Análise de Dados como ferramenta para Avaliar a Efetividade do Exame Papanicolau na Prevenção do Câncer do Colo do Útero

---

1. DISSERTAÇÃO SOBRE O PROBLEMA

1.1 Descrição detalhada do problema

O câncer do colo do útero é um dos tipos de câncer mais comuns entre mulheres no Brasil, especialmente entre os 25 e 64 anos. Mesmo sendo altamente prevenível, ainda registra milhares de novos casos e óbitos todos os anos.

Uma das principais formas de prevenção depois da vacinação correta contra o HPV(vírus causador da neoplasia estudada), é o **exame citopatológico do colo do útero** (Papanicolau), que identifica lesões precursoras antes de sua evolução para câncer invasivo.

Todavia, ainda com disponibilidade gratuita pelo SUS, **as taxas de cobertura do exame permanecem abaixo do recomendado pela OMS**, e as existentes desigualdades regionais, socioeconômicas e estruturais afetam diretamente a mortalidade. Mas apenas a realização de exames é o suficiente para impedir mortes?

A problemática central é:

> **“De que maneira a cobertura do exame Papanicolau influencia as taxas de mortalidade por câncer do colo do útero no Brasil, e quais estados apresentam maior risco por baixa cobertura?”**

Embora este trabalho busque responder a uma pergunta central, baseada nelas outros questionamentos surgem, ainda mais com resultados das análises e cruzando certas

métricas além de exames e número de óbitos. Cruzar esses dados é essencial para entender de forma mais precisa e aprofundada um problema complexo.

1.2 Importância e relevância social

- O câncer do colo do útero é evitável em uma porcentagem significativa quando a cobertura do exame é adequada.

- A mortalidade impacta diretamente mulheres em idade produtiva, causando efeitos sociais, familiares e econômicos.

- Políticas públicas mais eficazes dependem de diagnóstico situacional baseado em dados reais de cobertura, incidência e mortalidade.

- A análise de dados permite identificar:

    - regiões vulneráveis,

    - tendências temporais,

    - desigualdades de acesso,

    - falhas na oferta de exames,

    - oportunidade de intervenção.

Ou seja, compreender a correlação entre **cobertura do exame x mortalidade** é trivial para planejamento preventivo e redução de óbitos.

1.3 De que maneira a análise de dados ajuda a solucionar o problema?

A análise de dados permite sair do "achismo" e identificar exatamente quais Estados e regiões têm baixa cobertura e alta mortalidade, consequentemente, permitindo políticas públicas baseadas em evidências.A análise de dados permitirá:

- Verificar a correlação entre a cobertura de exames e número de óbitos.
- Identificar estados e municípios com **baixa cobertura populacional**.
- Mensurar o impacto da cobertura sobre a mortalidade (ex.: regressão, elasticidade).
- Analisar tendências temporais (crescimento ou queda de exames x mortes).
- Identificar regiões com maior risco epidemiológico.
- Criar dashboards que auxiliem profissionais e gestores na tomada de decisões.

Com isso, pode-se propor **intervenções direcionadas**, como campanhas regionais de conscientização sobre a importância da realização de exames, reforço da oferta, estratégias itinerantes e apoio à atenção primária.

1.4 Rastreamento e Mortalidade: Uma Relação Inversa, porém complexa

Para avaliar a efetividade do exame, fez-se necessário superar o viés do tamanho populacional, onde estados mais populosos naturalmente apresentam números absolutos maiores de óbitos e exames, a fim de dar a devida proporção. A etapa de engenharia de recursos (*Feature Engineering*) foi fundamental, criando métricas padronizadas: a *taxa_mortalidade* (óbitos por 100 mil mulheres) e a *razao_exames_pop* (um proxy para a cobertura de rastreamento). A transformação dos dados para o formato "long" e a integração em um df_master permitiram a análise temporal e comparativa entre as Unidades da Federação (UFs). Esta etapa é melhor detalhada na 3.2.

A análise estatística demonstrou que o rastreamento é funcional, porém sua relação com a mortalidade não é linear nem isolada. Isso foi verificado, observando que o cálculo do coeficiente de correlação de Pearson indicou uma correlação negativa (-0.3220) entre a cobertura de exames e a taxa de mortalidade, o que já era esperado. Isso confirma, de maneira estatística, a premissa clínica de que maior cobertura de rastreamento está associada a uma menor mortalidade a longo prazo.

