Descrevendo o perfil de pessoas com depressão nas regiões Sudeste e Centro-Oeste do Brasil por meio de Mineração de Dados

Glória Eleonor F. Alves[1], **Pedro Grojpen Couto**[1], **Luis Enrique Zárate**[1]

[1]Instituto de Ciências Exatas e Informática –
Pontifícia Universidade Católica de Minas Gerais (PUC Minas)

gloriaeleonorfa@gmail.com, pedrogrojpen@gmail.com, zarate@pucminas.br

Abstract. *This work aims to analyze relevant factors for the identification of individuals with depression in regions of Brazil with data extracted from the 2019 National Health Survey and application of the Decision Tree algorithm. The conclusion reinforces that emotional factors are efficient identifiers, while socioeconomic factors need to be investigated further.*

Resumo. *Esse trabalho busca analisar fatores relevantes para a identificação de indivíduos com depressão em regiões do Brasil com dados extraídos da Pesquisa Nacional de Saúde de 2019 e aplicação do algoritmo Árvore de Decisão. A conclusão reforça que fatores emocionais são identificadores eficientes, enquanto fatores socioecônomicos precisam ser investigados mais a fundo.*

1. Introdução

No ano de 2024 mais de 470 mil afastamentos por transtornos mentais foram registrados no Brasil pelo Instituto Nacional do Seguro Social (INSS), o que representa um aumento de 68% em comparação ao ano anterior. Dentre estes transtornos destaca-se a depressão, um distúrbio crônico que afeta mais de 11 milhões de brasileiros, segundo levantamento realizado pela Organização Mundial da Saúde (OMS) em 2017.

O aumento de afastamentos evidencia um problema crescente na sociedade brasileira. Uma pesquisa conduzida por [Lipp and Lipp 2020] constatou que a brusca mudança de estilo de vida infligida pela pandemia de COVID-19 agravou o número de indivíduos que relatam vivenciar sentimentos depressivos, incerteza e ansiedade. [Silva et al. 2023] destaca que a pandemia teve efeitos adversos não relacionados diretamente à doença, como perdas socioeconômicas, associados à piora da saúde mental da população. Por outro lado, [Meleiro et al. 2023] estima que a depressão é subdiagnosticada e subtratada no Brasil por uma série de motivos, como estigma social e despreparo de médicos da rede de atenção primária à saúde. O estudo também relata ter encontrado poucos dados relacionados à triagem, e nenhum relacionado à adesão e controle da doença no país.

A depressão é uma doença multifatorial recorrente. O diagnóstico pode estar relacionado a fatores psicológicos, ambientais, genéticos etc. Portanto, trata-se de uma condição complexa e diversa que requer acompanhamento e tratamento personalizado. Considerando este aspecto e as deficiências da jornada do paciente, infere-se que o sistema de saúde brasileiro não está preparado para lidar com o número crescente de casos.

A proposta desse trabalho é caracterizar o perfil do indivíduo que possui depressão nas regiões Sudeste e Centro-Oeste do Brasil por meio da aplicação de técnicas de descoberta de conhecimento (KDD) por meio de Árvore de Decisão. A base de dados empregada é resultado da Pesquisa Nacional de Saúde (PNS) realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE) no ano de 2019. Espera-se criar um panorama detalhado da doença com o intuito de entender sua manifestação no país a nível regional e prover informações para a elaboração de políticas públicas mais eficazes.

2. Trabalhos relacionados

O DSM-5 (2013) é a principal referência para diagnóstico psiquiátrico, enquanto as Diretrizes para o Tratamento da Depressão [Fleck et al. 2003] detalham métodos utilizados no Brasil, destacando desafios pelo subdiagnóstico.

Apenas 18,7% dos indivíduos com depressão possuem um diagnóstico formal, refletindo uma lacuna significativa no reconhecimento da doença [Meleiro et al. 2023]. No contexto da Atenção Primária à Saúde, [Aguiar et al. 2022] identificaram um alto número de tentativas de suicídio em indivíduos não diagnosticados com depressão.

Aprendizado de Máquina (AM) têm sido aplicada para identificar o risco de depressão em subgrupos específicos, como mulheres em estado puerperal. [Shin et al. 2020] demonstrou a eficácia de modelos preditivos na identificação de fatores de risco para depressão pós-parto, auxiliando em intervenções mais direcionadas.

[Lee and Ham 2022] conduziu uma revisão para avaliar o progresso do uso de algoritmos como Árvore de Decisão, Floresta Aleatória, entre outros, para realizar o diagnóstico precoce de depressão. O estudo concluiu que o aprendizado de máquina pode ser uma maneira eficaz e não-invasiva de detectar depressão, mas que é necessário realizar uma análise meticulosa da taxa de precisão dos modelos e que o uso de *big data*, bem como a combinação de diferentes métodos de aprendizado, poderia contribuir para melhorar os resultados e o diagnóstico precoce em sistemas regionais públicos de diagnóstico clínico.

