Research, Society and Development, v. 11, n. 1, e10511124547, 2022
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v11i1.24547

Depressão entre jovens brasileiros: uma investigação baseada em mineração de subgrupos

Depression among Brazilian youth: an investigation based on subgroup discovery

Depresión entre los jóvenes brasileños: una investigación basada en la minería de subgrupos

Recebido: 09/12/2021 | Revisado: 14/12/2021 | Aceito: 19/12/2021 | Publicado: 03/01/2022

Filipe Cordeiro de Medeiros Azevedo
ORCID: https://orcid.org/0000-0003-0821-8779
Universidade Católica de Pernambuco, Brasil
E-mail: filipe.2018113793@unicap.br
Tarcísio Daniel Pontes Lucas
ORCID: https://orcid.org/0000-0001-9843-5181
Universidade Católica de Pernambuco, Brasil
E-mail: tarcisio.lucas@unicap.br

Resumo
Este trabalho investigou os grupos de características socioeconômicas e de estilo de vida relacionadas à depressão entre jovens brasileiros utilizando uma abordagem baseada em mineração de dados. A depressão é o resultado da interação complexa entre um grande número de fatores ambientais e genéticos. No entanto, a interferência dos fatores ambientais contribuintes para depressão ainda é um desafio em aberto. Uma fonte de dados recente e volumosa com potencial para investigar estes fatores ambientais é a Pesquisa Nacional de Saúde (PNS), um estudo realizado periodicamente pelo Instituto Brasileiro de Geografia e Estatística (IBGE) que tem como objetivo produzir dados sobre a situação de saúde e os estilos de vida da população brasileira. Nesse sentido, utilizamos a mineração de subgrupos na PNS a fim de encontrar conjuntos de características que destacam um grupo alvo dos demais (ex.: pessoas com depressão das demais).
Palavras-chave: Depressão; Mineração de dados; Mineração de subgrupos.

Abstract
This paper investigated groups of socioeconomic and lifestyle characteristics related to depression among Brazilian youth using a data mining-based approach. Depression is the result of the complex interaction between a large number of environmental and genetic factors. However, environmental factors contributing to depression is still an open challenge. A recent and voluminous source of data with the potential to investigate these environmental factors is the National Health Survey (PNS), a study carried out periodically by the Brazilian Institute of Geography and Statistics (IBGE) which aims to produce data on the health situation and lifestyles of the Brazilian population. In this sense, we use Subgroup Discovery on the PNS in order to find sets of characteristics that make a target group stand out from others (e.g. people with depression from others).
Keywords: Depression; Data mining; Subgroup discovery.

Resumen
Este artículo investigó grupos de características socioeconómicas y de estilo de vida relacionadas con la depresión entre los jóvenes brasileños utilizando un enfoque basado en la minería de datos. La depresión es el resultado de la compleja interacción entre una gran cantidad de factores ambientales y genéticos. Sin embargo, la interferencia de factores ambientales que contribuyen a la depresión sigue siendo un desafío abierto. Una fuente reciente y voluminosa de datos con potencial para investigar estos factores ambientales es la Encuesta Nacional de Salud (PNS), un estudio realizado periódicamente por el Instituto Brasileño de Geografía y Estadística (IBGE) que tiene como objetivo producir datos sobre la situación de salud y los estilos de vida de la población brasileña. En este sentido, utilizamos la minería de subgrupos de PNS para encontrar conjuntos de características que hacen que un grupo objetivo se destaque de los demás (por ejemplo, personas con depresión de otros).
Palabras clave: Depresión; Minería de datos; Minería de subgrupos.

1. Introdução

Este trabalho investigou os grupos de características socioeconômicas e de estilo de vida relacionadas a depressão entre jovens brasileiros utilizando uma abordagem baseada em mineração de dados. A depressão afeta cerca de 322 milhões de pessoas

e é a causa de mais de 788 mil suicídios por ano em todo o mundo, de acordo com a Organização Mundial de Saúde (OMS). No Brasil, mais de 11,5 milhões de pessoas são atingidas pela doença, o que representa 5,8% do total da população (OMS, 2017).

