59); o tipo 68 é classificado como provavelmente causador de cancro.

### 2.3.2 Tipos de Papilomavírus Humano quanto as lesões genitais

Segundo Zur Hausen (2002), Os tipos de HPV encontrados preferencialmente no cancro do colo do útero e noutros cancros ano-genitais foram designados como tipos de "alto risco". Por outro lado, aqueles encontrados principalmente em verrugas genitais e lesões não malignas foram rotulados como tipos de "baixo risco". Posteriormente, foi demonstrado que apenas os genes E6 e E7 dos tipos de alto risco foram capazes de imortalizar células humanas em cultura de tecidos.

Tabela 2.2: Tipos de papilomavírus em lesões genitais

| Tipo de lesão genital | **Tipo de HPV** | |
|---|---|---|
| | menos prevalecente | **mais prevalecente** |
| Condiloma acuminado | 42,44,51,53,83 | **6,11** |
| Neoplasias intraepiteliais | 6,11,18,26,30,31,33,34,35,39,40,42,43, 45,51,52,53,54,55,56,57,58,59,61,62,64, 66,67,68,69,70,71,73,74,79,81,82,83,84 | **16** |
| Cancro do colo do útero e outros cancros ano-genitais | (6,11),18,31,33,35,39,45,51,52,54,56,58,59,66,68,69 | **16** |

**Fonte:** Zur Hausen (2002).

Os tipos de papilomavírus humano (HPV) entre parênteses indicam uma prevalência extremamente rara.

Segundo Mondiale de la Santé (2022), os tipos 16 e 18 do HPV foram os tipos mais frequentes em todo o mundo, com HPV16 sendo o tipo mais comum em todas as regiões.
O HPV18 e outros tipos de alto risco, como os tipos 31, 33, 39, 45, 51, 52, 56, 58 e 59, tiveram prevalência semelhante e foram os tipos de HPV de alto risco mais comuns depois do HPV16. Mulheres infectadas com um tipo de HPV pode ser reinfectado com o mesmo tipo ou coinfectado ou posteriormente infectado com outros tipos.

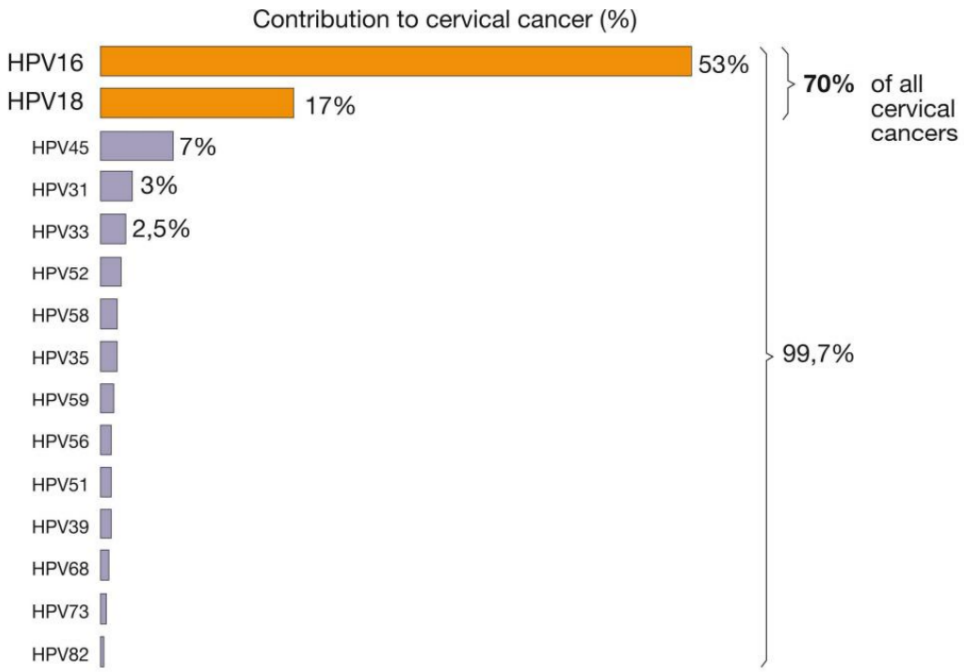


Figura 2.4: Frequência cumulativa de genótipos de HPV presentes no cancro do colo do útero
Fonte: Andersson (2009).

O gráfico na Figura 2.4, ilustra a contribuição relativa para o cancro do colo do útero dos diferentes tipos de HPV, os dois tipos de HPV16 e 18 descobertos por Zur Hausen tem uma maior contribuição em mais de 70%, enquanto que todos os demais agregam o restante da percentagem da contribuição total explicada para 99,7% .

### 2.3.3 Factores associados a presença do Papilomavírus Humano

Vários estudos têm apresentado alguns dos possíveis factores associados a prevalência do HPV no organismo humano.
De acordo com Andersson (2009), o aumento do desenvolvimento do carcinoma cervical é associado a factores como : idade precoce de início sexual; múltiplos parceiros sexuais; má higiene sexual; pobreza; estado imunossuprimido do hospedeiro; cofactores cancerígenos; antecedentes genéticos e dados epidemiológicos.

Segundo Jamison (2006), factores de risco como o aumento do número de gravidezes, a exposição a contraceptivos orais, o tabagismo e padrões alimentares específicos, constituem grandes itens para a presença do HPV.

Segundo Da Silva *et al.* (2023), factores como a idade, o nível de escolaridade, o estado civil e a multiplicidade de parceiros devem ser considerados cruciais para a implementação de medidas preventivas e educativas direcionadas à prevenção da infeção pelo HPV.
Segundo o mesmo autor, existe uma vasta literatura sobre o tema, com diferentes perspectivas

de abordagem, que remete para várias abordagens ao conceito de factores de risco associados ao HPV no cancro do colo do útero no Brasil, algumas das quais são apresentadas no quadro seguinte.

Tabela 2.3: Factores de risco associados ao HPV no cancro do colo do útero no Brasil

| Autores (ano) | Principais factores |
|---|---|
| Melo SCCS, *et al*. (2009) | Em entrevista com 25 mulheres, a maioria apresentou algum factor de risco como: tabagismo, doenças sexualmente transmissíveis, uso de anticoncepcional hormonal, número de parceiros, início precoce da atividade sexual. |
| Brito DMS e Galvão MTG (2010) | Os trabalhos foram categorizados em factores de risco atribuídos: 1) Determinante social que envolveu factores como: reduzida condição socioeconômica; tabagismo; higiene; desnutrição; estigma; défice do acompanhamento cervical e défice de conhecimento 2) Exposição sexual abrangeu coitarca precoce; múltiplos parceiros; contraceptivos orais e doenças sexualmente transmissíveis; 3) Condições clínicas envolveram contagem de células TCD4+ e uso de antirretrovirais. |
| Eduardo KGT, *et al*. (2012) | Factores de risco encontrados: idade, parceria eventual, classe econômica e escolaridade baixas, não realização do exame preventivo, baixo peso, tabagismo, uso de anticoncepcionais hormonais e factores sexuais e reprodutivos, sendo eles o número de gestações, idade na primeira gestação, coitarca, história de doenças sexualmente transmissíveis de repetição ou não (clamídia, tricomoníase, candidíase), uso de preservativo e realização de exame de Papanicolau. |
| Barasuol MEC e Schimidt DB (2014) | São conhecidos os diversos factores de risco para o desenvolvimento desse tumor, sendo este relacionado à infecção pelo papiloma vírus humano (HPV), tabagismo, iniciação sexual precoce, multiplicidade de parceiros, multiparidade, uso de contraceptivos orais, baixa ingestão de vitaminas e coinfecção por agentes infecciosos, como HIV, Chlamydia Trachomatis, tricomoníase e candidíase. |

**Fonte:** Da Silva (2023).

Em suma, os factores apresentados como sendo de risco para a infecção pelo HPV pelas literaturas apresentadas, foram: tabagismo, doenças sexualmente transmissíveis (HIV, clamídia, tricomoníase e candidíase), uso de anticoncepcional hormonal, número e características dos parceiros, início precoce da actividade sexual, infecção pelo papilomavírus humano.

Para categorizar os resultados, primeiramente identificou-se a presença de diversos determinantes de saúde relacionados ao papilomavírus. A categoria de determinantes sociais de saúde reflete, com variação no nível de detalhamento, o conceito de que as condições de vida e trabalho dos indivíduos e grupos populacionais estão associadas à sua situação de saúde. Assim, foram identificados como determinantes os factores socioeconómicos, pessoais, sexuais e reprodutivos, imunológicos e as ações preventivas, Descritos na Tabela 2.4.

Tabela 2.4: Categorização dos principais resultados relacionando com os principais determinantes clínicos identificados

| Principais determinantes | Autores (ano) | Principais factores |
|---|---|---|
| Factores socioeconómicos | Barasuol MEC e Schimidt DB (2014) França MCA, *et al*. (2013) Eduardo KGT, *et al*. (2012) Brito DMS e Galvão MTG (2010) Melo SCCS, *et al*. (2009) | Idade, condição de união, classe económica e escolaridade. |
| Factores pessoais | Andrade VRM e Brum JO (2020) Pancera TR e Santos GHN, (2018) Anjos SJSB, *et al.* (2013) França MCA, *et al*. (2013) Eduardo KGT, *et al*. (2012) Brito DMS e Galvão MTG (2010) Mendonça VG, *et al*. (2010) | baixo peso, tabagismo e uso de anticoncepcionais hormonais, higiene pós relação sexual e nutrição, conhecimento quanto às formas de transmissão do vírus. |
| Factores sexuais e reprodutivos | Barasuol MEC e Schimidt DB (2014) Eduardo KGT, *et al*. (2012) Brito DMS e Galvão MTG (2010) Mendonça VG, *et al.* (2010) | A idade da primeira relação sexual, o número e as características dos parceiros sexuais e a paridade. |
| Factores imunológicos | Barasuol MEC e Schimidt DB (2014) Eduardo KGT, *et al*. (2012) | Infecção pelo HIV, DST's de repetição (HPV, candidíase, tricomoníase e clamídia), contagem de células TCD4+ e o tratamento com drogas antiretrovirais. |
| Ações preventivas | Sousa ACO, *et al*. (2017) França MCA, *et al*. (2013) Diz MDPE e Medeiros RB (2009) | Atuação de profissionais de saúde no rastreio e prevenção do câncer de colo uterino e na assistência a mulheres com HPV ou com neoplasia uterina. Vacinação contra os tipos mais prevalente do HPV |
| Virulência do HPV | Andrade VRM e Brum JO (2020) Pancera TR e Santos GHN (2018) | Os tipos 6 e 11 causam verrugas e condilomas possuem menor associação com o carcinoma. Os tipos 16 e 18 possuem maior associação com carcinoma invasivo e pior prognóstico. |

**Fonte:** Da Silva (2023).

## 2.4 Relação entre o HPV e o Cancro do Colo do Útero

Segundo Pinto *et al*. (2002), a relação entre o papilomavírus humano (HPV) e o carcinoma escamoso cervical tem sido estudada há muitos anos. Actualmente, é reconhecido o papel central deste vírus na carcinogênese cervical, a ponto de se afirmar que o cancro do colo do útero não ocorre sem a presença do HPV.

Segundo a Nicolau (2003), a infecção do HPV foi reconhecida como a principal causa de cancro do colo do útero.

Segundo a OMS (2016), a infecção pelo HPV é transmitida através do contacto com a pele genital infectada, membranas mucosas ou fluidos corporais, e pode ser transmitida através de relações sexuais, incluindo sexo oral. A maioria das infecções por HPV 70% a 90% são assintomáticas e remitem espontaneamente dentro de 1 – 2 anos. A infecção persistente com tipos de

alto risco pode progredir para lesões pré-cancerosas que, se não forem detectadas e tratada adequadamente, pode evoluir para carcinoma invasivo no local da infecção. A infecção persistente por HPV, definida pela presença de ADN de HPV específico do tipo em amostras biológicas clínicas repetidas durante um período de tempo geralmente 6 meses é um precursor necessário do cancro do colo do útero.

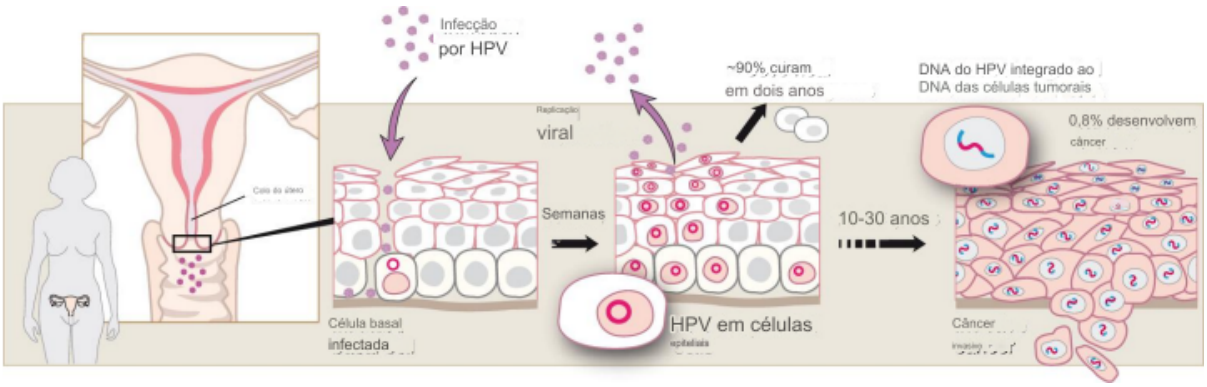


Figura 2.5: Infecção cervical por HPV e os eventos que levam à progressão para cancro do colo do útero
Fonte: Schwarz *et al*. (1985).

Segundo Schwarz *et al*. (1985), na Figura 2.5, O período de incubação é extremamente variável, de 2 semanas até cerca de 32 semanas, com média de 12 semanas. Em alguns casos, o período de latência pode chegar a anos ou indefinidamente, a infecção persistente por 10 a 30 anos permite o desenvolvimento de alterações genéticas adicionais e progressão de lesões de baixo, moderado e alto grau para cancro invasor.

Segundo Andersson (2009), infecção pelos tipos HPV16, 18, 31 e 45 é considerada de alto risco para o desenvolvimento de cancro cervical.

Segundo Zur Hausen (1989), mulheres com infecção persistente por HPV de "alto risco" ou seja, infectadas com um tipo oncogênico de HPV têm 300 vezes mais probabilidade de desenvolver neoplasia de alto grau. Muitas vezes, esses pacientes não possuem uma imunidade eficaz mediada por células, considerada importante para eliminar os infectados pelo HPV.

Ainda de acordo com Zur Hausen (1989), a identificação de tipos específicos de papilomavírus como agentes causadores do cancro do colo do útero e das suas lesões precursoras permitiu o desenvolvimento de um novo método de rastreio e diagnóstico precoce do cancro. Os genomas do HPV e as oncoproteínas virais, presentes em todas as células afectadas, devem representar marcadores convenientes para uma infecção transitória ou persistente. Da mesma forma, as proteínas celulares induzidas em resposta a estas infecções poderiam ter um papel análogo. A presença de marcadores de HPV de alto risco por si só não deve ser motivo de alarme, mas

antes deve levar à uma instigação clínica cuidadosa e de testes repetidos para a persistência do HPV, que é um factor de risco significativo para a desenvolvimento de lesões proliferativas e sua progressão.

## 2.5 Técnicas de Estatística Multivariada

De acordo com Hair *et al*. (2009), a análise multivariada refere-se a todos os métodos estatísticos que analisam simultaneamente múltiplas medidas em cada indivíduo ou objecto sob investigação.

Segundo Johnson & Wichern (2002), é difícil estabelecer um critério para técnicas multivariadas que seja amplamente aceite e indique a adequação das técnicas. Uma classificação distingue as técnicas destinadas a estudar as relações interdependentes das técnicas destinadas a estudar as relações dependentes. Outra classifica as técnicas de acordo com de acordo com o número de populações e o número de conjuntos de variáveis em estudo.

Ainda de acordo com Johnson & Wichern (2002), os objectivos das investigações científicas aos quais os métodos multivariados se prestam mais naturalmente incluem os seguintes:

1. Redução de dados ou simplificação estrutural;
2. Classificação e agrupamento;
3. Investigação da dependência entre variáveis;
4. Previsão;
5. Construção e teste de hipóteses.

A estatística multivariada, particularmente a regressão logística, é amplamente utilizada em estudos de dados transversais na área da saúde devido a várias razões como: permitir a identificação e quantificação de factores de risco associados a uma determinada condição de saúde; em estudos de saúde, muitas vezes o desfecho de interesse é binário ou seja presença ou ausência de uma condição e a regressão logística permite a modelagem de equações para este tipo de casos; os coeficientes estimados na regressão logística podem ser transformados em odds ratios, que são facilmente interpretáveis em termos de chances de ocorrência do evento de interesse (Dunne, *et al., 2007*)

Autores como D'Souza & Dempsey (2011), Dunne *et al.,* (2007), Maucort-Boulch *et al.,*

(2010), em seus estudos demonstram como a regressão logística pode ser utilizada para analisar dados transversais e identificar factores de risco significativos para a infecção por HPV, ajudando a direccionar intervenções de saúde pública e estratégias de prevenção.

Analogamente a estes estudos, na presente pesquisa, a análise de regressão logística será aplicada com intuito de produzir modelos para uma melhor classificação dos resultados positivos e negativos para o HPV.

## 2.6 Análise de Dados

Segundo Hair *et al*. (2009), as tarefas envolvidas na análise de dados podem parecer comuns e inconsequentes, mas são uma parte essencial de qualquer análise multivariada. Ao examinar os dados antes de aplicar uma técnica multivariada, o investigador adquire uma visão crítica das características dos dados.

De acordo com estes autores, a análise de dados pode ser separada em quatro fases distintas:

1. **Exame gráfico dos dados**
   Quando se consideram as análises univariadas, o nível de compreensão é relativamente simples. No entanto, quando o investigador passa para análises multivariadas mais complexas, a necessidade e o nível de compreensão aumentam significativamente.

   Quando queremos analisar a distribuição de uma variável, podemos utilizar o histograma, que pode ser utilizado para analisar qualquer tipo de variável métrica, desde os valores originais até aos resíduos de uma técnica multivariada. Para a relação entre duas variáveis, podemos avaliar relações bivariadas utilizando o diagrama de dispersão, um gráfico de pontos baseado em duas variáveis. Podemos examinar a diferença entre dois ou mais grupos para uma ou mais variáveis métricas utilizando um boxplot. Nestes casos, o investigador precisa de compreender como é que os valores se distribuem em cada grupo e se existem diferenças suficientes entre os grupos para suportar a significância estatística, e verificar se existem observações atípicas. Para responder à necessidade de apresentar um perfil multivariado de uma observação, foram criados vários métodos gráficos multivariados, como os meta-glifos ou glifos, sendo a técnica mais comum a transformação de Fourier de Andrew, cujo resultado é a capacidade de processamento inerente ao ser humano para a interpretação.

2. **Dados perdidos**
   Raramente o investigador evita algum tipo de problema com dados perdidos, pelo que o

grande desafio consiste em abordar as questões geradas pelos dados perdidos que afectam a generalização dos resultados. O impacto dos dados em falta é prejudicial não só devido ao seu enviesamento oculto nos resultados, mas também devido ao seu impacto prático na dimensão da amostra disponível. Se não forem aplicadas medidas correctivas aos dados em falta, qualquer observação com dados em falta em qualquer das variáveis será excluída da análise.

3. **Tratamentos para lidar com dados em falta**
   Os tratamentos para lidar com os dados em falta podem ser classificados com base na aleatoriedade do processo de dados em falta e no método utilizado para estimar os dados em falta.

   Se for identificado um processo de dados em falta não aleatório, o investigador deve aplicar apenas uma acção correctiva, o tratamento de modelação especificamente planeado. Se o investigador identificar um processo de dados em falta completamente aleatório, pode utilizar alguns dos tratamentos mencionados acima. Utilização de observações apenas com dados completos; ignorar ou eliminar caso(s) e/ou variável(eis) problemático(s); métodos de atribuição.

   No método de atribuição, podemos utilizar toda a informação disponível para substituir os dados em falta, por exemplo pela média ou desvio padrão, ou ainda pelo rácio entre todos os valores válidos disponíveis, podemos também substituir os dados em falta de forma aleatória, atribuir por letra marcada, atribuir por regressão, atribuição múltipla.

   Os procedimentos baseados em modelos incorporam explicitamente os dados em falta na análise, quer através de um processo especificamente concebido para a estimativa de dados em falta, quer como parte da análise multivariada padrão.

4. **Observações atípicas**
   As observações atípicas não podem ser caracterizadas categoricamente como benéficas ou problemáticas, mas devem ser vistas no contexto da análise e avaliadas pelos tipos de informação que podem fornecer. Seguem-se alguns métodos para detectar observações atípicas e depois decidir se devem ser mantidas ou excluídas, julgando não só com base nas suas características, mas também com base nos objectivos da análise.

   A identificação de *outliers* pode ser feita através da deteção univariada, aqueles que se encontram fora do intervalo da distribuição, da deteção bivariada, em que os pares de variáveis podem ser avaliados em conjunto através do diagrama de dispersão e, finalmente, da deteção multivariada, que envolve uma avaliação multivariada de cada observação num conjunto de variáveis, utilizando mais frequentemente a medida $D^2$ de

Mahalanobis.

**Testes das suposições da análise multivariada**

Esta é a última fase da análise de dados. A necessidade de testar os pressupostos estatísticos aumenta nas aplicações multivariadas devido a duas características. Em primeiro lugar, a complexidade das relações, devido à utilização rotineira de um grande número de variáveis, e, em segundo lugar, a complexidade da análise pode ocultar os sinais de violação dos pressupostos.