3. Metodologia

Etapa 1: Segmentação regional e etária

A fim de verificar a existência de pelo menos uma relação de equivalência proporcional entre diferentes populações, foi aplicado o teste de qui-quadrado. As regiões foram agrupadas em 10 pares, o nível de significância estabelecido foi de 0,1% com grau de liberdade 1, portanto $\chi^2_{1,\alpha=0.1\%} = 10,83$.

A hipótese nula (H_0) atesta que se não há diferença entre as populações das regiões quanto à distribuição de diagnósticos de depressão caso $Q < \chi^2$. Esse resultado foi verificado apenas para as regiões Sudeste e Centro-Oeste do Brasil, em que $Q = 9,9$ são regiões proporcionalmente equivalentes.

Em seguida foi realizado uma análise exploratória e construídos gráficos de frequência acumulada da quantidade de registros por idade do diagnóstico, com o intuito de detectar a idade em que acontece um aumento significativo de casos de depressão em cada região. Desse modo, a faixa etária identificada como ponto de inflexão foi de 23 a 55 anos.

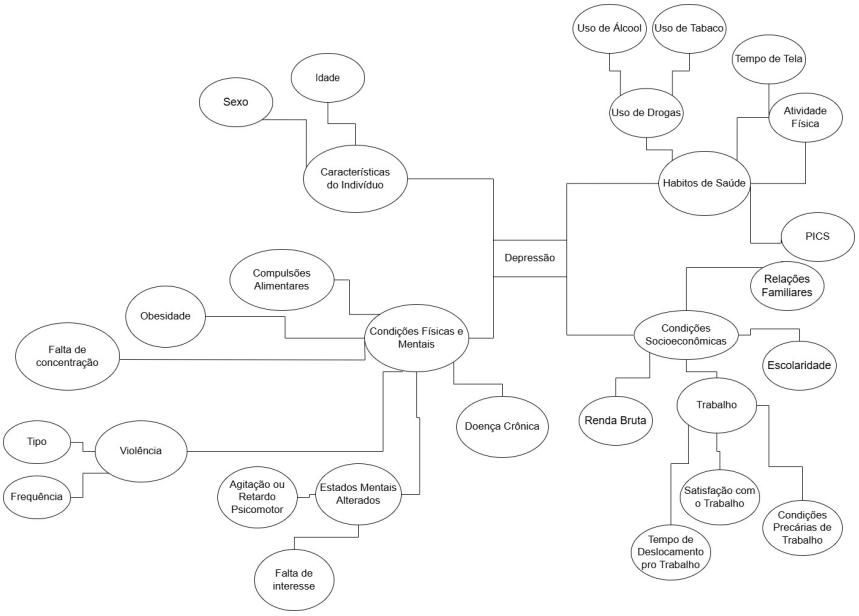

Figura 1. Mapa conceitual para o domínio do problema Depressão.

Etapa 2: Seleção de atributos

A seleção conceitual de atributos foi realizada seguindo a metologia CAPTO, proposta por [Gonçalves et al. 2024]. O CAPTO é um método para entendimento do domínio do problema, uma etapa essencial do processo de descoberta de conhecimento em base de dados (KDD) para criar modelos de aprendizado mais representativos do domínio de problema.

Com base em conhecimento extraído da literatura sobre o tema, foi construído um mapa conceitual composto pelas dimensões Hábitos de saúde; Condições socioecônomicas; Condições físicas e mentais; e Características do indivíduo. A partir disso, foram mapeados atributos disponíveis na base de dados da PNS 2019. A Tabela 1 mostra as referências de apoio das dimensões descritas anteriormente, assim como os atributos associados a cada uma e o que foi obtido da PNS. O significado de cada atributo e sua codificação original pode ser conferida no dicionário de microdados do IBGE.

