Codificação dos Atributos e sua Relação com a Interpretabilidade dos Modelos de Aprendizado de Máquina - Uma Análise em Base de Dados da Saúde Mental

Ludmila B. S. Nascimento[1], **Ana C. M. Gonçalves**[1], **Marcelo S. Balbino**[2], **Cristiane N. Nobre**[1]

[1]Pontifícia Universidade Católica de Minas Gerais,
Instituto de Ciências Exatas e Informática, Belo Horizonte, Brasil

[2]Centro Federal de Educação Tecnológica de Minas Gerais,
Departamento de Computação, Belo Horizonte, Brasil

{ludmila.bruna, ana.medeiros}@sga.pucminas.br, marcelobalbino@cefetmg.br, nobre@pucminas.br

Abstract. *This study examines data on mental disorders using machine learning techniques. The objective is to investigate six different methods of coding categorical attributes in three datasets related to mental disorders using Machine Learning algorithms and verify the interpretability of these methods. The results indicate that the choice of encoding is crucial for accurate results, varying according to the algorithm and data set. Hashing Encoding is the most effective in many situations, followed by Dummy Encoding in some cases. However, regarding interpretability, Dummy, One Hot, and Ordinal encodings offer greater interpretative clarity.*

Resumo. *Este estudo examina dados sobre transtornos mentais, utilizando técnicas de aprendizado de máquina. O objetivo é investigar seis diferentes métodos de codificação de atributos categóricos em três conjuntos de dados relacionados a transtornos mentais, utilizando algoritmos de aprendizado de máquina, e verificar a interpretabilidade desses métodos. Os resultados indicam que a escolha da codificação é crucial para resultados precisos, variando de acordo com o algoritmo e o conjunto de dados. A codificação Hashing destaca-se como a mais eficaz em muitas situações, seguida pela codificação Dummy em alguns casos. No entanto, quando se trata de interpretabilidade as codificações Dummy, One Hot e Ordinal oferecem maior clareza interpretativa.*

1. Introdução

A saúde mental refere-se ao estado geral do bem-estar psicológico e emocional de um indivíduo. É um elemento fundamental para a satisfação pessoal, relações familiares e interpessoais [Gamm et al., 2010].

Transtornos mentais são condições que prejudicam a cognição, emoção e controle comportamental, impactando o funcionamento socioeducacional. Geralmente, iniciam precocemente, são persistentes e comuns em todo o mundo. Devido à sua alta prevalência, esses transtornos contribuem significativamente para o impacto global de doenças [Hyman e Chisholm, 2006].

Segundo a OPAS [2022], em 2019, cerca de 1 bilhão de indivíduos em todo o mundo foram afetados por algum tipo de transtorno mental, sendo esse o motivo de principal causa de incapacidade. Ainda, segundo a OMS, os sistemas de saúde não conseguiram lidar adequadamente com a quantidade de pessoas com transtornos mentais, tendo como resultado uma lacuna considerável entre a demanda por tratamento e sua disponibilidade em todo o mundo. Em nações de baixa e média renda, entre 76% e 85% das pessoas afetadas por transtornos mentais não têm acesso ao tratamento. Já em países de alta renda, essa proporção varia entre 35% e 50%.

Intervenções precoces, no início do desenvolvimento de transtornos mentais, são cruciais e podem alterar o curso futuro do transtorno [Correll e Galling, 2018]. Além disso, a prevenção primária, direcionada a pessoas com alto risco clínico, é uma estratégia promissora, que pode alterar o curso do distúrbio e melhorar os resultados.

O progresso recente na pesquisa em saúde mental impulsionou uma colaboração entre psiquiatras, psicoterapeutas e psicólogos. Isso resultou no desenvolvimento de novas técnicas terapêuticas, como a Terapia Cognitiva-Comportamental (TCC) e a Terapia Interpessoal, melhorando o bem-estar dos pacientes [Gomes de Matos et al., 2005].

Simultaneamente, a evolução tecnológica, tem transformado a medicina e a saúde de modo revolucionário. Com o avanço das tecnologias, especialmente na esfera computacional, o uso de algoritmos de aprendizado de máquina (AM) está se tornando cada vez mais essencial em diversas áreas, incluindo medicina e saúde [Shailaja e Seetharamulu, 2018]. Esses algoritmos são fundamentais para análises preditivas e tomadas de decisão. Para utilizá-los efetivamente, é necessário contar com bases de dados para auxiliar na predição e na tomada de decisão nessa área. Neste contexto de saúde mental, é comum bases de dados contendo imagens e/ou textos, além das bases tradicionais, conhecidas como tabulares, compostas por instâncias (observações) e atributos (características).

No entanto, de acordo com Vishwarupe e Joshi [2022], além da eficácia dos algoritmos, é crucial considerar a qualidade dos dados e a interpretabilidade dos resultados, especialmente na área da saúde. Muitas bases de dados, quando tabulares, contêm dados nominais categóricos que precisam ser transformados em valores numéricos por meio de algum método de codificação, já que alguns algoritmos de AM não aceitam dados nominais. A codificação adequada dos dados é essencial para garantir a representação precisa das informações, reduzindo vieses e erros.