Nesta fase, devem ser verificados os pressupostos de normalidade relativos à distribuição dos dados para uma variável métrica individual e a sua correspondência com a distribuição normal, o que pode ser feito recorrendo à análise gráfica da normalidade, a testes estatísticos e à estatística $z$, deve ser verificada a homocedasticidade, em que a(s) variável(eis) dependente(s) têm níveis iguais de variância ao longo do domínio da(s) variável(eis) preditora(s), a linearidade, uma vez que as correlações apenas representam a associação linear entre variáveis, os efeitos não lineares não são representados no valor da correlação.

## 2.7 Modelos Lineares Generalizados e a Regressão Linear

Os modelos lineares generalizados correspondem a um grupo de modelos de regressão lineares e exponenciais não lineares, em que a variável dependente possui, por exemplo, a distribuição normal, Bernoulli, binomial, Poisson ou Poisson-Gama (Fávero e Belfiore, 2017).
Um Modelo Linear Generalizado é definido da seguinte forma:

$$\eta_i = \alpha + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \ldots + \beta_k \cdot X_{ki} \tag{2.1}$$

**Onde:**

- $\eta$ é conhecido por função de ligação canónica;
- $\alpha$ representa a constante;
- $\beta_j$ $(j = 1, 2, \ldots, k)$ são os coeficientes de cada variável explicativa e correspondem aos parâmetros a serem estimados;
- $X_j$ são as variáveis explicativas (métricas ou *dummies*) ;
- E os subscritos $i$ representam cada uma das observações da amostra em análise ($i = 1, 2, \ldots, n$, em que $n$ é o tamanho da amostra).

A regressão linear descreve uma relação linear entre uma variável independente, $X$, e uma variável dependente, $Y$ procurando estudar as distribuições de frequências de uma variável para

Tabela 2.5: Modelos lineares generalizados, características da variável dependente e funções de ligação canônica.

| **Modelo de Regressão** | **Característica da Variável Dependente** | **Distribuição** | **Função de Ligação Canónica** ($\eta$) |
|---|---|---|---|
| Linear | Quantitativa | Normal | $\hat{Y}$ |
| Com Transformação de Box-Cox | Quantitativa | Normal Após a Transformação | $\frac{\hat{Y}^{\lambda}-1}{\lambda}$ |
| Logística Binária | Qualitativa com 2 Categorias (*Dummy*) | Bernoulli | $\ln\left(\frac{p}{1-p}\right)$ |
| Logística Multinomial | Qualitativa $M(M>2)$ Categorias | Binomial | $\ln\left(\frac{p_m}{1-p_m}\right)$ |
| Poisson | Quantitativa com Valores Inteiros e Não Negativos (Dados de Contagem) | Poisson | $\ln(\lambda)$ |
| Binomial Negativo | Quantitativa com Valores Inteiros e Não Negativos (Dados de Contagem) | Poisson-Gama | $\ln(\mu)$ |

**Fonte:** Fávero & Belfiore (2017).

certos valores fixos de outra variável, dita controlada (Afonso & Nunes, 2019).

Um objectivo importante num problema de regressão é predizer ou estimar o valor mais provável da variável não controlada correspondente a um valor dado da outra variável.

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{2.2}$$

**Onde:**

- $\beta_0 e \beta_1$ são constantes;
- $X_i$ são conhecidos;
- $\epsilon_i$ representa o erro aleatório associado ao valor observado para $Y$.

**Regressão Linear Múltipla**

Segundo Fávero & Belfiore (2017), a regressão linear múltipla oferece a possibilidade de que seja estudada a relação entre uma ou mais variáveis explicativas, que se apresentam na forma linear e uma variável dependente quantitativa. Com isto, o modelo linear múltiplo é definido da seguinte forma :

$$Y_i = a + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_k X_{ki} + u_i \tag{2.3}$$

**Onde**:

- $Y_i$ é a variável dependente para a observação $i$;
- $a$ é a interceptação;
- $b_1, b_2, \ldots, b_k$ são os coeficientes das variáveis explicativas;
- $X_{1i}, X_{2i}, \ldots, X_{ki}$ são as variáveis explicativas para a observação $i$;
- $u_i$ é o termo de erro para a observação $i$.

## 2.8 Análise de Regressão Logística

Segundo Reis (2001), a análise de regressão é a metodologia estatística para prever valores de uma ou mais variáveis de resposta, ou seja, variáveis dependentes, a partir de um conjunto de valores de variáveis preditoras, ou seja, variáveis independentes.

Segundo Hair *et al*. (2009), os modelos de regressão logística, frequentemente referidos como análise logit, constituem uma combinação de regressão múltipla e análise discriminante múltipla. Esta técnica é similar à análise de regressão múltipla, uma vez que uma ou mais variáveis independentes são utilizadas para prever uma única variável dependente.

**Regressão múltipla**
Segundo Hair *et al*. (2009), a Regressão múltipla é o método apropriado quando o problema de pesquisa envolve uma única variável dependente métrica considerada relacionada a duas ou mais variáveis independentes.

**Análise discriminante múltipla**
Segundo Johnson & Wichern (2002), a discriminação e a classificação são técnicas multivariadas preocupadas em separar conjuntos distintos de objectos (ou observações) e em alocar novos objectos (observações) em grupos previamente definidos.
Na visão destes autores, os objectivos imediatos da discriminação resumem-se em descrever, quer graficamente (em três ou menos dimensões), quer algebricamente, as características diferenciais de objectos (observações) de diversas colecções conhecidas (populações). Em suma tentamos encontrar "discriminantes" cujos valores numéricos sejam tais que as coleções sejam separadas tanto quanto possível.
Segundo Hair *et al*.(2009), análise discriminante múltipla é a técnica multivariada adequada quando a única variável dependente é dicotómica como género masculino e feminino ou multicotômica como altura em alto, médio ou baixo.

Segundo Hair *et al*. (2009), o que distingue um modelo de regressão logística de regressão múltipla é que a variável dependente é não- métrica, como na análise discriminante. A escala não métrica da variável dependente requer diferenças no método de estimativa e nos pressupostos sobre o tipo de distribuição subjacente, mas na maioria dos outros aspectos é bastante semelhante à regressão múltipla.
De acordo com os mesmos autores, os modelos de regressão logística diferem da análise discriminante principalmente porque acomodam todos os tipos de variáveis independentes (métricas e não métricas) e não requerem o pressuposto de normalidade multivariada.

### 2.8.1 Estimação do modelo de regressão logística

Segundo Hair *et al.* (2009), o processo de estimação dos coeficientes logísticos é semelhante ao utilizado na regressão, embora neste caso sejam utilizados apenas dois valores reais para a variável dependente ou seja 0 e 1. Além disso, em vez de utilizar mínimos quadrados ordinários que minimiza a soma das diferenças quadradas entre os valores reais dos previstos como forma de estimar o modelo, utiliza-se o método da máxima verossimilhança devido a natureza não-linear resultante da transformação logística, com vista a obter estimativas mais prováveis para os coeficientes.

**Estimação do modelo de regressão logística por máxima verossimilhança**
Segundo Hair *et al.* (2009), os coeficientes estimados para as variáveis independentes são estimados usando o valor logit ou o valor de probabilidades como medida dependente.

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

$$\pi_j = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \tag{2.4}$$

O modelo logístico com a transformação logit é dado por :

$$\ln \pi_j = \ln \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \tag{2.5}$$

$$\ln(\frac{\pi_j}{1-\pi_j}) = \beta_0 + \beta_1 x$$

Segundo Fávero & Belfiore (2017), na regressão logística binária, a variável dependente segue uma distribuição de Bernoulli, ou seja, o facto de determinada observação $i$ ter incidido ou não no evento de interesse pode ser considerado como um ensaio de Bernoulli, em que a probabilidade de ocorrência do evento é $P_i$, e a probabilidade de ocorrência do não evento é (1 - $p_i$), com isto, podemos escrever que a probabilidade de ocorrência de $y_i$, podendo $y_i$ ser igual a 1 ou igual a 0, da seguinte forma:

$$p(y_i) = p_i^{y_i} * (1-p_i)^{1-y_i} \tag{2.6}$$

Para uma amostra com n observações, podemos definir a função de verossimilhança (likelihood function) como sendo:

$$L = \prod_{i=1}^{n} [p_i^{y_i} * (1-p_i)^{1-y_i}] \tag{2.7}$$

Como, na prática, é mais conveniente se trabalhar com o logaritmo da função de verossimilhança,

podemos chegar à seguinte função, também conhecida por log likelihood function:

$$LL = \sum_{i=1}^{n} [(y_i) * \ln(\frac{e^{z_i}}{1+e^{z_i}})] + [(1-y_i) * \ln(\frac{1}{1+e^{z_i}})] \tag{2.8}$$

os valores dos parâmetros do logito que fazem com que o valor de LL seja maximizado em modelos de regressão logística binária, podem ser respondidas com o uso de ferramentas de programação linear, a fim de que sejam estimados os parâmetros $\alpha$,$\beta_1$, $\beta_2$,...,$\beta_i$k com base na seguinte função-objectivo:

$$LL = \sum_{i=1}^{n} [(y_i) * \ln(\frac{e^{z_i}}{1+e^{z_i}})] + [(1-y_i) * \ln(\frac{1}{1+e^{z_i}})] = \text{máx} \tag{2.9}$$

### 2.8.2 Métodos de escolha de variáveis

Segundo Menard (2002), existem diversos métodos para selecionar o modelo óptimo, semelhantes aos utilizados na regressão linear múltipla, como os métodos forward, backward e stepwise, com ou sem efeitos de interação entre as variáveis, e com opções de escolha baseadas nos critérios de inclusão ou exclusão das variáveis.

Segundo Johnson & Wichern (2002), o método **Stepwise**, tentamos misturar os métodos forward e backward, ou seja, tenta selecionar preditores importantes sem considerar todas as possibilidades.
Este procedimento pode ser descrito listando as seguintes etapas básicas :

1. Todas as regressões lineares simples possíveis são consideradas. A variável preditora que explica a maior proporção significativa da variação em Y (a variável que tem a maior correlação com a resposta) é a primeira variável a entrar na função de regressão;

2. A próxima variável a inserir é aquela (das que ainda não foram incluídas) que dá a maior contribuição significativa para a soma dos quadrados da regressão.
   A significância da contribuição é determinada por um teste F. O valor da estatística F que deve ser excedido antes que a contribuição de uma variável seja considerada significativa é frequentemente chamado de F a ser inserido.

3. Uma vez incluída uma variável adicional na equação, as contribuições individuais para a soma dos quadrados da regressão das outras variáveis Os valores já presentes na equação são verificados quanto à significância usando testes F.
   Se a estatística F for menor que aquela (chamada F a remover) correspondente a um nível de significância prescrito, a variável é excluída da função de regressão;

4. Os passos 2 e 3 são repetidos até que todas as adições possíveis sejam não significativas e todas as exclusões possíveis sejam significativas. Neste ponto a seleção para.

Os procedimentos de **adição forward e eliminação backward** são processos de tentativa e erro para encontrar as melhores estimativas de regressão. O modelo de adição forward é semelhante ao procedimento stepwise descrito acima, enquanto a eliminação backward computa uma equação de regressão com todas as variáveis independentes e então elimina variáveis independentes que não contribuem significantemente. A principal distinção da abordagem stepwise em relação aos procedimentos adição forward e eliminação backward é sua habilidade em acrescentar ou eliminar variáveis em cada estágio. Uma vez uma variável adicionada ou eliminada nos esquemas de adição forward ou eliminação backward, não há como reverter a acção em um estágio posterior.

Normalmente, o critério de seleção utilizado é o AIC (*Critério de Informação de Akaike*). Dessa forma, após aplicar o método escolhido, o modelo com o menor valor de AIC será escolhido.
O AIC é definido da seguinte maneira:

$$AIC = -2\ln(L(\hat{\beta})) + 2q \tag{2.10}$$

Onde:
$\ln(L(\hat{\beta}))$ é a representação da função log-verossimilhança para o modelo com q variáveis explicativas que escolhemos (portanto, q + 1 parâmetros).

Segundo Langlotz (2003), ao comparar dois modelos, não podemos depender apenas da qualidade do ajuste, pois nesse caso, o melhor modelo seria aquele com todas as variáveis explicativas. É necessário penalizar a inclusão de uma variável preditora se sua adição resultar em uma melhoria pouco significativa no ajuste, em estatística, chamamos isso de "um zero", referindo-se a algo pouco significativo.

### 2.8.3 Avaliação do ajuste do modelo

Segundo Fávero & Belfiore (2017), a adequação de um modelo de regressão logística pode ser avaliada de duas maneiras a saber:

- Uma maneira é avaliar o ajuste da estimativa do modelo usando valores "pseudo" de $R^2$, semelhantes aos encontrados na regressão múltipla;
- A segunda abordagem é examinar a precisão preditiva (como a matriz de classificação na análise discriminante).

Ainda de acordo com estes autores, como a variável dependente é qualitativa, não faz sentido discutirmos o percentual de sua variância que é explicada pelas variáveis preditoras, ou seja, em modelos de regressão logística não há um coeficiente de ajuste $R^2$ como nos modelos tradicionais de regressão estimados pelo método de mínimos quadrados ordinários.

Segundo Hair *et al*. (2009), a medida geral do quão bem o modelo de regressão logística se ajusta, semelhante ao valor das somas dos quadrados do erro ou resíduos para a regressão múltipla é dada pelo valor de verossimilhança, ou seja é -2 vezes o logaritmo do valor da verossimilhança. Um modelo bem ajustado terá um valor pequeno para -2LL. O valor mínimo para -2LL é 0.

Segundo Menard (2002), na regressão logística, não há estatísticas equivalentes ao coeficiente de determinação utilizado na regressão linear, que indicam a percentagem de variação da variável dependente explicada pelo modelo.

Ainda de acordo com este autor, três medidas comparáveis à medida $R^2$ em Regressão Múltipla estão disponíveis, a saber:

1. Pseudo $R^2$;
2. $R^2$ de Cox e Snell;
3. $R^2$ Nagelkerke.

### 2.8.4 Medidas de diagnóstico

Segundo Fávero & Belfiore (2017), para a classificação depois de ter estimado o modelo de probabilidade de ocorrência do evento, usamos o **cutoff** que é um ponto de corte que o pesquisador escolhe, de modo a que sejam classificadas as observações em função das suas probabilidades e, desta, utilizamos quando queremos fazer previsões de ocorrência do evento para observações não presentes na amostra com base em suas probabilidades das observações presentes na amostra.
Assim, se determinada observação não presente na amostra apresentar uma probabilidade de incidir no evento maior do que o cutoff definido, espera-se que haja a incidência do evento e, portanto, será classificada como evento. Por outro lado, se a sua probabilidade for menor do que o cutoff definido, espera-se que haja a incidência do não evento e, portanto, será classificada como não evento.

De maneira geral, podemos estipular o seguinte critério:

- Se $P_i >$ cutoff$- >$ a observação i deverá ser classificada como sucesso ou evento.
- Se $P_i <$ cutoff$- >$ a observação i deverá ser classificada como insucesso ou não evento.

### 2.8.5 Validação da Regressão Logística

Segundo Hair *et al*. (2009), na RL, a validação é feita seguindo os mesmos passos da validação na análise discriminante, onde o principal meio de validação é pelo uso da amostra de validação e a avaliação da sua precisão preditiva, com isto, a validade é estabelecida se o modelo estimado classifica observações, em um nível aceitável, que não foram usadas no processo de estimação. Se a amostra de validação é obtida a partir da amostra original, então essa abordagem estabelece validade interna se uma outra amostra separada, talvez de uma outra população ou de outro segmento da população, forma a amostra de validação, isso correspondendo a uma validação externa dos resultados.

### 2.8.6 Curva ROC

A Curva Receiver Operating Characteristic (ROC) é uma ferramenta fundamental na análise estatística multivariada, especialmente em modelos de regressão logística. Ela fornece uma representação gráfica da capacidade de um modelo classificar ou distinguir entre classes positivas e negativas.

Segundo Hosmer e Lemeshow (2013), a curva ROC é um gráfico que traça a Taxa de Verdadeiros Positivos (True Positive Rate - TPR) contra a Taxa de Falsos Positivos (False Positive Rate - FPR) para diferentes pontos de corte para a tomada de decisão. Essencialmente, ela mostra o compromisso entre sensibilidade (capacidade de detectar corretamente os positivos) e especificidade (capacidade de detectar corretamente os negativos) de um modelo de classificação.

A área sob a curva ROC (AUC - Area Under the Curve) é uma medida utilizada para resumir a performance global do modelo. Quanto maior a AUC, melhor é a capacidade do modelo de distinguir entre as classes (Bradley, 1997).

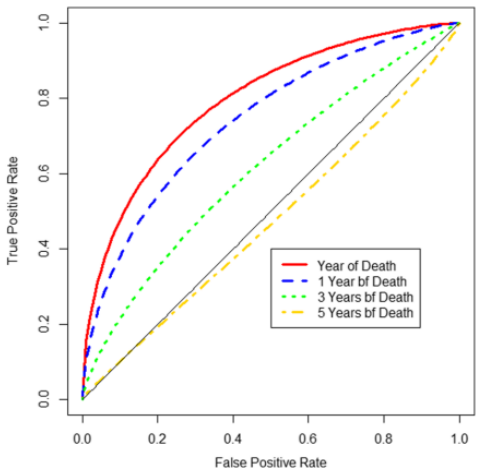


Figura 2.6: Exemplo de uma curva ROC

Fonte: Kamarudin, (2017).

### 2.8.7 Árvore de Classificação

Segundo Lemon *et al.* (2003), os métodos de árvore de decisão ou de classificação, também conhecidos como particionamento recursivo, são estratégias analíticas que foram desenvolvidas como uma ferramenta para classificar ou segmentar públicos-alvo para efeitos de uma melhor visualização e análise da informação.

#### Propriedades de um classificador ideal

De acordo com Loh (2011), um bom classificador deve ter as seguintes propriedades:

- Alta precisão preditiva;
- Estrutura intuitiva e compreensível;
- Inferência correta e imparcial;
- Tempo de treinamento rápido.

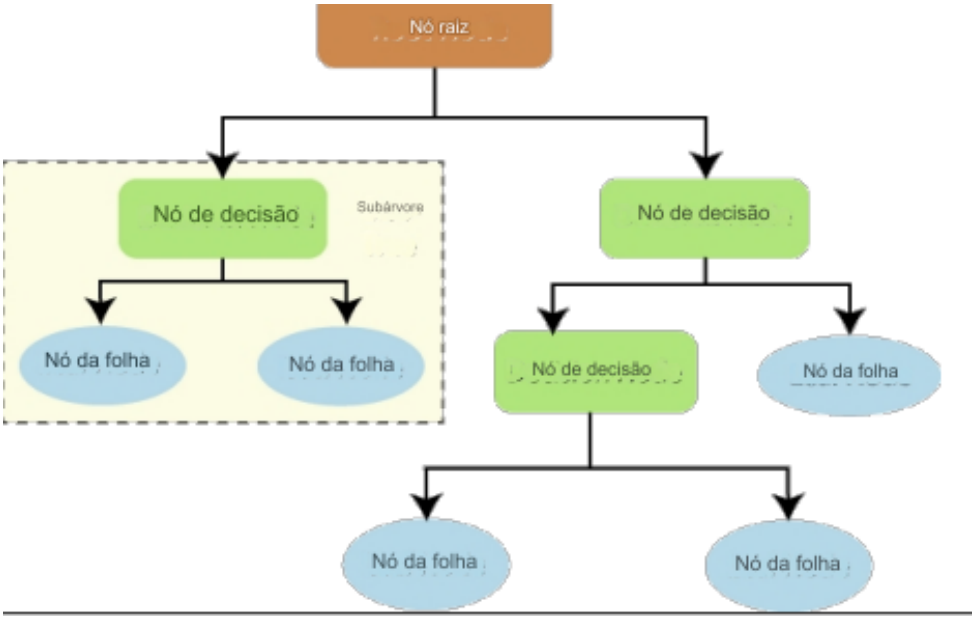


Figura 2.7: Exemplo de uma árvore de classificação

Fonte: Janikow (1988).

# Capítulo 3

# MATERIAL E MÉTODOS

No presente capítulo, serão descritos minuciosamente os materiais e métodos utilizados para alcançar os resultados que serão apresentados no capítulo 4. A secção 3.1 abordará os materiais, incluindo a fonte dos dados, o método de colecta dos dados e a relevância das informações. A secção 3.2 tratará dos métodos, apresentando a classificação da pesquisa e as técnicas empregadas para a obtenção dos resultados.

## 3.1 Classificação da pesquisa

O presente estudo de dados de corte transversal, adopta uma perspectiva combinatória entre a retrospectiva qualitativa bibliográfica e o estudo quantitativo. Na parte bibliográfica o foco do estudo é trazer para a pesquisa os principais conceitos da literatura sobre a temática a ser discutida e na parte quantitativa, são tomados em conta os principais achados na análise estatística dos dados para uma posterior comparação com os principais achados da literatura.

Nas diversas formas de classificação das pesquisas, esta apresenta as seguintes:

**Quanto à natureza**

Quanto à natureza a presente pesquisa é experimental pura pelo facto de partir da causa procurar-se estabelecer o efeito com base em dados secundários, ou seja, apartir do numero de mortes pelo cancro do colo do útero associada a presença do HPV procurar estabelecer o efeito dos factores de risco associados a infecção pelo HPV (Markoni & Lakatos, 1992).

**Quanto à abordagem**

Quanto à abordagem a pesquisa é mista onde combina abordagens qualitativas e quantitativas para proporcionar uma compreensão mais completa sobre os factores de risco a infecção peo HPV.

**Quanto aos objectivos**

Quanto aos objectivos a presente pesquisa é exploratória em que com o material agrupado, efectuam-se análises exploratórias conjuntas.