Descrição do Mapa Conceitual – Domínio de problema: Depressão		
Dimensão: Hábitos de saúde		
Aspectos	**Atributos associados**	**Atributos mapeados**
Uso de Drogas: O uso de drogas como tabaco e álcool podem intensificar a depressão, fazendo usuários que têm depressão mais identificáveis. [Beneton et al. 2021]	Uso diário de tabaco ou algum tipo de droga; Frequência de consumo de álcool.	Módulo P - Estilo de vida: P50 até P55; P27 até P33c.
Atividade física: A prática de exercício físico, bem como de Práticas Integrativas e Complementares em Saúde, está associada à melhora de pacientes com depressão. [Oliveira and Alves 2023] [Schwambach and Queiroz 2023]	Prática de algum tipo de exercício e/ou PIC; Frequência de exercício físico ou esporte.	Módulo P - Estilo de vida: P34, P35 e P37. Módulo J - Utilização de serviços de saúde: J53.
Tempo de tela: O tempo de tela em diferentes aparelhos eletrônicos pode influenciar a saúde mental. [Li et al. 2022]	Tempo diário de uso de aparelhos eletrônicos (celular, televisão etc.) para lazer.	Módulo P - Estilo de vida: P45a e P45b.
Dimensão: Condições socioeconômicas		
Aspectos	**Atributos associados**	**Atributos mapeados**
Condições precárias de trabalho: [Yang and Lee 2024] relata que trabalhadores em situação de instabilidade possuem um risco maior de desenvolver sintomas depressivos.	Ausência de benefícios; Poluição sonora; Carga de trabalho excessiva.	Módulo E - Características de trabalho das pessoas de 14 anos ou mais de idade: E14c, E19. Módulo M - Características do trabalho e apoio social: M5c até M11.
Tempo de deslocamento para o trabalho: Tempos de deslocamento mais longos estão associados a estresse e a uma saúde mental mais precária, aumentando o risco de depressão. [Wang et al. 2019]	Horas no trânsito; Uso de transporte público ou particular.	Módulo M - Características do trabalho e apoio social: M3b e M4a.
Dificuldade financeira: A perspectiva de dificuldade financeira e/ou desemprego é um fator estressante. [Cunha et al. 2012]	Desemprego; Renda bruta mensal; Tempo afastado do trabalho.	E1, E2. E3, E4, E5, E10a, E22.

Escolaridade: O nível de escolaridade está associado a melhores condições de vida. A baixa escolaridade é um fator prevalente em indivíduos com depressão. [Campos et al. 2021]	Nível de escolaridade.	Módulo D - Características de educação dos moradores: D1, D3a e D9a.
Relações familiares: Conflitos conjugais, divórcio e problemas com filhos estão associados ao desenvolvimento de depressão na população adulta. [Jorgetto and Marcolan 2021]	Estado civil; Relações com a família.	Módulo M - Características do trabalho e apoio social: M14a.
Isolamento social: Medidas de restrição impostas durante a pandemia de COVID-19 evidenciaram que o isolamento social contribui para o aumento de episódios depressivos. [Almeida et al. 2020]	Frequência de encontros sociais ou religiosos.	Módulo M - Características do trabalho e apoio social: M14a até M19a.
Condições de moradia: "Viver em moradias precárias é um estressor psicossocial que pode levar a problemas de saúde mental." [Kim et al. 2021]		Módulo A - Informações do Domicílio: A1 até A16a.
Dimensão: Condições físicas e mentais		
Aspectos	**Atributos associados**	**Atributos mapeados**
Critérios de diagnóstico de depressão: Apatia e falta de interesse, capacidade diminuída para se concentrar e pensar, agitação ou retardo psicomotor, percepção negativa sobre a própria saúde, relato de trauma violento. (American Psychiatrich Association)	Opinião sobre a própria saúde; Autoestima; Sintomas de depressão; Tipo de violência e frequência.	Módulo N - Percepção do estado de saúde: N11, N12, N13, N15, N1a, N16 até N18. Módulo V - Violência.
Doenças crônicas: Pessoas com doenças crônicas podem apresentar depressão como uma doença secundária. (American Psychiatrich Association)	Tem doença(s) crônica(s)	Módulo Q - Doenças Crônicas: Q079, Q03001 e Q06306
Obesidade e compulsão alimentar: Pessoas obesas e com compulsões alimentares têm maior chance de ter depressão. [Fusco et al. 2020]	IMC; Compulsões alimentares.	Módulo P - Estilos de vida: P1a e P4a. Módulo N - Percepção do estado de saúde: N14.
Bioquímica cerebral: Hipóteses indicam que a deficiência de neurotransmissores como a noradrenalina e serotonina pode levar à depressão. [Diniz et al. 2020]	Resultado de exames de sangue e urina.	Não está disponível na base de dados da PNS 2019.
Dimensão: Características do indivíduo		
Aspectos	**Atributos associados**	**Atributos mapeados**
Sexo: O diagnóstico de depressão é mais prevalente em mulheres no Brasil. [Meleiro et al. 2023]	Sexo	Módulo C – Características gerais dos moradores: C6.
Idade: O diagnóstico é mais comum em indivíduos acima de 40 anos. [Meleiro et al. 2023]	Idade	Módulo C – Características gerais dos moradores: C8.

Tabela 1. Tabela descritiva do mapa conceitual.

Etapa 3: Montagem e pré-processamento do conjunto de dados

Após a seleção de atributos, foram extraídas as instâncias da base de dados PNS 2019 que responderam se possuem diagnóstico de depressão (independente do resultado ser positivo ou negativo), residem nas regiões Sudeste e Centro-Oeste do Brasil, cuja idade está entre 23 e 55 anos e o informante reside no domicílio. O dataset resultante possui 129 atributos e 17394 instâncias. Destas, 1850 foram diagnosticadas com depressão.