Ao verificar mais a fundo essa relação através de uma Regressão Linear Simples, obteve-se um coeficiente angular de -0.2864. Isso significa que, para cada aumento de 1

unidade na cobertura do exame, a taxa média de mortalidade tende a cair aproximadamente 0,29 casos por 100 mil mulheres. Refinando este *insight* por meio da análise de elasticidade, demonstrou-se que um aumento de 1% na cobertura está associado a uma redução de 0,089% na mortalidade.

Todavia, o baixo coeficiente de determinação ($R2 \approx 0.10$) serve como um alerta crítico: a cobertura de exames explica apenas cerca de 10% da variação da mortalidade entre os estados. Isso sugere que, mesmo que o exame seja efetivo, a mortalidade é influenciada majoritariamente (quase 90%) por outros fatores estruturais não capturados apenas pela métrica de cobertura, como acesso ao tratamento oncológico, estadiamento da doença no diagnóstico e desigualdades socioeconômicas. Isso pode ser visto observando e isolando as métricas do estado de São Paulo por exemplo, onde existem diversos hospitais oncológicos e mesmo com uma cobertura de exames não tão alta, mantém uma baixa taxa de mortalidade em relação a outros estados.

A aplicação de testes de hipótese (ANOVA) revelou que existem diferenças estatisticamente significativas na cobertura de exames entre os estados ($p<0.001$). Curiosamente, a diferença entre as macro regiões não foi estatisticamente significativa ($p=0.14$), sugerindo que a desigualdade no acesso ao Papanicolau é um fenômeno intra-regional, variando drasticamente de estado para estado, e não apenas entre Norte/Sul. Embora a maioria dos estados mais ao norte sempre estejam com os piores índices.

Para mapear essas diferenças, utilizou-se o algoritmo de *Machine Learning* K-Means, agrupando os estados em perfis distintos. O "Cluster 1" foi identificado como o grupo de maior vulnerabilidade epidemiológica, caracterizado por baixa cobertura de exames e alta mortalidade. A Análise de Componentes Principais (PCA) fortalece essa visão, criando um "Eixo de vulnerabilidade" onde estados como Amazonas e Amapá se destacaram negativamente no decorrer dos últimos 11 anos.

A análise confirmou a alta eficácia do sistema em detectar a doença. Houve uma correlação muito forte e positiva (0.9414) entre o volume absoluto de exames e o volume de diagnósticos, indicando que a estratégia de rastreamento funciona: quanto mais exames, mais diagnósticos.

Por outro lado, o projeto identificou um artefato estatístico perigoso ao analisar a imunização. Observou-se uma correlação positiva extremamente forte (0.97) entre o total de doses de vacina HPV aplicadas e o total de óbitos. Isto significa, conforme destacado na análise, não implica que vacinas causam óbitos, mas sim revela um viés de confundimento

populacional: estados mais populosos (como SP) vacinam mais e também possuem mais óbitos em números absolutos. Isso reforça a necessidade de trabalhar com taxas padronizadas em vez de números absolutos para evitar conclusões espúrias. Quando analisado por meio de taxas a correlação permanece positiva, porém essa correlação enfraquece, fortalecendo a ideia de trabalhar com taxas ao invés de números absolutos. Além é claro dos números de vacinação estarem disponíveis apenas até o ano de 2022, enquanto os demais dados vão até os anos de 2025 ou 2024.

---

2. FONTES DE DADOS PÚBLICAS PARA O PROJETO

Aqui estão as bases públicas utilizadas e recomendadas — todas abertas, gratuitas e sem informação sensível(nomes e cpfs de pacientes ou quaisquer informação pessoal).

2.1 Datasus – SISCOLO / SISCAN (produção de exames citopatológicos)

- **Descrição**: Número de exames Papanicolau realizados no SUS.
- **Tipos de dados**: tabelas estruturadas (csv, xls, dbf).
- **Dado:** Exames citopatológicos (Papanicolau) e diagnósticos positivos.
- **Acesso**:
    - TabNet (consulta manual)
    - API / requisições HTTP
    - Download em lote
    - Scraping de tabelas (quando necessário)

2.2 SIM – Sistema de Informações sobre Mortalidade

- **Descrição**: Óbitos por câncer do colo do útero (CID-10 C53).
- **Tipo**: dados estruturados.
- **Dados**: Óbitos por neoplasia maligna do colo do útero (CID-10 C53).
- **Acesso**:
    - TabNet
    - Download DATASUS (arquivos DBC → convertidos para CSV)

2.3 IBGE – Estimativas Populacionais

- **Descrição**: População feminina por faixa etária e município.
- **Tipo**: CSV, API oficial..
- **Dados**: População feminina residente com faixa etária acima dos 9 anos.
- **Uso**: cálculo de taxas (ex.: mortalidade por 100 mil mulheres; cobertura por mulher-alvo).