**Quanto aos procedimentos técnicos**

Quanto aos procedimentos técnicos a presente pesquisa é qualiquantitativa em que faz-se a análise de acordo com o observado na literatura e o achado no processamento dos dados.

## 3.2 Materiais

A população do presente estudo será constituída por todas as mulheres que foram atendidas nos centros DREAM da comunidade de Sant'Egídio no primeiro trimestre de 2024, e para a recolha desta informação de fonte secundária, foi feito um pedido de autorização à direção da mesma instituição para o uso desta informação para fins académicos.

A dimensão da amostra será definida, compreendendo toda a população ou utilizadores destes centros no determinado intervalo de tempo acima referido.

A recolha da informação da fonte secundária e consequente organização em formato de dados prontos para serem manipulados foi feita apartir de livros clínicos de consultas para o rastreio do cancro do colo do útero e da mama.

Para o tratamento dos dados e análise estatística serão utilizados os softwares **Microsoft 365** para criação da base de dados e produção de gráficos; **R-studio** para limpeza de dados e processamento de resultados e **IBM SPSS Statistics** para análises descritivas.

A base de dados do presente estudo é composta por 161 Mulheres que foram atendidas nos centro DREAM da comunidade Sant'Egídio, e que tinham algumas características como as descritas na Tabela 3.3.

Tabela 3.1: Variáveis em estudo

| Variável | Descrição | Categorias das Variáveis | Classificação |
|---|---|---|---|
| Ciclo Menstrual | Verificação da menstruação na mulher | (irregular)<br>(NA)<br>(Regular) | Qualitativa nominal |
| Sangramento Vaginal | Referente a se a mulher tem ou não sangramento pela vagina | (Anormal)<br>(NA)<br>(Normal) | Qualitativa nominal |

Tabela 3.2: Variáveis em estudo (continuação)

| Variável | Descrição | Categorias das Variáveis | Classificação |
|---|---|---|---|
| Planeamento Familiar | Referente ao método de planeamento familiar usado pela paciente | (DIU) (implante); (injetável) (nenhum); (outros) (pílula); (preservativo) | Qualitativa nominal |
| Tipo de Rastreamento Realizado | Se a paciente ja realizou algum rastreio do cancro | (pela primeira vez) (para seguimento) | Qualitativa nominal |
| Menopausa | Se a paciente já terá atingido a menopausa | (Não) (sim) | Qualitativa categórica |
| ITS | Se a paciente ja teve alguma ITS | (não) (não sabe) (Sim) | Qualitativa categórica |
| ITS Sim | Se a paciente já teve alguma ITS qual foi | (Leucorreia) (NA) | Qualitativa nominal |
| HIV | Resultado do teste de HIV | (negativo) (positivo) | Qualitativa Categórica |
| História de Rastreamento | Se a paciente ja fez algum rastreio | (não) (sim) | Qualitativa categórica |
| História de Rastreamento se Sim | Se a paciente ja fez algum rastreio qual foi | (HPV) (NA) (VIA) | Qualitativa Nominal |
| Data do Último Rastreamento IC | Quanto tempo passa deste o rastreio | (NA) (menos de 3 meses) (entre 3 meses a 3 anos) (mais de 3 anos) | Qualitativa nominal |
| Resultado do Último Rastreamento | qual foi o resultado do ultimo rastreio | (NA) (não sabe) (negativo) (positivo) | Qualitativa Categórica |
| Se o Último Rastreamento Positivo | Se o último rastreio foi positivo, qual foi o método de tratamento | (crioterapia/termoablação) (NA) | Qualitativa Nominal |
| Exame Especular para Colo Sangramento | Exame para verificar se a paciente tinha algum sangramento vaginal | (com sangramento) (sem sangramento) (outros) (NA) | Qualitativa nominal |
| Exame Especular para Colo visualização | Exame para verificar o JEC da paciente | JEC parcialmente visualizada JEC totalmente visualizada NA | Qualitativa nominal |
| Rastreamento Uterino com VIA | Qual foi o resultado do Rastreamento Uterino com VIA | NA VIA Negativo via Positivo | Qualitativa categórica |
| HPV | Resultado do teste do HPV | Positivo Negativo | Qualitativa Categórica |
| HPV Positivo | Se a paciente foi positiva para o HPV que tipo de HPV tinha | HPV16 HPV16 + Outro HPV18/45 NA Outro | Qualitativa nominal |

Tabela 3.3: Variáveis em estudo (continuação)

| Variável | Descrição | Categorias das Variáveis | Classificação |
|---|---|---|---|
| Tratamento Feito | Qual foi o tratamento para HPV feito | Crioterapia<br>NA | Qualitativa nominal |
| Nível de Escolaridade | Nível máximo de escolaridade concluído pela paciente | Nenhuma<br>Primário<br>Secundário<br>Superior | Qualitativa Categórica |
| Estado Civil | Estado civil da paciente | Casado<br>Separado/Divorciado<br>Solteiro<br>Viuvá | Qualitativa nominal |
| Água na Casa | Se na casa da paciente têm agua canalizada | Não<br>Sim | Qualitativa categórica |
| Sexo | Género | Feminino | Qualitativa categórica |
| Tabagismo | Se a paciente fuma ou não | Não | Qualitativa categórica |
| Exame Pélvico | Resultado do exame pélvico | Normal | Qualitativa categórica |
| Unidade Sanitária | Em que unidade sanitária a paciente foi atendida | CC | Qualitativa categórica |
| Idade | Faixa etária das pacientes | [17,21); [21,25); [25,29); [29,33)<br>[33,37); [37,41); [41,45)<br>[45,49); [49,53); [53,57); [57,61)<br>[61,65); [65,69) | Quantitativa contínua |
| Numero de gestações | Quantas gravidezes a paciente ja teve | [0,1); [1,2); [2,3); [3,4)<br>[4,5); [5,6); [6,7); [7,8)<br>[8,9); [9,10); [10,11) | Quantitativa contínua |
| Idade da 1oRelação sexual | Com Quantos anos manteve a primeira relação sexual | [12,15); [15,18)<br>[18,21); [21,41) | Quantitativa contínua |
| Altura | Altura em cm da paciente | [144,147); [147,150); [150,153); [153,156)<br>[156,159);[159,162); [162,165); [165,168)<br>[168,171); [171,174); [174,177) | Quantitativa discreta |
| Peso | Peso em kg da paciente | [39.7,46.7); [46.7,53.7); [53.7,60.7); [60.7,67.7)<br>[67.7,74.7); [74.7,81.7); [81.7,88.7);[88.7,95.7)<br>[95.7,103); [103,110); [110,117); [117,131) | Quantitativa discreta |

## 3.3 Métodos

Para se atingir o objectivo traçado na presente pesquisa, seram aplicados testes e tecnicas estatísticas que estas por sua vez para a interpretação dos resultados obtidos será utilizado o nível de significância $\alpha$ de 5%.

### 3.3.1 Valores perdidos (Missing Values)

De acordo com Howell (2007), os dados perdidos ou missing values, correspondem a um conjunto de informações não disponíveis de uma variável ou caso sobre o qual outra informação está disponível. Os dados perdidos podem surgir por diversas razões.

Os dados perdidos podem ser responsáveis pela produção de viéis de estimação, redução da

precisão das estimativas e dificuldades nas análises a serem efetuadas.

**Tipos de dados perdidos**

Hair *et al.* (2009), classificam os dados perdidos em:

- **Perdidos ao Acaso (MAR- missing at random)**- isto acontece quando valores perdidos de Y dependem de X, mas não dependem de Y. A probabilidade de um valor ser ausente está relacionada com os dados observados, mas não com os dados ausente.
- **Completamente perdidos ao acaso (MCAR-missing completely at random)**- os valores perdidos de Y não dependem de X. Aqui, a ausência de dados não está relacionada com os dados observados ou não observados.

**Métodos de tratamento**

1. Uso de observações com dados completos somente.
   Também conhecidas com abordagem de caso completo, consiste em remover dos dados para a análise as observações que tenham algum valor perdido em qualquer das variáveis de interesse.

2. Desconsiderar variáveis e/ou observações
   Consiste na exclusão de variáveis ou observações com valores omissos. O pesquisador deve determinar a extensão dos dados perdidos em cada caso e variável e então eliminar aquelas que tiverem altas frequências de missing values. O pesquisador pode descobrir que os dados perdidos estão concentrados em um pequeno subconjunto de observações e/ou variáveis, sendo que a exclusão diminui substancialmente a extensão dos dados perdidos. Não existem orientações seguras sobre o nível necessário para a exclusão, mas qualquer decisão deve ser baseada em considerações empíricas e teóricas.

3. Imputação por um caso
   Observações com dados perdidos são substituídos por uma outra observação escolhida fora da amostra.

4. Imputação pela média/mediana/moda
   Consiste na troca de valores perdidos de uma variável com os valores médios, medianos e com maior frequência dessa mesma variável dependendo do tipo de variável.

5. Imputação por carta marcada
   O pesquisador substitui os valores perdidos por um valor constante obtido de fontes externas ou pesquisas anetriores.

6. Imputação por regressão
   Os vários métodos de regressão são usados para prever os valores perdidos de uma

variável com base na sua relação com as outras variáveis do conjunto de dados.
Este método pode ser mais preciso, porem como qualquer outro têm as suas desvantagens, tais como reforça as relações já existentes na amostra, a menos que termos estocásticos sejam acrescentados aos valores estimados, a variância da distribuição é subestimada e por fim este método pressupõe que a variável com dados perdidos tem correlações substanciais com as outras variáveis.

7. Imputação múltipla.
   Este método combina todos os demais métodos. Aqui dois ou mais métodos de imputação são usados para derivar uma estimativa mais composta.

8. Imputação baseada em Machine Learning
   Aqui a imputação pode ser feita pelo método de K-Nearest Neighbors(KNN) que utiliza os k vizinhos mais próximos com base em alguma amplitude para fazer a imputação dos valores perdidos ou pelos métodos da árvore de decisão, redes neurais entre outros.

Na presente pesquisa, para o tratamento de valores perdidos foram usados os métodos de imputação múltipla, onde foram combinados os metodos e imputação por regressão com os métodos de imputação pela média ou moda dos valores observados para cada variável.

### 3.3.2 Teste de Hawkins

O teste de normalidade e homocedasticidade de Hawkins é um método para verificar a normalidade dos resíduos e a homocedasticidade em modelos de regressão.

**Normalidade dos resíduos**: Refere-se à distribuição normal dos erros ou resíduos do modelo de regressão. É um pressuposto para a validade de muitos testes estatísticos (Montgomery *et al*., 2021).

**Homocedasticidade**: Significa que a variância dos resíduos é constante para todos os níveis dos preditores. É um pressuposto importante em modelos de regressão linear (Gujarati, 2000).

**Hipóteses**
Normalidade dos resíduos:
*Hipótese nula (H0)*: Os resíduos do modelo seguem uma distribuição normal.
*Hipótese alternativa (H1)*: Os resíduos do modelo não seguem uma distribuição normal.

Homocedasticidade:
*Hipótese nula (H0)*: A variância dos resíduos é constante (homocedasticidade).
*Hipótese alternativa (H1)*: A variância dos resíduos não é constante (heterocedasticidade).

**Cálculo da estatística de teste**

$$H(i) = \frac{e_i^2}{\sum_{j=1}^{n} e_j^2} \tag{3.1}$$

$$T = \sum_{i=1}^{k} \left( H(i) - \frac{1}{k} \right)^2 \tag{3.2}$$

**Onde:**

- $e_i$- são os resíduos do modelo de regressão ajustado.
- $K$- numero de grupos em que os resíduos foram agrupados.

**Regra de decisão**

*Usando a estatística de teste*:
$T >$ valor crítico, rejeita-se a hipótese nula.
$T <=$ valor crítico, não rejeita-se a hipótese nula.

*Usando o p-valor*:
Calcule o p-valor correspondente à estatística de teste $T$.
Se o $p - valor <=$ que o nível de significância $\alpha$, rejeita-se a hipótese nula.
Se o $p - valor >$ que o nível de significância $\alpha$, não rejeita-se a hipótese nula.

### 3.3.3 Teste de Kolmogorov-Smirnov e Shapiro-Wilk

O teste de Kolmogorov-Smirnov pode ser aplicado a dados que estejam pelo menos na escala ordinal, excepto nominais, e a distribuição a testar deve estar completamente especificada (Agastri, 2018).

**Hipóteses**

*Hipótese nula (H0)*: $X$ tem função de distribuição $F_0X$
*Hipótese alternativa (H1)*: $X$ não tem função de distribuição $F_0X$.

**Cálculo da estatística de teste**

$$D = \sup_{x \in \mathbb{R}} \left| F_n(x) - F_0(x) \right|. \tag{3.3}$$

**Regra de decisão**

*Pela Estatística de Teste*:
Se a estatística de teste $D$ for maior que o valor crítico, rejeita-se a hipótese nula (H0).
Se a estatística de teste $D$ for menor ou igual ao valor crítico, não se rejeita a hipótese nula (H0).

*Regra de Decisão com o p-valor:*
Se o p-valor for menor que $\alpha$, rejeita-se a hipótese nula (H0).
Se o p-valor for maior ou igual a $\alpha$, não se rejeita a hipótese nula (H0).

O teste de Shapiro-Wilk é utilizado apenas para avaliar se os dados quantitativos (não agrupados em classes) se ajustam a uma distribuição Normal (Afonso & Nunes, 2019).

**Hipóteses**
*Hipótese nula (H0)*: $X$ tem distribuição Normal
*Hipótese alternativa (H1)*: $X$ não tem distribuição Normal

**Cálculo da estatística de teste**

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{i:n}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3.4}$$

**Regra de decisão**
*Pela Estatística de Teste*:
Se a estatística de teste $W$ for menor que o valor crítico, rejeita-se a hipótese nula (H0).
Se a estatística de teste $W$ for maior ou igual ao valor crítico, não se rejeita a hipótese nula (H0).

*Regra de Decisão com o p-valor:*
Se o p-valor for menor que $\alpha$, rejeita-se a hipótese nula (H0).
Se o p-valor for maior ou igual a $\alpha$, não se rejeita a hipótese nula (H0).

### 3.3.4 Teste de independência do Qui-quadrado

O teste de independência do Qui-quadrado têm com objectivo testar a independência entre 2 variáveis, $X$ e $Y$, que se encontram agrupadas em classes mutuamente exclusivas e exaustivas (Afonso & Nunes, 2019 ).

**Hipóteses**
*Hipótese nula (H0)*: As variáveis $X$ e $Y$ são independentes
*Hipótese alternativa (H1)*:As variáveis $X$ e $Y$ não são independentes

**Cálculo da estatística de teste**

$$\chi^2 = \sum_{i=1}^{L}\sum_{j=1}^{C} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(L-1)(C-1)} \tag{3.5}$$

**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;
- $X^2_{(L-1)(C-1)}$- é distribuição qui-quadrado com (L-1)(C-1) graus de liberdade.

**Regra de decisão**
*Pela Estatística de Teste*:
Se a estatística de teste $X^2$ for maior que o valor crítico, rejeita-se a hipótese nula (H0).
Se a estatística de teste $X^2$ for menor ou igual ao valor crítico, não se rejeita a hipótese nula (H0).

*Regra de Decisão com o p-valor:*
Se o p-valor for menor que $\alpha$, rejeita-se a hipótese nula (H0).
Se o p-valor for maior ou igual a $\alpha$, não se rejeita a hipótese nula (H0).

### 3.3.5 Teste de Mann-Whitney $U$

Este teste também é conhecido por teste de Mann-Whitney-Wilcoxon ou teste Wilcoxon rank-sum, e têm como objectivo testar se duas amostras independentes têm a mesma mediana quando não se pode assumir que essas variáveis seguem uma distribuição normal nas populações (Bluman, 2017).

**Hipóteses**
*Hipótese nula (H0)*: As distribuições das duas amostras são idênticas.
*Hipótese alternativa (H1)*:As distribuições das duas amostras não são idênticas.

**Cálculo da estatística de teste**

$$U = R - \frac{n_1(n_1+1)}{2} \tag{3.6}$$

**Onde:**

- $R$ é o menor dos somatórios das classificações das duas amostras;
- $n_1$ é o tamanho da primeira amostra.

**Regra de decisão**

*Pela Estatística de Teste*:

Se $U <= U_\alpha$, onde $U_\alpha$ é o valor crítico para o nível de significância $\alpha$, rejeita-se a hipótese nula.

Se $U > U_\alpha$, onde $U_\alpha$ é o valor crítico para o nível de significância $\alpha$, não rejeita-se a hipótese nula.

*Valor p (p-value)*:

Se o valor p $<= \alpha$, onde $\alpha$ é o nível de significância, rejeita-se a hipótese nula.

Se o valor p $> \alpha$, onde $\alpha$ é o nível de significância, rejeita-se a hipótese nula.

### 3.3.6 Métodos de escolha de variáveis

Segundo Hair *et al.* (2009), existem vários métodos de busca sequencial e processos combinatórios para ajudar o pesquisador a encontrar o melhor modelo de regressão.

**Especificação confirmatória**

Neste método o pesquisador especifica completamente o conjunto de variáveis independentes a serem incluídas no modelo. Embora seja concetualmente simples, o pesquisador deve se assegurar que o conjunto de variáveis atinja a previsão máxima, mantendo o modelo parcimonioso.

**Estimação stepwise**

É um método que permite ao pesquisador examinar a contribuição de cada variável independente para o modelo de regressão. Cada variável é considerada para inclusão antes do desenvolvimento da equação. A variável independente com a maior contribuição é acrescentada em um primeiro momento. Variáveis independentes são então selecionadas para inclusão, com base em sua contribuição incremental sobre as variáveis já presentes na equação.

As questões específicas em cada estágio são as seguintes:

1. Comece com o modelo de regressão simples no qual apenas a variável independente, que é a mais fortemente correlacionada com a variável dependente, é usada.

2. Examine os coeficientes de correlação parcial para encontrar uma variável independente adicional que explique a maior parte estatisticamente significante do erro remanescente da primeira equação de regressão.

3. Recalcule a equação de regressão usando as duas variáveis independentes e examine o valor parcial F para a variável original no modelo para ver se esta ainda faz uma contribuição significante, dada a presença da nova variável independente. Se não for o caso, elimine a

variável. Essa habilidade de eliminar variáveis já no modelo diferencia o modelo stepwise dos modelos de adição forward / eliminação backward.

4. Continue esse procedimento examinando todas as variáveis independentes não-presentes no modelo para determinar se alguma deveria ser incluída na equação. Se uma nova variável independente é incluída, examine todas as variáveis independentes previamente no modelo para julgar se elas devem ser mantidas.

**Adição forward e eliminação backward**

**Forward**- iniciamos com o modelo nulo e verificamos, usando o critério, se há alguma variável independente ou predictora fora do modelo que valha a pena incluir, apenas uma variável pode entrar no modelo de cada vez (baseado no p-value e significância de entrada), se não houver, o processo termina, se houver, a variável independente é incluída, Depois duma variável entrar no modelo, não pode ser removida e o modelo é refeito, voltando a verificar se existe alguma outra variável independente fora que valha a pena incluir, e o procedimento continua até nenhuma outra variável explicativa possa ser adicionada ao modelo.

No método **backward**, começamos com o modelo que inclui todas as variáveis independentes e vamos excluindo essas variáveis enquanto for vantajoso, ou seja, começamos o modelo com todas as variáveis explicativas inclusas, e em cada etapa (fase), uma variável preditora é removido do modelo baseado no p-value e significância de remoção, em seguida uma variável removida não pode voltar a entrar no modelo, e o procedimento contínua até nenhuma outra variável poder ser removida do modelo.

Ápos a aplicação do método para seleção de variáveis, o *Critério de Informação Akaike (AIC)* e o *Critério de Informação Bayesiano de Schwarz (BIC)* são usados como critério de seleção do modelo que melhor se ajusta aos dados observados.

Embora ambos tenham a função de medir o grau de ajuste do modelo, e escolhendo-se dentre os modelos concorrentes o que tiver menor valor de AIC ou BIC, estes diferem pela sua penalidade, para o número de parâmetros do modelo, onde normalmente o BIC é considerado mais robusto.

$$AIC = n \ln\left(\frac{SQE_p}{n}\right) + 2p \tag{3.7}$$

$$BIC = n \ln\left(\frac{SQE_p}{n}\right) + p \ln(n) \tag{3.8}$$

No presente estudo, tods os métodos para escolha de variáveis foram aplicados, embora somente apresentados os resultados da especificação confirmatória e da estimação stepwise, em que foram considerados somente os modelos finais com menor valor do AIC e BIC.

### 3.3.7 Regressão Logística

A regressão logística é uma técnica estatística que tem como objectivo produzir, a partir de um conjunto de observações, um modelo que permita a predição de valores tomados por uma variável categórica, frequentemente binária, em função de uma ou mais variáveis independentes contínuas e/ou binárias. Então, a partir desse modelo gerado e possível calcular ou prever a probabilidade de um evento ocorrer, dado uma observação aleatória (Gujarati, 2000).

A regressão Logística será usada para a construção de uma equação que possa ajudar na explicação com 95% de confiança dos factores de risco para o aparecimento de novos casos de cancro do colo do útero e cancro da mama nas mulheres atendidas nos centros DREAM da comunidade Sant'Egídio na cidade de Maputo.

**Pressupostos para a Regressão Logística Binária**

Segundo Hair *et al.* (2009), a normalidade não é uma exigência para a aplicação da regressão logística binária, porem a outros cuidados a ter em conta.

**Multicolinearidade**

A multicolinearidade ocorre quando qualquer variável independente é altamente correlacionada com uma ou outras variáveis independentes.