Tratamento de dados ausentes e criação de atributos

Cerca de 60% dos atributos no dataset apresentavam dados nulos. Foi realizado um processo minucioso de imputação manual com o uso da técnica produto cartesiano, que consiste em multiplicar e analisar as combinações de valores entre dois ou mais atributos. Desse modo, verificamos a possibilidade de inferir o valor do dado ausente com base nos valores de atributos relacionados.

A Tabela 2 detalha as transformações feitas na base de dados e a codificação dos atributos categóricos. Os tópicos em itálico são atributos que foram criados a partir de atributos da PNS, enquanto os outros são referentes a atributos que permaneceram com o código da PNS como nome.

Foi decidido eliminar instâncias quando os atributos não permitiam imputação sem uma distorção significativa da realidade. Esse critério foi aplicado em relação às instâncias em que o entrevistado não soube informar se tinha diagnóstico de diabetes; qual o seu peso e altura; e quantos dias da semana se deslocava para ir ao trabalho.

Detecção e eliminação de outliers

Concomitante ao tratamento de dados ausentes, foi realizada a remoção de outliers. Para isso, foram criados gráficos boxplot para todos os atributos de tipo contínuo do conjunto

Tópico	Atributos	Novos valores
Está em um relacionamento	C014 = [1,2]	Sim (1)
	C014 = null	Não (2)
Escolaridade	D00901 != null	Valor permanece o mesmo
	D00901 = null e D00301 != null	D00301 -= 1
	D00901 = null e D00301 = null	Não frequentou escola (0)
Posse de carteira assinada	E01403 != null	Valor permanece o mesmo
	E01403 = null e E01401 != 0	Sim (1)
	E01403 = null e E01401 = 0	Não (2)
Salario	E01602 + E01604 + E01802 + E01804	
Carga_Horaria_Semanal	E017 + E019	
Tempo_Deslocamento_Min	M00401*60+M00402	
Categoria_IMC	P00104/(P00404/100)^2	
Frequencia_Alcool	P027 = 1	Nunca bebe (1)
	P027 = 2	< de 1 vez por mês (2)
	P027 = 3 e P02801 = 0	Min. 1 vez por mês (3)
	P027 = 3 e P02801 > 0	Min. 1 vez por semana (4)
Tempo_Minimo_Ezercicio	(P03701*60+P03702)≥ 150	Sim (1)
	(P03701*60+P03702)< 150	Não (2)
TempoDeTela	P04501 e P04502 = 6	Nada (1)
	P04501 e P04502 = [1,2]	Menos de 2 h. (2)
	P04501 ou P04502 = [3,4,5]	Mais de 2 h. (3)
Frequencia_Fumo	P050 = 3 e P052 = 3	Nunca fumou (1)
	P050 = 3 e P052 = [1,2]	No passado com alguma frequência (2)
	P050 = 2	Ocasionalmente (4)
	P050 = 1 e P055 >1	Diariamente, pelo menos 5 min após acordar (5)
	P050 = 1 e P055 = 1	Diariamente, menos 5 min após acordar (6)
Violencia_Psicologica (V00201 até V00205)	Todos = 2	Nenhuma vez (1)
	V0020x = 1 e V003 = 1	Uma vez (2)
	V0020x = 1 e V003 = 2	Algumas vezes (3)
	V0020x = 1 e V003 = 3	Muitas vezes (4)
Violencia_Sexual (V02701 até V02801)	Algum V02xxx = 1	Sim (1)
	Todos V02xxx = 2	Não (2)

Atributo	Codificação
Sexo (C006)	Masculino = 1; Feminino = 2
Realiza PIC (J05301)	Sim = 1; Não = 2
Binge drinking no último mês (P03201)	
Diagnóstico de depressão (Classe, Q092)	
Insônia (N010)	(1) Nenhum dia (2) Menos da metade dos dias (3) Mais da metade dos dias (4) Quase todo dia
Apatia (N012)	
Falta de perspectiva (N016)	
Sensação de fracasso (N017)	
Ideação suicida (N018)	
Tipo de trabalho (E01401)	(0) Desempregado (1) Trabalhador doméstico (2) Militar (3) Empregado do setor privado (4) Empregado do setor público (5) Empregador (6) Conta própria (7) Trabalhador não remunerado
N° de familares com quem pode contar (M01401)	(0) Nenhum; (1) Um ou dois; (3) Três ou mais
Idade (C008)	(1) 23-29; (2) 30-39; (3) 40-49; (4) 50-55
Deslocamento_Trabalho_Dias	(0) Nenhum; (1) 1-5 dias; (3) 6-7 dias

Tabela 2. Tabela de atributos.

de dados, porém, não houve remoção de instâncias no caso de *Categoria_IMC*. Ademais, para atributos relacionados a tempo dedicado ao trabalho ou atividades domésticas, foram incluídas informações como posse de carteira assinada, quantidade de empregos, principal tipo de trabalho informado etc. de forma a realizar remoções mais assertivas.