Além disso, a interpretabilidade dos resultados é fundamental para compreender como as decisões são tomadas e confiar nas recomendações do sistema [Ribeiro et al., 2016], especialmente na área da saúde, onde as consequências das decisões podem ser significativas. Portanto, assegurar a interpretabilidade dos resultados é essencial para garantir a confiabilidade das análises e a confiança na aplicação das soluções de AM na saúde.

Diante do exposto, o objetivo deste trabalho é investigar seis diferentes tipos de codificação para atributos categóricos nominais em três diferentes bases de dados relacionadas a transtornos mentais (depressão, transtorno do pânico e autismo). As codificações *One Hot* (OHE) e *Dummy* oferecem representações diretas das categorias, enquanto *Frequency* e *Count* fornecem a contagem das opções de atributos. Por outro lado, *Ordinal* preserva a ordem e a relação entre as categorias, enquanto *Hashing* reduz a dimensionali-

dade. O intuito é analisar se existe uma codificação que apresenta resultados mais eficazes e se essa codificação auxilia na interpretabilidade do modelo.

O desafio científico na codificação e interpretabilidade é evidente quando ocorre a dificuldade de interpretar os resultados dos algoritmos de AM, sem afetar o desempenho; o que usualmente é priorizado. Desse modo, temos como propósito investigar distintos métodos de codificação, buscando conciliar desempenho e interpretabilidade, proporcionando soluções mais transparentes.

Nosso trabalho visa contribuir com a análise de seis codificações usuais da literatura, e verificar se são interpretáveis e eficientes para aplicação na área da saúde. A implementação de um sistema de codificação interpretável pode melhorar significativamente a eficácia do atendimento de saúde e impulsionar a pesquisa médica.

2. Referencial Teórico

2.1. Transtornos Mentais

Os transtornos mentais apresentam uma ampla variedade de manifestações, afetando os aspectos psicológicos de um indivíduo. Essas condições representam alterações na mente do indivíduo, resultando em impactos significativos no convívio do dia-a-dia. Embora não tenham uma causa específica, podem ser ocasionadas por uma combinação de fatores biopsicossociais [Amorim Cruz et al., 2021].

Problemas de saúde mental altamente frequentes na população são denominados como Transtornos Mentais Comuns (TMC). Entre eles estão, a depressão, o transtorno do pânico e transtornos de desenvolvimento, incluindo o autismo [Ribeiro e Lemos, 2020].

A depressão, também conhecida como Transtorno Depressivo, é um transtorno comum, mas sério, que interfere na vida diária. É causada por uma combinação de fatores biológicos [Thase et al., 2002], ambientais [Fragelli e Fragelli, 2021] e psicológicos [Guedes et al., 2022]. A depressão caracteriza-se pela melancolia, baixa da autoestima e desinteresse. Os quadros mais graves podem levar ao suicídio.

Um diagnóstico preciso é crucial para garantir um tratamento adequado desde o início, sendo que a identificação precoce desempenha um papel fundamental para iniciar o tratamento da patologia no estágio inicial [Rufino et al., 2018]. Esses dois elementos são essenciais para melhorar a qualidade de vida do paciente.

O Transtorno do Pânico (TP) é uma condição neuropsiquiátrica caracterizada por ataques de pânico espontâneos e recorrentes, acompanhados por sintomas como dispneia, palpitações, medo intenso, evitação fóbica e tremores. Esses episódios têm início abrupto, gerando preocupações persistentes e alterações comportamentais em relação à possibilidade de futuros episódios de ansiedade [Vasconcelos Filho et al., 2023].

O TP é afetado por fatores hereditários e do ambiente, sendo suas características associadas a condições médicas concomitantes, como depressão e ansiedade. A presença de Transtorno Depressivo Maior (TDM) junto ao TP agrava significativamente o bem-estar do paciente, sendo necessária a compreensão e o tratamento desses transtornos para promover a qualidade de vida e da saúde mental [Vasconcelos Filho et al., 2023].

O transtorno do espectro do autismo (TEA) é uma condição de neurodesenvolvimento caracterizada por dificuldades na comunicação social, interação social e padrões

de comportamento restritos e repetitivos. Embora os sintomas apareçam precocemente, o diagnóstico muitas vezes ocorre apenas alguns anos depois [Fuentes e Bakare, 2012].

O TEA tem um grande impacto emocional nas famílias e exige tratamento especializado, mas a maioria das pessoas, especialmente em países em desenvolvimento, não recebe tratamento adequado. Embora os tratamentos tenham melhorado a qualidade de vida, o TEA ainda possui cura [Fuentes e Bakare, 2012].

2.2. Técnicas de Codificação de Atributos Categóricos

O pré-processamento de dados é crucial para melhorar a qualidade dos dados brutos antes de serem utilizados em análises e modelagem dos algoritmos de AM. Uma etapa importante é lidar com atributos categóricos, já que muitos algoritmos de aprendizado de máquina requerem dados numéricos. Existem várias técnicas de transformação de variáveis categóricas, incluindo seis métodos comumente usados na literatura.

No contexto de variáveis categóricas, é importante considerar um conceito que se refere ao número de valores únicos que um atributo pode assumir, chamado Cardinalidade. Segundo Moeyersoms e Martens [2015], a cardinalidade é considerada alta quando os atributos contêm mais de 100 valores únicos. No entanto, valores de cardinalidade bem menores que 100 normalmente já afetam o desempenho dos algoritmos.

