Relatório – Aplicação de Técnicas de Machine Learning na Análise de Depressão em Mulheres

1. Introdução

A depressão é um importante problema de saúde pública, especialmente entre mulheres, estando frequentemente associada a fatores sociodemográficos, condições de saúde e experiências de violência. Nesse contexto, a aplicação de técnicas de Machine Learning permite não apenas identificar associações estatísticas, mas também avaliar a capacidade preditiva desses fatores e explorar relações não lineares entre variáveis.

Este trabalho tem como objetivo analisar os fatores associados à depressão em mulheres por meio de uma abordagem integrada, combinando métodos inferenciais tradicionais e modelos de Machine Learning. Busca-se, assim, compreender o comportamento dos dados, identificar variáveis relevantes e comparar o desempenho de diferentes modelos preditivos.

2. Estratégias de Pré-processamento dos Dados

O pré-processamento dos dados foi uma etapa fundamental para garantir a qualidade das análises e o desempenho adequado dos modelos.

Inicialmente, foi realizada uma análise exploratória dos dados, com o objetivo de compreender a distribuição das variáveis, identificar padrões, verificar a presença de valores ausentes e avaliar possíveis inconsistências. Variáveis categóricas foram analisadas por meio de frequências e proporções, enquanto variáveis numéricas foram avaliadas utilizando medidas descritivas e visualizações gráficas.

Os principais passos de pré-processamento incluíram:

- Seleção de variáveis: Foram selecionadas variáveis sociodemográficas, de saúde e relacionadas à experiência de violência sexual, com base na relevância teórica e nos padrões observados na análise exploratória.

- Tratamento de valores ausentes: Observações com valores ausentes em variáveis essenciais foram tratadas conforme a estratégia adotada nos notebooks, garantindo consistência entre os modelos.

- Codificação de variáveis categóricas: Para a regressão logística e o Random Forest, variáveis categóricas foram transformadas em variáveis indicadoras (one-hot encoding ou codificação equivalente). No caso do CatBoost, as variáveis categóricas foram utilizadas diretamente, explorando uma das principais vantagens do algoritmo.

- Desbalanceamento da variável alvo: A variável resposta (presença ou ausência de depressão) apresentou desbalanceamento entre as classes. Esse fator foi considerado na escolha dos modelos e nas métricas de avaliação, priorizando medidas como ROC AUC e Precision–Recall AUC.

- Divisão dos dados: O conjunto de dados foi dividido em treino e teste, garantindo que a comparação entre os modelos fosse realizada de forma justa, utilizando o mesmo conjunto de teste.

3. Modelos Utilizados e Justificativa

Foram utilizados três modelos principais, cada um com objetivos distintos dentro da análise.

3.1 Regressão Logística Multivariada

A regressão logística multivariada foi empregada como modelo inferencial, permitindo estimar associações ajustadas entre as variáveis explicativas e a ocorrência de depressão. Esse modelo foi escolhido por permitir interpretação direta dos coeficientes, ser amplamente utilizado em estudos epidemiológicos e por servir como referência comparativa para os modelos de Machine Learning.

Apesar de suas limitações preditivas, a regressão logística é fundamental para compreender a direção e a magnitude das associações entre as variáveis.

3.2 Random Forest

O Random Forest foi utilizado como um modelo de Machine Learning baseado em árvores de decisão, escolhido por conseguir capturar relações não lineares entre as variáveis, ser relativamente robusto a ruído e multicolinearidade e por permitir análise de importância das variáveis.

Esse modelo atuou como uma abordagem intermediária entre a regressão logística e métodos mais sofisticados de boosting.

3.3 CatBoost

O CatBoost foi escolhido como modelo principal de Machine Learning devido a características específicas do conjunto de dados:

- Capacidade nativa de lidar com variáveis categóricas;
- Bom desempenho em bases tabulares;
- Robustez em cenários de desbalanceamento entre classes.

O objetivo principal do CatBoost foi maximizar a capacidade preditiva e avaliar se os padrões identificados pela regressão logística se mantinham em um modelo não linear.

4. Resultados e Interpretação

4.1 Análise Exploratória

A análise exploratória revelou padrões consistentes entre a ocorrência de depressão e variáveis como idade, escolaridade, estado nutricional e, principalmente, experiências de violência sexual e suas consequências psicológicas. Esses achados embasaram a escolha das variáveis incluídas nos modelos preditivos e inferenciais.

4.2 Desempenho dos Modelos

A comparação dos modelos foi realizada utilizando o mesmo conjunto de teste e métricas adequadas ao problema.

Regressão Logística:

Apresentou desempenho preditivo inferior aos modelos de Machine Learning, o que era esperado devido à suposição de linearidade e ao seu foco inferencial. No entanto, forneceu estimativas interpretáveis e associações estatisticamente relevantes.

Random Forest:

Obteve desempenho intermediário, superando a regressão logística em termos preditivos, mas apresentando limitações na identificação consistente da classe minoritária (casos de depressão), especialmente no equilíbrio entre sensibilidade e precisão.

CatBoost:

Apresentou o melhor desempenho geral, com maior ROC AUC e melhor desempenho na curva Precision–Recall. O modelo demonstrou elevada sensibilidade para detectar casos de depressão, característica particularmente relevante em contextos de saúde pública.

4.3 Interpretação Integrada

A análise integrada mostrou convergência entre os modelos quanto às variáveis mais relevantes. Fatores como experiências de violência sexual, consequências psicológicas associadas, idade, escolaridade e estado nutricional apareceram consistentemente como importantes preditores da depressão.

Embora os modelos de Machine Learning apresentem maior capacidade preditiva, eles não permitem inferência causal direta. Dessa forma, seus resultados devem ser interpretados de maneira complementar à regressão logística, que fornece o embasamento estatístico das associações observadas.

5. Conclusão

Os resultados deste estudo demonstram que a combinação de métodos inferenciais e modelos de Machine Learning é uma abordagem eficaz para a análise de fenômenos complexos como a depressão. A regressão logística permitiu compreender as associações entre os fatores analisados, enquanto os modelos de Machine Learning, especialmente o CatBoost, apresentaram desempenho preditivo superior e maior sensibilidade na identificação de indivíduos em risco.

Essa abordagem integrada reforça a robustez dos achados e evidencia o potencial do Machine Learning como ferramenta complementar em estudos na área da saúde, especialmente para apoio à identificação de grupos vulneráveis e ao planejamento de estratégias de prevenção e cuidado.