Aprendizado de máquina aplicado no diagnóstico de transtorno depressivo e transtorno bipolar: um estudo com pacientes reais

Larissa Mitie Curi Hirai[1], **Alexandre Tadeu Rossini da Silva**[1]

[1]Curso de Ciência da Computação – Universidade Federal do Tocantins (UFT)

{larissa.hirai, arossini}@uft.edu.br

Abstract. *This paper presents a comparison between three machine learning algorithms applied to the diagnosis of mental disorders (depression, bipolar type I and II) based on data from 120 real patients. The methods used were random forest, KNN, and neural network, evaluated using accuracy, confusion matrix, and AUC metrics. The neural network with OneHotEncoder encoding showed the best performance, with accuracy above 95%, highlighting the potential of using AI to support the clinical diagnosis of mental diseases.*

Resumo. *Este artigo apresenta uma comparação entre três algoritmos de aprendizado de máquina aplicados ao diagnóstico de transtornos mentais (depressão, bipolar tipo I e II) com base em dados de pacientes reais. Foram utilizados os métodos random forest, KNN e rede neural, avaliados por métricas de acurácia, matriz de confusão e AUC. A rede neural com codificação OneHotEncoder apresentou o melhor desempenho, com acurácia superior a 95%, destacando o potencial da IA para apoiar o diagnóstico de transtornos mentais.*

1. Introdução

Os transtornos mentais, como depressão e transtorno bipolar, afetam milhões de pessoas, impactando significativamente sua qualidade de vida [Dattani et al. 2023]. No Brasil, a incidência do transtorno bipolar foi estimada em 1,08% da população em 2021, tornando-o o segundo país mais afetado no mundo, atrás apenas da Nova Zelândia [IHME 2024]. O diagnóstico clínico tradicional utiliza critérios padronizados, como os dos manuais DSM-5 [APA 2014] e CID-11 [OMS 2023b], mas pode apresentar limitações decorrentes da subjetividade dos avaliadores e da complexidade dos sintomas. Com o Aprendizado de Máquina (AM), surgem oportunidades promissoras para o desenvolvimento de modelos que complementem e aprimorem a precisão diagnóstica [OMS 2023a]. Destaca-se o papel desses modelos como componentes de sistemas colaborativos, nos quais a inteligência artificial atua de forma integrada com profissionais da saúde, oferecendo suporte à tomada de decisão clínica. Tais sistemas não visam substituir o julgamento médico, mas sim ampliar sua capacidade analítica, ao processar grandes volumes de dados e identificar padrões sutis que podem não ser evidentes na análise humana. Este estudo explora e compara diferentes algoritmos de AM aplicados ao diagnóstico de transtornos mentais, contribuindo para a melhoria dos métodos diagnósticos.

2. Procedimento metodológico

Este estudo, de natureza exploratória e quantitativa, utiliza AM para o diagnóstico de transtornos mentais. Para sua execução, foi adotado o método CRisp-DM (*CRoss-*

Industry Standard Process for Data Mining) [Shearer 2000], uma abordagem flexível e independente de domínio para orientar projetos de análise de dados.

2.1. CRisp-DM: Entendimento do negócio

O método CRISP-DM é caracterizado por uma fase inicial de entendimento do negócio. Para isso, deve-se definir o objetivo, que, neste estudo, é comparar a precisão de diagnóstico de algoritmos de aprendizado de máquina (AM) para a classificação de pacientes com depressão e transtorno bipolar tipo I ou tipo II. Também é necessário determinar as perguntas específicas que se pretende responder com a análise dos dados, a fim de entender como os resultados poderão ser utilizados:

1. Existem padrões ou associações entre certos sintomas comportamentais e diagnósticos específicos (depressão, transtorno bipolar tipo I ou tipo II)?
2. Qual é a eficácia dos algoritmos de AM na detecção de transtornos mentais?
3. Quais sintomas comportamentais são mais importantes na predição do diagnóstico de depressão e no de transtorno bipolar tipo I versus tipo II?
4. Quais características ou combinações de características são mais discriminativas na distinção entre pacientes com depressão e pacientes com transtorno bipolar?

Além das perguntas, também se busca identificar os *stakeholders* envolvidos. A pesquisa é voltada para profissionais de saúde, como psicólogos e médicos, com o objetivo de auxiliar no diagnóstico precoce de transtorno depressivo e transtorno bipolar.

