

Análise Preditiva de Saúde Mental: Uma Abordagem de Machine Learning para Previsão de Risco de Depressão

Guilherme de Souza Teodoro

Fatec Rubens Lara

Guilhermest1125@gmail.com

Jose Augusto Theodosio Pazetti

Fatec Rubens Lara

Jose.pazetti@cps.sp.gov.br

Resumo

A depressão constitui um importante problema de saúde pública, destacando-se a necessidade de métodos capazes de auxiliar na identificação de indivíduos em situação de risco. Este estudo teve como objetivo desenvolver e avaliar modelos de Machine Learning supervisionados para previsão de risco de depressão a partir de variáveis psicossociais e comportamentais. Foi aplicado um questionário digital a 62 participantes, utilizando o PHQ-9 como referência para avaliação dos sintomas. Os dados foram submetidos a pré-processamento, análise exploratória e modelagem por Regressão Logística, Random Forest e XGBoost, com validação cruzada K-Fold. Na análise por níveis de severidade, o Random Forest apresentou o melhor desempenho geral, destacando-se nas classes leve (ROC-AUC = 0,78) e moderadamente severa (ROC-AUC = 0,92). O nível de estresse diário e a idade apresentaram maior relevância preditiva. Entretanto, o tamanho reduzido da amostra limita a generalização dos resultados e reforça a necessidade de estudos futuros com amostras maiores e mais diversificadas.

Palavras-Chave

Machine Learning, Depressão, Saúde Mental, Ciência de Dados, Aprendizado Supervisionado.

1. Introdução

A depressão constitui um importante problema de saúde pública e está associada a impactos significativos na qualidade de vida e na capacidade funcional. Segundo a Organização Mundial da Saúde, aproximadamente 332 milhões de pessoas vivem com depressão no mundo (WORLD HEALTH ORGANIZATION, 2025). Nesse contexto, técnicas de Ciência de Dados e Inteligência Artificial apresentam potencial para auxiliar na identificação de padrões relacionados à saúde mental (WORLD HEALTH ORGANIZATION, 2021). Estudos anteriores demonstram a aplicação de Machine Learning na predição e identificação da depressão, inclusive a partir de registros eletrônicos de saúde (NICKSON et al., 2023). Entretanto, permanece relevante investigar essas técnicas utilizando dados psicossociais e comportamentais obtidos diretamente por questionário, considerando fatores cotidianos como estresse, sono, atividade física e relações sociais.

Assim, este trabalho objetiva desenvolver e avaliar modelos supervisionados de Regressão Logística, Random Forest e XGBoost, utilizando o PHQ-9 como referência, para identificar padrões associados ao risco de depressão e comparar o desempenho dos modelos.

2. Metodologia

O estudo possui abordagem quantitativa e caráter observacional, utilizando dados primários coletados por meio de questionário digital entre 21 de agosto e 21 de outubro de 2025, totalizando 62 participantes. O instrumento reuniu questões sociodemográficas e relacionadas a hábitos de vida, qualidade do sono, nível de estresse, relações sociais, atividade física e sintomas depressivos. O Patient Health Questionnaire-9 (PHQ-9) foi utilizado como referência para avaliação da severidade dos sintomas depressivos (DAMIANO et al., 2023). Após a coleta, os dados foram organizados e analisados utilizando Python. Na etapa de pré-processamento, variáveis categóricas foram transformadas em representações numéricas por meio de One-Hot Encoding, enquanto variáveis numéricas foram padronizadas com o StandardScaler. Essas etapas são importantes para adequar os dados aos algoritmos de aprendizado de máquina utilizados (GÉRON, 2023).

Foram avaliados os modelos supervisionados Regressão Logística, Random Forest e XGBoost. Os dados foram divididos aproximadamente em 70% para treinamento e 30% para teste. Adicionalmente, foi empregada validação cruzada K-Fold para avaliar a estabilidade dos modelos e reduzir a dependência de uma única divisão dos dados.

O desempenho foi analisado por métricas de classificação, incluindo Precision, Recall, F1-Score, Accuracy e ROC-AUC. Também foram empregadas técnicas de interpretação e análise exploratória, incluindo importância das variáveis, correlação de Spearman e Partial Dependence Plots (PDP). Adicionalmente, K-Means e Análise de Componentes Principais (PCA) foram utilizados para explorar agrupamentos de participantes segundo características psicossociais e comportamentais.