O atributo *Tempo_Deslocamento_Min* foi discretizado de acordo com as categorias utilizadas na publicação "Indicadores de Efetividade da Política Nacional de Mobilidade Urbana", elaborada pelo Ministério das Cidades. *Tempo_Minimo_Exercicio* foi transformado em binário com base no tempo mínimo de atividade física para adultos estabelecido pela OMS. Demais atributos de tipo contínuo foram discretizados tendo como prioridade criar conjuntos com quantidades similares de instâncias.

Redução de dimensionalidade com base em análise de proporcionalidade

Alguns atributos categóricos ordinais passaram por uma redução de valores com o objetivo de melhorar a interpretabilidade da árvore e tornar atributos com muitas opções menos suscetíveis à eliminação durante a análise de entropia. O processo consistiu em criar histogramas onde as abscissas correspondem aos valores do atributo e as ordenadas à distribuição por classes para cada opção de resposta; valores adjacentes com proporções semelhantes (máximo de 3% de diferença) foram agrupados, contanto que a união tivesse sentido semântico.

De forma a aumentar a interpretabilidade, atributos categóricos foram pré-processados para no máximo 4 valores distintos. Exceções a essa regra são o atributo D00901 (Escolaridade) e *Frequencia_Fumo*, ambos com 5 valores. O primeiro foi dividido em: não frequentou escola; ensino infantil; ensino fundamental; ensino médio; e ensino superior. Já o segundo, como visto na Tabela 2, contém uma distinção entre pessoas que fumam diariamente baseada na pesquisa de [Bainter et al. 2020]. Segundo a qual, pessoas que fumam pela primeira vez menos de 5 minutos após acordar tendem a experimentar sintomas de depressão com mais intensidade.

Seleção de atributos baseado na análise de entropia

A partir da análise da entropia, foi observado que atributos com valor muito baixo continham distribuição desbalanceada (induzindo enviesamento) portanto foi estabelecido um

limite mínimo de 5% para a frequência relativa do valor para eliminação do atributo. Dentro desse filtro, foi decidido preservar alguns atributos, seja por ganho relativamente alto, importância explícita na literatura, ou número possível de valores maior, balanceando a distribuição.

Após essa filtragem, foi analisada a entropia condicional com relação à classe. Considerando que, quanto mais próxima da entropia da classe estiver a entropia condicional de um atributo, menor é sua capacidade de diminuir a incerteza da classe, portanto, foi decidido um corte máximo de 0,5 bits para a entropia condicional dos atributos da base, 0,012 bits a menos que o valor da entropia da classe. Novamente, alguns atributos foram preservados por importância explícita na literatura.

Por fim, ao analisar os 29 atributos restantes da base, foi observado uma concentração de atributos que avaliam a frequência de sintomas associados ao estado emocional do entrevistado nas duas últimas semanas. Para evitar *overfitting* e reduzir a redundância informacional, optamos por remover os atributos N00101, N011, N013, N014 e N015, que tratam respectivamente da opinião do indivíduo sobre seu bem estar físico e mental, problemas por não se sentir disposto durante o dia, dificuldade de concentração, falta de apetite ou alimentação excessiva, e inquietamento. A escolha foi feita com base no menor ganho de informação desses itens em comparação com os demais da mesma seção.

Eliminação de dados redundantes

Com o intuito de certificar que não existiam instâncias repetidas que poderiam comprometer o processo de treino e teste, foi utilizada a função *drop_duplicates* da biblioteca *pandas* para eliminar duplicatas da base. Como resultado, 943 instâncias foram eliminadas.

Terminada a etapa do pré-processamento, a base de dados apresentava 23 atributos e 15249 instâncias, das quais 1780 possuem diagnóstico de depressão. A base de dados está disponível em https://github.com/licapLaboratory/Database-Depressao-sudeste

4. Aplicação do algoritmo *Decision Tree*

Para a construção do modelo foi utilizado o software de análise de dados KNIME (ver Figura 2). O nó *Decision Tree Learner*, responsável pelo treinamento, foi configurado com a medida de qualidade Gain Ratio, poda, mínimo de registros por nó equivalente a 10 e divisão binária para atributos categóricos nominais.