2.2. CRisp-DM: Entendimento dos dados

Foi utilizada uma base de dados pública "*A Collection of 120 Psychology Patients with 17 Essential Symptoms to Diagnose Mania Bipolar Disorder, Depressive Bipolar Disorder, Major Depressive Disorder, and Normal Individuals*" [Karbalaeipour et al. 2023], disponível no repositório Harvard Dataverse. A base reúne informações de 120 pacientes reais de uma clínica psicológica particular e contém 19 variáveis, sendo 17 sintomas comportamentais, uma variável de identificação e uma variável alvo, que representa o diagnóstico clínico (normal, depressão, transtorno bipolar tipo I ou tipo II). Essa base foi escolhida por sua confiabilidade e por conter registros completos e organizados. Ademais, não foram encontrados trabalhos científicos publicados de AM que a utilizaram.

Foi realizada uma análise exploratória da base de dados, utilizando a matriz de correlação (Figura 1) para identificar relações entre sintomas e diagnósticos. A análise revelou que, embora não existam correlações fortes (acima de 0,80), foram observadas correlações moderadas entre algumas variáveis. No caso do transtorno depressivo, as correlações mais destacadas foram com os sintomas *"mood swing"* (0,46), *"optimism"* (-0,39), *"suicidal thoughts"* (0,24), *"overthinking"* (0,24) e *"aggressive response"* (-0,23). Para o transtorno bipolar tipo I, observou-se correlação com *"sexual activity"* (0,49), *"optimism"* (0,49), *"mood swing"*(0,46), *"aggressive response"* (0,41) e *"authority respect"* (-0,20). Já para o transtorno bipolar tipo II, a maior correlação registrada foi com *"mood swing"* (0,62), seguida por *"suicidal thoughts"* (0,32), *"nervous breakdown"* (0,23), *"optimism"* (-0,37) e *"sexual activity"* (-0,39). Em contrapartida, os indivíduos classificados como normais apresentaram correlação com *"optimism"* (0,29), *"concentration"* (0,28), *"mood swing"* (-0,51), *"suicidal thoughts"* (-0,47) e *"nervous breakdown"* (-0,37).

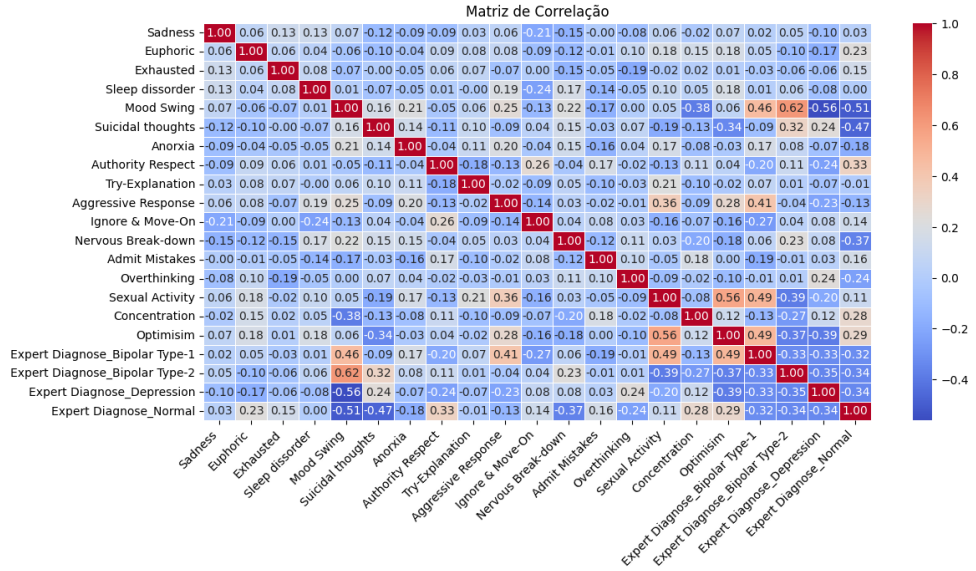

Figura 1. Matriz de correlação

Ademais, foi realizada uma análise dos dados de cada classe diagnóstica. Observou-se que, na base de dados, os diagnósticos estão bem balanceados, com 31 registros de pacientes com transtorno depressivo, 28 registros de pacientes com transtorno bipolar tipo I, 31 registros com transtorno bipolar tipo II e 30 registros com ausência de transtorno mental (normal). Dessa forma, os dados estão bem distribuídos e balanceados, apresentando consistência entre si.