2.4 INCA – Estatísticas de câncer

- **Descrição**: séries históricas e estimativas.
- **Tipo**: relatórios PDF, tabelas, planilhas..

- **Dados:** Laudo histopatológico: Carcinoma Epidermóide , Adenocarcinoma invasor , Adenocarcinoma in situ , NIC III / Carc. in situ

- **Uso**: contextualização, comparação nacional.

---

3. ANÁLISE EXPLORATÓRIA DE DADOS (EDA)

Realizada com uma série de investigações estatísticas e visuais para compreender a distribuição dos dados, identificar relações entre as variáveis e preparar os dados para a modelagem mais aprofundada.

Em um projeto de ciência e análise de dados, a EDA é a etapa crucial para entender o que pode ser digno de nota ou interessante nos dados a serem destacados, antes de comunicá-lo formalmente (análise explanatória).

A seguir, detalho as principais abordagens de EDA utilizadas, que se concentraram em estatísticas descritivas, comparação de grupos e modelagem não supervisionada:

3.1 Limpeza e pré-processamento

Embora a preparação de dados não seja estritamente EDA, ela é um pré-requisito que molda o que pode ser explorado. A **Extração e Limpeza:** Os dados foram coletados utilizando **Web Scraping** do DATASUS (SIM/SISCAN). Após a coleta, os DataFrames foram **limpos** (remover colunas desnecessárias, garantir tipagem correta) e convertidos para o formato "*long*" para facilitar a análise de séries temporais.

Feature Engineering: Métricas-chave foram criadas, agregando os dados no *df_master* para permitir comparações que não fossem dominadas apenas pelo tamanho populacional.

- Conversão de arquivos DBC → CSV
- Padronização de nomes de colunas
- Transformação de valores nulos
- Criação de novas variáveis:
    - *taxa_mortalidade = óbitos / população feminina* × 100.000
    - *cobertura = n_exames / população feminina alvo*
    - *razão exames/população*
    - *log-transformações* para regressões

3.2 Análise descritiva

Esta etapa concentrou-se em quantificar as associações entre as variáveis utilizando dados agrupados por UF no período de 2013-2021.

- **Cálculo de Correlações (Pearson, Spearman e Kendall):** Foram calculados os coeficientes de correlação para avaliar a força e a direção da associação entre pares de variáveis, como:
    ◦ Cobertura de Exames vs. Mortalidade (Correlação esperada: negativa no longo prazo, sugerindo eficácia do rastreamento).
    ◦ Exames vs. Diagnósticos e Diagnósticos vs. Óbitos (Correlação esperada: positiva forte, indicando que mais exames levam a mais detecção).
    ◦ Imunização HPV vs. Diagnósticos/Óbitos (Correlação esperada: fraca devido à latência da doença).
    ◦ O uso de Spearman e Kendall forneceu uma **validação robusta** da direção da correlação, sendo menos sensível a *outliers* do que Pearson.
- **Ranking Top 5:** Foram identificadas as 5 Unidades da Federação (UFs) com a **maior mortalidade** e **menor cobertura de exames**, fornecendo um resumo de desempenho crucial para focar a análise em áreas críticas.

3.3 Identificação de variáveis importantes

- mortalidade
- população feminina
- número de exames
- número de diagnósticos positivos
- número de imunizações

3.4 Principais técnicas usadas além da análise descritiva

- ***Análise de Variação entre Grupos (ANOVA):***

A Análise de Variância (ANOVA) foi utilizada para explorar se as diferenças observadas em métricas-chave entre grupos eram estatisticamente significativas.

- **Diferença entre UFs:** Foi testado se havia diferença significativa na **cobertura média de exames** entre as 27 UFs. O resultado confirmou uma diferença estatística.
- **Diferença entre Regiões:** Foi testado se havia diferença significativa na cobertura média de exames entre as cinco grandes regiões brasileiras. Este teste ajuda a entender as desigualdades regionais e o agrupamento dos dados.

- ***Análise de Tendências e Modelagem Exploratória:***

Modelos estatísticos e técnicas de visualização foram aplicados para identificar padrões não óbvios e agrupar estados com perfis semelhantes.

- • **Análise de Séries Temporais (Decomposição):** A taxa média nacional de mortalidade foi decomposta utilizando *seasonal_decompose* para isolar a **Tendência**

dos dados anuais ao longo do tempo. Esta é uma ferramenta trivial para de forma visual identificar a trajetória de longo prazo da mortalidade.

