57º SBPO
SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL
Gramado • RS
DE 5 A 9 DE OUTUBRO DE 2025

Visualização de Dados no Apoio à Modelagem Preditiva da Depressão Infantojuvenil com Aprendizado de Máquina

Ludmila Bruna Santos Nascimento, Cristiane Neri Nobre
Pontifícia Universidade Católica de Minas Gerais
Dom José Gaspar, 500, Belo Horizonte - Minas Gerais, Brasil
ludmila.bruna@sga.pucminas.br,nobre@pucminas.br

RESUMO

Este estudo investiga o uso da visualização de dados (VD) como ferramenta de apoio na construção de modelos preditivos voltados ao diagnóstico da depressão infantojuvenil. A VD foi utilizada para orientar decisões no pré-processamento, como limpeza, codificação e análise de correlação, e também na interpretabilidade das previsões por meio da técnica *SHAP*. A análise foi aplicada a dados de 378 indivíduos entre 10 e 16 anos. Os resultados mostram que a VD contribuiu para identificar padrões, inconsistências e variáveis relevantes, tornando o modelo mais compreensível e transparente.Além disso, a VD pode auxiliar a tomada de decisão clínica e técnica ao facilitar a interpretação dos resultados, destacar elementos críticos associados aos sintomas de depressão e tornar o processo decisório mais acessível, mesmo em contextos complexos. O estudo reforça a importância de práticas interpretáveis no uso da inteligência artificial em saúde mental.

PALAVRAS CHAVE. Visualização de Dados. Aprendizado de Máquina. Tomada de Decisão. Interpretabilidade. Depressão em Crianças e Adolescentes.

Tópicos: PO na Área de Saúde. Análise e Ciência de Dados. PO Analytics em Estatística e Aprendizado de Máquina.

ABSTRACT

This study investigates data visualization (VD) to support the construction of predictive models to diagnose childhood and adolescent depression. VD was used to guide decisions in pre-processing, such as cleaning, coding, and correlation analysis, and also in the interpretability of predictions using the SHAP technique. The analysis was applied to data from 378 individuals between the ages of 10 and 16. The results show that VD helped identify patterns, inconsistencies, and relevant variables, making the model more understandable and transparent. In addition, VD can assist clinical and technical decision-making by facilitating the interpretation of results, highlighting critical elements associated with symptoms of depression, and making the decision-making process more accessible, even in complex contexts. The study reinforces the importance of interpretable practices in using artificial intelligence in mental health.

KEYWORDS. Data Visualization. Machine Learning. Decision-Making. Interpretability. Depression in Children and Adolescents.

Topics: OR in Health. Data Science and Analytics. OR Analytics in Statistics and Machine Learning.

The authors would like to thank the National Council for Scientific and Technological Development of Brazil (CNPq - Code: 311573/2022-3), the Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES - Grant PROAP 88887.842889/2023-00 - PUC/MG, Grant PDPG 88887.708960/2022-00 - PUC/MG - Informatics and Finance Code 001), the Foundation for Research Support of Minas Gerais State (FAPEMIG - Codes: APQ-03104-24, APQ-03076-18, APQ-05058-23 and PCE-00344-25).

DOI: 10.59254/sbpo-2025-212508
https://proceedings.science/p/212508?lang=pt-br

1. Introdução

A afirmação da OMS — "Não há saúde sem saúde mental" [OMS, 2022] — evidencia que o bem-estar psicológico é parte indissociável da saúde geral. A saúde mental envolve mais do que ausência de transtornos, incluindo equilíbrio emocional e relações sociais saudáveis. Em 2022, estimou-se que uma em cada oito pessoas no mundo vivia com algum transtorno mental, sendo a depressão um dos mais prevalentes, com aumento para 8% entre crianças de 6 a 12 anos na última década [APA, 2014; OMS, 2022]. A detecção precoce é determinante para o sucesso terapêutico.

Avanços em Aprendizado de Máquina (AM) vêm possibilitando identificar padrões complexos e apoiar diagnósticos de depressão [Liu e et al., 2025]. Entretanto, a qualidade da base de dados (BD) é decisiva para a confiabilidade dos resultados. O pré-processamento, que prepara dados brutos para análise, influencia diretamente o desempenho e a interpretabilidade dos modelos [Zelaya, 2019]. Nesse contexto, a Visualização de Dados (VD) assume papel central: ao ser incorporada ao pré-processamento, permite identificar com maior agilidade *outliers*, valores ausentes e inconsistências, além de orientar decisões sobre limpeza, transformação e seleção de variáveis. Dessa forma, evita-se que o pré-processamento seja realizado "às cegas", reduzindo o risco de ajustes inadequados.

Na análise exploratória, representações visuais simplificam a compreensão de padrões e relações complexas [Brynne et al., 2013], sendo amplamente utilizadas em áreas como o sequenciamento de DNA [Burkhardt, 2022]. Gráficos de dispersão evidenciam *outliers*, *heatmaps* revelam correlações e histogramas mostram distribuições, ajudando a evitar redundâncias e perdas de informação. Essa etapa visual não apenas otimiza a preparação da base de dados, mas também aumenta a transparência das decisões tomadas na modelagem Oberoi e Chauhan [2019].

A visualização de dados também favorece a interpretabilidade dos resultados, especialmente na saúde mental, onde múltiplos fatores biopsicossociais interagem [Xu, 2023]. Explicações visuais, como as fornecidas pela técnica *SHAP*, tornam previsões mais compreensíveis e fortalecem a confiança dos profissionais de saúde [Shinde e Shivthare, 2024]. No âmbito das políticas públicas, a VD sintetiza grandes volumes de dados em evidências acionáveis, auxiliando na definição de estratégias e alocação de recursos.