Os principais métodos de codificações são os seguintes:

1) *One Hot Encoding* (OHE): Nesta codificação, cada opção de resposta de uma característica categórica é representada por uma nova variável, onde '0' representa a ausência e '1' representa a presença dessa categoria. O resultado retornado desse codificador é uma matriz numérica binária, com a quantidade de colunas igual ao número de variações de valores no atributo. Ou seja, para bases de dados que possuem atributos de alta cardinalidade, retorna um vetor de características de alta dimensão [Dahouda e Joe, 2021]; 2) *Dummy Encoding*: Esta é uma versão aperfeiçoada do OHE. Este método, também retorna uma matriz binária, onde '0' e '1' são utilizados como valores simbólicos de ausência e presença, respectivamente. No entanto, ao contrário do OHE, este método utiliza $N - 1$ colunas para representar um atributo com N opções de resposta [Dahouda e Joe, 2021]; 3) *Frequency Encoding*: Esta codificação atribui valores numéricos às opções de resposta de um atributo categórico com base em suas frequências em relação ao total de ocorrências na coluna dos dados. As opções mais frequentes recebem valores mais altos, e as menos frequentes, valores mais baixos. Entretanto, segundo Roy [2019], é possível perder informações importantes se existirem duas categorias distintas com a mesma taxa de ocorrência; 4) *Count Encoding*: A codificação por contagem substitui os nomes das opções de resposta em um atributo categórico pelo número de ocorrências de cada categoria. No entanto, assim como no *Frequency Encoding*, pode ocorrer perda de informações se duas categorias distintas tiverem a mesma taxa de ocorrência; 5) *Ordinal Encoding*: Esta codificação converte as opções de um atributo categórico em uma única coluna composta de números inteiros a partir do conhecimento do número de categorias existentes [Potdar e Pardawala, 2017]. Pode-se opcionalmente fornecer um dicionário de mapeamento para definir uma ordem específica. Na ausência de mapeamento, os valores inteiros são atribuídos aleatoriamente, resultando em uma coluna que varia de 1 a N, representando o número de categorias; 6) *Hashing Encoding*: Essa técnica utiliza funções de *Hashing* para representar dados categóricos como valores numéricos, mapeando con-

juntos complexos em uma estrutura menor de tamanho fixo . No entanto, as funções *Hash* são unidirecionais, impedindo a reversão para os valores originais. A colisão, quando chaves diferentes geram o mesmo valor de *Hash*, é um problema que pode comprometer a interpretabilidade dos dados [Kuhn e Johnson, 2019].

2.3. Interpretabilidade

Segundo Adadi e Berrada [2018], sistemas interpretáveis são aqueles cujas operações podem ser entendidas por humanos, sendo fundamental na área da saúde devido à sensibilidade do trabalho envolvido. A interpretabilidade dos modelos de AM é essencial já que muitos deles são complexos e faltam transparência. Esta falta de explicação das decisões pode representar uma lacuna crítica nos processos de tomada de decisão.

As explicações SHAP *(SHapley Additive exPlanations)* são uma abordagem popular para atribuir importância às características em Inteligência Artificial Explicável [Van den Broeck e Lykov, 2022].O SHAP visa explicar as predições de um modelo, calculando a contribuição de cada característica para essas predições usando os Valores de *Shapley* da teoria dos jogos. Ele determina a média da contribuição marginal de uma característica em todas as possíveis combinações, representadas por vetores de coalizão. Ao mapear esses vetores no espaço das características, o SHAP destaca as contribuições relevantes para as predições do modelo[Nguyen e Cao, 2021].

3. Trabalhos Relacionados

Udilă [2023] compara cinco métodos de codificação para dados categóricos, mostrando que o *One Hot* é mais preciso, mas mais lento, enquanto o *catboost* oferece um equilíbrio entre precisão e tempo de treinamento. A combinação de métodos não melhorou significativamente os resultados, e investir em codificação personalizada foi mais eficaz do que soluções automatizadas, como o AutoML.

Johnson e Khoshgoftaar [2022] abordam a codificação de códigos de procedimentos médicos para detecção de fraudes. Comparando técnicas tradicionais de codificação com uma abordagem agregada, constatou-se que o *Hcpcs2Vec* e as técnicas incorporadas de *CatBoost* e *LightGBM* superaram o método OHE. O *Hcpcs2Vec* obteve o melhor desempenho geral, destacando-se pela significativa importância da variável de código de procedimento.

Vishwarupe e Joshi [2022] destacam a importância dos frameworks de XAI (Inteligência Artificial Explicável), como *SHAP*, *LIME* e *ELI5*, ao interpretar modelos de aprendizado de máquina em conjunto com fatores clínicos, fornecendo benefícios sobre modelos tradicionais do tipo caixa preta. Segundo os autores, a análise de tendências por meio de *Partial Dependence Plots* possibilita diagnósticos clínicos mais precisos.