- • **Clusterização *K-Means*:** Este algoritmo de aprendizado não supervisionado (*clustering*) foi aplicado a recursos padronizados (*media_cobertura*, *media_mortalidade*, *total_doses_hpv*, etc.) para **agrupar as UFs em 3 perfis distintos**. A identificação desses *clusters* é uma forma exploratória de sugerir políticas direcionadas.
- • **Análise de Componentes Principais (PCA):** O PCA foi utilizado para **reduzir a dimensionalidade** dos dados multivariados das UFs para dois componentes principais (PC1 e PC2), auxiliando na visualização e interpretação de fatores estruturais subjacentes que explicam a variação entre os estados.
- • **Visualização Exploratória:** Gráficos foram gerados para examinar distribuições e relações, incluindo:

    ◦ **Heatmap de Correlação:** Usado para visualizar rapidamente todos os coeficientes de correlação (Pearson/Spearman) entre as métricas agregadas.

    ◦ **Boxplot por Região:** Usado para visualizar a **dispersão e a variação interna** da cobertura de exames entre as regiões.

Este projeto não se concentrou apenas em calcular médias (estatística descritiva), mas também empregou **análise de correlação avançada, testes de hipótese (ANOVA) e métodos de aprendizado de máquina (K-Means e PCA)** para transformar dados brutos em *insights* estruturados, prontos para serem comunicados ao público de maneira cirúrgica e eficaz.

4. RELATÓRIO DE INSIGHTS

4.1 Desempenho e Vulnerabilidade por UF (Análise de Ranking e Clusters)

O ranking e a clusterização revelaram uma forte desigualdade na qualidade dos indicadores entre os estados, permitindo a identificação de grupos de risco prioritário para intervenção.

Insights de Risco (Ranking Top 5):

- **Maior Mortalidade:** O ranking de mortalidade média no período é liderado pelo **Amazonas** (19.77/100k) e **Amapá** (16.22/100k), seguidos por Maranhão, Roraima e Rio de Janeiro.
- **Menor Cobertura de Exames (Rastreamento):** Os estados com pior desempenho em cobertura de exames (proxy para rastreamento) são **Rio de Janeiro**, Distrito Federal e Piauí. A baixa cobertura de exames (*razao_exames_pop*) apresenta uma correlação negativa fraca a moderada com a mortalidade (*taxa_mortalidade*) (-0.3220 Pearson), indicando que mais rastreamento está associado a menos mortalidade.
- **Maior Taxa de Positividade:** Tocantins, Amapá e Acre lideram a taxa média de positividade. Uma alta taxa de positividade (*taxa_positividade*) sugere que, embora poucos exames sejam realizados, a proporção de diagnósticos positivos por exame é alta, o que pode indicar que os casos são detectados tardiamente.

Insights de Agrupamento (K-Means e PCA):

- **Clusterização (K-Means):** O algoritmo agrupou as UFs em três perfis, sendo o **Cluster 1** composto por estados com o **pior cenário epidemiológico** (baixa cobertura e alta mortalidade), o que é essencial para o direcionamento de políticas.
- **Análise de Componentes Principais (PCA):** O **PC1** foi interpretado como o **"Eixo de Vulnerabilidade Epidemiológica"**. Estados com **PC1 muito positivo** (ex: Acre, Amapá, Amazonas) apresentam as maiores fragilidades estruturais (baixa cobertura, alta mortalidade, baixa vacinação), enquanto estados com **PC1 muito negativo** (ex: São Paulo, Minas Gerais, Santa Catarina) demonstram melhor desempenho e maior proteção populacional.

4.2 Relevância dos Achados

- Evidenciam desigualdade regional severa.
- Baseiam decisões como alocação de recursos, campanhas ou expansão do atendimento itinerante.

- Permitem priorizar regiões críticas para intervenção imediata.
- Comprovam estatisticamente que aumentar a cobertura reduz mortes.

4.3 Eficácia e Associação das Variáveis

Relação entre Rastreamento e Doença:

- **Exames vs. Detecção:** Foi encontrada uma **correlação muito forte e positiva** (Pearson: 0.9414, Spearman: 0.9274) entre o volume absoluto de exames realizados (total_exames) e o volume absoluto de diagnósticos positivos (total_diagnosticos). Isso confirma que o **rastreamento está funcionando conforme o esperado:** mais exames de papanicolau levam a mais detecção de casos. Vale ressaltar que aqui não há uma causalidade direta, mas sim uma correlação onde outras variáveis interferem na taxa de mortalidade além da taxa de cobertura.
- **Detecção vs. Mortalidade:** A correlação entre diagnósticos absolutos e óbitos absolutos foi **moderada e positiva** (Pearson: 0.5947, Spearman: 0.6850). O fato de essa correlação não ser tão forte quanto a de Exames vs. Diagnósticos sugerem que o **tratamento e a gravidade dos casos detectados podem variar entre os estados**. Assim como questões socioeconômicas e estruturais.