A depressão é o resultado da interação complexa entre um grande número de fatores ambientais e genéticos. Os fatores ambientais são foco de algumas aplicações de técnicas de Inteligência Artificial no contexto da depressão. Em (Hullam et al., 2019), por exemplo, os autores aplicaram redes Bayesianas para identificar fatores ambientais que estejam correlacionados com a depressão, incluindo características sociais e de estilo de vida na população britânica.

Já no trabalho de (Daimi & Banitaan, 2014), foi desenvolvido um modelo de classificação para auxiliar o diagnóstico de transtornos depressivos. Há também estudos que tentam classificar a depressão a partir da análise de sentimento extraída de textos em redes sociais, como por exemplo (Islam et al., 2018) e (Gonçalves, Ferreira, Neto, Abelha & Machado, 2020).
Outros pesquisadores ainda utilizaram regressão de Poisson para investigar a influência de diferentes características na depressão. (Barger, Messerli-Bürgy & Barth, 2014) investigaram a população suíça com uma abordagem de relacionamentos sociais em termos de frequência e qualidade. Por fim, (Carvalho, 2016) avaliou a população nordestina brasileira economicamente ativa com relação a doença.

Alguns trabalhos investigaram a associação da depressão e fatores ambientais, mas com uma abordagem em temas específicos. Por exemplo, (da Rocha, Myva & Almeida, 2020) investigaram o papel da alimentação no tratamento da depressão. (Pasini et at., 2020) e (Brito, 2011) analisaram fatores de risco e estratégias de prevenção da depressão na adolescência. Já (Azuelo et at., 2020) investigou a associação da depressão com a violência. Em (Justo & Calil, 2006) é feita uma análise da diferença da depressão entre os gêneros.

No entanto, mesmo com os esforços citados, a interferência dos fatores ambientais contribuintes para depressão ainda é um desafio em aberto (Hullam et al., 2019). Além disso, no contexto brasileiro poucos trabalhos tem se dedicado a investigar a relação entre fatores ambientais e a doença utilizando grande volume de dados recentes e técnicas mais modernas de mineração de dados.

Uma fonte de dados recente e volumosa com potencial para investigar fatores ambientais relacionados a depressão é a Pesquisa Nacional de Saúde (PNS), um estudo realizado periodicamente pelo Instituto Brasileiro de Geografia e Estatística (IBGE) que tem como objetivo produzir dados sobre a situação de saúde e os estilos de vida da população brasileira (IBGE, 2021).

A mineração de subgrupos é uma área da mineração de dados com o objetivo de encontrar, a partir de dados, conjuntos de características que destacam um grupo alvo dos demais (ex.: pessoas com depressão das demais). A mineração de subgrupos têm sido utilizada em diversos domínios, como na mineração de fatores relevantes no sucesso acadêmico (Helal et al., 2019) e na investigação descritiva de pacientes de uma emergência psiquiátrica (Carmona, González, Jesus, Navío-Acosta & Jiménez-Trevino, 2011). A Aplicação desse tipo de técnica tem alto custo computacional, principalmente em bases de dados com grande número de atributos e exemplos. No entanto, em 2018 foi proposto o SSDP+ (Lucas, Vimieiro & Ludermir, 2018), um algoritmo heurístico com foco em bases de dados com esse perfil.

Assim, esse trabalho tem o objetivo de minerar os grupos de características ambientais associadas à pessoas jovens diagnosticadas com depressão a partir da aplicação do algoritmo SSDP+ num recorte dos dados na PNS de 2019, considerando aspectos socioeconômicos e estilo de vida.
As demais seções deste trabalho foram divididas da seguinte forma. Na Seção 2 nós apresentamos a área de mineração de subgrupos e as principais características do algoritmo que utilizamos na pesquisa. Na Seção 3 nós apresentamos a metodologia aplicada na análise. Por fim, na Seção 4 nós apresentamos os resultados e na Seção 5 as conclusões e trabalhos futuros.