Diante disso, este estudo tem como objetivo investigar o uso de técnicas de AM no diagnóstico de transtornos depressivos, destacando o papel da VD tanto no pré-processamento dos dados quanto na interpretabilidade dos resultados dos modelos. A hipótese central é que a VD tem o papel de auxiliar na definição das etapas de preparação dos dados e na interpretação dos modelos, contribuindo na identificação de indicadores principais, aumentando a transparência das decisões e, por consequência, promovendo diagnósticos mais precisos e tratamentos mais eficazes. Nesse contexto, utiliza-se a técnica *SHAP* (*SHapley Additive exPlanations*), que contribui para a interpretabilidade ao explicar, de forma visual e quantitativa, a influência de cada variável nas previsões dos modelos. Com isso, a análise busca aprimorar a compreensão dos dados clínicos, orientar escolhas mais adequadas de pré-processamento e, em última instância, favorecer diagnósticos mais precisos e tratamentos mais eficazes.

Ao contrário de abordagens que restringem a visualização de dados a um papel secundário, neste estudo ela foi integrada como elemento estruturante do processo decisório em todas as etapas de pré-processamento e interpretabilidade. Cada visualização foi cuidadosamente alinhada a critérios estatísticos, possibilitando a validação visual de resultados de testes formais e a detecção de padrões ou anomalias que poderiam passar despercebidos em análises exclusivamente numéricas. Essa integração acelerou o diagnóstico de problemas na base, reduziu retrabalho e promoveu maior consistência entre decisões técnicas e interpretações clínicas. A principal contribuição técnica consiste na sistematização de um fluxo metodológico que integra visualização de dados, métricas

DOI: 10.59254/sbpo-2025-212508
https://proceedings.science/p/212508?lang=pt-br

estatísticas e técnicas de imputação, codificação e seleção de atributos, avaliando o impacto de cada decisão na interpretabilidade final via SHAP. Embora componentes isolados desse processo já tenham sido explorados em trabalhos anteriores [Nascimento et al., 2024], sua formalização e aplicação a dados de saúde mental infantojuvenil configuram um avanço, permitindo reprodução e adaptação a outros contextos complexos com múltiplas variáveis biopsicossociais.

Para isso, utilizou-se uma base de dados para classificação de depressão em crianças e adolescentes brasileiros de 10 a 16 anos, obtida através de uma parceria com o Programa de Pós-Graduação em Psicologia: Cognição e Comportamento da Universidade Federal de Minas Gerais (UFMG). A base contém informações de 378 indivíduos, sendo 158 do gênero masculino e 220 do gênero feminino, além de 72 características demográficas, sociais e de saúde mental.

2. Referencial Teórico

2.1. Depressão

A depressão é um transtorno de saúde mental caracterizado por sentimentos contínuos de tristeza, desesperança e irritabilidade e outros sintomas como fadiga, culpa, desinteresse por atividades, alterações no sono e no apetite, além de dores físicas. O transtorno depressivo maior (TDM) é um tipo comum de depressão, influenciado por fatores biopsicossociais, com impacto direto na estrutura e função neuronal [Khune et al., 2023].

O diagnóstico geralmente envolve uma combinação de métodos incluindo questionários, entrevistas clínicas e exames laboratoriais para descartar comorbidades. A detecção precoce e o diagnóstico preciso são essenciais para garantir intervenções eficazes, considerando os impactos negativos da depressão sobre a qualidade de vida e produtividade [Wang et al., 2024].

O tratamento pode incluir intervenções farmacológicas, como inibidores seletivos da recaptação da serotonina e da serotonina-noradrenalina [Stachowicz e Sowa-Kućma, 2022], além de abordagens não farmacológicas, como terapia cognitivo-comportamental, estimulação magnética transcraniana e neuropsicoterapia [Preval e Gietzen, 2024].

2.2. Visualização, Pré-processamento e Interpretabilidade de Dados

A análise de dados em AM envolve etapas fundamentais que se complementam: VD, pré-processamento e interpretabilidade. Em conjunto, essas etapas são especialmente relevantes no contexto da saúde mental, onde múltiplas variáveis biopsicossociais estão interligadas [Xu, 2023], exigindo análises transparentes e acessíveis para aplicação clínica.

A VD é a etapa inicial da análise e permite explorar padrões, identificar anomalias e orientar decisões subsequentes de limpeza e transformação dos dados. Ferramentas visuais como gráficos de dispersão, histogramas e *heatmaps* revelam tendências, ciclos e anomalias, fornecendo as primeiras percepções sobre a estrutura dos dados [Saraswathi e Kasyap, 2024]. O uso de atributos visuais (como cor, forma e tamanho) também facilita a compreensão rápida de estruturas complexas, promovendo extração de conhecimento desde as etapas iniciais [Ware, 2019]. Essa análise inicial é fundamental para orientar as decisões no processo subsequente de pré-processamento, garantindo que as transformações aplicadas sejam apropriadas para o contexto dos modelos de AM [Chaudhari e Choudhari, 2018].

O pré-processamento de dados é responsável por tratar os dados brutos, tornando-os limpos, padronizados e prontos para a modelagem. Isso inclui tratamento de ruídos, valores ausentes e discrepantes e variáveis categóricas, por meio de técnicas como imputação, remoção, normalização e codificação, assegurando que o modelo de AM processe informações de maneira confiável e estruturada [Bala e Behal, 2024].

A etapa seguinte é a interpretabilidade, essencial para garantir que as previsões geradas pelos modelos de AM sejam compreensíveis ao usuário, além de confiáveis. No contexto da saúde,

DOI: 10.59254/sbpo-2025-212508
https://proceedings.science/p/212508?lang=pt-br

a interpretabilidade permite que médicos e especialistas entendam as decisões do modelo, aumentando a transparência e a confiança no uso dessas tecnologias [Imans et al., 2024]. Ela pode ser dividida em duas abordagens: local, que explica previsões individuais, e global, que fornece uma visão geral do comportamento do modelo. O método *SHapley Additive exPlanations (SHAP)* aplica-se a ambas, oferecendo *insights* sobre a influência de cada variável nas previsões do modelo [Dafali et al., 2023]. De maneira geral, o valor *SHAP* é calculado somando os efeitos de todos os atributos de entrada, o que resulta em uma aproximação da predição que se deseja explicar.