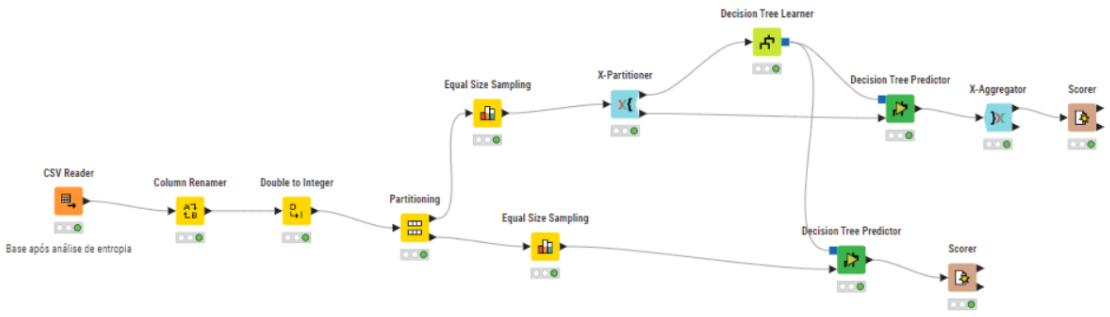

Figura 2. Fluxograma da ferramenta Knime.

Como ajustes finais antes do processo de indução do modelo, os atributos que ainda possuíam códigos da base original da PNS 2019 como nome foram renomeados para refletir seus significados e foi realizada a padronização de todos os atributos que apresentam valores numéricos para tipo *int*.

O conjunto de dados foi separado em 70% para treino e 30% para teste, conforme a metodologia *hold-out*. Esse particionamento preservou a proporção original das classes, portanto 88.33% das instâncias de cada conjunto pertencem à classe 2 (sem diagnóstico) e 11.67% à classe 1 (possui diagnóstico de depressão).

O conjunto de treino então passou por um processo de balanceamento undersampling e foi reduzido a 2492 instâncias. Esse novo conjunto passou por um processo de validação cruzada. Optamos por realizar 10 dobras com amostragem aleatória. Desse modo, 2243 instâncias foram utilizadas no processo de treinamento do algoritmo de *Decision Tree* e 249 foram reservadas para validar o resultado do modelo.

O nó *Scorer* superior na Figura 2 apresenta a média geral das métricas apresentadas pelos 10 modelos, ver Tabela 3, onde é possível observar uma F1-score média de 72,5%.

Classe	**Recall**	**Precisão**	**F1-Measure**
Sem diagnóstico (2)	0.729	0.724	0.726
Com diagnóstico (1)	0.722	0.727	0.724

Tabela 3. Resultado da Árvore de Decisão após cross-validação.

Após o processo de treinamento, o modelo foi testado com o conjunto de teste separado previamente. A performance foi afetada negativamente pelo grande desbalanceamento das classes, em razão disso optou-se por aplicar balanceamento undersampling no conjunto de teste e avaliar o número de acertos e erros do modelo ao invés de priorizar métricas de precisão e recall. Observou-se que a árvore classificou erroneamente 164 instâncias como pertencentes à classe 2 e 157 instâncias como pertencentes à classe 1, totalizando 321 erros (cerca de 30%). A taxa de acertos foi de 70% para ambos os casos.

	Sem diagnóstico (2)	**Possui diagnóstico (1)**
Sem diagnóstico (2)	377	157
Possui diagnóstico (1)	164	370

Tabela 4. Matriz de confusão da Árvore de Decisão.

Análise de regras geradas pela árvore

O nó *Decision Tree to Ruleset* foi incluído no fluxograma do KNIME após o treino e teste do modelo para extração das regras utilizadas pela árvore para classificar as instâncias. As três regras com maior número de acertos foram separadas e analisadas.

(IF RealizaPIC $>$ 1.5 AND Insonia $<=$ 1.5 AND Apatia $<=$ 2.5 AND Insonia $<=$ 3.5 AND FaltaDePerspectiva $<=$ 1.5) THEN class = Sem Depressão Indivíduos que não têm insônia frequentemente, não tiveram sentimentos negativos como apatia e falta de perspectiva recentemente e não realizam atividades complementares a saúde, tais como ioga, acupuntura etc. tendem a não ter diagnóstico de depressão. Essa regra classificou 777 instâncias e acertou 633.

(IF FaltaDePerspectiva $>$ 2.5 AND PensamentosSuicidas $<=$ 3.5 AND FaltaDePerspectiva $>$ 1.5) THEN class = Com Depressão Altos níveis de falta de perspectiva e a existência de pensamentos suicidas, ainda que infrequentes, indicaram diagnóstico positivo para depressão. Acertou 354 de 423 instâncias.