Se no conjunto de dados existirem variáveis que resultam de combinações lineares de outras variáveis, isto faz com que haja elevada correlação entre elas, logo estas variáveis ditas co-variáveis devem ficar de fora do modelo (Hair *et al., 2009*).

A análise de multicolinearidade é feita baseando-se no cálculo do Factor de inflação da variância que mede o efeito de outras variáveis independentes sobre o erro padrão de um coeficiente de regressão.

$$VIF = \frac{1}{1 - R_j^2} \tag{3.9}$$

Habitualmente valores de VIF$>$10 indicam a existência de multicolinearidade, logo a solução passa por retirar uma dessas variáveis e refazer a análise.

Os valores da tolerância que é o inverso do valor do VIF, indicam a existência de multicolinearidade se forem inferiores que 0.1.

**Observações influentes (Outliers)**

São observações que exercem uma influência desproporcional sobre um ou mais aspectos das estimativas da regressão. Estas observações podem ser boas ou não, tendo que se verificar a sua

influência para os dados e desempenhando uma influência negativa remover essas observações do conjunto de dados.

**Teste de Omnibus ou (Teste da razão de verossimilhança)**

O teste de Omnibus avalia se os coeficientes de um modelo de regressão logística são significativos, ou seja, se o modelo como um todo é melhor do que um modelo sem variáveis preditoras ou modelo nulo.

**Hipóteses**
*Hipótese nula (H0)*: Todos os coeficientes do modelo de regressão são iguais a zero.
*Hipótese alternativa (H1)*: Pelo menos um dos coeficientes do modelo de regressão é diferente de zero.

**Cálculo da estatística de teste**

$$G = -2\left(\ln(L_0) - \ln(L_M)\right) \tag{3.10}$$

**Regra de decisão**
*Pela Estatística de Teste*:
Se $G$ for maior que o valor crítico, rejeita-se $H0$
Se $G$ for menor que o valor crítico, não rejeita-se $H0$

*Pelo Valor-p*:
Se o valor-p for menor que o nível de significância $\alpha$, rejeita-se $H0$
Se o valor-p for maior que o nível de significância $\alpha$, não rejeita-se $H0$

**Pseudo $R^2$ de McFadden, de Cox e Snell e $R^2$ de Nagelkerke**

**$R^2$ de Cox e Snell**
Muitos pesquisadores apresentam, em seus trabalhos, um coeficiente conhecido por pseudo $R^2$ de McFadden, cuja expressão é dada por:

$$\text{pseudo } R^2 = \frac{-2LL_0 - (-2LL_{\text{Max}})}{-2LL_0} \tag{3.11}$$

**$R^2$ de Cox e Snell**
Cox e Snell desenvolveram um coeficiente com certas semelhanças por isso é chamado de pseudo, que também se encontra no mesmo intervalo paramétrico (0 a 1), mas com uma interpretação distinta, pois não pode alcançar o valor máximo de um neste intervalo.

$$R^2_{\text{CS}} = 1 - \left(\frac{L(0)}{L(\hat{\theta})}\right)^{\frac{2}{n}} \tag{3.12}$$

onde:
L- refere-se ao valor da função de verossimilhança calculado para o modelo nulo e para o modelo atual ($\theta$) que estamos a avaliar a sua relevância.

Este coeficiente é um indicador da associação entre a variável dependente e o conjunto de variáveis preditoras. Se o valor for inferior a 1, isso sugere uma relação estatisticamente significativa entre a variável dependente e as variáveis preditoras. Além disso, pode ser utilizado para comparar o desempenho de vários modelos concorrentes.

**$R^2$ de Nagelkerke**

O pseudo $R^2$ de Nagelkerke é uma versão ajustada do $R^2$ de Cox e Snell, que também assume valores entre 0 e 1. No entanto, o seu objectivo é tornar mais fácil a interpretação dos resultados, pois é projetado para ser mais preciso e transparente, pois o seu domínio alcança os extremos 0 e 1, e valores acima de 0.30 são considerados tradutores de boa qualidade de ajustamento do modelo.
O coeficiente é dado pela seguinte formula matemática:

$$R^2_{Nagelkerke} = \frac{1 - (\frac{LL}{L0})^{1/n}}{1 - (\frac{L0}{Lmax})^{1/n}} \tag{3.13}$$

**Teste de Hosmer e Lemeshow**

O Teste de Hosmer-Lemeshow verifica o ajuste dos modelos de regressão logística. Ele verifica se as previsões do modelo ajustam-se bem aos dados observados.

**Hipóteses**

*Hipótese nula (H0)*: O modelo de regressão logística ajusta-se bem aos dados.
*Hipótese alternativa (H1)*: O modelo de regressão logística não ajusta-se bem aos dados.

**Cálculo da estatística de teste**

$$\chi^2 = \sum_{g=1}^{G} \frac{(O_g - E_g)^2}{E_g(1 - \pi_g)} \tag{3.14}$$

**Onde**:

- $G$ é o número de grupos;
- $O_g$ é o número observado de eventos no grupo $g$;

- $E_g$ é o número esperado de eventos no grupo $g$;
- $\pi_g$ é a probabilidade média predita para o grupo $g$.

**Regra de decisão**
*Pela Estatística de Teste*:
Se $X^2$ for maior que o valor crítico, rejeita-se $H0$;
Se $X^2$ for menor ou igual ao valor crítico, não rejeita-se $H0$.

*Pelo Valor-p*:
Se o valor-p for menor que o nível de significância $\alpha$, rejeita-se $H0$;
Se o valor-p for maior ou igual ao o nível de significância $\alpha$, não rejeita-se $H0$.

**Teste de Wald**

O Teste de Wald é utilizado para avaliar a significância individual dos coeficientes no modelo de regressão logística. Ele testa se cada coeficiente $\beta_j$ é significativamente diferente de zero, o que indica que a variável preditora associada tem um impacto significativo na variável dependente.

**Hipóteses**
*Hipótese nula (H0)*:$\beta_j = 0$;
*Hipótese alternativa (H1)*: $\beta_j \neq 0$.

**Cálculo da estatística de teste**

$$W = \left( \frac{\hat{\beta}_j}{\text{SE}(\hat{\beta}_j)} \right)^2 \tag{3.15}$$

**Onde**:
$\hat{\beta}_j$ é a estimativa do coeficiente $\beta_j$;
$SE(\hat{\beta}_j)$ é o erro padrão da estimativa do coeficiente $\beta_j$.

**Regra de decisão**
*Pela Estatística de Teste*:
Se $W$ for maior que o valor crítico, rejeita-se $H_0$;
Se $W$ for menor ou igual ao valor crítico, não rejeita-se $H_0$;

*Pelo Valor-p*:
Se o valor-p for menor ou igual ao o nível de significância $\alpha$, rejeita-se $H_0$;
Se o valor-p for maior que o nível de significância $\alpha$, não rejeita-se $H_0$;

### Modelo Logístico

#### Interpretação dos coeficientes do modelo

De acordo com Reis (2001), uma das principais estatísticas utilizadas na análise de dados binários é a razão de chances (RC) ou odds ratio, é definida como a razão entre a chance de um evento ocorrer em um grupo e a chance de ocorrer em outro grupo. Chance é a probabilidade de ocorrência deste evento dividida pela probabilidade da não ocorrência do mesmo evento que é definida pela seguinte formula:

$$RC(X_i) = e^{\beta_i}; \qquad i = 0, 1, \ldots, 13.$$

Se$\beta_i > 0$ as chances aumentam;
Se $\beta_i < 0$ as chances diminuem.
Ou:
Se $RC(X_i) > 1$ as chances aumentam;
Se $RC(X_i) < 1$ as chances diminuem.

#### Capacidade predictiva do modelo

A **eficiência global** do modelo corresponde ao percentual de acerto da classificação para um determinado cutoff.

A **sensibilidade** diz respeito ao percentual de acerto, para um determinado cutoff, considerando-se apenas as observações que de facto são sucesso ou evento.

$$\text{Sensibilidade} = \frac{\text{Verdadeiros Positivos}}{\text{Verdadeiros Positivos} + \text{Falsos Negativos}} \tag{3.16}$$

A **especificidade**, por outro lado, refere-se ao percentual de acerto, para um dado cutoff, considerando-se apenas os insucessos ou as observações que não são evento.

$$\text{Especificidade} = \frac{\text{Verdadeiros Negativos}}{\text{Verdadeiros Negativos} + \text{Falsos Positivos}} \tag{3.17}$$

Para mensurar a capacidade preditiva do modelo, foi fixado um cutoff de 0.5 em que todos os valores de probabilidade maior ou igual a este foram classificados como sucesso, ou seja probabilidade do individuo estar realmente infectado pelo HPV.

# Capítulo 4

# RESULTADOS E DISCUSSÃO

## 4.1 Análise Exploratória dos Dados

Após arrolados os dados, procede-se com o próximo passo da análise estatística que consiste em descrever os dados, ou seja, apresentar e descrever as suas principais características, com o objectivo de representar de uma forma compreensível a informação contida nos dados e estabelecer bases para inferências estatísticas mais detalhadas apartir da previsão.

### 4.1.1 Análise de Dados Perdidos (missing values)

Valores ausentes podem surgir por diversas razões, como erros na recolha de dados, respostas não fornecidas, problemas no registo, entre outros, e seu impacto é prejudicial não só devido ao seu enviesamento oculto nos resultados, mas também devido ao seu impacto prático na dimensão da amostra disponível.


Figura 4.1: Padrões dos valores perdidos

O gráfico na Figura 4.1, tem como objectivo verificar a existência ou não de algum padrão

entre os valores perdidos dentro da base de dados, após uma breve análise gráfica, podemos constatar que das 35 variáveis aqui analisadas, a maioria apresenta um padrão de dados perdidos aleatórios, 4 destas apresentam um padrão elevado de dados perdidos em mais de 90%, nomeadamente as variáveis exame especular para colo formato, colo sangramento e outros, rastreio com HPV-DNA e VIA para HPV16/18 tratamento, e também para algumas observações as variáveis sócio-demográficas apresentam um padrão repetido de valores perdidos.

Tabela 4.1: Principais características das variáveis em estudo

| Variável | N | Média | SD | Missing Values | | Número dos Extremos | |
|---|---|---|---|---|---|---|---|
| | | | | n | % | Min. | Max. |
| Idade | 161 | 41.1 | 9.1 | 0 | 0% | 17 | 68 |
| Numero_de_gestações | 161 | 3.49 | 2.11 | 4 | 2.5% | 0 | 10 |
| Idade_da_1a Relação_sexual | 161 | 17.4 | 2.64 | 11 | 6.8% | 12 | 42 |
| Altura | 161 | 162.2 | 6.34 | 12 | 7.5% | 144 | 174 |
| Peso | 161 | 72.3 | 16.0 | 15 | 9.3% | 39.7 | 128 |
| Sexo | 161 | — | — | 0 | 0% | — | — |
| Ciclo_menstrual | 161 | — | — | 6 | 3.7% | — | — |
| Sangramento_vaginal | 161 | — | — | 9 | 5.6% | — | — |
| Planeamento_Familiar | 161 | — | — | 3 | 1.9% | — | — |
| Tipo_de_rastreio_realizado | 161 | — | — | 2 | 1.2% | — | — |
| Menopausa | 161 | — | — | 1 | 0.6% | — | — |
| ITS | 161 | — | — | 1 | 0.6% | — | — |
| ITS_Sim | 161 | — | — | 1 | 0.6% | — | — |
| Tabagismo | 161 | — | — | 0 | 0% | — | — |
| HIV | 161 | — | — | 3 | 1.9% | — | — |
| Historia_de_rastreio | 161 | — | — | 1 | 0.6% | — | — |
| Historia_de_rastreio_Se_Sim | 161 | — | — | 1 | 0.6% | — | — |
| Data_do_ultimo_rastreio_IC | 161 | — | — | 8 | 5.0% | — | — |
| Resultado_do_ultimo_rastreio | 161 | — | — | 3 | 1.9% | — | — |
| Se_Ultimo_rastreio_Positivo | 161 | — | — | 3 | 1.9% | — | — |
| Exame_pélvico | 161 | — | — | 0 | 0% | — | — |
| Exame_Especular_para_colo_Formato | 161 | — | — | 160 | 99.4% | — | — |
| Exame_Especular_para_colo_Sangramento | 161 | — | — | 2 | 1.2% | — | — |
| Exame_Especular_para_colo_Sangramento_E_outros | 161 | — | — | 156 | 96.9% | — | — |
| Exame_Especular_para_colo_Visualização | 161 | — | — | 0 | 0% | — | — |
| Rastreio_uterino_com_VIA | 161 | — | — | 2 | 1.2% | — | — |
| Rastreio_com_HPV-DNA | 161 | — | — | 146 | 90.7% | — | — |
| HPV | 161 | — | — | 0 | 0% | — | — |
| HPV_Positivo | 161 | — | — | 0 | 0% | — | — |
| VIA_P_HPV16/18_Tratamento | 161 | — | — | 156 | 96.9% | — | — |
| Tratamento_feito | 161 | — | — | 3 | 1.9% | — | — |
| unidade_Sanitária | 161 | — | — | 0 | 0% | — | — |
| Nível_de_escolaridade | 161 | — | — | 24 | 14.9% | — | — |
| Estado_civil | 161 | — | — | 10 | 6.2% | — | — |
| Agua_na_casa | 161 | — | — | 17 | 10.6% | — | — |

A Tabela 4.1, mostra que a amostra analisada no presente estudo é composta por 161 pacientes, submetidas ao rastreio do cancro do colo do útero, com idades compreendidas entre os 17 aos 68 anos, com uma média de 41 anos. Quanto ao numero de gestações, estas apresentaram entre 0 a 10, onde a média foi de 3 e com cerca de 2.5% de valores omissos. A idade da pri-

meira relação variou entre 12 aos 42 anos, com uma média de 17 anos e com valores omissos em 6.8%. No que diz respeito a altura das mesmas, esta variou de 144 a 174cm, com uma media de 162.2cm e com valores omissos de 7.5%. Em relação ao peso, estas apresentavam valores entre 39.7 a 128Kg, com um peso médio de 72.3Kg, e uma percentagem de valores omissos de 9.3%. As variáveis como o sexo, tabagismo, exame pélvico, exame especular para colo visualização, HPV, HPV positivo e unidade sanitária não registaram valores ausentes. As demais variáveis categóricas com valores omissos tiveram as seguintes percentagens, ciclo menstrual 3.7%, sangramento vaginal 5.6%, planeamento familiar 1.9%, tipo de rastreio realizado 1.2%, Menopausa 0.6%, ITS 0.6%, ITS sim 0.6%, HIV 1.9%, historia de rastreio 0.6%, historia de rastreio se sim 0.6%, data do ultimo rastreio em IC 5%, resultado do ultimo rastreio 1.9%, se o ultimo rastreio foi positivo 1.9%, exame especular para o formato 99,4%, exame especular para o sangramento 1.2%, exame especular para o sangramento e outros 96.9%, rastreio com VIA 1.2%, rastreio com HPV-DNA 90.7%, VIA para HPV 16/18 96.9%, tratamento feito 1.9%, nível de escolaridade 14.9%, estado civil 6.2% e agua na casa 10.6% respectivamente.

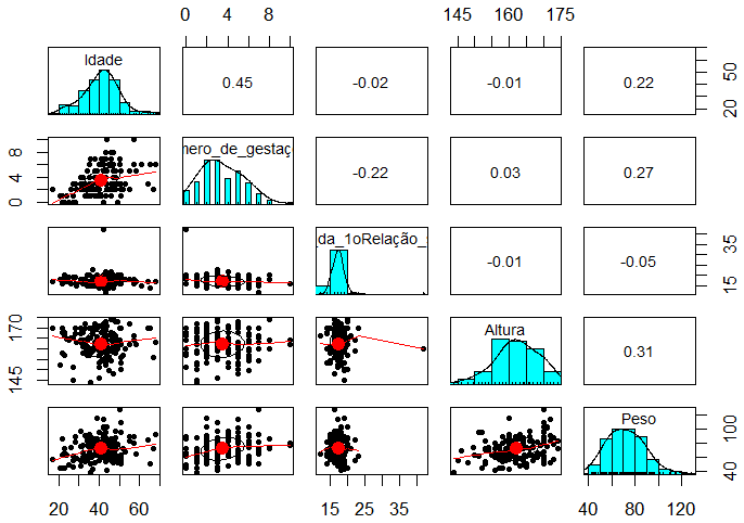


Figura 4.2: Matriz de Correlação e teste de normalidade e homoscedasticidade de Hawkins

Pelo teste de Hawkins e o teste não paramétrico de homoscedasticidade para verificar a normalidade e homoscedasticidade e se os missing values existentes na base de dados são completamente aleatórios ou não, rejeitamos a hipótese nula de que os missing values são perdidos completamente ao acaso (MCAR- missing completely at random), ou seja, rejeitamos a hipótese de que os missing values são completamente aleatórios, com um nível de significância de 5%, com um P-value de 0.51% e de 1.1% respectivamente, o que indica que para usar todos os dados existentes na base de dados, deve-se usar métodos de tratamentos para lidar com dados em falta com destaque para a imputação para variáveis com dados perdidos superiores a 5%.

No presente estudo para a atribuição por remoção de linhas ou colunas, foram consideradas todas as variáveis com a percentagem de valores omissos abaixo de 20% e desconsideradas todas as demais cuja percentagem estava acima da fixada.

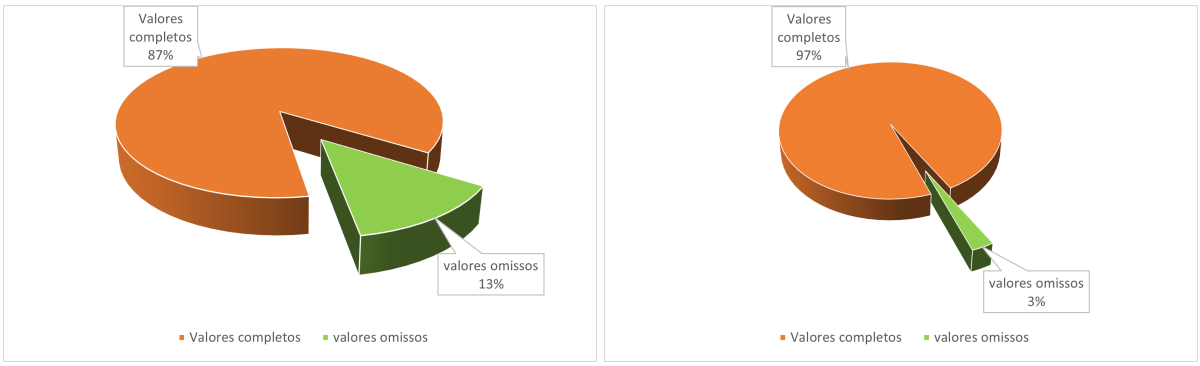


Figura 4.3: Gráficos Circulares dos valores omissos para a remoção e imputação

Os Gráficos na Figura B.2, mostram que após a remoção das variáveis com mais de 20% de valores omissos, a percentagem de missing values diminuiu significativamente de 13.48% para 2.84% percentagem esta que segundo a literatura ja nos permite fazer análises com os dados existentes pois esta é menor que 5% fixado como limite, porem com o intuito de possibilitar a comparação futura de modelos estimados, procede-se com outros metodos de imputação com vista a salvaguardar o tamanho inicial da amostra.

Tabela 4.2: Métodos de imputação

| Variável | Tipo de modelo | Número de valores omissos | Valores imputados |
|---|---|---|---|
| Número de gestações | Regressão Linear | 4 | 12 |
| Idade da primeira relação sexual | Regressão Linear | 11 | 33 |
| Ciclo menstrual | Regressão Logística | 6 | 18 |
| Sangramento vaginal | Regressão Logística | 9 | 27 |
| Planeamento familiar | Regressão Logística | 3 | 9 |
| Tipo de rastreio realizado | Regressão Logística | 2 | 6 |
| Menopausa | Regressão Logística | 1 | 3 |
| ITS | Regressão Logística | 1 | 3 |
| ITS Sim | Regressão Logística | 1 | 3 |
| HIV | Regressão Logística | 3 | 9 |
| História de rastreio | Regressão Logística | 1 | 3 |
| História de rastreio se Sim | Regressão Logística | 1 | 3 |
| Data do último rastreio IC | Regressão Logística | 8 | 24 |
| Resultado do último rastreio | Regressão Logística | 3 | 9 |
| Se o último rastreio positivo | Regressão Logística | 3 | 9 |
| Exame especular para colo Sangramento | Regressão Logística | 2 | 6 |
| Rastreio uterino com VIA | Regressão Logística | 2 | 6 |
| Tratamento feito | Regressão Logística | 3 | 9 |
| Nível de escolaridade | Regressão Logística | 24 | 72 |
| Altura | Regressão Linear | 12 | 36 |
| Peso | Regressão Linear | 15 | 45 |
| Estado civil | Regressão Logística | 10 | 30 |
| Água na casa | Regressão Logística | 17 | 51 |

De acordo com a Tabela 4.2, para as variáveis numéricas, o método de imputação usado foi a regressão linear múltipla e para as variáveis categóricas com duas e com mais de duas

categorias foram a regressão Logística múltipla e multinomial respetivamente, onde para cada variável, foram aplicados 3 diferentes métodos de imputação por regressão com vista a garantir a consistência dos valores imputados.