Dessa forma, a VD atua como a conexão entre a exploração inicial dos dados e a compreensão das decisões algorítmicas, ampliando a confiança nos modelos e contribuindo para uma aplicação clínica mais eficaz. Ao tornar as informações acessíveis e compreensíveis, ela se consolida como uma ferramenta estratégica no apoio à tomada de decisão médica.

3. Trabalhos Relacionados

A VD e o pré-processamento são etapas essenciais para o sucesso de modelos de AM na área da saúde, especialmente ao lidar com grandes volumes de dados e problemas complexos, como epidemias e transtornos mentais. Estudos recentes, como o de Wang [2024], destacam a importância da VD na identificação ágil de padrões e tendências, evidenciando seu amplo uso na análise de dados da COVID-19. Da mesma forma, a necessidade de um pré-processamento rigoroso é ressaltada por Chahid et al. [2023], que analisaram técnicas específicas para o setor médico, incluindo limpeza, transformação e normalização dos dados. Esses cuidados não apenas aprimoram a precisão dos modelos, como também reduzem vieses, promovendo decisões médicas mais confiáveis.

Na área da saúde mental, a Descoberta de Conhecimento em Bases de Dados tem se destacado como uma abordagem eficaz para a exploração de grandes volumes de informações. No estudo de [Farid e Donyatalab, 2023], dados da Open Sourcing Mental Illness[1] foram analisados para identificar padrões de saúde mental entre funcionários do setor de tecnologia, revelando uma maior prevalência de transtornos mentais entre homens e sua associação com a queda de produtividade. Esses achados reforçam a importância de processos estruturados de análise e interpretação de dados para embasar estratégias eficazes de mitigação.

Além disso, modelos explicáveis têm ganhado destaque na detecção e avaliação de transtornos mentais, contribuindo para aumentar a transparência e facilitar sua adoção na prática clínica. No estudo de Imans et al. [2024], um modelo baseado em *Dynamic Ensemble Learning* foi aplicado a dados do National Social Life, Health, and Aging Project[2], alcançando 88,33% de precisão na detecção da depressão e 83,68% na previsão de sua gravidade. Nesse contexto, técnicas de Inteligência Artificial Explicável desempenharam um papel essencial ao identificar fatores-chave, ampliando a interpretabilidade dos resultados e oferecendo um suporte mais robusto à tomada de decisão médica.

4. Metodologia

Este estudo utilizou um conjunto de dados fornecido pela UFMG, composto por informações de 378 crianças e adolescentes com idades entre 10 e 16 anos, sendo 158 meninos e 220 meninas. O banco de dados contempla 72 atributos relacionados a determinantes sociais e à saúde mental. A variável de interesse (classe), denominada CDI, indica o nível de depressão dos participantes, categorizado em “baixo”(63 casos) ou “alto”(315 casos), com base no percentil 85 da distribuição dos escores. A Tabela 1 descreve os atributos considerados e seus respectivos tipos.

Nesta análise, o uso da VD é fundamental para orientar as decisões em cada etapa do pré-processamento. Visualizações específicas são geradas conforme o tipo de dado, antes da aplicação de cada técnica, e sempre acompanhadas por análises estatísticas correspondentes — como o teste

[1] Organização que apoia a saúde mental na comunidade de tecnologia.
[2] Estudo longitudinal dos EUA sobre conexões sociais, saúde e envelhecimento em idosos.

DOI: 10.59254/sbpo-2025-212508
https://proceedings.science/p/212508?lang=pt-br

Tabela 1: Classificação dos atributos do conjunto de dados

Tipo	**Atributo**
Binário	Sexo, Pais vivem juntos ou separados, Mãe trabalha, Atendimento psicológico mãe, Atendimento psicológico familiar mãe, Medicamento mãe, Pai trabalha, Atendimento psicológico pai, Atendimento psicológico familiar pai, Medicamento pai, CDI (classe)
Numérico	Idade, Tempo com mãe durante semana, Tempo com mãe final semana, Tempo com pai durante semana, Tempo com pai final semana, Afetividade, Problemas ansiedade, Problemas somáticos, Problemas hiperatividade/déficit de atenção, Problemas de comportamento opositor-desafiador, Problemas conduta, Problemas obsessivo-compulsivo, Problemas stress pós-traumático, Qualidades positivas, Idade mãe, Tempo com filhos durante semana mãe, Tempo com filhos final semana mãe, Idade pai, Tempo com filho durante semana pai, Tempo com filho final semana pai, Questionário YSR (10 questões)
Nominais	Mora com quem, Atendimento psicológico, Escolaridade pai, Escolaridade mãe, Questionário CDI (27 questões)

qui-quadrado para atributos nominais, a correlação de Pearson para variáveis numéricas e o cálculo de p-valores para significância. Essa integração permitiu validar visualmente os resultados dos testes formais e identificar padrões, comportamentos e possíveis anomalias que poderiam passar despercebidos em abordagens exclusivamente numéricas, direcionando com mais precisão as ações subsequentes.

A Figura 1 ilustra o fluxo do processo, que começa com a coleta da base. Em seguida, utiliza-se a VD para a verificação de padrões e incoerências. A próxima etapa é a limpeza, que elimina inconsistências e valores irrelevantes. No tratamento dos dados, realizam-se transformações e imputações de dados ausentes para análise. A modelagem aplica algoritmos para extrair padrões, e o processo é finalizado com a interpretabilidade, explicando os resultados dos modelos de aprendizado. Toda a metodologia foi realizada utilizando a linguagem *Python*[3].