(IF Insonia $>$ 2.5 AND E01401 IN (”Empregado do setor público”, ”Empregado do setor privado”, ”Conta própria”, ”Desempregado”, ”Trabalhador doméstico”, ”Militar”) AND FaltaDePerspectiva $<=$ 2.5 AND PensamentosSuicidas $<=$ 3.5 AND FaltaDePerspectiva $>$ 1.5) THEN class = Com Depressão Apresenta os mesmos atributos descritos nas regras anteriores, exceto pelo Tipo de Emprego (E01401). Nessa regra é interessante notar a ausência das categorias ‘Empregador’e ‘Trabalhador não remunerado em auxílio a domicílio ou parente’, o que pode indicar que indivíduos nessas posições não estão tão sujeitos a estresse. Ao mesmo tempo, a aparição dos atributos Insônia; FaltaDePerspectiva; e PensamentosSuicidas reforça a capacidade desses atributos identificarem indivíduos com depressão. Essa regra classificou corretamente 132 instâncias de 167.

Os atributos mais influentes tratam de sensações vivenciadas nas últimas duas semanas. O grau elevado de recorrência desses sentimentos em pessoas que foram diagnosticadas há mais tempo podem ser vistos como sinais de que a pessoa está em um quadro mais grave da doença. Tratando-se de pessoas que não têm diagnóstico, podem ser interpretados como indicativos de desenvolvimento de algum problema emocional. Desse modo, habilitar médicos da rede de atenção primária para reconhecerem possíveis pacientes com depressão e incentivar a realização de check-ups com foco em saúde mental poderia ajudar a minimizar o subdiagnóstico no país.

5. Conclusão

Embora o algoritmo baseado em árvore de decisão apresente métricas com resultados satisfatórios, o conhecimento extraído das regras obtidas e a possibilidade de interpretação das mesmas deixa maiores desafios. Os atributos mais influentes tratam de sensações tipicamente vividas por pessoas com depressão, e permitem traçar iniciativas para auxiliar na identificação dos mesmos. Ainda assim, as regras por si só não possibilitam criar grupos com características bem definidas para auxiliar na identificação de pessoas que têm ou não a doença sob outras perspectivas.

Acredita-se que pensar em estratégias para aumentar o peso de outros atributos seria uma forma de mudar esse cenário, pois a presença constante de E01401 nas demais regras é um indicativo de que motivos socioeconômicos estão relacionados ao problema. A inclusão de informações como o número de filhos, rendimento domiciliar, condição de ocupação do domicílio entre outros fatores poderia enriquecer as regras geradas pela árvore. Em retrospectiva, a adição de atributos que falam sobre restrições ou nível de incômodo causado por uma doença crônica também poderia ser mais informativo do que utilizar somente o diagnóstico de doenças relacionadas à depressão.

Além dessas estratégias, acreditamos que seria positivo ter mais dados sobre possíveis eventos traumáticos e o impacto deles na vida dos entrevistados. Como visto na tabela de atributos, *Violencia_Psicologica* e *Violencia_Sexual* passaram pelo teste de entropia e compõem a base final. Porém, a base da PNS é escassa nesse sentido. Outra limitação é a ausência de dados relativos à bioquímica cerebral, um aspecto mapeado na

etapa de seleção conceitual de atributos.

Após ajustar o peso dos atributos de outros segmentos, consideramos utilizar um algoritmo de agrupamento para investigar grupos de indivíduos entre aqueles diagnosticados corretamente com o intuito de gerar *insights* mais detalhados e úteis, tendo em vista o objetivo inicial desse estudo.

A PNS 2019 é uma base de dados transversal. Uma vez que a depressão é uma doença multifatorial e influenciada por mudanças diversas ao longo da vida, destacamos que estudos com bases de dados longitudinais sobre o assunto poderiam obter resultados eficientes. O impacto de razões socioeconômicas, traumas entre outros eventos a priori da aplicação da pesquisa poderiam ser estudados mais a fundo e apresentar mais relevância nas decisões do algoritmo de classificação.

Referências

Aguiar, R. A., Riffel, R. T., Acrani, G. O., and Lindemann, I. L. (2022). Tentativa de suicídio: prevalência e fatores associados entre usuários da atenção primária à saúde. *Jornal Brasileiro de Psiquiatria*, 71(2):133–140.

Almeida, W. d. S. d., Szwarcwald, C. L., Malta, D. C., Barros, M. B. d. A., Souza Júnior, P. R. B. d., Azevedo, L. O., Romero, D., Lima, M. G., Damacena, G. N., Machado, E., Gomes, C. S., Pina, M. d. F. d., Gracie, R., Werneck, A. O., and Silva, D. R. P. d. (2020). Mudanças nas condições socioeconômicas e de saúde dos brasileiros durante a pandemia de covid-19. *Revista Brasileira de Epidemiologia*, 23.

Bainter, T., Selya, A. S., and Oancea, S. C. (2020). A key indicator of nicotine dependence is associated with greater depression symptoms, after accounting for smoking behavior. *PLOS ONE*, 15(5):1–11.

Beneton, E. R., Schmitt, M., and Andretta, I. (2021). Sintomas de depressão, ansiedade e estresse e uso de drogas em universitários da área da saúde. *Revista da SPAGESP*, 22(1):145–159.