### 4.1.2 Análise exploratória univariada

Tabela 4.3: Principais Características Amostrais das Variáveis Númericas

| Variável | N | Média | Mediana | Moda | SD | Missing Values | | Número dos Extremos | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | n | % | Min. | Máx. |
| Idade | 161 | 41.1 | 42 | 22 & 23 | 9.1 | 0 | 0% | 17 | 68 |
| Numero_de_gestações | 161 | 3.49 | 3 | 2 & 3 | 2.11 | 0 | 0% | 0 | 10 |
| Idade_da_1a Relação_sexual | 161 | 17.4 | 17 | 17 | 2.64 | 0 | 0% | 12 | 42 |
| Altura | 161 | 162.2 | 163 | 170 | 6.34 | 0 | 0% | 144 | 174 |
| Peso | 161 | 72.3 | 72.3 | 45.6 | 16.0 | 0 | 0% | 39.7 | 128 |

Na Tabela 4.3, após a imputação de valores na base de dados, a média, a variância e os valores extremos das variáveis numéricas não alterou, porem a idade cujo maior numero de pacientes têm 22 ou 23 anos. Quanto ao numero de gestações, estas também na maioria dos casos entre 2 e 3 gestações. A idade da primeira relação muitas delas responderam ter iniciado com 17 anos. No que diz respeito a altura das mesmas, 170cm foi a altura que registou-se com mais frequência. Em relação ao peso, estas apresentavam com maior frequência 45.6Kg.

Tabela 4.4: Testes de Normalidade de Kolmogorov-Smirnov e Shapiro-Wilk

| **Variável** | **Kolmogorov-Smirnov** | | | **Shapiro-Wilk** | | |
|---|---|---|---|---|---|---|
| | **Estatística D** | **gl** | **P-value** | **Estatística W** | **gl** | **p-value** |
| Idade | 0.06 | - | 0.45 | 0.98 | - | 0.02 |
| Numero de gestações | 0.14 | - | 0.002 | 0.95 | - | 0.00< |
| Idade da 1oRelação sexual | 0.21 | - | 0.00 < | 0.62 | - | 0.00< |
| Peso | 0.06 | - | 0.51 | 0.98 | - | 0.03 |
| Altura | 0.06 | - | 0.45 | 0.97 | - | 0.01 |

Na Tabela 4.4, são apresentados os testes de normalidade de Kolmogorov-Smirnov e Shapiro-Wilk para as varáveis numéricas, onde a um nível de significância de 5%, para o teste de kolmogorov rejeitamos apenas a hipótese nula para as variáveis numero de gestações e idade da primeira relação sexual, e pelo teste de Shapiro-Wilk rejeitamos a hipótese nula de que a variável x segue a distribuição normal para todas as variáveis pois o valor de p é menor que o nível de significância em todos os casos, o que podemos assumir pelo exame dos gráficos de normalidade em anexo, e em censo comum rejeitamos a hipótese nula pelo teste de normalidade de Shapiro-Wilk.

### 4.1.3 Análise exploratória bivariada

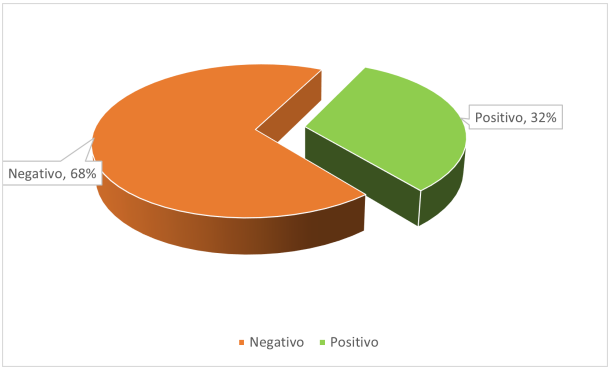


Figura 4.4: HPV

Os resultados ilustrados na Figura 4.4, e na Tabela A.1 em anexo, mostram que para as 161 mulheres que foram submetidas ao rastreio do cancro do colo do útero, 109 (67.7%) tiveram como resultado do teste de HPV negativo, e as restantes 52 (32.3%), tiveram um resultado positivo para este teste.

Deste mesmo numero, 6 delas com o ciclo menstrual irregular tiveram igual percentagem de resultados nos testados ao HPV em relação ao total 1.86%, as 131 com um ciclo regular, 54% foram negativas e 27% positivas, e as restantes 19 mulheres que ja haviam atingido a menopausa, 11.8% tiveram resultado negativo e 3.11% resultado positivo para o HPV.

Das 125 mulheres que não usavam nenhum método de planeamento familiar, 54.6% destas testaram negativo e as restantes 22,98% testaram positivo e as demais estiveram dividas entre os demais métodos entre o DIU, implante, pílula, preservativo e outros.

Quanto ao tipo de rastreio, das 101 que realizaram o rastreio pela primeira vez, 43.48% testaram negativo para o HPV e 19.25% positivo, as 60 que realizavam o rastreio para seguimento, 24.22% testaram negativo e as restantes 13% testaram positivo.

Ao todo, as 135 que ainda não haviam atingido a menopausa, 55.9% tiveram um resultado negativo e 27.95% tiveram um resultado positivo, para as 26 que ja haviam atingido a menopausa, 11.8% tiveram um resultado negativo e as restantes 4.35% um resultado positivo.

Para as 147 mulheres que não apresentaram nenhuma ITS, 61.49% foram negativas para o HPV e 29.81% foram positivas, e das 12 com alguma ITS a leucorreia especificamente 4.97% foram negativas e 2.48% foram positivas para o HPV.

Num total de 148 mulheres positivas para o HIV, 62.11% foram negativas para o HPV e 29.81% foram positivas para o HPV, e das 13 negativas, 5.59% foram negativas e 2.48% foram positivas para o HPV.

As 104 mulheres que nunca haviam realizado nenhum rastreio para o cancro do colo do útero, 44.72% foram negativas para o HPV e 19.88% foram positivas, e para as 57 que ja haviam realizado algum rastreio, 22.98% foram negativas e 12.42% foram positivas e destas somente 2 que testaram para o HPV representando 1.24% testaram negativo para o HPV e ainda para estas, 1.86% havia feito o ultimo rastreio a menos de 3 meses, 27.33% havia feito a mais de 3 meses e menos de 1 ano, e 3.73% a mais de 3 anos. Das que ja haviam realizado algum tipo de rastreio, somente 1.24% foram positivas para o cancro do colo do útero e as demais 36% negativas.

No exame especular das 158 sem sangramento, 65.84% foram negativas e 32.3% foram positivas para o HPV, a apenas 1 com sangramento (0.62%) foi negativa para o HPV, as 2 com JEC parcialmente visualizada foram negativas, e as 158 com JEC totalmente visualizada, 65.84% foram negativas, e 32.3% foram positivas.

As pacientes com o HPV positivo, 4.97% tinham 0 HPV16, 3.73% tinham o HPV16 ou outro, 0.62% tinham o HPV18/45, e 23.6% tinham outro tipo de HPV.

Das 65 mulheres com o nível primário de escolaridade, 27.33% foram negativas ao HPV e 13% foram positivas, a única sem nenhum nível de escolaridade teve resultado positivo para o HPV, as 80 com o nível secundário 33.54% tiveram resultado negativo e 16.15% tiveram resultado positivo, e as 15 com nível superior, 6.83% foram negativas e 2.48% foram positivas ao HPV.

Para as 21 mulheres casadas 11.18% foram negativas para o HPV e 1.86% positivas, as 3 separadas ou divorciadas 1.24% foram negativas e 0.62% foram positivas, as 135 solteiras, 54% foram negativas e 29.81% foram positivas e 1.24% das viúvas foram negativas para o HPV.

Das 28 que afirmaram não ter água em casa, 13.66% tiveram como resultado do teste negativo e 3.73% tiveram resultado positivo, e as 133 que afirmaram ter água em casa 54% foram negativas e 28.57% foram positivas para o HPV.

Pelo teste de independência de qui-quadrado de pearson, a um nível de significância de 5%, há fortes evidências para rejeitar a hipótese nula de que o HPV e as demais variáveis são independentes ou seja não são associadas. Em relação as variáveis sangramento vaginal, rastreamento uterino com via, HPV positivo, tratamento feito, sexo, tabagismo, exame pélvico. Portanto,

podemos afirmar que existe associação entre a variável resposta e as variáveis mencionadas , com a exceção das variáveis não mencionadas.

Tabela 4.5: Distribuição das frequências absolutas, relativas e o teste de teste de Mann-Whitney

| Variável | Categorias das Variáveis | HPV | | | Mann Whitney | |
|---|---|---|---|---|---|---|
| | | Negativo | Positivo | Total | $V$ | p-value |
| Idade | [17,21) | 1 (0.62%) | 0 (0%) | 1 (0.62%) | | |
| | [21,25) | 4 (2.48%) | 4 (2.48%) | 8 (4.97%) | | |
| | [25,29) | 6 (3.73%) | 1 (0.62%) | 7 (4.35%) | | |
| | [29,33) | 6 (3.73%) | 4 (2.48%) | 10 (6.21%) | | |
| | [33,37) | 5 (3.11%) | 12 (7.45%) | 17 (10.56%) | | |
| | [37,41) | 17 (10.56%) | 11 (6.83%) | 28 (17.39%) | | |
| | [41,45) | 27 (16.77%) | 10 (6.21%) | 37 (22.98%) | 300 | 0.00< |
| | [45,49) | 24 (14.91%) | 3 (1.86%) | 27 (16.77%) | | |
| | [49,53) | 11 (6.83%) | 4 (2.48%) | 15 (9.32%) | | |
| | [53,57) | 2 (1.24%) | 1 (0.62%) | 3 (1.86%) | | |
| | [57,61) | 2 (1.24%) | 1 (0.62%) | 3 (1.86%) | | |
| | [61,65) | 2 (1.24%) | 0 (0%) | 2 (1.24%) | | |
| | [65,69) | 2 (1.24%) | 1 (0.62%) | 3 (1.86%) | | |
| | **Total** | 109 (67.7%) | 52 (32.3%) | 161 (100%) | | |
| Numero de gestações | [0,1) | 5 (3.11%) | 6 (3.73%) | 11 (6.83%) | | |
| | [1,2) | 11 (6.83%) | 5 (3.11%) | 16 (9.94%) | | |
| | [2,3) | 19 (11.8%) | 12 (7.45%) | 31 (19.25%) | | |
| | [3,4) | 21 (13.04%) | 10 (6.21%) | 31 (19.25%) | | |
| | [4,5) | 17 (10.56%) | 4 (2.48%) | 21 (13.04%) | | |
| | [5,6) | 15 (9.32%) | 8 (4.97%) | 23 (14.29%) | | |
| | [6,7) | 11 (6.83%) | 4 (2.48%) | 15 (9.32%) | 171 | 0.00< |
| | [7,8) | 5 (3.11%) | 3 (1.86%) | 8 (4.97%) | | |
| | [8,9) | 3 (1.86%) | 0 (0%) | 3 (1.86%) | | |
| | [9,10) | 0 (0%) | 0 (0%) | 0 (0%) | | |
| | [10,11) | 2 (1.24%) | 0 (0%) | 2 (1.24%) | | |
| | **Total** | 109 (67.7%) | 52 (32.3%) | 161 (100%) | | |
| Idade da 1oRelação sexual | [12,15) | 5 (3.11%) | 1 (0.62%) | 6 (3.73%) | | |
| | [15,18) | 58 (36.02%) | 30 (18.63%) | 88 (54.66%) | | |
| | [18,21) | 44 (27.33%) | 18 (11.18%) | 62 (38.51%) | 45 | 0.009 |
| | [21,41) | 2 (1.24%) | 3 (1.86%) | 4 (2.48%) | | |
| | **Total** | 109 (67.7%) | 52 (32.3%) | 161 (100%) | | |
| Altura | [144,147) | 1 (0.62%) | 2 (1.24%) | 3 (1.86%) | | |
| | [147,150) | 2 (1.24%) | 0 (0%) | 2 (1.24%) | | |
| | [150,153) | 6 (3.73%) | 0 (0%) | 6 (3.73%) | | |
| | [153,156) | 4 (2.48%) | 5 (3.11%) | 9 (5.59%) | | |
| | [156,159) | 12 (7.45%) | 6 (3.73%) | 18 (11.18%) | | |
| | [159,162) | 23 (14.29%) | 10 (6.21%) | 33 (20.5%) | 190 | 0.00< |
| | [162,165) | 25 (15.53%) | 8 (4.97%) | 33 (20.5%) | | |
| | [165,168) | 12 (7.45%) | 8 (4.97%) | 20 (12.42%) | | |
| | [168,171) | 16 (9.94%) | 7 (4.35%) | 23 (14.29%) | | |
| | [171,174) | 6 (3.73%) | 6 (3.73%) | 12 (7.45%) | | |
| | [174,177) | 2 (1.24%) | 0 (0%) | 2 (1.24%) | | |
| | **Total** | 109 (67.7%) | 52 (32.3%) | 161 (100%) | | |
| Peso | [39.7,46.7) | 4 (2.48%) | 2 (1.24%) | 6 (3.73%) | | |
| | [46.7,53.7) | 7 (4.35%) | 3 (1.86%) | 10 (6.21%) | | |
| | [53.7,60.7) | 12 (7.45%) | 9 (5.59%) | 21 (13.04%) | | |
| | [60.7,67.7) | 24 (14.91%) | 8 (4.97%) | 32 (19.88%) | | |
| | [67.7,74.7) | 18 (11.18%) | 8 (4.97%) | 26 (16.15%) | | |
| | [74.7,81.7) | 13 (8.07%) | 7 (4.35%) | 20 (12.42%) | | |
| | [81.7,88.7) | 16 (9.94%) | 6 (3.73%) | 22 (13.66%) | 253 | 0.00< |
| | [88.7,95.7) | 6 (3.73%) | 7 (4.35%) | 13 (8.07%) | | |
| | [95.7,103) | 5 (3.11%) | 0 (0%) | 5 (3.11%) | | |
| | [103,110) | 2 (1.24%) | 1 (0.62%) | 3 (1.86%) | | |
| | [110,117) | 1 (0.62%) | 1 (0.62%) | 2 (1.24%) | | |
| | [117,131) | 1 (0.62%) | 0 (0%) | 1 (0.62%) | | |
| | **Total** | 109 (67.7%) | 52 (32.3%) | 161 (100%) | | |

Após não tendo sido comprido o pressuposto de normalidade pelo teste de Shapiro wilk para as variáveis numéricas na Tabela 4.4 e pelos Histogramas da Figura B.2 em anexo, pelo teste não paramétrico Mann-Whitney U na Tabela 4.5, podemos concluir que como o valor de p é menor que o nível significância de 5% estabelecido, as variáveis numéricas têm diferenças estatisticamente significativas nos seus valores médios entre os grupos. Conclui-se que no grupo dos positivos para o HPV a idade média é superior à do grupo dos negativos para o HPV como o numero médio de gestações, idade da primeira relação sexual, altura e peso das mulheres.

## 4.2 Regressão Logística

Tabela 4.6: VIF e Tolerância

| Variável | $\beta$ | std. Error | $t$ value | $Pr(>\|t\|)$ | VIF | Tolerância |
|---|---|---|---|---|---|---|
| Idade | 1.25 | 0.68 | 1.82 | 0.07 | 1.43 | 0.70 |
| Numero_de_gestações | -0.44 | 0.29 | -1.52 | 0.13 | 1.36 | 0.73 |
| Idade_da_1a Relação_sexual | 0.46 | 0.38 | 1.22 | 0.22 | 1.32 | 0.75 |
| Ciclo_menstrual | -0.16 | 0.24 | -0.66 | 0.50 | 2.09 | 0.47 |
| Sangramento_vaginal | -1.09 | 0.57 | -1.87 | 0.06 | 2.05 | 0.48 |
| Planeamento_Familiar | 0.48 | 0.51 | 0.94 | 0.34 | 1.28 | 0.77 |
| Tipo_de_rastreio_realizado | -0.17 | 0.39 | -0.44 | 0.65 | 4.59 | 0.18 |
| Menopausa | 0.17 | 0.19 | 0.89 | 0.37 | 2.00 | 0.49 |
| ITS | 0.13 | 0.57 | 0.23 | 0.81 | 1.62 | 0.61 |
| HIV | 0.06 | 0.21 | 0.31 | 0.75 | 1.60 | 0.62 |
| Historia_de_rastreio | 0.25 | 0.56 | 0.45 | 0.65 | 7.42 | 0.13 |
| Data_do_ultimo_rastreio_IC | 0.42 | 0.38 | 1.10 | 0.27 | 2.14 | 0.46 |
| Resultado_do_ultimo_rastreio | -0.53 | 0.53 | -0.99 | 0.32 | 2.31 | 0.43 |
| Exame_Especular_para_colo_Sangramento | -0.39 | 1.33 | -0.29 | 0.77 | 1.82 | 0.54 |
| Tratamento_feito | -0.47 | 0.25 | -1.88 | 0.06 | 1.40 | 0.71 |
| Nivel_de_escolaridade | -1.30 | 0.58 | -2.23 | 0.02 | 1.36 | 0.73 |
| Altura | -0.63 | 0.37 | -1.69 | 0.09 | 1.32 | 0.75 |
| Peso | -0.79 | 0.81 | -0.98 | 0.33 | 1.40 | 0.71 |
| Estado_cívil | 0.62 | 0.16 | 1.48 | 0.14 | 1.56 | 0.64 |
| Agua_na_casa | -0.00 | 0.15 | -0.03 | 0.97 | 1.59 | 0.62 |

A Tabela 4.6 apresenta os resultados da estimação dos parâmetros do modelo de regressão linear múltipla com o intuito de análisar a multicolinearidade pelo cálculo do Factor de Inflação da Variância (VIF) e a tolerância, e os resultados apresentados sugerem que as variáveis candidatas a entrarem no modelo de regressão Logística não apresentam problemas de multicolinearidade pois os valores do VIF estão abaixo do limite comum (10), e os valores da tolerância estão todos acima do limite inferior permitido (0.1), com isto podendo proceder para a estimação dos parâmetros deste modelo.

### 4.2.1 Regressão Logística para dados sem imputação pelo método de stepwise

Tabela 4.7: Teste de Omnibus para dados sem imputação

| Etapas | $\chi^2$ | P-value |
|---|---|---|
| Passo | 119.25 | 0.00 |
| Bloco | 60.70 | 0.00 |
| Modelo | 58.24 | 0.00 |

Na Tabela 4.7 apresenta-se o teste de Omnibus para as etapas: passo, bloco, e modelo com objectivo de verificar se os parâmetros a serem estimados no modelo são todos nulos, e pelos valores de p, rejeitamos essa hipótese pois todos são inferiores que o nível de significância de 5%, concluindo assim que o modelo com as variáveis independentes é significativo, sugerindo que pelo menos uma destas variáveis contribui para explicar a variável resposta HPV.

Tabela 4.8: Razão de Verossimilhança, Pseudo R de McFadden, Pseudo R Cox e Snell, Pseudo de Nagelkerke

| Etapas | Probabilidade de **-2** log | Pseudo $R^2$ de McFadden | $R^2$ de Cox e Snell | $R^2$ de Nagelkerke |
|---|---|---|---|---|
| 1 | -50.02 | 0.04 | 0.05 | 0.07 |

Na Tabela 4.8, o valor da razão de verossimilhança indica que o modelo estimado ajusta-se adequadamente aos dados, pois o seu valor tende a ser muito reduzido, e devido a fraca robustez dos $R^2$ estimados para a regressão logística comparativamente ao $R^2$ ajustado obtido na estimação de outros tipos de regressão, a noção sobre a adequação dos modelos, sugerem pelos pseudos coeficientes de determinação de Nagelkerke e de Cox e snell que o modelo é capaz de explicar 7% e 5% respectivamente das variações registadas entre os positivos e negativos para o HPV.

Tabela 4.9: Teste de Hosmer e Lemeshow

| Etapas | $\chi^2$ | P-value |
|---|---|---|
| 1 | 7.27 | 0.49 |

Pelo valor de P 49% na Tabela 4.9 obtido pelo teste de Hosmer Lemeshow, indica que não existe uma diferença entre as frequências observadas e as esperadas pois o valor é maior que o nível de significância de 5%, com isto não rejeitando a hipotese nula e contudo concluindo que o modelo de regressão logística estimado para os dados sem imputação tem um bom ajuste.

Tabela 4.10: Classificação do modelo

| | **Previsão HPV Positiva** | **Previsão HPV Negativa** | Percentagem correcta |
|---|---|---|---|
| **Observado HPV Positivo** | VP = 4 | FN = 0 | 100% |
| **Observado HPV Negativo** | FP = 8 | VN = 0 | 0% |
| percentagem Total | | | 33.3% |

Na Tabela 4.10, constatamos que o nosso modelo de regressão logística estimado, tem uma sensibilidade em identificar corretamente os resultados positivos de HPV de 100%, e uma especificidade de identificar os resultados negativos de HPV de 0%. com isto, concluindo que globalmente, o modelo teve uma taxa de sucesso geral de 33.3%.

### Modelo de Regressão Logística Binário Para variáveis sem imputação

Pelo método de stepwise, a variável Resultado do ultimo rastreio têm um impacto significativo no resultado positivo para o HPV, esta significância da variável a 5%, é descrita pelo modelo

abaixo.

$$\ln\left(\frac{\pi_j}{1-\pi_j}\right) = 1.122 - 0.747 \cdot \text{Resultado do ultimo rastreio Negativo}$$

Os resultados da Tabela B.1 em anexo, indicam que o aumento de uma unidade para o resultado do ultimo rastreio negativo, fazem com que as chances de ocorrência do evento de interesse (HPV positivo) diminua em 53%, pois os seu valores das razões de chances 0.47 são menores que 1, mantendo todas as demais variáveis explicativas constantes.

Para o nosso modelo estimado sem imputação de dados, as mulheres que tiveram o resultado negativo no ultimo rastreio negativo têm uma probabilidade 0.74 vezes menor de terem resultado positivo para o HPV, comparativamente as que nunca realizaram o rastreio com um p valor de 0.004 e uma RC de 0.47.