3. Resultados e discussão

A análise dos modelos demonstrou diferenças de desempenho entre os algoritmos e entre os níveis de severidade dos sintomas depressivos. Na avaliação das curvas ROC, a Regressão Logística apresentou valores de ROC-AUC de 0,75 para a classe mínima, 0,71 para leve, 0,35 para moderada e 0,39 para moderadamente severa. O Random Forest obteve valores de 0,72, 0,78, 0,35 e 0,92, respectivamente, enquanto o XGBoost apresentou 0,62, 0,73, 0,28 e 0,67.

Entre os modelos avaliados, o Random Forest apresentou o melhor desempenho geral, destacando-se principalmente na classe moderadamente severa, com ROC-AUC de 0,92, e na classe leve, com 0,78. Entretanto, os três algoritmos apresentaram desempenho reduzido na classe moderada, evidenciando que a capacidade preditiva não foi uniforme entre os diferentes níveis de severidade. A comparação das métricas Accuracy, Precision, Recall e F1-Score também indicou desempenho geral superior do Random Forest.

A análise de importância das variáveis indicou o nível de estresse diário e a idade como os fatores de maior relevância preditiva. Os gráficos de dependência parcial também demonstraram aumento da probabilidade prevista de risco associado à elevação do nível de estresse.

Os resultados devem, entretanto, ser interpretados considerando as limitações do estudo. A amostra reduzida de 62 participantes constitui uma limitação importante para modelos de Machine Learning, podendo aumentar a variabilidade das métricas, o risco de sobreajuste e restringir a generalização para outras populações. A utilização da validação cruzada K-Fold permitiu avaliar a estabilidade dos modelos, mas não elimina as limitações decorrentes do tamanho da amostra. Estudos com amostras maiores e mais diversificadas são necessários para avaliar a capacidade de generalização dos resultados.

Também foram utilizados Partial Dependence Plots (PDP), K-Means e PCA para interpretação dos modelos e exploração de perfis de participantes.

De forma geral, os resultados demonstram o potencial da aplicação de técnicas de Ciência de Dados e Machine Learning na análise de informações relacionadas à saúde mental. Além do tamanho reduzido da amostra, devem ser considerados o caráter não probabilístico da coleta e a utilização de dados obtidos por questionário. Dessa forma, os resultados não devem ser interpretados como diagnóstico clínico, mas como uma análise exploratória e preditiva dentro das características da amostra estudada.

As Figuras 1, 2 e 3 apresentam, respectivamente, a análise de desempenho multiclasse, a comparação das métricas de desempenho dos modelos e as variáveis de maior relevância.