2.3. CRisp-DM: Preparação dos dados

Na etapa de preparação dos dados, foram realizadas transformações com o objetivo de adequar a base para aplicação dos modelos. Inicialmente, os dados passaram por um processo de codificação das variáveis categóricas, utilizando dois métodos distintos: *OneHotEncoder* e *OrdinalEncoder*. O *OneHotEncoder* converteu as categorias em colunas binárias (0 ou 1), gerando aumento significativo na dimensionalidade da base de dados, com 62 variáveis resultantes. Já o *OrdinalEncoder* atribuiu valores numéricos inteiros às categorias, mantendo a estrutura com 17 variáveis, o que reduziu a complexidade dimensional.

Além das codificações, foram aplicadas técnicas de seleção e redução de dimensionalidade para avaliar o impacto no desempenho dos modelos. A técnica de RFE (*Recursive Feature Elimination*) foi usada para selecionar as variáveis mais relevantes, eliminando gradualmente as menos importantes com base na performance de um modelo de classificação. Já a técnica de SVD (*Singular Value Decomposition*) foi utilizada para realizar a redução de dimensionalidade, testando diferentes quantidades de variáveis reduzidas, e observando a porcentagem de perda de informação em cada cenário.

Foi identificado que o atributo '*suicidal thoughts*', que é uma variável categórica binária, apresenta três valores únicos, o que sugere a presença de um possível '*outlier*. Para tratar esse problema, constatou-se que um registro apresenta uma escrita divergente dos demais, com um espaço extra no final do valor "Yes ", o que o torna único em comparação com os outros valores "Yes". Para resolver isso, foi removido o espaço fi-

nal do valor, o que foi suficiente para corrigir e tratar o *outlier*. As demais variáveis não exibem *outliers* evidentes.

2.4. CRisp-DM: Modelagem

Foram utilizados três modelos supervisionados amplamente empregados em problemas de classificação e com diferentes características: rede neural (*Multilayer Perceptron*), *random forest* e *K-Nearest Neighbors* (KNN). As rede neural são eficazes na identificação de padrões não lineares e complexos; o *random forest* é robusto contra *overfitting* e útil para avaliar a importância das variáveis; e o KNN é um método simples, baseado na similaridade entre amostras, que pode oferecer bons resultados quando há uma estrutura clara de agrupamento nos dados.

Na etapa de modelagem, os três algoritmos foram treinados separadamente com os dados transformados pelas duas técnicas de codificação — *OneHotEncoder* e *OrdinalEncoder* — com o objetivo de comparar seu desempenho em diferentes estruturas de dados. A rede neural foi implementada utilizando o *MLPClassifier* da biblioteca *scikit-learn*, com camadas ocultas compostas por 256, 128 e 64 neurônios, utilizando a função de ativação ReLU e uma taxa de aprendizado inicial de 0,0005, visando uma convergência mais robusta. O modelo de *random forest* também foi configurado com a biblioteca *scikit-learn*, utilizando 300 árvores de decisão com profundidade limitada a 7. A métrica escolhida para avaliar a qualidade das divisões foi o índice Gini. Já o algoritmo KNN foi implementado com o *KNeighborsClassifier*, também da *scikit-learn*, considerando os 55 vizinhos mais próximos. Os pesos foram definidos com base nas distâncias entre os vizinhos, e a métrica adotada para calcular a proximidade entre as amostras foi a distância de *Manhattan*, que leva em conta as diferenças absolutas entre as coordenadas.

A pesquisa utilizou validação cruzada para dividir os dados em subconjuntos de treinamento e teste, escolhendo esse método devido ao tamanho reduzido da base de dados. A abordagem adotada foi a validação cruzada estratificada com 120 *folds*, implementada pelo *StratifiedKFold* do *scikit-learn*, configurando um processo equivalente ao LOOCV. Essa estratégia garantiu o uso eficiente dos dados, minimizando viés e proporcionando estimativas robustas do desempenho dos modelos.

2.5. CRisp-DM: Avaliação

Para avaliar a eficácia desses modelos, foi necessário definir métricas de desempenho adequadas ao contexto do aprendizado supervisionado. As métricas escolhidas foram: acurácia, que mede a proporção de classificações corretas; matriz de confusão, que detalha os acertos e erros de cada classe, permitindo compreender como o modelo se comporta diante dos diferentes diagnósticos; e a área sob a curva ROC (AUC), que avalia a capacidade do modelo de distinguir corretamente entre as classes.