### 4.2.2 Regressão Logística para dados com imputação pelo método de stepwise

Tabela 4.11: Teste de Omnibus para os dados Sem e com imputação

| Imputação | Etapas | $\chi^2$ | P-value |
|---|---|---|---|
| **Sem** | Passo | 119.25 | 0.00 |
| | Bloco | 60.70 | 0.00 |
| | Modelo | 58.24 | 0.00 |
| **Com** | Passo | 97.19 | 0.00 |
| | Bloco | 999.64 | 0.00 |
| | Modelo | 902.44 | 0.00 |

Na Tabela 4.11, ao comparar os dois modelos, ambos rejeitam a hipótese nula e indicam que as variáveis independentes têm uma contribuição significativa para explicar a variável resposta HPV. No entanto, o modelo com imputação mostra valores de $\chi^2$ muito mais elevados nas etapas de bloco e modelo, sugerindo uma explicação mais robusta e abrangente da variabilidade na variável resposta. Portanto, o teste de Omnibus, sugere que o modelo com imputação é preferível, pois fornece uma evidência mais forte da significância das variáveis independentes, melhorando a capacidade do modelo de explicar a variável resposta HPV.

Tabela 4.12: Razão de Verossimilhança, Pseudo R de McFadden, Pseudo R Cox e Snell, Pseudo de Nagelkerke para dados sem e com imputação

| Imputação | Etapas | Probabilidade de **-2** log | Pseudo $R^2$ de McFadden | $R^2$ de Cox e Snell | $R^2$ de Nagelkerke |
|---|---|---|---|---|---|
| **Sem** | 1 | -50.02 | 0.04 | 0.05 | 0.07 |
| **Com** | 3 | -41.88 | 0.53 | 0.48 | 0.68 |

A Tabela 4.12 apresenta resultados que sugerem que o modelo com imputação se ajusta

melhor aos dados, como indicado pelo aumento na razão de verossimilhança, que é um indicativo de adequação do modelo. Além disso, os pseudo coeficientes de determinação (McFadden, Cox e Snell, Nagelkerke) mostram que o modelo com imputação é capaz de explicar uma maior percentagem da variação na variável resposta HPV (53% pelo pseudo $R^2$ de McFadden, 48% pelo $R^2$ de Cox e Snell, e 68% pelo $R^2$ de Nagelkerke), quando comparado com o modelo sem imputação.

Tabela 4.13: Teste de Hosmer e Lemeshow para dados sem e com imputação

| Imputação | Etapas | $\chi^2$ | P-value |
|---|---|---|---|
| **Sem** | 1 | 7.27 | 0.49 |
| **Com** | 3 | 3,62 | 0.89 |

A partir dos resultados do teste de Hosmer e Lemeshow na Tabela 4.13, podemos concluir que tanto o modelo sem imputação quanto o modelo com imputação ajustam-se bem aos dados. No entanto, o modelo com imputação mostra um ajuste ligeiramente melhor, como indicado pelo maior valor de p 89% quando comparado ao modelo sem imputação, sugerindo que a imputação pode melhorar a adequação do modelo na previsão dos valores observados para a variável resposta HPV.

Tabela 4.14: Classificação do modelo para dados sem e com imputação

| Imputação | | **Previsão HPV Positiva** | **Previsão HPV Negativa** | Percentagem correcta |
|---|---|---|---|---|
| **Sem** | **Observado HPV Positivo** | VP = 4 | FN = 0 | 100% |
| | **Observado HPV Negativo** | FP = 8 | VN = 0 | 0% |
| | percentagem Total | | | 33.3% |
| **Com** | **Observado HPV Positivo** | VP = 3 | FN = 3 | 50% |
| | **Observado HPV Negativo** | FP = 4 | VN = 9 | 69% |
| | percentagem Total | | | **63%** |

A Tabela 4.14, apresenta os resultados da classificação dos modelos sem e com imputação de dados, e os resultados mostram que a imputação dos dados melhora especificidade ou a capacidade do modelo em classificar corretamente os casos de HPV negativo, aumentando a percentagem correta de 0% para 69%. Embora a sensibilidade ou percentagem correta para a previsão de HPV positivo tenha reduzido de 100% para 50%, a melhoria na classificação dos casos negativos contribui para uma percentagem total de classificação correta mais alta de 63% com imputação em comparação a 33.3% sem imputação.

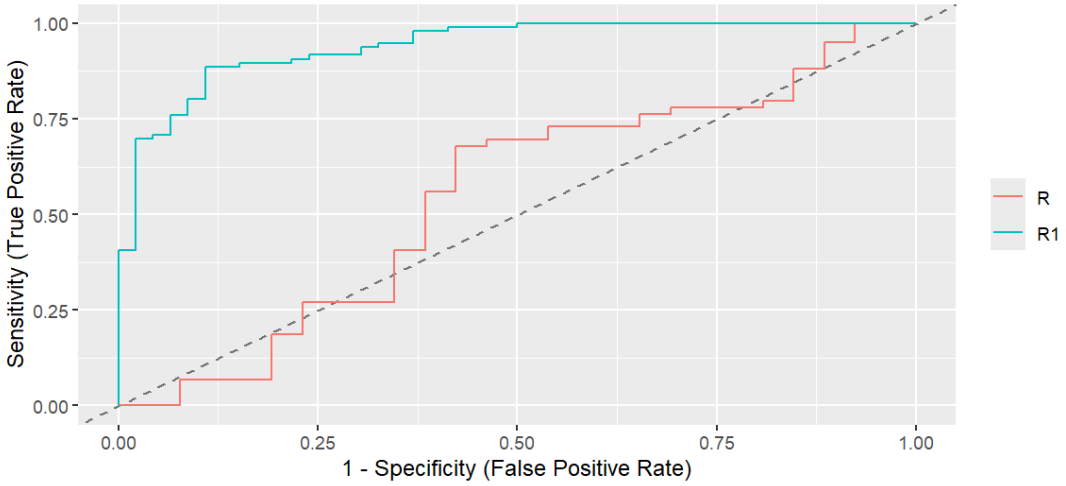


Figura 4.5: Gráfico das curvas ROC para os dados sem e com imputação

De acordo com a Figura 4.5, a comparação das áreas abaixo da curva ROC entre os modelos sem e com imputação revela que ambos os modelos são altamente eficazes na discriminação entre os casos positivos de HPV embora não se verifique o mesmo para os casos negativos de HPV. No entanto, o modelo com imputação é subitamente superior, com uma área abaixo da curva ROC de 94% em comparação com 54.43% do modelo sem imputação. Portanto, a inclusão de imputação nos dados não apenas melhora a precisão geral do modelo, mas também sua capacidade discriminativa, tornando-o mais robusto na previsão dos resultados observados para a variável resposta HPV.

**Modelo de Regressão Logística Binário Para variáveis com imputação**

Pelo método de stepwise, as variáveis idade da primeira relação sexual, ciclo menstrual, HIV, Resultado do ultimo rastreio, altura e estado civil têm um impacto significativo no resultado positivo para o HPV, com diferentes faixas e condições apresentando maior ou menor probabilidade de um resultado positivo. Esta relação é descrita pelo modelo abaixo.

$$\begin{aligned}\ln\left(\frac{\pi_j}{1-\pi_j}\right) = 3.32 &- 6.45 \cdot \text{Idade da primeira relação}[21,24[ + 10.42 \cdot \text{ciclo menstrual NA} \\ &+ 3.94 \cdot \text{HIV Positivo} - 1.69 \cdot \text{Resultado do ultimo rastreio Negativo} \\ &+ 7.8 \cdot \text{Altura}[156,159[ + 8.0 \cdot \text{Altura}[159,162[ \\ &+ 7.9 \cdot \text{Altura}[162,165[ + 5.09 \cdot \text{Altura}[165,168[ \\ &+ 7.90 \cdot \text{Altura}[168,171[ - 3.92 \cdot \text{Estado civil solteiro}\end{aligned}$$

Os resultados da Tabela B.2 em anexo, indicam que para a maioria das variáveis independentes, o aumento de uma unidade, fazem com que as chances de ocorrência do evento de interesse (HPV positivo) aumente também, pois os seus valores das razões de chances são maiores que 1, mantendo todas as demais variáveis explicativas constantes.

Tabela 4.15: Comparação dos modelos logísticos dos dados sem imputação e com imputação

| | | | | | | IC 95% para RC (Sem Imputação) | | | | | | | | IC 95% para RC (Com Imputação) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Sem Imputação | | | | | | | Com Imputação | | | | | | | |
| **Variável** | $\beta$ | **Erro Padrão** | **Wald** | **P-valor** | **RC%** | **Inferior** | **Superior** | $\beta$ | **Erro Padrão** | **Wald** | **P-valor** | **RC%** | **Inferior** | **Superior** |
| (Intercepto) | 1.1221 | 0.3077 | 3.647 | 0.000266 *** | 3.07142 | 1.68042 | 5.61386 | 5.61386 3.322e+00 | 1.075e+00 | 3.092 | 0.0382 | 2.621579e+01 | 2.355603e-01 | 2.917587e+03 |
| **Resultado_do_ultimo_rastreioNA** | — | — | — | — | — | — | — | — | — | — | — | — | — | — |
| Resultado_do_ultimo_rastreioNão_sabe | -16.6882 | 1455.3976 | -0.011 | 0.990851 | 5.654602e-08 | 0.0000000 | $\infty$ | -3.0965 | 4.2266 | -0.733 | 0.46378 | $4.520528e-02$ | $1.141628e-05$ | $1.790002e+02$ |
| Resultado_do_ultimo_rastreioNegativo | -0.7474 | 0.4981 | -1.501 | 0.043451* | 0.4735 | 0.1784071 | 1.257076 | -1.6921 | 0.8705 | -1.944 | 0.03194 | $1.841414e-01$ | $3.343051e-02$ | $1.014285e+00$ |
| Resultado_do_ultimo_rastreioPositivo | — | — | — | — | — | — | — | 40.4225 | 9804.0674 | 0.004 | 0.99671 | $3.591306e+17$ | 0.000000e+00 | $\infty$ |
| **Idade[17,25) (Ref)** | — | — | — | — | — | — | — | — | — | — | — | — | — | — |
| Idade[21,25) | — | — | — | — | — | — | — | -25.5728 | 10754.0133 | -0.002 | 0.99810 | $7.832402e-12$ | 0.000000e+00 | $\infty$ |
| Idade[25,29) | — | — | — | — | — | — | — | -20.5117 | 10754.0128 | -0.002 | 0.99848 | $1.235609e-09$ | 0.000000e+00 | $\infty$ |
| Idade[29,33) | — | — | — | — | — | — | — | -27.2875 | 10754.0129 | -0.003 | 0.99798 | $1.409913e-12$ | 0.000000e+00 | $\infty$ |
| Idade[33,37) | — | — | — | — | — | — | — | -27.8956 | 10754.0129 | -0.003 | 0.99793 | $7.675461e-13$ | 0.000000e+00 | $\infty$ |
| Idade[37,41) | — | — | — | — | — | — | — | -25.1582 | 10754.0128 | -0.002 | 0.99813 | $1.185533e-11$ | 0.000000e+00 | $\infty$ |
| Idade[41,45) | — | — | — | — | — | — | — | -21.6326 | 10754.0128 | -0.002 | 0.99839 | $4.027758e-10$ | 0.000000e+00 | $\infty$ |
| Idade[45,49) | — | — | — | — | — | — | — | -22.4729 | 10754.0128 | -0.002 | 0.99833 | $1.738327e-10$ | 0.000000e+00 | $\infty$ |
| Idade[49,53) | — | — | — | — | — | — | — | -27.4317 | 10754.0129 | -0.003 | 0.99796 | $1.220587e-12$ | 0.000000e+00 | $\infty$ |
| Idade[53,57) | — | — | — | — | — | — | — | -36.3617 | 10754.0138 | -0.003 | 0.99730 | $1.615603e-16$ | 0.000000e+00 | $\infty$ |
| Idade[57,61) | — | — | — | — | — | — | — | 12.4233 | 16713.9898 | 0.001 | 0.99941 | $2.485253e+05$ | 0.000000e+00 | $\infty$ |
| Idade[61,65) | — | — | — | — | — | — | — | -2.7522 | 11773.5335 | 0.000 | 0.99981 | $6.378982e-02$ | 0.000000e+00 | $\infty$ |
| Idade[65,69) | — | — | — | — | — | — | — | -9.4286 | 13130.5941 | -0.001 | 0.99943 | $8.039060e-05$ | 0.000000e+00 | $\infty$ |
| **Numero_de_gestações[0,1) (Ref)** | — | — | — | — | — | — | — | — | — | — | — | — | — | — |
| Numero_de_gestações[1,2) | — | — | — | — | — | — | — | 2.9310 | 3.2354 | 0.906 | 0.36498 | $1.874627e+01$ | $3.303217e-02$ | $1.063880e+04$ |
| Numero_de_gestações[2,3) | — | — | — | — | — | — | — | 0.6698 | 3.2623 | 0.205 | 0.83733 | $1.953847e+00$ | $3.265874e-03$ | $1.168911e+03$ |
| Numero_de_gestações[3,4) | — | — | — | — | — | — | — | 1.3069 | 3.2322 | 0.404 | 0.68597 | $3.694605e+00$ | $6.550567e-03$ | $2.083805e+03$ |
| Numero_de_gestações[4,5) | — | — | — | — | — | — | — | 2.9335 | 3.3550 | 0.874 | 0.38191 | $1.879397e+01$ | $2.619690e-02$ | $1.348302e+04$ |
| Numero_de_gestações[5,6) | — | — | — | — | — | — | — | 0.8232 | 3.2994 | 0.249 | 0.80298 | $2.277716e+00$ | $3.540120e-03$ | $1.465484e+03$ |
| Numero_de_gestações[6,7) | — | — | — | — | — | — | — | 1.2438 | 3.5746 | 0.348 | 0.72787 | $3.468883e+00$ | $3.143791e-03$ | $3.827591e+03$ |
| Numero_de_gestações[7,8) | — | — | — | — | — | — | — | -3.0816 | 3.5666 | -0.864 | 0.38757 | $4.588614e-02$ | $4.224976e-05$ | $4.983551e+01$ |
| Numero_de_gestações[8,9) | — | — | — | — | — | — | — | 20.1789 | 5089.5644 | 0.004 | 0.99684 | $5.802253e+08$ | 0.000000e+00 | $\infty$ |
| Numero_de_gestações[10,11) | — | — | — | — | — | — | — | 16.6435 | 10754.0136 | 0.002 | 0.99877 | $1.691084e+07$ | 0.000000e+00 | $\infty$ |
| **Idade_da_1oRelação_sexual[12,15) (Ref)** | — | — | — | — | — | — | — | — | — | — | — | — | — | — |
| Idade_da_1oRelação_sexual[15,18) | — | — | — | — | — | — | — | -3.5777 | 1.8914 | -1.892 | 0.05855 | $2.794083e-02$ | $6.859137e-04$ | $1.138176e+00$ |
| Idade_da_1oRelação_sexual[18,21) | — | — | — | — | — | — | — | -3.1195 | 1.8623 | -1.675 | 0.09391 | $4.417991e-02$ | $1.148347e-03$ | $1.699717e+00$ |
| Idade_da_1oRelação_sexual[21,24) | — | — | — | — | — | — | — | -6.4519 | 2.5449 | -2.535 | 0.01124 | $1.577519e-03$ | $1.075808e-05$ | $2.313207e-01$ |
| Idade_da_1oRelação_sexual[42,45) | — | — | — | — | — | — | — | -26.2216 | 10754.0138 | -0.002 | 0.99805 | $4.093678e-12$ | 0.000000e+00 | $\infty$ |
| **Ciclo_menstrualIrregular (Ref)** | — | — | — | — | — | — | — | — | — | — | — | — | — | — |
| Ciclo_menstrualNA | — | — | — | — | — | — | — | 10.4198 | 3.3571 | 3.104 | 0.00191 | $3.351535e+04$ | $4.652003e+01$ | $2.414612e+07$ |
| Ciclo_menstrualRegular | — | — | — | — | — | — | — | 1.8976 | 1.5663 | 1.212 | 0.22570 | $6.669772e+00$ | $3.096546e-01$ | $1.436628e+02$ |
| **HIVNegativo (Ref)** | — | — | — | — | — | — | — | — | — | — | — | — | — | — |
| HIVPositivo | — | — | — | — | — | — | — | 3.9436 | 1.6678 | 2.365 | 0.01805 | $5.160189e+01$ | $1.963553e+00$ | $1.356090e+03$ |
| **Nível_de_escolaridadeNenhuma (Ref)** | — | — | — | — | — | — | — | — | — | — | — | — | — | — |
| Nível_de_escolaridadePrimário | — | — | — | — | — | — | — | 27.8091 | 10754.0132 | 0.003 | 0.99794 | $1.194899e+12$ | 0.000000e+00 | $\infty$ |
| Nível_de_escolaridadeSecundário | — | — | — | — | — | — | — | 29.4839 | 10754.0133 | 0.003 | 0.99781 | $6.377868e+12$ | 0.000000e+00 | $\infty$ |
| Nível_de_escolaridadeSuperior | — | — | — | — | — | — | — | 28.6919 | 10754.0133 | 0.003 | 0.99787 | $2.888840e+12$ | 0.000000e+00 | $\infty$ |
| **Altura[144,147) (Ref)** | — | — | — | — | — | — | — | — | — | — | — | — | — | — |
| Altura[147,150) | — | — | — | — | — | — | — | 24.5054 | 7429.2649 | 0.003 | 0.99737 | $4.390865e+10$ | 0.000000e+00 | $\infty$ |
| Altura[150,153) | — | — | — | — | — | — | — | 30.1085 | 2969.8410 | 0.010 | 0.99191 | $1.191077e+13$ | 0.000000e+00 | $\infty$ |
| Altura[153,156) | — | — | — | — | — | — | — | 3.6053 | 2.5356 | 1.422 | 0.15507 | $3.679306e+01$ | $2.555311e-01$ | $5.297708e+03$ |
| Altura[156,159) | — | — | — | — | — | — | — | 7.8690 | 3.0212 | 2.605 | 0.00920 | $2.614978e+03$ | $7.012156e+00$ | $9.751792e+05$ |
| Altura[159,162) | — | — | — | — | — | — | — | 8.0808 | 2.8398 | 2.846 | 0.00443 | $3.231757e+03$ | $1.236625e+01$ | $8.445775e+05$ |
| Altura[162,165) | — | — | — | — | — | — | — | 7.9428 | 2.7549 | 2.883 | 0.00394 | $2.815291e+03$ | $1.272141e+01$ | $6.230337e+05$ |
| Altura[165,168) | — | — | — | — | — | — | — | 5.0970 | 2.5312 | 2.014 | 0.04405 | $1.635297e+02$ | $1.145497e+00$ | $2.334530e+04$ |
| Altura[168,171) | — | — | — | — | — | — | — | 7.9022 | 2.6977 | 2.929 | 0.00340 | $2.703157e+03$ | $1.366528e+01$ | $5.347169e+05$ |
| Altura[171,174) | — | — | — | — | — | — | — | 3.8878 | 2.7971 | 1.390 | 0.16454 | $4.880467e+01$ | $2.030318e-01$ | $1.173164e+04$ |
| Altura[174,177) | — | — | — | — | — | — | — | 8.8273 | 6030.7317 | 0.001 | 0.99883 | $6.817950e+03$ | 0.000000e+00 | $\infty$ |
| **Estado_civilCasado (Ref)** | — | — | — | — | — | — | — | — | — | — | — | — | — | — |
| Estado_civilSeparado/Divorciado | — | — | — | — | — | — | — | -30.6964 | 6932.5232 | -0.004 | 0.99647 | $4.663431e-14$ | 0.000000e+00 | $\infty$ |
| Estado_civilSolteiro | — | — | — | — | — | — | — | -3.9213 | 1.4913 | -2.629 | 0.00855 | $1.981455e-02$ | $1.065480e-03$ | $3.684875e-01$ |
| Estado_civilViuvá | — | — | — | — | — | — | — | 23.2787 | 4792.4217 | 0.005 | 0.99612 | $1.287709e+10$ | 0.000000e+00 | $\infty$ |

Na Tabela 4.15, são apresentados dois modelos de dados sem imputação e com imputação respectivamente, para aferir o impacto da imputação de dados na resposta dos modelos estimados em relação a variável de interesse HPV.

Mulheres que ja realizaram o ultimo rastreio foi significativo em ambos os modelos. Em ambos os casos, mulheres com o resultado negativo para o ultimo rastreio apresentam uma menor probabilidade de resultado positivo para HPV, indicando consistência na importância desta variável.

As demais variáveis foram somente significativas no modelo com a imputação de dados, e a sua significância é descrita da seguinte forma:

A Idade da primeira relação sexual [21,24[ não apareceu como significativa na análise sem imputação de dados, mas mostrou significância quando os dados foram imputados, onde mulheres nesta que iniciaram a sua actividade sexual nesta faixa etária apresentavam 6.45 vezes a menos a probabilidade de testarem positivo ao HPV, a um nível de 5% de significância. Isso

sugere que a idade da primeira relação sexual pode ser um factor relevante que foi melhor capturado com a imputação de dados.

O Ciclo menstrual NA foi significativo apenas na análise com imputação, onde todas as mulheres que ja haviam atingido a menopausa apresentavam 10.42 vezes maior possibilidade de ter o resultado positivo para o HPV.

As mulheres que vivem com o HIV, apresentam uma possibilidade 3.94 vezes maior de terem resultado positivo para o HPV quando comparadas as mulheres sem HIV.

Mulheres entre 156 a 171cm mostraram significância apenas na análise com imputação de dados, onde estas apresentavam uma maior possibilidade de terem o resultado positivo para o HPV quando comparadas com categoria de referência.