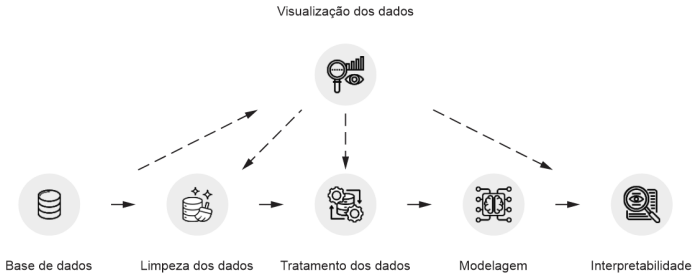


Figura 1: Fluxo da metodologia aplicada: desde a obtenção da BD, seguido pela VD, limpeza, tratamento e modelagem, até a interpretação dos resultados.

A limpeza dos dados é essencial para garantir a qualidade da base utilizada. Técnicas de VD, como *heatmaps* e *dotplots*, foram aplicadas para identificar padrões de correlação entre os atributos e a variável alvo, além de detectar valores ausentes. Essa análise permitiu identificar atributos e instâncias com muitos dados faltantes, assim como atributos altamente colineares, levando à eliminação desses elementos da base.

A transformação dos dados foi realizada com o objetivo de preparar a base para a modelagem. A distribuição das atributos categóricos foi analisada por meio de gráficos de barras e *Proportional Area Charts (Half Circle)*, possibilitando a percepção da frequência e das opções de resposta disponíveis. Com base nas análises realizadas, os dados nominais foram convertidos para formato numérico utilizando técnicas de codificação específicas. Atributos binários nominais foram transformados nos valores 0 e 1, atributos ordinais foram transformados com *Ordinal Encoding*, preservando a hierarquia, e atributos categóricos não ordinais foram tratados com *Dummy Encoding*,

[3]Documentação disponível em: `https://www.python.org/doc/`

DOI: 10.59254/sbpo-2025-212508
https://proceedings.science/p/212508?lang=pt-br

criando atributos indicadores para cada categoria. Em seguida, os dados foram imputados com o método *KNN Imputer*.

Na etapa de modelagem, a base de dados foi dividida em 80% para treino e 20% para teste. Foi utilizada a validação cruzada com 10 dobras e, para lidar com o desbalanceamento das classes, aplicou-se *SMOTE*. Na sequência, procedeu-se à otimização de hiperparâmetros do *Random Forest* via *BayesSearchCV* (*scikit-optimize*), com *n_iter*=30, *cv*=3, *scoring*=*'recall'* e *random_state*=42, explorando *n_estimators*, *criterion*, *max_depth*, *min_samples_split*, *min_samples_leaf*, *max_features* e *bootstrap*. O melhor conjunto obtido foi: *bootstrap=True*, *criterion='entropy'*, *max_depth=8*, *max_features='sqrt'*, *min_samples_split=7* e *n_estimators=22*. Todo o fluxo foi implementado em *Python* 3.11, com as bibliotecas `pandas`, `numpy`, `scikit-learn`, `imbalanced-learn`, `scikit-optimize`, `scipy`, `matplotlib`, `seaborn` e `shap`.

Com essa configuração, o *Random Forest* superou outros algoritmos avaliados — como *XGBoost*, Árvore de Decisão e Rede Neural — todos treinados e validados com o mesmo procedimento e ajuste de hiperparâmetros. A escolha final baseou-se no desempenho no conjunto de teste, adotando o *recall* como principal métrica, por se tratar de um cenário em que minimizar falsos negativos é essencial.

Por fim, a etapa de interpretabilidade teve como objetivo compreender os fatores que influenciaram as decisões do método de aprendizado. Para isso, foi empregada a técnica *SHAP*, que atribui valores de importância a cada variável em relação à saída do modelo. A análise foi conduzida por meio do *Summary Plot*, destacando as 10 variáveis mais influentes na predição.

5. Resultados e Discussões

Nesta seção, são apresentadas as visualizações das etapas de pré-processamento realizadas, bem como os resultados obtidos com o método de AM aplicado. Após a preparação da base, o algoritmo *Random Forest* foi implementado para a classificação dos casos de depressão infantojuvenil. O desempenho do modelo foi avaliado por meio de métricas amplamente utilizadas em problemas de classificação, incluindo precisão, *recall* e *F1-score*.

5.1. Resultados da visualização e das principais etapas de pré-processamento da BD

A VD foi fundamental para auxiliar na compreensão das características da base de dados e apoiar a tomada de decisão durante o pré-processamento. A aplicação de técnicas gráficas facilitou a identificação de padrões, inconsistências e estratégias adequadas para o tratamento dos atributos, contribuindo para decisões mais informadas e criteriosas. Com isso, foi possível garantir maior qualidade nos dados utilizados na análise da depressão infantojuvenil.