Reiter [2020] investigou métodos interpretáveis para classificação de esquizofrenia. Uma rede neural profunda é desenvolvida para classificação. Diferentes iterações do *DeepSHAP* são exploradas como métodos interpretáveis, destacando o *DeepSHAP* Padrão e o *DeepSHAP* Consolidado por Rótulo. Os resultados indicam a eficácia do *DeepSHAP* para interpretabilidade na classificação da esquizofrenia, com potencial para uma abordagem centrada no paciente.

A partir dos trabalhos supracitados, é possível perceber que as pesquisas geralmente abordam a codificação de dados categóricos e a interpretabilidade de forma sepa-

rada, criando a necessidade de estudos que avaliam estes dois aspectos conjuntamente. Assim, esta pesquisa busca combinar ambos os aspectos para entender como diferentes métodos de codificação podem afetar a interpretabilidade de modelos de aprendizado de máquina.

4. Materiais e Métodos

4.1. Descrição da Base de Dados

Neste trabalho foram utilizadas três bases de dados públicas relacionadas a saúde mental, com variações em relação ao tamanho das instâncias e atributos. Durante a análise, verificamos a cardinalidade de cada atributo categórico nominal não ordinal presente nos conjuntos de dados. As bases escolhidas foram: 1) *Autism Screening Adult*[1] , 2) *Depression Detection*[2], e 3) *Panic Disorder Detection*[3].

A Tabela 1 apresenta a descrição completa das bases de dados, incluindo seus respectivos nomes, quantidade de instâncias, número de dados ausentes, tipo de atributos de entrada, quantidade total de atributos, e cardinalidade máxima e mínima dos atributos categóricos não ordinais.

Tabela 1. Descrição das bases de dados quanto ao número de instâncias e tipos de atributos

			Atributos de Entrada								
Base de Dados	**Inst.**	**NA%**	**Cont.**	**Disc.**	**C. Ord.**	**C. N. Ord.**	**Bin.**	**Classes**	**Total**	**C. Min.**	**C. Max.**
Link											
Autism Screening Adult	704	1,29	1	1	-	4	14	2	21	5	67
Depression Detection Data Set	1000	0,14	-	1	3	2	18	2	25	3	-
Panic Disorder Detection Dataset	120000	-	-	2	4	6	4	2	17	3	5

Para esta tabela considere o seguinte: *Inst.*: Instâncias; *NA%*: Porcentagem de Dados Ausentes; *Cont.*: Contínuo; *Disc.*: Discreto; *C. Ord.*: Atributos Categóricos Ordinais; *C. N. Ord.*: Atributos Categóricos Não Ordinais; *Bin.*: Atributos Binários; *Classes*: Quantidade de Classes; *Total*: Quantidade Total de Atributos; *C. Min.*: Cardinalidade Mínima; e *C. Max.*: Cardinalidade Máxima.

4.2. Pré-processamento

Para avaliar as diferentes técnicas de codificação dos atributos categóricos, criamos um *pipeline* na linguagem *Python*, empregando várias bibliotecas amplamente utilizadas em projetos de ciência de dados, como *Scikit-learn*[4], *Pandas*[5] e *Category Encoders*[6].

Quanto às etapas de pré-processamento realizadas, efetuamos o seguinte:

[1]Disponível em: https://archive.ics.uci.edu/dataset/426/autism+screening+adult

[2]Disponível em: https://www.kaggle.com/datasets/andrewmvd/autism-screening-on-adults

[3]Disponível em: https://www.kaggle.com/datasets/muhammadshahidazeem/panic-disorder-detection-dataset

[4]A Documentação oficial pode ser encontrada em: https://scikit-learn.org/0.21/documentation.html *Scikit-learn*

[5]A Documentação oficial pode ser encontrada em: https://pandas.pydata.org/docs/reference/index.html *Pandas*

[6]A Documentação oficial pode ser encontrada em: https://contrib.scikit-learn.org/category_encoders *Category Encoders*

1. *Remoção de Atributos*: na base de dados relacionada à depressão, o atributo *'Time'* continha valores inconsistentes, enquanto na base de dados relacionada ao autismo, o atributo *'Participant ID'* continha valores irrelevantes para a classificação do modelo.
2. *Remoção de instâncias com dados ausentes*: uma vez que as bases de dados escolhidas apresentaram uma baixa taxa de dados ausentes, menos de 1%, optamos por remover estas instâncias.
3. *Remoção de instâncias duplicadas*: ao lidar com bases de dados para algoritmos de AM, é comum encontrar instâncias duplicadas em algumas delas. Para diminuir o risco de vazamento de dados, foram removidas as instâncias duplicadas.
4. *Codificação de atributos categóricos nominais ordinais*: a função *map* foi aplicada com o objetivo de converter atributos categóricos nominais ordinais em atributos numéricos. Para realizar essa transformação, estabelecemos um mapeamento entre os valores nominais e seus respectivos valores numéricos correspondentes. Em seguida, aplicamos esse mapeamento utilizando a função *map*.
5. *Codificação de atributos categóricos nominais não ordinais*: Tendo em vista que este estudo visa comparar diferentes técnicas de codificação de atributos categóricos nominais, todas as bases de dados selecionadas continham atributos que requeriam esse tipo de transformação. Para realizar essa etapa de codificação, foram utilizados vários codificadores específicos para lidar com esses atributos. A Tabela 2 mostra os codificadores e as bibliotecas em *Python* empregadas para implementá-los.