Ainda para os dados com imputação, as mulheres solteiras, apresentaram menos chances de testarem positivo para o HPV quando comparadas com as mulheres casadas, sugerindo mais uma vez a necessidade da imputação dos valores perdidos.

## 4.3 Árvore de Classificação

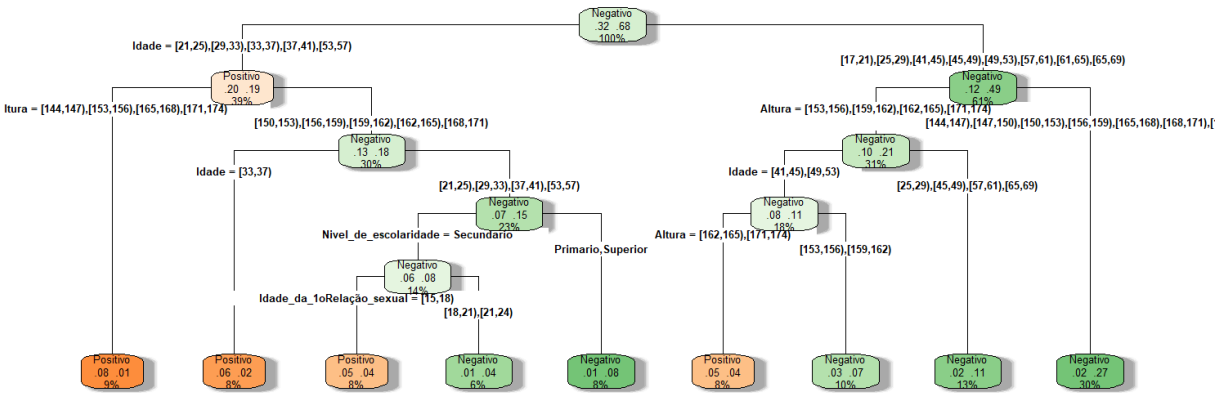


Figura 4.6: Gráfico da árvore de classificação para os dados com imputação

A árvore de classificação na Figura 4.3, começa com a variável idade. A idade é dividida em vários intervalos de classes ou faixas etárias, indicando que esta é uma variável importante para a classificação.

- A árvore divide os dados inicialmente pela variável idade em intervalos específicos.

- Dependendo do intervalo da idade, a árvore faz divisões adicionais com base na altura, idade, Nivel de escolaridade e Idade da primeira Relação sexual.
- Mulheres com idades entre os 21 a 25, cuja a altura varia entre 144 a 147cm levam a um nó final classificando como ”Positivo”com 9% de probabilidade.
- Mulheres com 19% de terem resultado positivo com idades compreendidas entre os 21 a 25, 33 a 37 etc. com 156 a 159cm e com Nivel de escolaridade Secundário resultam em um nó final com classificação Negativo com 6% de probabilidade.
- Mulheres potencialmente negativas ao HPV com faixas etárias entre os 17 aos 21 anos, 25 a 29 etc. e com alturas entre os 153 aos 174 cm tem probabilidades nos nós finais de serem 8% e 13% positivas e negativas respectivamente para o HPV.
- Mulheres potencialmente negativas ao HPV com faixas etárias entre os 17 aos 21 anos, 25 a 29 etc. e com alturas entre os 144 aos 153cm tem probabilidades nos nós finais de serem 30% negativas para o HPV.

## 4.4 Discussão dos Resultados

A aplicação da regressão logística binária múltipla para a análise dos factores de risco a infecção pelo papilomavírus humano no rastreio do cancro do colo do útero nas mulheres atendidas nos centros DREAM da comunidade Sant'Egídio atendendo a probabilidade destas terem resultados positivos ou negativos possibilitou a verificação de todos os factores e tendo se constatado que apenas alguns destes têm um contributo estatisticamente significativo para a variável de interesse.

Para as variáveis envolvidas na estimação dos parâmetros do modelos sem e com imputação de dados, as variáveis significativas foram: a idade da primeira relação sexual, ciclo menstrual, HIV, Resultado do ultimo rastreio, altura e estado civil.

A idade da primeira relação sexual como factor de risco é defenida por autores como Andersson (2009), Barasuol M. E. C, Schimidt D. B (2014), Eduardo K. G. T, *et al.* (2012) e Mendonça V. G, *et al.* (2010) como sendo um factor sexual e reprodutivo capaz de influenciar significativamente de forma crescente ou decrescente na presença do HPV em homens e mulheres. Para estes autores mulheres com menor idades de inicio da actividade sexual tendem a desenvolver com maior facilidade a presença do HPV comparado com mulheres que tem o inicio mais tardio da actividade sexual, o que de certa forma é verificado nos resultados do presente estudo.

No que tange ao ciclo menstrual não existente por terem atingido a menopausa, mulheres nestas condições têm maior possibilidade de terem o resultado positivo para o HPV associado a altas faixas etárias condição esta necessária para o surgimento da menopausa. Este factor como verificado no estudo, terá sido referido por autores como (Andersson, 2009), (Da silva *et al.* 2023).

Na óptica de Barasuol M. E. C e Schimidt D. B (2014) e Eduardo K. G. T., *et al.* (2012), o HIV é um factor de risco altamente associado a presença do HPV onde indivíduos positivos para o HPV tem uma alta possibilidade de desenvolver a presença do HPV no organismo como verificado nos resultados em estudo.

De acordo com Da Silva *et al.* (2023), o estado civil é um factor influente para a presença do HPV, onde indivíduos com múltiplos parceiros devem ser considerados cruciais para a implementação de medidas preventivas a prevenção da infecção pelo HPV.

Factores como o resultado do ultimo rastreio realizado e a altura dos indivíduos, mostraram serem factores significativamente a terem uma especial atenção para a identificação

dos resultados positivos ou negativos nos indivíduos, pois estes tem uma contribuição relativamente alta para a classificação destes.

De acordo com Andersson (2009), Zur Hausen (1989), atendendo aos factores de risco como o tipo de HPV de alto risco os HPV's 16, 18, e 45, indivíduos com estes tipos de HPV têm alta probabilidade de desenvolver neoplasia de alto grau e consequentemente o cancro do colo do útero, razão esta verificada no estudo em que das mulheres positivas para o HPV representando 32.3% da população pelo menos 4.35% das mulheres tiveram este tipo de HPV nos resultados.

# Capítulo 5

# CONCLUSÕES E RECOMENDAÇÕES

## 5.1 Conclusões

Os factores de risco associados a infecção pelo papilomavírus humano (HPV) embora possam apresentar algumas diferenças de um grupo social para o outro, estes tendem a apresentar a mesma raiz do problema relacionado a factores sócio económicos, sexuais e reprodutivos, imonológicos entre outros.

Com base nos objectivos traçados, verificou-se que :

- Ao descrever-se o perfil das pacientes, constatou-se com maior relevância que a maioria das mulheres atendidas tinham ciclo menstrual regular (81.4%) e eram HIV positivas (91.9%).
- Grande parte das mulheres não utilizavam métodos de planeamento familiar (77.6%).
- Entre as variáveis estudadas, a idade da primeira relação sexual, o ciclo menstrual, o HIV positivo, resultado do último rastreio, altura e estado civil foram identificados como factores de risco para infecção por HPV onde mulheres com início precoce da vida sexual e aquelas com HIV positivo tende a apresentar maior probabilidade de testar positivo para HPV.
- A imputação de valores perdidos terá demonstrado ser relevante pois algumas das variáveis que numa primeira fase não constituíam um factor de risco, tornaram-se relevantes e foram de acordo com o sugerido pela literatura mostrando uma grande importância no combate ao cancro do colo do útero em particular ao HPV.

## 5.2 Recomendações

Com os resultados obtidos e com o intuito de reduzir os casos do cancro do colo do útero associados a existência do HPV, propõem-se as seguintes actividades preventivas e de combate:

- Implementar programas educacionais para informar sobre os riscos de infecção por HPV e a importância de práticas sexuais seguras;
- Melhorar o acesso aos serviços de saúde através da melhoria do acesso ao rastreamento regular do câncro do colo do útero e tratamentos adequados para mulheres em risco;
- Fornecer suporte específico e contínuo para mulheres HIV positivas, incluindo monitoramento regular e aconselhamento preventivo;
- Efectuar-se pesquisas de monitoramento regular e continuo para identificar outros possíveis factores de risco e monitorar a eficácia das intervenções implementadas;
- Aos orgãos responsáveis, em especial ao MISAU devem desenvolver e implementar políticas de saúde pública que abordem os factores de risco identificados e promovam a saúde reprodutiva das mulheres na comunidade com segurança, eficácia e eficiência.

# Referências

[1] Afonso, A., & Nunes, C. (2019). *Probabilidades e Estatística. Aplicações e Soluções em SPSS*. Versão revista e aumentada.

[2] Agresti, A. (2018). *Statistical Methods for the Social Sciences (5th ed.).* Pearson.

[3] Almeida, C. M. C., Souza, A. N., Bezerra, R. S., Lima, F. L. O., & Santa Izabel, T. D. S. (2021). *Principais fatores de risco associados ao desenvolvimento do câncer de colo do útero, com ênfase para o Papilomavírus humano (HPV): um estudo de revisão*. Research, Society and Development.

[4] Andersson, J. (2009). *The discoveries of human papilloma viruses that cause cervical cancer and of human immunodeficiency virus*. The Nobel Assembly at Karolinska Institutet.

[5] Barasuol M. E. C., Schimidt D. B. (2014). *Neoplasia do colo do útero e seus fatores de risco: revisão integrativa*. Revista Saúde e desenvolvimento.

[6] Bluman, A. G. (2017). *Elementary Statistics: A Step by Step Approach (10th ed.)*. McGraw-Hill Education.

[7] Bosch FX *et al*. (2002). *The causal relation between human papillomavirus and cervical cancer*.

[8] Bradley, A. P. (1997). *The use of the area under the ROC curve in the evaluation of machine learning algorithms.* Pattern Recognition, 30(7), 1145-1159.

[9] D'Souza, G., & Dempsey, A. (2011). The role of HPV in head and neck cancer and review of the HPV vaccine. Preventive Medicine, 53(S1), S5-S11.

[10] da Silva, M. L. L. G., de Morais, A. M. B., & de Sousa, M. N. A. (2023). *Papilomavírus humano e fatores de risco no câncer de colo uterino*. Revista Eletrônica Acervo Saúde.

[11] De Oliveira, A. D. T., de Castro, C. E. R., Trindade Filho, J. O., de Souza Amaro, K. D., Trajano, V. N., & Costa, H. F. (2019). *Análise histopatológica do adenocarcinoma invasivo de colo uterino*. Revista de Ciências da Saúde Nova Esperança.

[12] Dunne, E. F., Unger, E. R., Sternberg, M., McQuillan, G., Swan, D. C., Patel, S. S., ... & Markowitz, L. E. (2007). Prevalence of HPV infection among females in the United States. JAMA, 297(8), 813-819.

[13] Eduardo K. T. G. *et al*. (2012). *Conhecimento e mudanças de comportamento de mulheres junto a fatores de risco para câncer de colo uterino*.

[14] Estrela, F. M., da Cruz, M. A., Gomes, N. P., Da Silva Oliveira, M. A. , dos Santos Santos, R., Magalhães, J. R. F., *et al*.(2020) *COVID-19 and chronic diseases: Impacts and developments before the pandemic*. Rev Baiana Enferm.

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

[16] Figueiredo, A. E. B., Ceccon, R. F., Figueiredo, J. H. C.(2021). *Doenças crônicas não transmissíveis e suas implicações na vida de idosos dependentes*. Ciên Saúde Coletiva. Disponivel em :https://doi.org/10.1590/1413- 81232020261.33882020

[17] Geisser, S. (1977). *Discrimination, allocatory and separatory, linear aspects. In Classification and clustering*.

[18] Gorsuch, R. L. (2014). *Factor analysis: Classic edition*. Routledge.

[19] Goulart, F. A. A. (2011). *Doenças crônicas não transmissíveis: estrategias de controle e desafios e para os sistemas de saúde*.

[20] Gujarati, D. N. (2000). *Econometria basica (tradução)*. 3ª edição, Campus. São Paulo.

[21] Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2009). *Análise multivariada de dados*. Bookman editora.

[22] Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2009). *Multivariate Data Analysis 7th Edition Pearson Prentice Hall*.

[23] Hawkins, D. M. (1980). *Identification of Outliers*. Chapman and Hall.

[24] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons

[25] Howell, D. C. (2007). *The treatment of missing data*. The Sage handbook of social science methodology, 208-224.

[26] Jamison, D. T. (Ed.). (2006). *Disease and mortality in sub-Saharan Africa*.

[27] Janikow, C. Z. (1998). *Fuzzy decision trees: issues and methods. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*.

[28] Johnson, R. A., & Wichern, D. W. (2002). *Applied multivariate statistical analysis*.

[29] Kamarudin, A. N., Cox, T., & Kolamunnage-Dona, R. (2017). *Time-dependent ROC curve analysis in medical research: current methods and applications*. BMC medical research methodology.

[30] Lalkhen, A. G., & McCluskey, A. (2008). *Clinical tests: sensitivity and specificity. Continuing education in anaesthesia, critical care & pain*.

[31] Langlotz, C. P. (2003). *Fundamental measures of diagnostic examination performance: usefulness for clinical decision making and research. Radiology*.

[32] Lemon, S. C., Roy, J., Clark, M. A., Friedmann, P. D., & Rakowski, W. (2003). *Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. Annals of behavioral medicine*.

[33] Loh, W. Y. (2011). *Classification and regression trees. Wiley interdisciplinary reviews: data mining and knowledge discovery*.

[34] Matos, C. S. (2008). *Plano estratégico nacional de prevenção e controlo das doenças não transmissíveis para o período 2008-2014*. Maputo: Ministério da Saúde.

[35] Markoni, M. d., Lakatos, E. M. (1992). *Metodologia do trabalho científico*. São Paulo: Atlas.

[36] Maucort-Boulch, D., Plummer, M., Castle, P. E., Demuth, F., Safaeian, M., & Wheeler, C. M. (2010). *Predictors of human papillomavirus persistence among women undergoing concurrent testing for HPV16/18 infections and cervical cytology*. International Journal of Cancer, 126(1), 111-116.

[37] Menard, S. (2002). *Applied logistic regression analysis*. Sage.

[38] Mendonça VG, *et al*. (2010). *Infecção cervical por papilomavírus humano: genotipagem viral e fatores de risco para lesão intraepitelial de alto grau e câncer de colo do útero*. Revista Brasileira de Ginecologia e Obstetrícia.

[39] Ministério da Saúde. (2013). *Plano Estratégico do Sector da Saúde, PESS 2014–2019*.

[40] Ministerio da saúde. (2020). *Plano Estratégico Multissetorial de Prevenção e Controlo das Doenças Não Transmissíveis 2020 – 2029*.

[41] Mondiale de la Santé, O., & World Health Organization. (2022). *Human papillomavirus vaccines: WHO position paper (2022 update)–Vaccins contre les papillomavirus humains: note de synthèse de l'OMS (mise à jour de 2022). Weekly Epidemiological Record= Relevé épidémiologique hebdomadaire*.

[42] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to Linear Regression Analysis (6th ed.). Wiley*.

[43] Morgan, J. (2014). *Classification and regression tree analysis*. Boston: Boston University.

[44] Murray, G. D. (1977). *A cautionary note on selection of variables in discriminant analysis. Journal of the Royal Statistical Society Series C: Applied Statistics*.

[45] Nicolau, S. M. (2003). *Existe câncer do colo uterino sem HPV?* Revista da Associação Médica Brasileira.

[46] Organização Mundial da Saude. [OMS]. (2011). *Doenças Crônicas. Acesso em 26/04/2024*. Disponível em http://www.who.int/topics/chronic_diseases/en/

[47] Pinto, A. P. *et al*. (2002). *Co- fatores do HPV na oncogênese cervical*. Revista da Associação Médica Brasileira.

[48] Reis, E. (2001). *Estatística multivariada aplicada*. Edições Sílabo.

[49] Schwarz, E., Freese, U. K., Gissmann, L., Mayer, W., Roggenbuck, B., Stremlau, A., & Hausen, H. Z. (1985). *Structure and transcription of human papillomavirus sequences in cervical carcinoma cells*. Nature.

[50] Victorino, M. T. A. (2012). *Currículo de Formação de Líderes Comunitários facilitadores para a promoção do apoio às COV´s e da Continuidade de Cuidados nas suas comunidades*.

[51] Weinstein, R. (2005). *RFID: a technical overview and its application to the enterprise*. IT professional.

[52] WHO. World Health Organization (2024a). *Câncer Cervical*. Geneva, Disponível em:
https://www.who.int/news-room/fact-sheets/detail/cervical-cancer?gad_source=1&gclid=CjwKCAjwtqmwBhBVEiwAL-WAYSD2zHadMGFmqn1hvjr-U_4JobcSf36uoCEeddLm20M8fTg-NL3hbBoC93AQAvD_BwE;

[53] WHO. World Health Organization (2024b). *Papilomavírus humano e câncer*. Geneva, Disponível em:
https://www.who.int/news-room/fact-sheets/detail/human-papilloma-virus-and-cancer

[54] WHO. World Health Organization. (2020). *Recomendações para garantir a qualidade, segurança e eficácia das vacinas recombinantes de partículas semelhantes ao vírus do papilomavírus humano*. Série de Relatórios Técnicos da OMS, disponivel em: https://www.who.int/publications/m/item/recombinant-hpv-like-particle-vaccines-annex-4-trs-no-999, acesso 28 de maio e 2024.

[55] Zur Hausen, H. (2002). *Papillomaviruses and cancer: from basic studies to clinical application*. Nature reviews cancer.

[56] Zur Hausen H. (1989). *Papillomaviruses in anogenital cancer as a model to understand the role of viruses in human cancers*. Cancer Res.

# Apêndice A

# Apêndice 1

Tabela A.1: Distribuição das frequências absolutas, relativas e o teste de independência do Qui-quadrado

| Variável | Categorias das Variáveis | HPV | | | Qui Quadrado | | |
|---|---|---|---|---|---|---|---|
| | | Negativo | Positivo | Total | $\chi^2$ | gl | p-value |
| Ciclo Menstrual | (irregular) | 3 (1.86%) | 3 (1.86%) | 6 (3.73%) | | | |
| | (NA) | 19 (11.8%) | 5 (3.11%) | 24 (14.91%) | 2.40 | 2 | 0.30 |
| | (Regular) | 87 (54.04%) | 44 (27.33%) | 131 (81.37%) | | | |
| | **Total** | 109 (67.7%) | 52 (32.3%) | 161 (100%) | | | |
| Sangramento Vaginal | (Anormal) | 0 (0%) | 1 (0.62%) | 1 (0.62%) | | | |
| | (NA) | 22 (13.66%) | 4 (2.48%) | 26 (16.15%) | 5.97 | 2 | 0.05 |
| | (Normal) | 87 (54.04%) | 47 (29.19%) | 134 (83.23%) | | | |
| | **Total** | 109 (67.7%) | 52 (32.3%) | 161 (100%) | | | |
| Planejamento Familiar | (DIU) | 2 (1.24%) | 1 (0.62%) | 3 (1.86%) | | | |
| | (implante) | 5 (3.11%) | 5 (3.11%) | 10 (6.21%) | | | |
| | (injetável) | 5 (3.11%) | 5 (3.11%) | 10 (6.21%) | | | |
| | (nenhum) | 88 (54.66%) | 37 (22.98%) | 125 (77.64%) | 4.37 | 6 | 0.62 |
| | (outros) | 1 (0.62%) | 1 (0.62%) | 2 (1,24%) | | | |
| | (pílula) | 5 (3.11%) | 1 (0.62%) | 6 (3.73%) | | | |
| | (preservativo) | 3 (1.86%) | 2 (1.24%) | 5 (3.11%) | | | |
| | **Total** | 109 (67.7%) | 52 (32.3%) | 161 (100%) | | | |
| Tipo de Rastreamento Realizado | (pela primeira vez) | 70 (43.48%) | 31 (19.25%) | 101 (62.73%) | | | |
| | (para seguimento) | 39 (24.22%) | 21 (13.04%) | 60 (37.27%) | 0.15 | 1 | 0.69 |
| | **Total** | 109 (67.7%) | 52 (32.3%) | 161 (100%) | | | |
| Menopausa | (Não) | 90 (55.9%) | 45 (27.95%) | 135 (83.85%) | | | |
| | (sim) | 19 (11.8%) | 7 (4.35%) | 26 (16.15%) | 0.16 | 1 | 0.68 |
| | **Total** | 109 (67.7 %) | 52 (32.3%) | 161 (100%) | | | |
| ITS | (não) | 99 (61.49%) | 48 (29.81%) | 147 (91.3%) | | | |
| | (não sabe) | 2 (1.24%) | 0 (0%) | 2 (1.24%) | 0.96 | 2 | 0.62 |
| | (Sim) | 8 (4.97%) | 4 (2.48%) | 12 (7.45%) | | | |
| | **Total** | 109 (67.7%) | 52 (32.3%) | 161 (100%) | | | |
| ITS Sim | (Leucorreia) | 8 (4.97%) | 4 (2.48%) | 12 (7.45%) | | | |
| | (NA) | 101 (62.73%) | 48 (29.81%) | 149 (92.55%) | 0.00 | 1 | 1 |
| | **Total** | 109 (67.7%) | 52 (32.3%) | 161 (100%) | | | |
| HIV | (negativo) | 9 ( 5.59%) | 4 (2.48%) | 13 (8.07%) | | | |
| | (positivo) | 100(62.11%) | 48 (29.81%) | 148 (91.93%) | 0.00 | 1 | 1 |
| | **Total** | 109 (67.7%) | (52.3%) | 161 (100%) | | | |
| História de Rastreamento | (não) | 72 (44.72%) | 32 (19.88%) | 104 (64.6%) | | | |
| | (sim) | 37 (22.98%) | 20 (12.42%) | 57 (35.4%) | 0.14 | 1 | 0.70 |
| | **Total** | 109 (67.7%) | 52 (32.3%) | 161 (100%) | | | |
| História de Rastreamento se Sim | (HPV) | 2 (1.24%) | 0 (0%) | 2 (1.24%) | | | |
| | (NA) | 72 (44.72%) | 32 (19.88%) | 104 (64.6%) | 1.48 | 2 | 0.47 |
| | (VIA) | 35 (21.74%) | 20 (12.42%) | 55 (34.16%) | | | |
| | **Total** | 109 (67.7%) | 52 (32.3%) | 161 (100%) | | | |
| Data do Último Rastreamento IC | (NA) | 75 (46.58%) | 33 (20.5%) | 108 (67.08%) | | | |
| | (menos de 3 meses) | 2 (1.24%) | 1 (0.62%) | 3 (1.86%) | | | |
| | (entre 3 meses a 3 anos) | 28 (17.39%) | 16 (9.94%) | 44 (27.33%) | 0.48 | 3 | 0.92 |
| | (mais de 3 anos) | 4 (2.48%) | 2 (1.24%) | 6 (3.73%) | | | |
| | **Total** | 109 (67.7%) | 52 (32.3%) | 161(100%) | | | |
| Resultado do Último Rastreamento | (NA) | 69 (42.86%) | 30(18.63% ) | 99 (61.49%) | | | |
| | (não sabe) | 1 (0.62%) | 1 (0.62%) | 2 (1.24%) | | | |
| | (negativo) | 37 (22.98%) | 21 (13.04%) | 58 (36.02%) | 1.82 | 3 | 0.60 |
| | (positivo) | 2 (1.24%) | 0 (0%) | 2 (1.24%) | | | |
| | **Total** | 109 (67.7%) | 52 (32.3%) | 161 (100%) | | | |
| Se o Último Rastreamento Positivo | (crioterapia/termoablação) | 1 (0.62%) | 0 (0%) | 1 (0.62%) | | | |
| | (NA) | 108 (67.08%) | 52 (32.3%) | 160 (99.38%) | 0.00 | 1 | 1 |
| | **Total** | 109 (67.7%) | 52 (32.3%) | 161 (100%) | | | |