5.1.1. Correlação entre os atributos

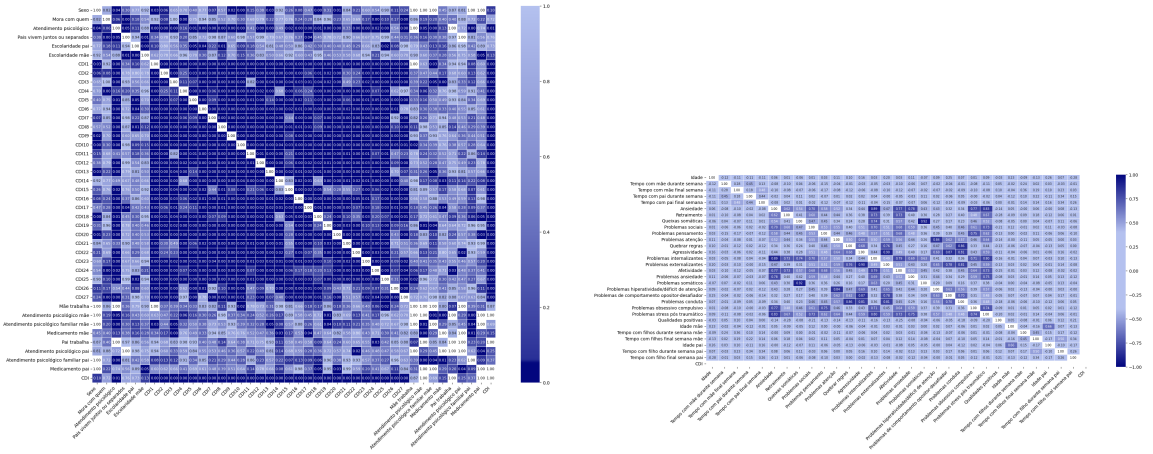
A análise de correlação entre os atributos foi conduzida com o apoio de dois *heatmaps*, ilustrados na Figura 2, um voltado para atributos numéricos e outro para atributos nominais. Essas visualizações atuaram como suporte na tomada de decisão sobre a manutenção ou remoção de variáveis, facilitando a interpretação das relações entre atributos e contribuindo para uma seleção mais criteriosa.

A Figura 2(a) apresenta o *heatmap* gerado a partir da matriz de p-valores do teste qui-quadrado, permitindo identificar associações estatisticamente significativas entre os atributos nominais e a classe (baixa ou alta depressão). Foi adotado um valor de corte de 0,05 para essa avaliação. A visualização revelou correlações relevantes nos atributos 'Atendimento psicológico' e nas variáveis do questionário CDI, destacados em azul escuro. Com base nessa evidência visual e estatística, optou-se por remover esses atributos da modelagem a fim de evitar redundância e a criação de relações artificiais ou não confiáveis.

Já a Figura 2(b) apresenta o *heatmap* produzido através da matriz de correlação de Pearson para os atributos numéricos. Utilizou-se um valor de corte de $\pm 0,7$ para identificar correlações

DOI: 10.59254/sbpo-2025-212508
https://proceedings.science/p/212508?lang=pt-br

fortes. A análise demonstrou que não havia atributos altamente correlacionados com a classe, o que sustentou a decisão de preservar todos esses atributos para a modelagem.



(a) Correlação entre os atributos nominais (b) Correlação entre os atributos numéricos

Figura 2: Correlação entre os atributos e sua relação com a identificação de atributos redundantes.

O uso dessas representações visuais foi útil para esclarecer, de forma intuitiva, as dependências entre atributos e auxiliar na identificação de redundâncias. Além disso, a VD contribuiu para selecionar atributos mais relevantes, aprimorando a qualidade do conjunto de dados e a interpretabilidade dos modelos. Esse processo se mostra especialmente importante em contextos de multicolinearidade, onde correlações elevadas podem distorcer a análise do impacto individual de cada variável nas previsões.

5.1.2. Verificação de dados ausentes

A VD é uma ferramenta útil na análise de valores ausentes, pois permite observar a quantidade de ausência total e de cada atributo. Técnicas como *heatmaps* de valores nulos, *dotplots*, gráficos de dispersão e histogramas facilitam a identificação de atributos com maiores proporções de ausência, possibilitando uma melhor compreensão da distribuição desses valores e contribuindo nas decisões sobre o tratamento adequado dos dados incompletos.

Além de auxiliar na identificação de ausências, a VD também contribui na tomada de decisão sobre a preservação ou exclusão de atributos, considerando não apenas a taxa de ausência, mas também sua relevância para o conjunto. Ao visualizar correlações, é possível avaliar se a exclusão de certos atributos comprometeria a diversidade informacional ou se há redundância suficiente para justificar essa remoção sem prejuízo significativo.

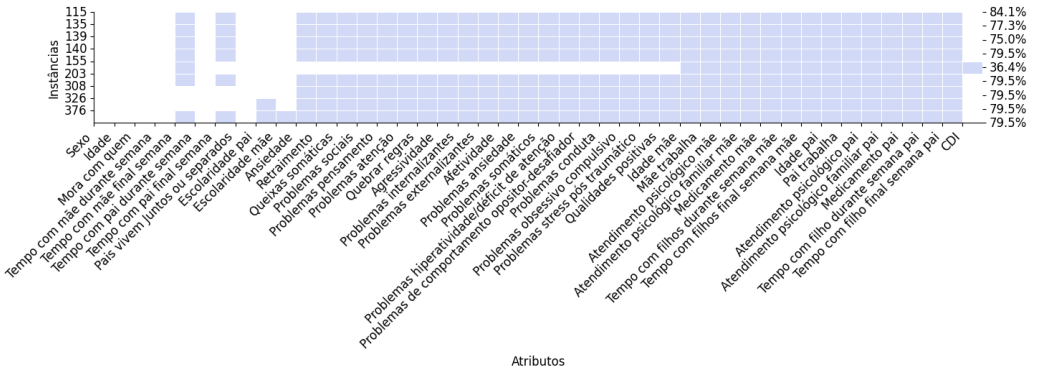
A Figura 3 ilustra os *dotplots* e *heatmaps* utilizados para analisar a distribuição dos valores ausentes, contribuindo para definir estratégias de imputação. Como parte desse processo, instâncias, Figura 3(a), e atributos, Figuras 3(b) e 3(c), com mais de 70% de valores ausentes foram removidos para reduzir ruídos e preservar a consistência da base.

Dessa forma, a VD auxilia na identificação do percentual de ausência, o que pode orientar a escolha de métodos de tratamento mais adequados. Embora não defina diretamente a técnica a ser aplicada, a visualização fornece subsídios importantes para decisões mais fundamentadas, contribuindo para a qualidade dos dados e, consequentemente, para a interpretabilidade dos modelos.