Tabela 2. Bibliotecas e métodos utilizados para implementar as codificações

Codificação	Biblioteca	Métodos
One Hot Encoding	Pandas	get_dummies
Dummy Encoding	Pandas	get_dummies
Frequency Encoding	Pandas	-
Count Encoding	Category_Encoders	CountEncoder
Ordinal Encoding	Category_Encoders	OrdinalEncoder
Hashing Encoding	Category_Encoders	HashEncoder

Durante o processo de codificação, foram utilizadas as configurações padrões dos métodos de codificação exceto para a codificação *Dummy Enconding*. Para essa técnica, foi alterado o padrão *drop_first* que remove o primeiro nível de cada coluna.

4.3. Algoritmos de Aprendizagem e métricas de avaliação de qualidade

Neste trabalho, foram utilizados quatro algoritmos de AM para comparar e determinar o melhor método de codificação dentre os apresentados. Os algoritmos utilizados foram Árvore de Decisão (*DecisionTreeClassifier*), *Random Forest* (*RandomForestClassifier*), *XGBoost* (*XGBClassifier*) e Rede Neural (*MLPClassifier*). A implementação desses algoritmos e a obtenção dos resultados para comparação foram realizadas em *Scikit-learn* na linguagem *Python*. Em todos os algoritmos de AM, foram utilizados os valores *default*.

Para avaliar a capacidade de generalização do modelo, empregamos a técnica de validação cruzada de 10 dobras e reservamos 20% das instâncias para teste.

Utilizamos a métrica *F1-score* para avaliar o desempenho do modelo, que é uma média harmônica da precisão e sensibilidade ($F1 = \frac{2 \times Precisão \times Sensibilidade}{Precisão + Sensibilidade}$).

Além disso, neste estudo utilizamos o *Teste-t* para comparar as diferentes codificações em pares nas bases de dados. O nível de significância utilizado em todos os testes foi estabelecido em 0,05.

4.4. Avaliação da interpretabilidade dos modelos obtidos

A interpretabilidade de um modelo é uma característica crucial, especialmente em contextos como a saúde, onde a transparência e a compreensão das decisões tomadas são fundamentais. Utilizamos a técnica SHAP com a intenção de contribuir para a interpretabilidade do modelo, fornecendo percepções sobre a importância das variáveis no processo de tomada de decisão do modelo, permitindo uma compreensão mais detalhada de como os atributos afetam as previsões finais.

5. Resultados e Discussões

A Tabela 3 apresenta os resultados dos testes com seis métodos de codificação de atributos categóricos e os algoritmos de AM utilizados (Árvore de Decisão, *Random Forest*, *XGBoost* e Rede Neural). A escolha adequada da codificação é fundamental para obter resultados precisos, considerando a importância dos atributos nominais e a necessidade de dados numéricos para os algoritmos de aprendizado de máquina.

Tabela 3. Resultados das métricas dos algoritmos de aprendizado de máquina para cada codificação

	***Árvore de Decisão* (F1-Score Médio /Desvio Padrão)**					
	Codificação					
Base de Dados	***Count***	***Dummy***	***Frequency***	***Hashing***	***One Hot***	***Ordinal***
Autism Screening Adult	**0,864/0,03**	**0,860/0,06**	**0,864/0,03**	0,848/0,04	0,847/0,04	0,842/0,02
Depression Detection Data Set	0,865/0,02	**0,869/0,03**	0,865/0,02	**0,869/0,02**	0,865/0,02	0,863/0,02
Panic Disorder Detection Dataset	0,644/0,01	**0,678/0,02**	0,644/0,01	0,673/0,01	0,641/0,01	0,643/0,01

	***Random Forest* (F1-Score Médio /Desvio Padrão)**					
	Codificação					
Base de Dados	***Count***	***Dummy***	***Frequency***	***Hashing***	***One Hot***	***Ordinal***
Autism Screening Adult	0,908/0,06	**0,916/0,06**	0,908/0,06	**0,915/0,06**	**0,910/0,06**	**0,913/0,06**
Depression Detection Data Set	**0,919/0,03**	0,916/0,03	**0,919/0,03**	**0,918/0,03**	0,913/0,03	0,913/0,03
Panic Disorder Detection Dataset	0,517/0,01	0,585/0,01	0,517/0,01	**0,615/0,01**	0,523/0,01	0,514/0,01

	***XGBoost* (F1-Score Médio /Desvio Padrão)**					
	Codificação					
Base de Dados	***Count***	***Dummy***	***Frequency***	***Hashing***	***One Hot***	***Ordinal***
Autism Screening Adult	**0,935/0,03**	0,919/0,04	**0,935/0,03**	0,926/0,03	0,921/0,04	0,913/0,05
Depression Detection Data Set	0,906/0,03	0,913/0,03	0,906/0,03	0,912/0,03	**0,916/0,03**	0,903/0,03
Panic Disorder Detection Dataset	0,570/0,01	0,644/0,01	0,570/0,01	**0,656/0,01**	0,559/0,01	0,571/0,01

	***Rede Neural* (F1-Score Médio /Desvio Padrão)**					
	Codificação					
Base de Dados	***Count***	***Dummy***	***Frequency***	***Hashing***	***One Hot***	***Ordinal***
Autism Screening Adult	0,896/0,04	0,894/0,04	**0,930/0,04**	0,923/0,04	0,913/0,05	0,917/0,03
Depression Detection Data Set	0,894/0,04	0,909/0,03	0,912/0,02	**0,916/0,03**	0,903/0,02	0,908/0,02
Panic Disorder Detection Dataset	0,489/0,01	**0,623/0,04**	0,489/0,01	0,598/0,04	0,534/0,02	0,529/0,03

*Em negrito, estão as melhores codificações para cada algoritmo de AM, levando em consideração o Teste-t. Todos os testes realizados com 95% de confiança.