O Impacto (Fraco) da Cobertura e Renda:

- **Regressão Linear (Taxa de Cobertura -> Taxa de Mortalidade):** O modelo linear simples mostrou que a cobertura (media_cobertura) tem uma **relação inversa esperada** (coeficiente angular -0.2864) com a mortalidade: maior cobertura tende a reduzir a mortalidade.

    - **R^2 Baixo (0.1037):** O modelo explica **apenas 10.37% da variação na mortalidade**. Isso é um insight crucial, pois indica que a **cobertura por si só é insuficiente** para explicar a mortalidade; a maior parte da variação (quase 90%) depende de outros fatores não incluídos no modelo, como acesso e qualidade do tratamento, fatores socioeconômicos e desigualdade.

- **Elasticidade (Impacto Percentual):** A regressão log-log confirmou que o impacto é fraco: um aumento de **+1% na cobertura está associado a uma redução de apenas -0.0890% na mortalidade**.

4.4 Alertas Críticos de Artefato Estatístico (Vacinação HPV)

A análise exploratória dos dados absolutos de imunização (vacina HPV) gerou alertas estatísticos graves devido a um forte viés de confundimento por tamanho populacional.

- **Imunização vs. Óbitos**: Foi observada uma correlação extremamente forte e positiva (Pearson: 0.9721; Spearman: 0.9878) entre o volume absoluto de doses aplicadas (total_doses_hpv) e o volume absoluto de óbitos (total_obitos).

    ◦ Interpretação (Alerta): Esse resultado é um artefato estatístico e não implica que a vacinação cause óbitos. Ele apenas reflete que estados com maior população (como São Paulo e Minas Gerais) aplicam mais doses absolutas e registram mais óbitos absolutos, criando uma correlação adulterada em dados não padronizados (totais absolutos). Há também o fato de que a vacinação de hpv no Brasil só começou em 2014, e os danos do vírus ao corpo podem levar anos até gerar o câncer e em alguns casos mais anos para levar a óbito. Logo, são muitos fatores a serem levados em consideração, todo um contexto que implica em uma interpretação correta dos insights.

4.5 Desigualdades Geográficas (ANOVA e Boxplots)

- **Diferença entre UFs:** Por meio da análise ANOVA demonstrou-se uma **diferença estatisticamente significativa (p ≈ 0.000000)** na cobertura média de exames entre os 27 estados. A diferença entre estados é 6.81x maior que a variação dentro de um estado ao longo do tempo.
- **Diferença entre Regiões:** O teste ANOVA subsequente comparando as cinco grandes regiões do Brasil resultou em **não significância estatística (p = 0.1423)**. O *boxplot* (visualização) reforça este insight ao mostrar que, mesmo o Sul tendo a maior mediana de cobertura, há **grande variação interna dentro de cada região**, o

que sugere que as diferenças se manifestam mais no nível estadual do que no agrupamento macrorregional.

4.6 Ações e Soluções Propostas

1. **Aumentar a oferta do exame em estados e municípios de baixa cobertura**, especialmente no interior.

2. **Investir em equipes itinerantes**, vans de saúde da mulher e mutirões.

3. **Campanhas educativas regionais** com foco em mulheres de 25 a 64 anos.

4. **Integração entre ESF e rastreamento ativo de mulheres faltosas**.

5. **Dashboard contínuo para monitoramento** das metas por município.

A EDA demonstrou que, mesmo que o rastreamento seja funcional (detectando mais casos quando há mais exames) e a cobertura tenha o efeito esperado (redução da mortalidade), a **capacidade preditiva de um modelo baseado apenas em cobertura é baixa ($R^2 \approx 0.10$)**. O principal desafio está na **desigualdade inter-estadual** e na necessidade de incorporar **fatores estruturais e socioeconômicos** (além da cobertura) para explicar e diminuir a mortalidade, além é claro incluir outras variáveis a fim de verificar o espectro como um todo. A análise com dados absolutos (Vacinação) reforçou a necessidade de **padronizar todas as variáveis** por população para evitar correlações falsas.