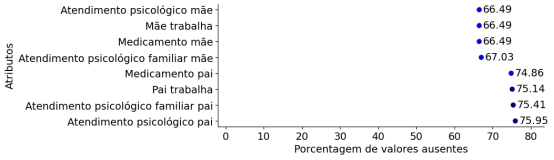
5.1.3. Codificação de atributos nominais

A VD desempenha um papel importante como ferramenta de apoio no processo de codificação de atributos, especialmente após a etapa de limpeza da BD. No caso desta BD, que

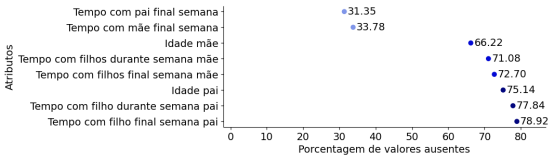
DOI: 10.59254/sbpo-2025-212508
https://proceedings.science/p/212508?lang=pt-br



(a) Valores ausentes para as instâncias com taxa de ausência de 50% ou mais.



(b) Valores ausentes em atributos nominais



(c) Valores ausentes em atributos numéricos

Figura 3: Distribuição de valores ausentes em instâncias e atributos e sua relação com a qualidade das informações.

inclui atributos nominais binários, categóricos ordinais e categóricos não ordinais, as visualizações contribuíram significativamente para a compreensão da distribuição e das relações entre as variáveis. Essas informações auxiliaram na escolha e validação das técnicas de codificação mais adequadas, tornando o processo mais fundamentado e transparente.

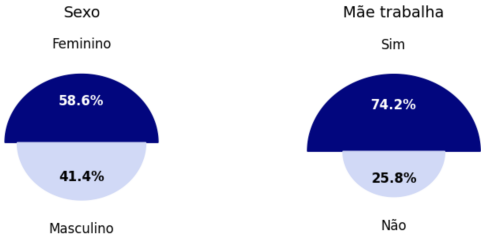
Para os atributos nominais binários, a Figura 4(a) exibe a distribuição em um *Proportional Area Chart (Half Circle)*, facilitando a visualização das proporções de cada categoria. Esses atributos foram binarizados, substituindo uma categoria por 0 e a outra por 1, mantendo a simplicidade e a interpretação direta dos dados.

Na Figura 4(b), a distribuição dos atributos categóricos ordinais é exibida em um gráfico de barras, evidenciando a hierarquia entre as categorias. Com base nessa hierarquia, foi aplicada a codificação ordinal, atribuindo números inteiros sequenciais que respeitam a ordem natural, como no atributo ‘Escolaridade Mãe’, que foi mapeado de 0 a 7 para refletir diferentes níveis de escolaridade. A VD reforça a coerência dessa escolha ao representar claramente a ordem dos valores.

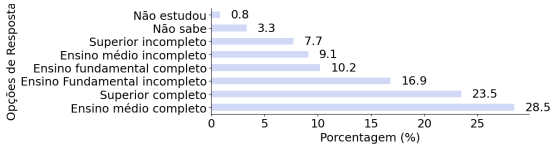
Para o atributo categórico nominal ‘Mora com quem’, a Figura 4(c) mostra a distribuição de frequências por meio de um gráfico de barras. Como não há hierarquia entre as categorias, a codificação ordinal não se aplicaria adequadamente. Inicialmente, considerou-se a técnica de *Frequency Encoding*, mas a repetição de frequências idênticas entre diferentes categorias, como ‘pai e irmão e outros’, ‘pai’ e ‘pai e madrasta e outros’, todas com frequência de 0,3, poderia comprometer a interpretabilidade dos dados. Diante disso, optou-se pela codificação *Dummy Encoding*, que transforma cada categoria em uma variável binária independente. Essa abordagem oferece maior clareza e interpretabilidade, especialmente em contextos onde a distinção entre categorias é essencial.

A VD, nesse contexto, foi particularmente útil para identificar padrões nas frequências e fundamentar essa decisão, atuando como um suporte visual à análise técnica, sem, no entanto, substituir a necessidade de critérios estatísticos e conceituais na escolha das codificações.

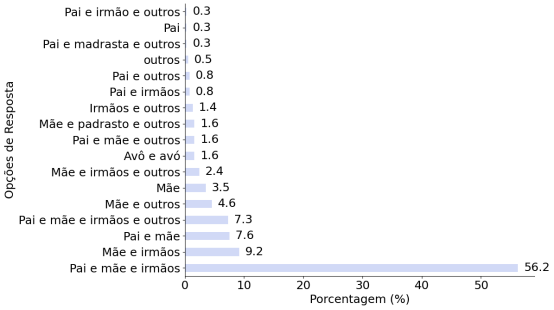
DOI: 10.59254/sbpo-2025-212508
https://proceedings.science/p/212508?lang=pt-br



(a) Proporção de respostas de dois atributos binários.



(b) Frequência de respostas de um atributo ordinal.



(c) Frequência de resposta do atributo categórico.

Figura 4: Visualização dos atributos nominais e sua relação com a codificação.

5.1.4. Imputação de dados ausentes

A VD, conforme ilustrado na Figura 3, serve como recurso complementar relevante na compreensão da distribuição das ausências e na definição da melhor estratégia de imputação. Embora o processo de imputação possa ser conduzido com base em critérios estatísticos, a VD oferece uma perspectiva complementar que facilita a identificação de padrões de ausência e possíveis vieses. Mesmo após a exclusão de atributos, ainda havia variáveis com menos de 70% de valores ausentes, tornando necessário um tratamento adequado para evitar perdas de informação e distorções na análise.

Inicialmente, considerou-se a imputação pela moda para atributos nominais e pela média para numéricos. No entanto, a análise dos gráficos de frequência revelou que esses métodos poderiam super-representar determinados grupos, distorcendo padrões e comprometendo a interpretabilidade do modelo. Por exemplo, no atributo ‘Mãe trabalha’, veja Figura 4(a), a imputação pela moda resultaria em ‘Sim’, o que poderia ocultar variações reais na distribuição dos dados. Com base nessa exploração visual, foi possível reconhecer limitações dos métodos simples de imputação. Como alternativa, optou-se pelo uso do *KNN Imputer* para imputá-los.