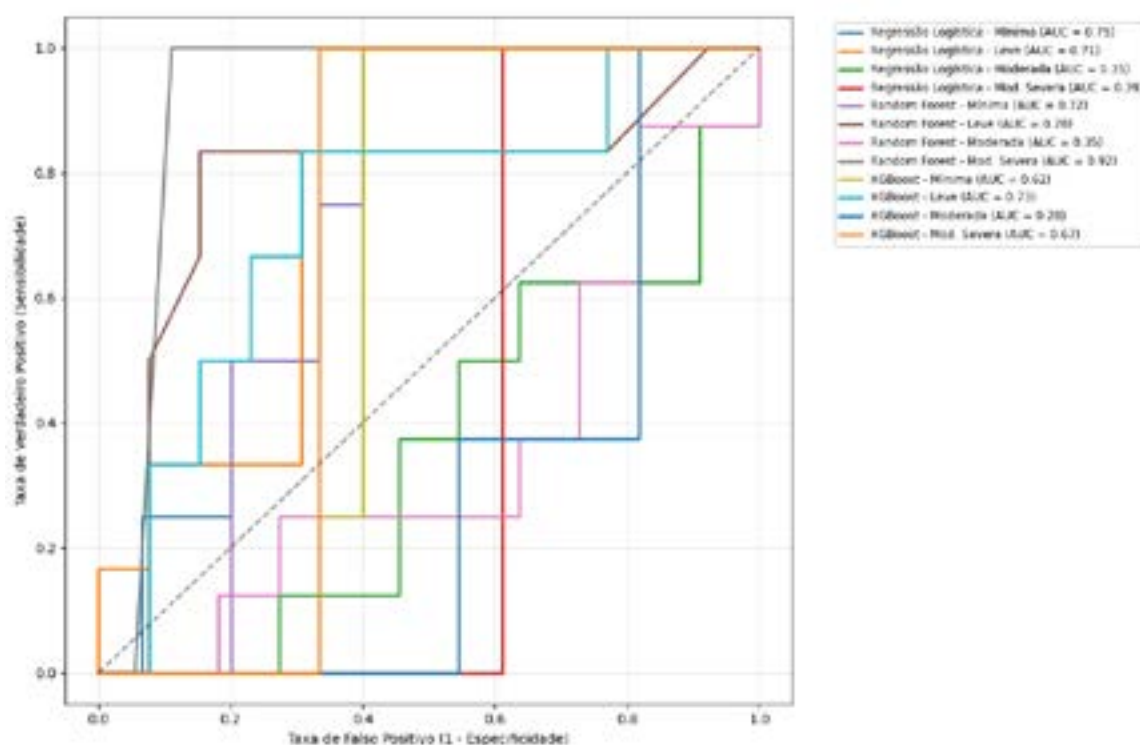


Figura 1: Análise de Performance Multiclasse por Nível de Severidade (PHQ-9)

Fonte: Elaboração do autor, 2026

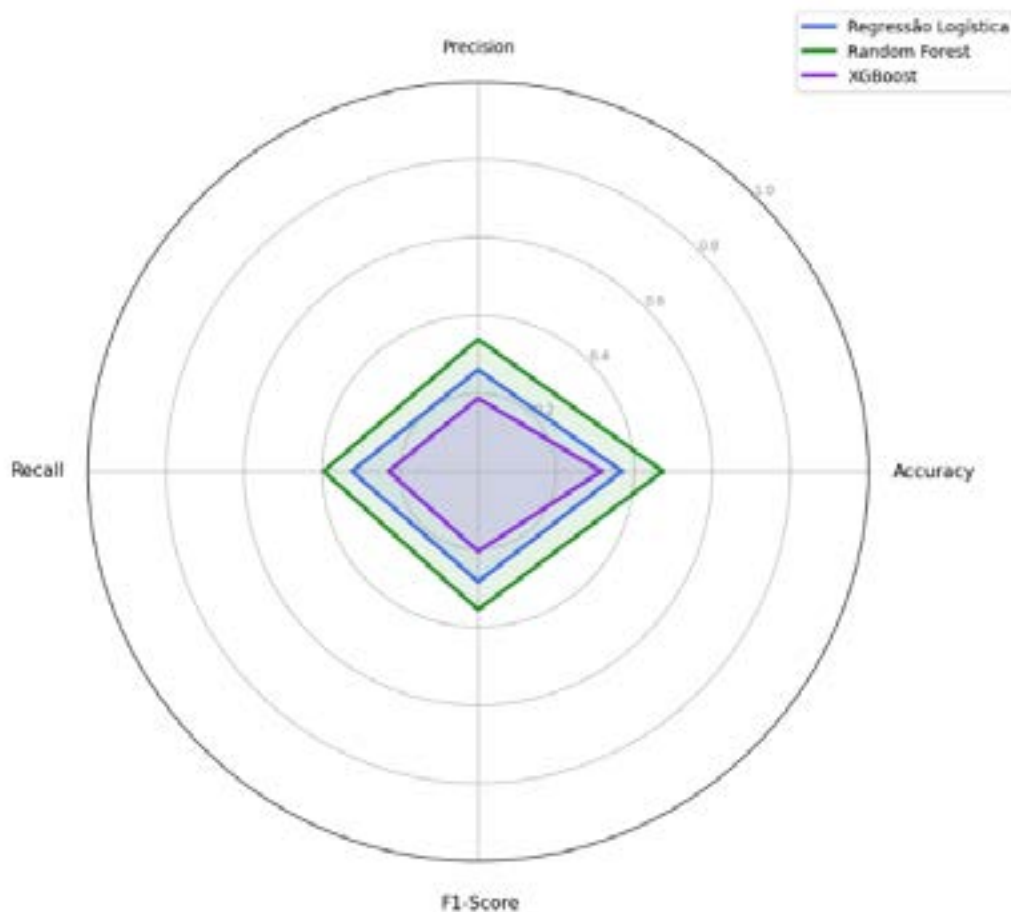


Figura 2: Comparação Multidimensional de Performance dos Modelos
Fonte: Elaboração do autor, 2026