Campos, I. d. O., Cruz, D. M. C. d., Magalhães, Y. B., and Rodrigues, D. d. S. (2021). Escolaridade, trabalho, renda e saúde mental: um estudo retrospectivo e de associação com usuários de um centro de atenção psicossocial. *Physis: Revista de Saúde Coletiva*, 31(3).

Cunha, R. V. d., Bastos, G. A. N., and Duca, G. F. D. (2012). Prevalência de depressão e fatores associados em comunidade de baixa renda de porto alegre, rio grande do sul. *Revista Brasileira de Epidemiologia*, 15(2):346–354.

Diniz, J. P., Neves, S. A. d. O., and Vieira, M. L. (2020). Ação dos neurotransmissores envolvidos na depressão. *Ensaios e Ciência C Biológicas Agrárias e da Saúde*, 24(4):437–443.

Fleck, M. P. d. A., Lafer, B., Sougey, E. B., Del Porto, J. A., Brasil, M. A., and Juruena, M. F. (2003). Diretrizes da associação médica brasileira para o tratamento da depressão (versão integral). *Rev. Bras. Psiquiatr.*, 25(2):114–122.

Fusco, S. d. F. B., Amancio, S. C. P., Pancieri, A. P., Alves, M. V. M. F. F., Spiri, W. C., and Braga, E. M. (2020). Ansiedade, qualidade do sono e compulsão alimentar em adultos com sobrepeso ou obesidade. *Revista da Escola de Enfermagem da USP*, 54.

Gonçalves, L., Franca, D., and Zarate, L. (2024). Relevância do entendimento do domínio de problema na construção de modelos computacionais de aprendizado. In *Anais do XVIII Brazilian e-Science Workshop*, pages 135–142, Porto Alegre, RS, Brasil. SBC.

Jorgetto, G. V. and Marcolan, J. F. (2021). Risk and protective factors for depressive symptoms and suicidal behavior in the general population. *Revista Brasileira de Enfermagem*, 74(suppl 3).

Kim, S., Jeong, W., Jang, B. N., Park, E.-C., and Jang, S.-I. (2021). Associations between substandard housing and depression: insights from the korea welfare panel study. *BMC Psychiatry*, 21(1).

Lee, K.-S. and Ham, B.-J. (2022). Machine learning on early diagnosis of depression. *Psychiatry Investigation*, 19(8):597–605.

Li, L., Zhang, Q., Zhu, L., Zeng, G., Huang, H., Zhuge, J., Kuang, X., Yang, S., Yang, D., Chen, Z., Gan, Y., Lu, Z., and Wu, C. (2022). Screen time and depression risk: A meta-analysis of cohort studies. *Frontiers in Psychiatry*, 13.

Lipp, M. E. N. and Lipp, L. M. N. (2020). Stress e transtornos mentais durante a pandemia da COVID-19 no Brasil. *Boletim - Academia Paulista de Psicologia*, 40:180 – 191.

Meleiro, A., Teng, C. T., Demetrio, F. N., Batista, V. C., Vieira, L. F., and Elorza, P. M. (2023). Understanding the journey of patients with depression in brazil: A systematic review. *Clinics*, 78:100192.

Oliveira, J. S. and Alves, S. F. d. S. (2023). Impacto da prÁtica de exercÍcio fÍsico na saÚde mental dos indivÍduos acometidos pela depressÃo: Revisão integrativa. *REVISTA FOCO*, 16(8):e1616.

Schwambach, L. B. and Queiroz, L. C. (2023). Uso de práticas integrativas e complementares em saúde no tratamento da depressão. *Physis: Revista de Saúde Coletiva*, 33:e33077.

Shin, D., Lee, K. J., Adeluwa, T., and Hur, J. (2020). Machine learning-based predictive modeling of postpartum depression. *Journal of Clinical Medicine*, 9(9):2899.

Silva, P. C. e., Pereira Filho, C. H. d. M., Marques, G. L. P., Melo, I. Q. d. S., Wagner, E. R., Santos, K. d. S. d., Costa, A. C. M. d. S. F. d., Romeiro, E. T., Pereira, H. B., and Colnaghi, B. S. (2023). Efeitos da pandemia na saÚde mental da populaÇÃo. *Revista Ibero-Americana de Humanidades, Ciências e Educação*, 9(7):1281–1291.

Wang, X., Rodríguez, D. A., Sarmiento, O. L., and Guaje, O. (2019). Commute patterns and depression: Evidence from eleven latin american cities. *Journal of Transport amp; Health*, 14:100607.

Yang, Y.-J. and Lee, J. (2024). Association between depressive symptoms and employment type of korean workers: the fifth korean working conditions survey. *BMC Public Health*, 24(1).