Essa verificação visual complementa os métodos estatísticos e contribui para manter a coerência dos dados, reforçando a confiabilidade dos modelos de AM. Ao possibilitar a análise integrada das estratégias de pré-processamento aplicadas, a VD mostrou-se relevante no apoio às decisões tomadas durante suas fases fundamentais. Sob essa perspectiva, a integração entre representações visuais e análises quantitativas assegurou que a base fosse adequadamente preparada para os algoritmos de AM, preservando tanto a informação contida nos atributos quanto a interpretabilidade dos resultados.

5.2. Resultados do modelo de aprendizado

Com precisão de 95% e *recall* de 90%, o modelo treinado classifica corretamente a maioria dos casos da classe baixa. O *F1-score* de 93%, uma média harmônica entre precisão e *recall*, sugere

DOI: 10.59254/sbpo-2025-212508
https://proceedings.science/p/212508?lang=pt-br

baixa ocorrência de erros ao identificar indivíduos com baixa depressão, reforçando a eficácia do modelo neste grupo.

Em contraste, a detecção de indivíduos com alta depressão apresenta um desempenho moderado. A precisão de 60% indica uma taxa considerável de falsos positivos, enquanto o *recall* de 75% sugere que a maioria dos casos reais de alta depressão foi corretamente identificada. O *recall* é crucial nesse contexto, pois a falha em detectar casos de depressão severa pode resultar no tratamento inadequado e agravamento do quadro. Embora o *F1-score* de 67% indique um certo desequilíbrio entre precisão e *recall*, a priorização do *recall* é essencial para garantir que a maioria desses indivíduos seja identificada. Um dos possíveis fatores para esse desempenho pode ser o desbalanceamento de classes, indicando que o algoritmo *SMOTE* pode não ter sido eficaz na tarefa de balanceamento. Além disso, a elevada quantidade de valores ausentes reduziu o número de atributos disponíveis, o que pode ter impactado o desempenho dos algoritmos de AM.

Quanto à interpretabilidade, a Figura 5 apresenta o *Summary Plot*, representação da distribuição e variabilidade dos valores de *SHAP* para cada variável. A visualização exibe as 10 variáveis mais influentes nas predições do modelo. Cada ponto representa uma observação individual (criança ou adolescente) e mostra o valor *SHAP* correspondente ao impacto daquele atributo na saída do modelo. A posição no eixo X indica se a variável contribui para aumentar (valores positivos) ou reduzir (valores negativos) a probabilidade de classificação como depressão alta. As cores refletem a magnitude do valor original da variável: vermelho para valores altos e azul para valores baixos. Vale destacar que nenhuma variável, isoladamente, determina a classe do indivíduo.

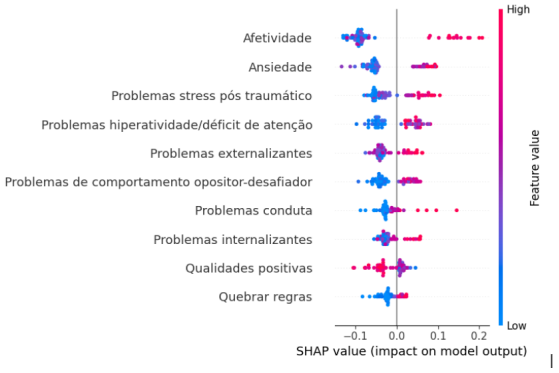


Figura 5: 10 atributos mais relevantes na detecção de depressão

Ao utilizar o algoritmo *Random Forest* em conjunto com a técnica *SHAP* para análise interpretativa, observou-se que as variáveis mais relevantes estão associadas a escalas orientadas no DSM-5 [APA, 2014], a qualidades positivas e ao questionário YSR, destacando a importância desses instrumentos na identificação de sintomas depressivos em jovens de 10 a 16 anos.

A análise indica que atributos como afetividade, ansiedade, problemas de estresse pós-traumático e problemas de atenção/hiperatividade estão fortemente associados ao aumento da probabilidade de depressão alta, especialmente quando seus valores são elevados. Por outro lado, variáveis como qualidades positivas e, em menor grau, comportamentos de quebra de regras, apresentam valores de *SHAP* negativos quando seus valores são altos, sugerindo um efeito protetor ou redutor no risco de depressão.

Esses resultados destacam a importância de características emocionais e comportamentais como preditores da condição depressiva em crianças e adolescentes. Além disso, evidenciam o valor da técnica *SHAP* como ferramenta de interpretabilidade, contribuindo para decisões clínicas e pedagógicas mais informadas com base em modelos de AM.

DOI: 10.59254/sbpo-2025-212508
https://proceedings.science/p/212508?lang=pt-br

Importante ressaltar que as visualizações em tamanho original estão disponíveis no material suplementar[4]. O uso do *SHAP* trouxe transparência às predições do modelo, facilitando sua aplicação prática em contextos reais. Além disso, a análise reforça a relevância da VD no pré-processamento, especialmente na transformação de variáveis e exploração gráfica, aprimorando a interpretabilidade e a identificação de padrões. Técnicas como *heatmaps*, *dotplots* e gráficos de barras foram fundamentais para detectar padrões e anomalias, contribuindo para a construção de um modelo mais robusto e garantindo resultados mais claros e consistentes.

6. Conclusões

Este estudo destacou a importância da visualização de dados nas etapas de pré-processamento para criar modelos de AM mais interpretáveis e confiáveis. Utilizando uma base de dados de crianças e adolescentes de 10 a 16 anos, a análise explorou a relação entre aspectos psicossociais e de saúde na classificação da depressão infantojuvenil.