Tabela A.2: Distribuição das frequências absolutas, relativas e o teste de independência do Qui-quadrado (continuação)

| Variável | Categorias das Variáveis | HPV | | | Qui Quadrado | | |
|---|---|---|---|---|---|---|---|
| | | Negativo | Positivo | Total | $\chi^2$ | gl | p-value |
| Exame Especular para Colo Sangramento | (com sangramento) | 1 (0.62%) | 0 (0%) | 1 (0.62%) | | | |
| | (sem sangramento) | 106 (65.84%) | 52 (32.3%) | 158 (98.14%) | | | |
| | (outros) | 1 (0.62%) | 0 (0%) | 1 (0.62%) | 1.45 | 3 | 0.69 |
| | (NA) | 1 (0.62%) | 0 (0%) | 1 (0.62%) | | | |
| | **Total** | 109 (67.7%) | 52 (32.3%) | 161 (100%) | | | |
| Exame Especular para Colo visualização | JEC parcialmente visualizada | 2 (1.24%) | 0 (0%) | 2 (1.24%) | | | |
| | JEC totalmente visualizada | 106 (65.84%) | 52 (32.3%) | 158 (98.14%) | 1.48 | 2 | 0.48 |
| | NA | 1 (0.62%) | 0 (0%) | 1 (0.62%) | | | |
| | **Total** | 109 (67.7%) | 52 (32.3%) | 161 (100%) | | | |
| Rastreamento Uterino com VIA | NA | 1 (0.62%) | 0 (0%) | 1 (0.62%) | | | |
| | VIA Negativo | 107 (66.46%) | 47 (29.19%) | 154 (95.65%) | | | |
| | via Positivo | 1 (0.62%) | 5 (3.11%) | 6 (3.73%) | 7.84 | 2 | 0.01 |
| | **Total** | 109 (67.7%) | 52 (32.3%) | 161 (100%) | | | |
| HPV Positivo | HPV16 | 0 (0%) | 8 (4.97%) | 8 (4.97%) | | | |
| | HPV16 + Outro | 0 (0%) | 6 (3.73%) | 6 (3.73%) | 156.55 | 4 | 0.00< |
| | HPV18/45 | 0 (0%) | 1 (0.62%) | 1 (0.62%) | | | |
| | NA | 108 (67.08%) | 0 (0%) | 108 (67.08%) | | | |
| | Outro | 1 (0.62%) | 37 (22.98%) | 38 (23.6%) | | | |
| | **Total** | 109 (67.7%) | 52 (32.3%) | 161 (100%) | | | |
| Tratamento Feito | Crioterapia | 0 (0%) | 7 (4.35%) | 7 (4.35%) | | | |
| | NA | 109 (67.7%) | 45 (27.95%) | 154 (95.65%) | 12.27 | 1 | 0.00< |
| | **Total** | 109 (67.7%) | 52 (32.3%) | 161 (100%) | | | |
| Nível de Escolaridade | Nenhuma | 0 (0%) | 1 (0.62%) | 1 (0.62%) | | | |
| | Primário | 44 (27.33%) | 21 (13.04%) | 65 (40.37%) | 2.31 | 3 | 0.50 |
| | Secundário | 54 (33.54%) | 26 (16.15%) | 80 (49.69%) | | | |
| | Superior | 11 (6.83%) | 4 (2.48%) | 15 (9.32%) | | | |
| | **Total** | 109 (67.7%) | 52 (32.3%) | 161 (100%) | | | |
| Estado Civil | Casado | 18 (11.18%) | 3 (1.86%) | 21 (13.04%) | | | |
| | Separado/Divorciado | 2 (1.24%) | 1 (0.62%) | 3 (1.86%) | 4.72 | 3 | 0.19 |
| | Solteiro | 87 (54.04%) | 48 (29.81%) | 135 (83.85%) | | | |
| | Viuvá | 2 (1.24%) | 0 (0%) | 2 (1.24%) | | | |
| | **Total** | 109 (67.7%) | 52 (32.3%) | 161 (100%) | | | |
| Água na Casa | Não | 22 (13.66%) | 6 (3.73%) | 28 (17.39%) | | | |
| | Sim | 87 (54.04%) | 46 (28.57%) | 133 (82.61%) | 1.27 | 1 | 0.25 |
| | **Total** | 109 (67.7%) | 52 (32.3%) | 161 (100%) | | | |
| Sexo | Feminino | 109 (67.7%) | 52 (32.3%) | 161 (100%) | | | |
| | **Total** | 109 (67.7%) | 52 (32.3%) | 161 (100%) | 20.18 | 1 | 0.00< |
| Tabagismo | Não | 109 (67.7%) | 52 (32.3%) | 161 (100%) | | | |
| | **Total** | 109 (67.7%) | 52 (32.3%) | 161 (100%) | 20.18 | 1 | 0.00< |
| Exame Pélvico | Normal | 109 (67.7%) | 52 (32.3%) | 161 (100%) | | | |
| | **Total** | 109 (67.7%) | 52 (32.3%) | 161 (100%) | 20.18 | 1 | 0.00< |
| Unidade Sanitária | CC | 109 (67.7%) | 52 (32.3%) | 161 (100%) | | | |
| | **Total** | 109 (67.7%) | 52 (32.3%) | 161 (100%) | 20.18 | 1 | 0.00< |

# Apêndice B

# Apêndice 2

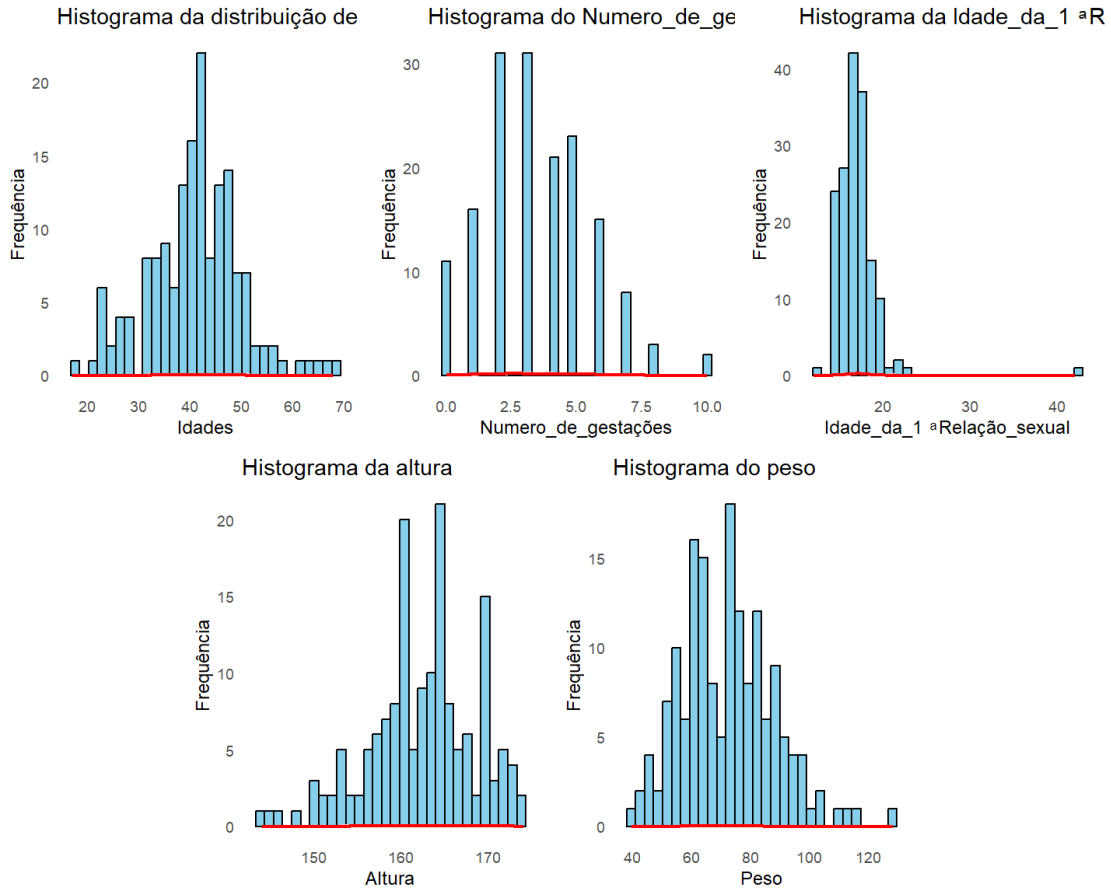


Figura B.1: Histogramas das variáveis numéricas

Tabela B.1: Variáveis na equação do modelo logístico dos dados sem imputação

| | | | | | | IC 95% para RC | |
|---|---|---|---|---|---|---|---|
| | $\beta$ | Erro Padrão | Wald | P-valor | RC | Inferior | Superior |
| (Intercepto) | 1.1221 | 0.3077 | 3.647 | 0.000266 *** | 3.07142 | 1.68042 | 5.61386 |
| **Resultado do ultimo rastreio NA** | — | — | — | — | — | — | — |
| Resultado do ultimo rastreioNão sabe | -16.6882 | 1455.3976 | -0.011 | 0.990851 | 5.654602e-08 | 0.0000000 | ∞ |
| Resultado do ultimo rastreioNegativo | -0.7474 | 0.4981 | -1.501 | 0.043451* | 0.4735 | 0.1784071 | 1.257076 |

Tabela B.2: Variáveis na equação do modelo logístico dos dados com imputação

| | | | | | | IC 95% para RC | |
|---|---|---|---|---|---|---|---|
| | $\beta$ | Erro Padrão | Wald | P-valor | RC% | Inferior | Superior |
| (Intercepto) | 3.322e+00 | 1.075e+00 | 3.092 | 0.0382 | 2.621579e+01 | 2.355603e-01 | 2.917587e+03 |
| **Idade[17,25) (Ref)** | — | — | — | — | — | — | — |
| Idade[21,25) | -25.5728 | 10754.0133 | -0.002 | 0.99810 | $7.832402e-12$ | 0.000000e+00 | $\infty$ |
| Idade[25,29) | -20.5117 | 10754.0128 | -0.002 | 0.99848 | $1.235609e-09$ | 0.000000e+00 | $\infty$ |
| Idade[29,33) | -27.2875 | 10754.0129 | -0.003 | 0.99798 | $1.409913e-12$ | 0.000000e+00 | $\infty$ |
| Idade[33,37) | -27.8956 | 10754.0129 | -0.003 | 0.99793 | $7.675461e-13$ | 0.000000e+00 | $\infty$ |
| Idade[37,41) | -25.1582 | 10754.0128 | -0.002 | 0.99813 | $1.185533e-11$ | 0.000000e+00 | $\infty$ |
| Idade[41,45) | -21.6326 | 10754.0128 | -0.002 | 0.99839 | $4.027758e-10$ | 0.000000e+00 | $\infty$ |
| Idade[45,49) | -22.4729 | 10754.0128 | -0.002 | 0.99833 | $1.738327e-10$ | 0.000000e+00 | $\infty$ |
| Idade[49,53) | -27.4317 | 10754.0129 | -0.003 | 0.99796 | $1.220587e-12$ | 0.000000e+00 | $\infty$ |
| Idade[53,57) | -36.3617 | 10754.0138 | -0.003 | 0.99730 | $1.615603e-16$ | 0.000000e+00 | $\infty$ |
| Idade[57,61) | 12.4233 | 16713.9898 | 0.001 | 0.99941 | $2.485253e+05$ | 0.000000e+00 | $\infty$ |
| Idade[61,65) | -2.7522 | 11773.5335 | 0.000 | 0.99981 | $6.378982e-02$ | 0.000000e+00 | $\infty$ |
| Idade[65,69) | -9.4286 | 13130.5941 | -0.001 | 0.99943 | $8.039060e-05$ | 0.000000e+00 | $\infty$ |
| **Numero_de_gestações[0,1) (Ref)** | — | — | — | — | — | — | — |
| Numero_de_gestações[1,2) | 2.9310 | 3.2354 | 0.906 | 0.36498 | $1.874627e+01$ | $3.303217e-02$ | $1.063880e+04$ |
| Numero_de_gestações[2,3) | 0.6698 | 3.2623 | 0.205 | 0.83733 | $1.953847e+00$ | $3.265874e-03$ | $1.168911e+03$ |
| Numero_de_gestações[3,4) | 1.3069 | 3.2322 | 0.404 | 0.68597 | $3.694605e+00$ | $6.550567e-03$ | $2.083805e+03$ |
| Numero_de_gestações[4,5) | 2.9335 | 3.3550 | 0.874 | 0.38191 | $1.879397e+01$ | $2.619690e-02$ | $1.348302e+04$ |
| Numero_de_gestações[5,6) | 0.8232 | 3.2994 | 0.249 | 0.80298 | $2.277716e+00$ | $3.540120e-03$ | $1.465484e+03$ |
| Numero_de_gestações[6,7) | 1.2438 | 3.5746 | 0.348 | 0.72787 | $3.468883e+00$ | $3.143791e-03$ | $3.827591e+03$ |
| Numero_de_gestações[7,8) | -3.0816 | 3.5666 | -0.864 | 0.38757 | $4.588614e-02$ | $4.224976e-05$ | $4.983551e+01$ |
| Numero_de_gestações[8,9) | 20.1789 | 5089.5644 | 0.004 | 0.99684 | $5.802253e+08$ | 0.000000e+00 | $\infty$ |
| Numero_de_gestações[10,11) | 16.6435 | 10754.0136 | 0.002 | 0.99877 | $1.691084e+07$ | 0.000000e+00 | $\infty$ |
| **Idade_da_1oRelação_sexual[12,15) (Ref)** | — | — | — | — | — | — | — |
| Idade_da_1oRelação_sexual[15,18) | -3.5777 | 1.8914 | -1.892 | 0.05855 | $2.794083e-02$ | $6.859137e-04$ | $1.138176e+00$ |
| Idade_da_1oRelação_sexual[18,21) | -3.1195 | 1.8623 | -1.675 | 0.09391 | $4.417991e-02$ | $1.148347e-03$ | $1.699717e+00$ |
| Idade_da_1oRelação_sexual[21,24) | -6.4519 | 2.5449 | -2.535 | 0.01124 | $1.577519e-03$ | $1.075808e-05$ | $2.313207e-01$ |
| Idade_da_1oRelação_sexual[42,45) | -26.2216 | 10754.0138 | -0.002 | 0.99805 | $4.093678e-12$ | 0.000000e+00 | $\infty$ |
| **Ciclo_menstrualIrregular (Ref)** | — | — | — | — | — | — | — |
| Ciclo_menstrualNA | 10.4198 | 3.3571 | 3.104 | 0.00191 | $3.351535e+04$ | $4.652003e+01$ | $2.414612e+07$ |
| Ciclo_menstrualRegular | 1.8976 | 1.5663 | 1.212 | 0.22570 | $6.669772e+00$ | $3.096546e-01$ | $1.436628e+02$ |
| **HIVNegativo (Ref)** | — | — | — | — | — | — | — |
| HIVPositivo | 3.9436 | 1.6678 | 2.365 | 0.01805 | $5.160189e+01$ | $1.963553e+00$ | $1.356090e+03$ |
| **Resultado_do_ultimo_rastreioNA (Ref)** | — | — | — | — | — | — | — |
| Resultado_do_ultimo_rastreioNão_sabe | -3.0965 | 4.2266 | -0.733 | 0.46378 | $4.520528e-02$ | $1.141628e-05$ | $1.790002e+02$ |
| Resultado_do_ultimo_rastreioNegativo | -1.6921 | 0.8705 | -1.944 | 0.03194 | $1.841414e-01$ | $3.343051e-02$ | $1.014285e+00$ |
| Resultado_do_ultimo_rastreioPositivo | 40.4225 | 9804.0674 | 0.004 | 0.99671 | $3.591306e+17$ | 0.000000e+00 | $\infty$ |
| **Nível_de_escolaridadeNenhuma (Ref)** | — | — | — | — | — | — | — |
| Nível_de_escolaridadePrimário | 27.8091 | 10754.0132 | 0.003 | 0.99794 | $1.194899e+12$ | 0.000000e+00 | $\infty$ |
| Nível_de_escolaridadeSecundário | 29.4839 | 10754.0133 | 0.003 | 0.99781 | $6.377868e+12$ | 0.000000e+00 | $\infty$ |
| Nível_de_escolaridadeSuperior | 28.6919 | 10754.0133 | 0.003 | 0.99787 | $2.888840e+12$ | 0.000000e+00 | $\infty$ |
| **Altura[144,147) (Ref)** | — | — | — | — | — | — | — |
| Altura[147,150) | 24.5054 | 7429.2649 | 0.003 | 0.99737 | $4.390865e+10$ | 0.000000e+00 | $\infty$ |
| Altura[150,153) | 30.1085 | 2969.8410 | 0.010 | 0.99191 | $1.191077e+13$ | 0.000000e+00 | $\infty$ |
| Altura[153,156) | 3.6053 | 2.5356 | 1.422 | 0.15507 | $3.679306e+01$ | $2.555311e-01$ | $5.297708e+03$ |
| Altura[156,159) | 7.8690 | 3.0212 | 2.605 | 0.00920 | $2.614978e+03$ | $7.012156e+00$ | $9.751792e+05$ |
| Altura[159,162) | 8.0808 | 2.8398 | 2.846 | 0.00443 | $3.231757e+03$ | $1.236625e+01$ | $8.445775e+05$ |
| Altura[162,165) | 7.9428 | 2.7549 | 2.883 | 0.00394 | $2.815291e+03$ | $1.272141e+01$ | $6.230337e+05$ |
| Altura[165,168) | 5.0970 | 2.5312 | 2.014 | 0.04405 | $1.635297e+02$ | $1.145497e+00$ | $2.334530e+04$ |
| Altura[168,171) | 7.9022 | 2.6977 | 2.929 | 0.00340 | $2.703157e+03$ | $1.366528e+01$ | $5.347169e+05$ |
| Altura[171,174) | 3.8878 | 2.7971 | 1.390 | 0.16454 | $4.880467e+01$ | $2.030318e-01$ | $1.173164e+04$ |
| Altura[174,177) | 8.8273 | 6030.7317 | 0.001 | 0.99883 | $6.817950e+03$ | 0.000000e+00 | $\infty$ |
| **Estado_cívilCasado (Ref)** | — | — | — | — | — | — | — |
| Estado_cívilSeparado/Divorciado | -30.6964 | 6932.5232 | -0.004 | 0.99647 | $4.663431e-14$ | 0.000000e+00 | $\infty$ |
| Estado_cívilSolteiro | -3.9213 | 1.4913 | -2.629 | 0.00855 | $1.981455e-02$ | $1.065480e-03$ | $3.684875e-01$ |
| Estado_cívilViuvá | 23.2787 | 4792.4217 | 0.005 | 0.99612 | $1.287709e+10$ | 0.000000e+00 | $\infty$ |

# ANEXOS 1

UNIVERSIDADE EDUARDO MONDLANE

FACULDADE DE CIÊNCIAS

*Departamento de Matemática e Informática*

CREDENCIAL

Para efeitos de pedido de dados na Instituição - **Centro Dream "Sant'Egidio"**, no âmbito da sua formação académica, credencia - se o Sr. **Jorge Sidumo Patrício**, estudante do Curso de Licenciatura em Estatística, regime laboral, do Departamento de Matemática e Informática, da Faculdade de Ciências, Universidade Eduardo Mondlane.

Maputo, 20 de Outubro de 2023

O Director do Curso

Doutor Miranda Albino Martins Muaualo

Figura B.2: Credencial