A codificação *Hashing* mostrou-se a mais eficaz, mostrando-se presente no algoritmo de classificação *Random Forest* para todas as bases de dados, e também adotada pelos demais algoritmos em pelo menos uma das bases. Outra codificação que se destacou foi a *Dummy* destacando-se em todas as bases no algoritmo Árvore de Decisão, apesar

de não se apresentar como uma das melhores codificações apenas no algoritmo *XGBoost*. Olhando individualmente para cada algoritmo temos:

- Para a Árvore de Decisão, as melhores codificações variam de acordo com a base de dados, mas a codificação *Dummy* ganha destaque.
- Para o *Random Forest*, todas as codificações se destacam em alguma das bases. Entretanto, a codificação *Hashing* tende a fornecer os melhores resultados em todas as bases de dados.
- Para o *XGBoost*, as codificações *Count*, *Frequency*, *Hashing* e *One Hot* são consistentemente boas escolhas, variando de acordo com as bases de dados.
- Para a Rede Neural, as codificações *Dummy*, *Frequency* e *Hashing* são as melhores opções, cada uma para uma base de dados.

A codificação *Ordinal* apresentou o pior desempenho, sendo identificada como a melhor opção apenas em um dos testes realizados.

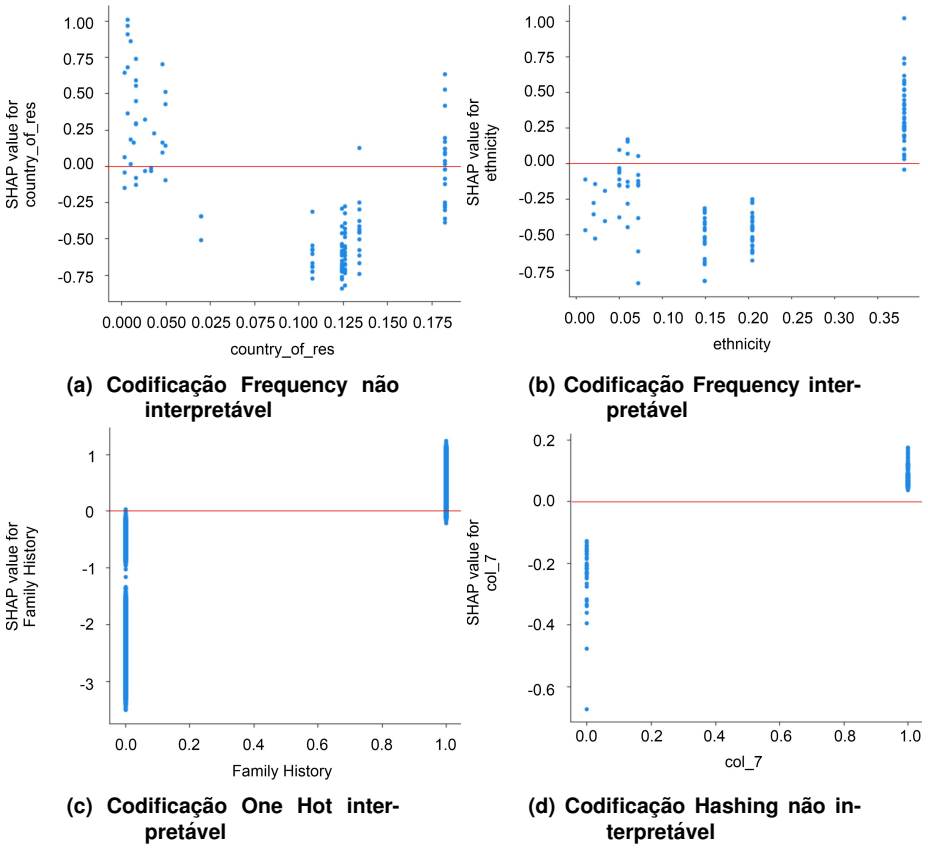
No entanto, em contextos relacionados à área da saúde, é crucial não apenas considerar a codificação que foi mais frequente ou aquela que apresentou a maior métrica de desempenho, já que a interpretabilidade dos resultados é de suma importância.