A VD demonstrou ter papel complementar no suporte ao desenvolvimento de modelos interpretáveis para diagnóstico de depressão em crianças e adolescentes. No pré-processamento, a VD auxiliou na detecção de padrões, inconsistências e correlações, auxiliando decisões sobre tratamento de atributos e balanceamento. Na fase de interpretação, técnicas como o *SHAP* permitiram visualizar o impacto das variáveis nas previsões, promovendo maior transparência.

Embora métodos como *SMOTE* e *KNN Imputer* tenham sido úteis para lidar com desbalanceamento e dados ausentes, também introduzem limitações à interpretabilidade, exigindo avaliação criteriosa, já que ambos geram dados artificiais. Nesse contexto, surge um *trade-off* entre eficiência e interpretabilidade, onde a busca por melhores resultados em termos de performance pode comprometer a clareza e a transparência nas decisões tomadas pelo AM. Assim, é fundamental avaliar cuidadosamente a necessidade dessas técnicas em cada aplicação, ponderando os ganhos em acurácia em relação à potencial perda de confiabilidade na explicação dos resultados.

Desta forma, este trabalho ressalta a importância de incorporar práticas éticas e transparentes no uso de IA em contextos de saúde. A integração dessas práticas tem o potencial de melhorar a confiança dos usuários e expandir a sua aplicação em setores críticos, como a saúde mental.

Pesquisas futuras devem explorar novas visualizações para auxiliar em etapas adicionais do pré-processamento que não foram consideradas neste estudo. Além disso, é fundamental avaliar a aplicabilidade e a validação dos métodos em bases de dados de saúde de diferentes dimensões. Outro aspecto relevante é o uso de visualizações interativas e o desenvolvimento de ferramentas acessíveis que integrem essas técnicas de forma prática para profissionais da saúde.

Referências

APA (2014). *DSM-5: Manual diagnóstico e estatístico de transtornos mentais*.

Bala, B. e Behal, S. (2024). A brief survey of data preprocessing in machine learning and deep learning techniques. In *I-SMAC 2024*, p. 1755–1762. IEEE.

Brynne, L., Bresell, A., e Sjögren, N. (2013). Effective visualization of integrated knowledge and data to enable informed decisions in drug development and translational medicine. *Journal of translational medicine*, 11:1–11.

Burkhardt, J. (2022). *Statistics and Data Visualization*, p. 117–128. Facet.

Chahid, I., Elmiad, A. K., e Badaoui, M. (2023). Data preprocessing for machine learning applications in healthcare: A review. In *SITA 2023*, p. 1–6.

[4]Disponível em: `https://github.com/ludbruna/SBPO_2025`

DOI: 10.59254/sbpo-2025-212508
https://proceedings.science/p/212508?lang=pt-br

Chaudhari, M. e Choudhari, N. (2018). Data visualization as a preprocessing step in designing of data mining tools visualizing time series pattern of rainfall data. In *SMART 2018*, p. 99–103. IEEE.

Dafali, S. M., Kissi, M., e El Beggar, O. (2023). Comparative study between global and local explainable models. In *SITA 2023*, p. 1–8. IEEE.

Farid, F. e Donyatalab, Y. (2023). Unlocking insights into mental health and productivity in the tech industry through kdd and data visualization. In *INFUS 2023*, p. 90–98. Springer.

Imans, D., Abuhmed, T., Alharbi, M., et al. (2024). Explainable multi-layer dynamic ensemble framework optimized for depression detection and severity assessment. *Diagnostics*, 14.

Khune, A., Rathod, H., Deshmukh, S., e Chede, S. (2023). Mental health, depressive disorder and its management: A review. *GSC Biol. Pharm. Sci*, 25:1–13.

Liu, X. e et al. (2025). Identification of depressive symptoms in adolescents using machine learning combining childhood and adolescence features. *BMC Public Health*, 25(1):264.

Nascimento, L., Gonçalves, A., Balbino, e Nobre (2024). Codificaçao dos atributos e sua relaçao com a interpretabilidade dos modelos de aprendizado de máquina-uma análise em base de dados da saude mental. In *SBCAS*, p. 483–494. SBC.

Oberoi, A. e Chauhan, R. (2019). Visualizing data using matplotlib and seaborn libraries in python for data science. *IJSRP*, 9(3):p8733.

OMS (2022). *World mental health report: Transforming mental health for all*. World Health Organization.

Preval, H. e Gietzen, L. (2024). Treating depression using non-pharmacological interventions. *PJH*, 7(1):33.

Saraswathi, D. e Kasyap, S. (2024). Data visualization: From raw data to actionable insight. *IJSREM*, 08(12):1—-6.

Shinde, B. G. e Shivthare, S. (2024). Impact of data visualization in data analysis to improve the efficiency of machine learning models. *Journal of Advanced Zoology*, 45.

Stachowicz, K. e Sowa-Kućma, M. (2022). The treatment of depression—searching for new ideas. *Frontiers in pharmacology*, 13:988648.

Wang, T., Zhang, Z., Zhang, X., et al. (2024). Identification of depressive symptoms: A cause-and-effect based machine learning study. *Chinese Medical Journal*, 137(10):1258–1260.

Wang, Y. (2024). Machine learning in data analysis and visualization in healthcare. *TCSISR*, 7: 356—-362.

Ware, C. (2019). *Information visualization: perception for design*. Morgan Kaufmann.

Xu, B. (2023). The causes of major depressive disorder. In *ICBioMed 2022*, volume 12611, p. 934–939. SPIE.

Zelaya, C. V. G. (2019). Towards explaining the effects of data preprocessing on machine learning. In *ICDE 2019*, p. 2086–2090. IEEE.

DOI: 10.59254/sbpo-2025-212508
https://proceedings.science/p/212508?lang=pt-br