Os gráficos de impacto do número de variáveis, apresentados nas Figuras 2a e 2b, destacam que a rede neural foi o modelo que melhor explorou as informações disponíveis, exibindo um aumento na acurácia à medida que mais variáveis eram incluídas.

Os resultados indicaram que o modelo de rede neural apresentou o melhor desempenho geral quando utilizado com *OneHotEncoder*, atingindo acurácia superior a 95% e AUC de 0,99. Esse desempenho se deve ao bom aproveitamento da alta dimensionalidade

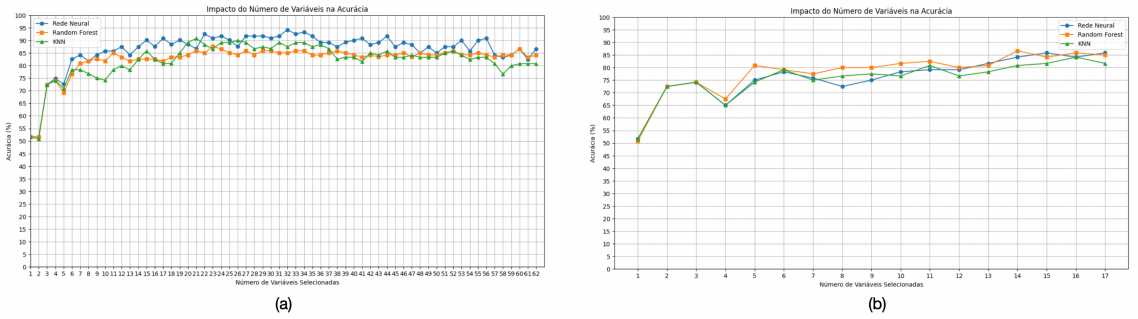

Figura 2. Relatório do impacto do número de variáveis

que o *OneHotEncoder* proporciona. Já com *OrdinalEncoder*, o desempenho da rede neural foi inferior, com acurácia em torno de 86% e AUC de 0,97, o que pode ser atribuído à introdução de uma ordem artificial entre as variáveis categóricas.

O algoritmo *random forest* também apresentou bons resultados com ambas as técnicas de codificação. Com *OneHotEncoder*, alcançou acurácia acima de 85% e AUC de 0,98 com 23 variáveis. Com *OrdinalEncoder*, obteve acurácia próxima de 85% e AUC de 0,98 com 14 variáveis, demonstrando robustez e estabilidade no desempenho.

O algoritmo KNN teve desempenho inferior quando comparado aos demais no *OrdinalEncoder* e um desempenho relativamente bom com o *OneHotEncoder*, com *OneHotEncoder*, atingindo acurácia acima de 90% e AUC de 0,96 com 21 variáveis. No entanto, ao utilizar *OrdinalEncoder*, o KNN obteve resultados mais satisfatórios, com acurácia em torno de 85% utilizando 16 variáveis, beneficiando-se da menor dimensionalidade.

De modo geral, os resultados demonstraram que a rede neural com *OneHotEncoder* foi a combinação mais eficaz, enquanto o KNN foi o mais sensível às transformações nos dados, evidenciando a importância da escolha da técnica de codificação conforme o algoritmo utilizado. A Figura 3 apresenta a matriz de confusão da rede neural com codificação *OneHotEncoder*, que foi a que obteve o melhor resultado nos testes.

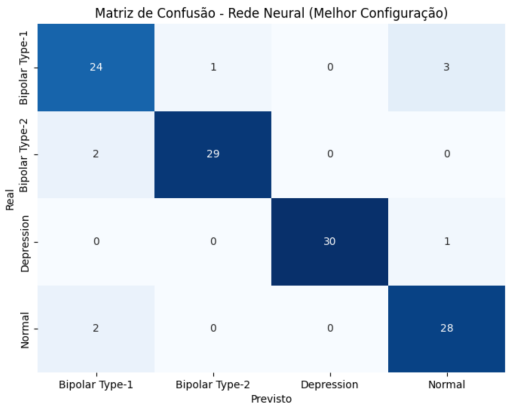

Figura 3. Matriz de confusão da rede neural com *OneHotEncoder*

3. Resultados obtidos

Como resultados, foram obtidas respostas para as perguntas formuladas na etapa de entendimento do negócio, as quais esta pesquisa se propôs a responder.