Cada uma dessas codificações possui suas vantagens e desvantagens e devem ser levadas em consideração ao serem selecionadas para aplicações reais:

- As codificações *One Hot* e *Dummy* mantém as representações explícitas das categorias, facilitando a interpretabilidade. No entanto, quando aplicadas a conjuntos de dados de alta dimensionalidade e cardinalidade, podem gerar aumento na complexidade computacional, no tempo de execução e no consumo de memória.
- As codificações *Frequency* e *Count* garantem a contagem das ocorrências de cada opção de atributo, sendo útil quando a frequência de ocorrência é relevante. Por outro lado, podem atrapalhar na interpretabilidade quando um atributo possuir duas opções de resposta ou mais com a mesma frequência.
- A codificação *Ordinal* preserva a ordem e hierarquia entre as categorias, sendo útil para atributos que possuem relação natural de ordem, além disso, facilita a interpretabilidade dos resultados. Contudo, pode não retornar bons resultados para algoritmos que possuem dificuldade para lidar com a informação de ordem ou que requerem independência entre as categorias.
- A codificação *Hashing* reduz a dimensionalidade, beneficiando conjuntos de dados com muitos atributos e atributos de alta cardinalidade. Apesar disso, essa codificação impossibilita a interpretabilidade por ser unidirecional (não é possível retornar aos valores originais) e possui a possibilidade de colisões, onde diferentes categorias são mapeadas para o mesmo valor *Hashing*.

A Figura 1 mostra como algumas dessas codificações podem afetar a interpretabilidade dos modelos. Esses gráficos foram gerados utilizando o SHAP com o gráfico *Dependency Plot*. Essa é uma ferramenta visual utilizada na análise de modelos de AM para compreender como um atributo de entrada específico influencia as saídas do modelo. Ele mostra a relação entre uma variável de entrada e a saída do modelo. Cada ponto representa uma instância do conjunto de dados, com seu impacto medido pelo valor SHAP. Consideramos nesse estudo que valores SHAP positivos indicam contribuição para o transtorno mental estudado no gráfico, enquanto negativos indicam impacto negativo.

Na codificação *Frequency*, o atributo *'country_of_res'* (Figura 1a), com valor de 0,125, pode influenciar a classificação do indivíduo negativamente, mas não especifica como cada valor afeta. Esse valor pode se referir a várias opções de resposta, como *'New Zealand'* e *'United Kingdom'*. Já o atributo *'ethnicity'* (Figura 1b), com valor de 0,15, também pode influenciar negativamente, sendo possível identificar o impacto desse valor, como *'Middle Eastern'*, já que nesse atributo os valores de frequência são únicos para cada opção de resposta.



(a) Codificação Frequency não interpretável

(b) Codificação Frequency interpretável

(c) Codificação One Hot interpretável

(d) Codificação Hashing não interpretável

Figura 1. Codificações interpretáveis e não interpretáveis

Na codificação *One Hot* no conjunto de dados *Depression Detection*, o atributo *'Family History'* (Figura 1c) influencia a classificação do indivíduo quanto à presença ou ausência do transtorno depressivo. Os valores 0 e 1 indicam a ausência e a presença de histórico familiar, respectivamente, tornando o modelo mais interpretável. Assim, conclui-se que indivíduos com histórico familiar tendem a ser diagnosticados com o transtorno depressivo, enquanto aqueles sem histórico familiar tendem a ser diagnosticados como não possuindo o transtorno.

Na codificação *Hashing* no conjunto de dados *Panic Disorder Detection*, o atributo *'col_7'* (Figura 1d) pode influenciar positiva ou negativamente a classificação do indivíduo quanto à presença ou ausência do transtorno do pânico. No entanto, devido à natureza unidirecional desta codificação, não é possível identificar o significado específico desse atributo, limitando a interpretabilidade do modelo.

6. Considerações Finais

Este estudo investigou seis diferentes tipos de codificações para atributos categóricos nominais não ordinais, *Count*, *Dummy*, *Frequency*, *Hashing*, *One Hot* e *Ordinal*, em três bases de dados relacionadas a transtornos mentais, avaliando seu desempenho em quatro algoritmos de AM e utilizando o algoritmo SHAP para interpretabilidade dos modelos. Observou-se que a codificação *Hashing* foi a mais eficaz em dois dos algoritmos; porém, não é interpretável. Os resultados revelaram que a escolha adequada da codificação é fundamental para obter resultados melhores e interpretáveis.

Portanto, ao selecionar uma codificação para atributos nominais não ordinais, é essencial atentar sobre a complexidade computacional, a interpretabilidade dos resultados e as características específicas do conjunto de dados.

Entretanto, nossa pesquisa tem algumas limitações. Primeiro, ela pode não ser adequada para cenários que focam apenas na eficácia do modelo sem se preocupar com a interpretabilidade, podendo ser utilizadas outras abordagens mais simples. Além disso, a eficácia das codificações testadas pode não ser generalizável, pois a pesquisa utilizou um número limitado de bases de dados.

Para estender este estudo, várias direções podem ser consideradas, tais como utilizar variações nos parâmetros das codificações e dos algoritmos de aprendizado para otimizar o desempenho, explorar outros métodos de codificação incluindo aqueles que são supervisionados, além de avaliar o desempenho das codificações em conjuntos de dados adicionais e em diferentes contextos clínicos para validar e generalizar os resultados encontrados. Esta pesquisa apontou também, uma oportunidade de converter o conhecimento adquirido em um produto que possa ser amplamente utilizável. A criação de uma biblioteca em *Python* poderá permitir que profissionais de diversas áreas, com interesse na interpretabilidade de atributos nominais utilizem essas técnicas de forma automatizada.