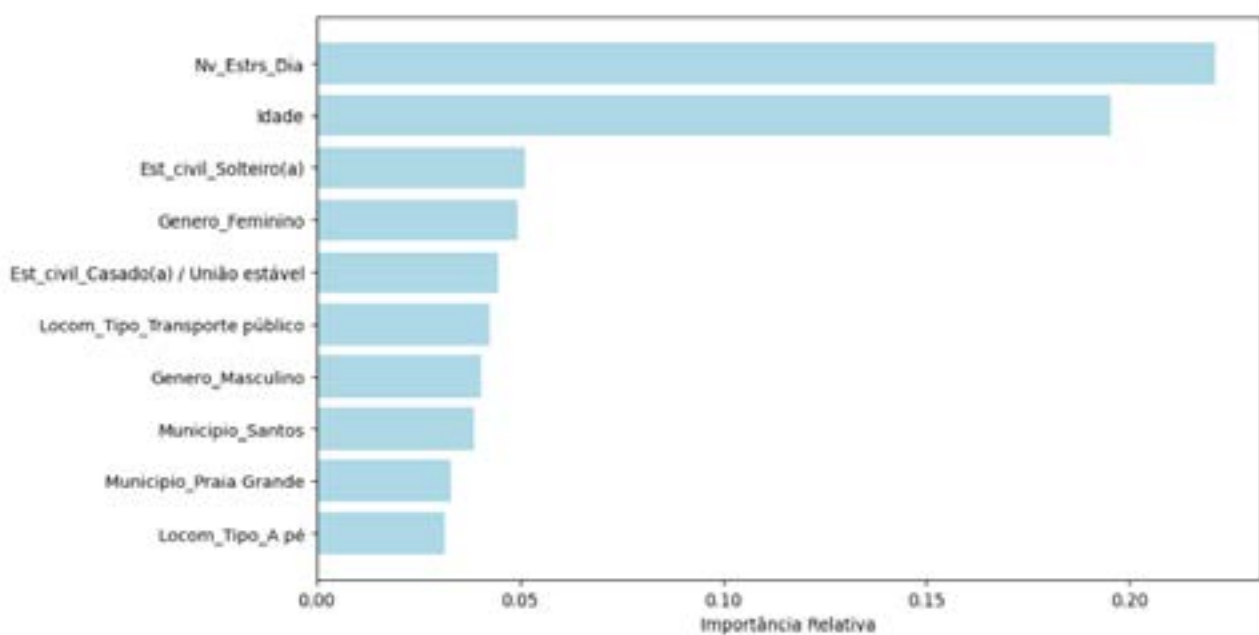


Figura 3: Top 10 Variáveis Predictoras de Risco de Depressão
Fonte: Elaboração do autor, 2026

4. Conclusão OU Considerações finais

Os resultados permitiram desenvolver e comparar modelos de Machine Learning para previsão de risco de depressão a partir de dados psicossociais e comportamentais. Entre os algoritmos avaliados, o Random Forest apresentou o melhor desempenho geral, destacando-se nas classes leve e moderadamente severa, com valores de ROC-AUC de 0,78 e 0,92, respectivamente. O nível de estresse diário e a idade foram identificados entre as variáveis de maior relevância preditiva.

Entretanto, o desempenho não foi uniforme entre os níveis de severidade, especialmente na classe moderada. Além disso, a amostra reduzida de 62 participantes limita a generalização dos resultados. Dessa forma, os modelos devem ser compreendidos como uma aplicação exploratória de técnicas de Ciência de Dados à saúde mental, e não como instrumentos de diagnóstico clínico.

5. Agradecimentos

Agradeço ao meu orientador, Professor José Augusto Theodosio Pazetti, pela orientação, acompanhamento e apoio durante o desenvolvimento deste projeto. Agradeço também ao Robert Richard das Neves Correia dos Santos, pela colaboração na elaboração dos modelos de Machine Learning e pela contribuição na análise dos dados. Por fim, agradeço a todos que contribuíram, direta ou indiretamente, para a realização e conclusão deste trabalho.

6. Referências

DAMIANO, Rodolfo Furlan et al. Translating measurement into practice: Brazilian norms for depressive symptom assessment with the Patient Health Questionnaire (PHQ-9). *Brazilian Journal of Psychiatry*, v. 45, n. 4, p. 310-317, 2023. DOI: 10.47626/1516-4446-2022-2945.

NICKSON, David; MEYER, Caroline; WALASEK, Lukasz; TORO, Carla. Prediction and diagnosis of depression using machine learning with electronic health records data: a systematic review. *BMC Medical Informatics and Decision Making*, v. 23, art. 271, 2023. DOI: 10.1186/s12911-023-02341-x.

WORLD HEALTH ORGANIZATION. *Comprehensive mental health action plan 2013–2030*. Geneva: World Health Organization, 2021.

WORLD HEALTH ORGANIZATION. *Depressive disorder (depression)*. Geneva: World Health Organization, 2025. Disponível em: <https://www.who.int/news-room/fact-sheets/detail/depression>.