1) Existem padrões ou associações entre certos sintomas comportamentais e diagnósticos específicos (depressão, transtorno bipolar tipo I ou tipo II)?

Sim. A matriz de correlação revelou que sintomas como *Overthinking* (0,24) e *Suicidal Thoughts* (0,24) estão positivamente correlacionados com a depressão, enquanto *Mood Swing* (-0,56) apresentou correlação negativa com esse transtorno, indicando que é mais característico dos transtornos bipolares.

2) Qual é o eficácia dos algoritmos de AM na detecção de transtornos mentais?

Sim. Os modelos de AM implementados foram capazes de prever com boa precisão os diagnósticos. A rede neural, com codificação OneHotEncoder, alcançou acurácia superior a 95%, mostrando ser a abordagem mais eficaz. *Random Forest* também obteve resultados consistentes (acima de 85%), enquanto o KNN teve desempenho aceitável, mas foi mais sensível ao tipo de codificação.

3) Quais sintomas comportamentais são mais importantes na predição do diagnóstico de depressão e no de transtorno bipolar tipo I versus tipo II?

O ranking dos sintomas indicou que *Mood Swing*, *Suicidal Thoughts*, *Optimism*, *Overthinking* e *Sleep Disorder* foram os mais relevantes para diferenciar entre os transtornos. *Mood Swing*, por exemplo, ficou em primeiro lugar em todos os rankings dos transtornos bipolares.

4) Quais características ou combinações de características são mais discriminativas na distinção entre pacientes com depressão e pacientes com transtorno bipolar?

As características mais discriminativas foram *Mood Swing*, *Optimism*, *Suicidal Thoughts*, *Sexual Activity* e *Overthinking*. *Mood Swing*, por exemplo, apresentou correlação positiva com os transtornos bipolares e negativa com a depressão, sendo o sintoma mais marcante. Esses sintomas se destacaram tanto na análise de correlação quanto na seleção por RFE, evidenciando padrões emocionais intensos e pensamentos recorrentes como fatores-chave na distinção entre os diagnósticos.

4. Considerações finais

Esta pesquisa avaliou a precisão de algoritmos de AM no diagnóstico de transtorno depressivo e transtorno bipolar. Utilizando o método CRISP-DM e uma base pública, foram treinados três modelos: rede neural, *random forest* e KNN. A rede neural, com codificação *OneHotEncoder*, obteve a melhor acurácia, acima de 95%. Ressalta-se que os modelos de AM não substituem o trabalho clínico, mas atuam como agentes colaborativos que apoiam profissionais da saúde em suas decisões diagnósticas.

Essa colaboração ocorre uma vez que os algoritmos processam grandes volumes de dados e identificam padrões comportamentais que podem não ser evidentes em avaliações humanas, contribuindo para a redução da subjetividade e do viés nas análises. A integração desses modelos em sistemas colaborativos, que unem capacidades humanas e computacionais, promove decisões mais precisas. Em especial nas áreas sensíveis como a saúde mental, o uso de tecnologias colaborativas potencializa o diagnóstico precoce e contribui para uma prática clínica fundamentada em dados. Desse modo, este trabalho apresenta contribuições para o fortalecimento de sistemas colaborativos em saúde mental.

Referências

APA, A. P. A. (2014). *DSM-5: Manual diagnóstico e estatístico de transtornos mentais*. Artmed Editora.

Dattani, S., Rodés-Guirao, L., Ritchie, H., and Roser, M. (2023). Mental health. *Our World in Data*. https://ourworldindata.org/mental-health.

IHME (2024). Global burden of disease. Disponível em: `https://www.healthdata.org/research-analysis/gbd`. Acessado em: 5 mai. 2024.

Karbalaeipour, H., Damari, S., Zolfagharnasab, M. H., and Haghdadi, A. (2023). A collection of 120 psychology patients with 17 essential symptoms to diagnose mania bipolar disorder, depressive bipolar disorder, major depressive disorder, and normal individuals.

OMS, O. (2023a). Depressive disorder (depression). Disponível em: `https://www.who.int/news-room/fact-sheets/detail/depression`. Acessado em: 09 jun. 2024.

OMS, O. (2023b). *International Classification of Diseases, 11th Revision (ICD-11)*. Retrieved from https://icd.who.int/en.

Shearer, C. (2000). The crisp-dm model: the new blueprint for data mining. *Journal of data warehousing*, 5(4):13–22.