Agradecimentos

Os autores agradecem ao Conselho Nacional de Desenvolvimento Científico e Tecnológico do Brasil (CNPq, Código: 311573/2022-3), à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), à Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG, Código: APQ-03076-18), ao Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG) e à Pontifícia Universidade Católica de Minas Gerais (Código: FIP-2023/29184-1S).

Referências

A. Adadi e e. Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 2018.

E. M. Amorim Cruz, A. B. Callou Sampaio Neves, A. Gomes da Rocha, R. Arrais Macário, J. W. Araújo dos Anjos, T. A. Alencar Lima, A. M. Duarte de Melo, e P. Amorim Cruz Nascimento. Assistência humanizada a pessoa com transtornos mentais. *Id on Line. Revista de Psicologia*, 2021.

C. U. Correll e e. Galling. Comparison of early intervention services vs treatment as usual for early-phase psychosis: a systematic review, meta-analysis, and meta-regression. *JAMA psychiatry*, 2018.

M. K. Dahouda e I. Joe. A deep-learned embedding technique for categorical features encoding. *IEEE Access*, 2021.

T. B. O. Fragelli e R. R. Fragelli. Por que estudantes universitários apresentam estresse, ansiedade e depressão? uma rapid review de estudos longitudinais. *Revista docência do ensino superior*, 11:1–21, 2021.

J. Fuentes e e. Bakare. Autism spectrum disorders. *IACAPAP e-textbook of child and adolescent mental health. Geneva*, 1:27, 2012.

L. Gamm, S. Stone, e S. Pittman. Mental health and mental disorders—a rural challenge: A literature review. *Rural healthy people*, 2(1):97–114, 2010.

E. Gomes de Matos, T. M. Gomes de Matos, e G. M. Gomes de Matos. A importância e as limitações do uso do DSM-IV na prática clínica. *Revista de Psiquiatria do Rio Grande do Sul*, 27:312–318, 2005.

D. R. Guedes, E. dos Santos Bispo, e L. M. A. F. Nobre. Depressão, o mal do século: Prevalência de depressão e os fatores associados em mulheres-uma revisão de literatura. *Recisatec-Revista Científica Saúde e Tecnologia. ISSN 2763-8405*, 2(2):e2277–e2277, 2022.

S. Hyman e e. Chisholm. Mental disorders. *Disease control priorities related to mental, neurological, developmental and substance abuse disorders*, 2006.

J. M. Johnson e T. M. Khoshgoftaar. Encoding high-dimensional procedure codes for healthcare fraud detection. *SN Computer Science*, 2022.

M. Kuhn e K. Johnson. *Feature Engineering and Selection: A Practical Approach for Predictive Models*. CRC Press, 2019.

J. Moeyersoms e D. Martens. Including high-cardinality attributes in predictive models: A case study in churn prediction in the energy sector. *Decision Support Systems*, 72:72–81, 2015.

H. T. T. Nguyen e e. Cao. Evaluation of explainable artificial intelligence: Shap, lime, and cam. In *Proceedings of the FPT AI Conference*, pages 1–6, 2021.

OPAS. Oms destaca necessidade urgente de transformar saúde mental e atenção, jul 2022. URL `https://www.paho.org/pt/noticias/17-6-2022-oms-destaca-necessidade-urgente-transformar-\saude-mental-e-atencao\`.

K. Potdar e T. Pardawala. A comparative study of categorical variable encoding techniques for neural network classifiers. *International Journal of Computer Applications*, 175, 2017.

J. Reiter. Developing an interpretable schizophrenia deep learning classifier on fMRI and smri using a patient-centered DeepSHAP. In *NeurIPS 2018*, 2020.

C. F. Ribeiro e e. Lemos. Prevalence of and factors associated with depression and anxiety in brazilian medical students. *Revista Brasileira de Educação Médica*, 2020.

M. T. Ribeiro, S. Singh, e C. Guestrin. "why should i trust you?"explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

B. Roy. All about categorical variable encoding, jun 2019.

S. Rufino, R. S. Leite, L. Freschi, V. K. Venturelli, E. d. Oliveira, e D. A. M. Mastrorocco Filho. Aspectos gerais, sintomas e diagnóstico da depressão. *Revista Saúde em foco*, 10(1):837–843, 2018.

K. Shailaja e e. Seetharamulu. Machine learning in healthcare: A review. 2018.

M. E. Thase, R. Jindal, e R. H. Howland. Biological aspects of depression. 2002.

A. Udilă. Encoding methods for categorical data: A comparative analysis for linear models, decision trees, and support vector machines. 2023.

G. Van den Broeck e e. Lykov. On the tractability of shap explanations. *Journal of Artificial Intelligence Research*, 74:851–886, 2022.

J. C. Vasconcelos Filho, J. O. Rocha, H. N. Curto, M. H. D. Barbosa, e T. S. Miranda. *Aspectos Clínicos e Diagnósticos em Saúde Mental*. 2023. Ebook Acadêmico.

V. Vishwarupe e e. Joshi. Explainable ai and interpretable machine learning: A case study in perspective. *Procedia Computer Science*, 2022.