2. Mineração de subgrupos

Mineração de subgrupos é uma técnica de mineração de dados para análise exploratória e descritiva de base de dados (Atzmueller, 2015). Essa técnica tem como objetivo descobrir relações interessantes entre os diferentes objetos de um conjunto, com respeito a uma propriedade específica de interesse. Esses padrões extraídos são normalmente representados em forma de regras e chamados de subgrupos (Herrera, Carmona, González & Del Jesus, 2011).

Dado que se está interessado em investigar uma propriedade específica, as relações não buscam necessariamente uma exatidão nos resultados, mas encontrar regras independentes que sobressaem (Herrera et al., 2011). Nesse sentido, um subgrupo interessante é definido como uma distribuição anormal dos dados em relação ao alvo de investigação (Wrobel, 1997).

Outras técnicas não foram capazes de atingir esse objetivo. Por exemplo, técnicas de predição tentam maximizar a precisão para classificar novos objetos. Já as técnicas de abordagens descritivas, apesar de conseguirem extrair padrões dos dados, falham em focar em propriedades de interesses específicos (Helal et al., 2019).

É importante entender que diferentes abordagens para mineração de subgrupos têm sido implementadas. Usualmente, todas as propriedades da base de dados devem conter possibilidades finitas, com alvo de investigação binário e nominal. Entretanto, autores têm desenvolvido abordagens utilizando multi-alvos e numéricos (Atzmueller, 2015).

Ao final do processo de mineração de subgrupos, obtém-se a descrição dos melhores k subgrupos que são altamente interessantes conforme a métrica adotada. Cada subgrupo é definido como $S = \{a_1, a_2, ..., a_n\}$, onde S é um conjunto de n variáveis chamadas de atributos. Um atributo a_i é categórico se possui valor em domínio finito $\{v_{i1}, v_{i2}, ..., v_{im}\}$ e contínuo se assume qualquer valor dentro de um intervalo $[min,max]$ (Lucas, 2019).

Dessa forma, as descrições dos subgrupos representam a conjunção de atributos, como na Equação (1), o qual S é um conjunto com n pares de atributo-valor (a_i,v_i) e implicam em uma distribuição anormal para uma propriedade $Alvo_{valor}$. Consequentemente, os subgrupos podem ser facilmente interpretados por usuários e especialistas do domínio para gerar hipóteses e extrair conhecimentos não triviais da base de dados (Atzmueller, 2015).

$$S : Condiçao \rightarrow Alvo_{valor} \quad (1)$$

2.1 Definição formal

Seja D um data set rotulado, com um conjunto A de atributos categoricos/discretos. É possível particiona-lo em $D^+ = \{e^+_1, e^+_2, ..., e^+_{|D+|}\}$, representando o atributo alvo (positivo) e $D^- = \{e^-_1, e^-_2, ..., e^-_{|D-|}\}$ os demais exemplos (negativo). Seja $dom(A_i)$ o domíinio de possíveis valores dos atributos $A_i \in A$. E definido o item I o par (atributo, valor), tal que $I = \cup A_i \times dom(A_i)$. Dessa forma, um exemplo d possui um item $x = (A_i , v) \in I$ se d possui o valor v para o atributo A_i (Lucas et al., 2018).

Um subgrupo é definido pelo conjunto $s \subseteq I$ tal que cada item possui uma proporção $c^+(s) = \{D^+ \cap s\}$ e $c^-(s) = \{D^- \cap s\}$ desbalanceada em relação ao atributo alvo, isto é, a frequencia entre os exemplos positivos e negativos. A soma destes valores definem o suporte do subgrupo (Lucas et al., 2018).

2.2 Métricas de avaliação

A definição da relevância e interesse de subgrupos é medido através de métricas (Lavrač, Flach & Zupan, 1999). Uma das métricas mais utilizadas é o Weighted Relative Accuracy (WRAcc), dado pela Equação (2), O qual $TP = |c^+(s)|$ é o suporte positivo e $FP = |c^-(s)|$ é o suporte negativo.

$$WRAcc(s) = \frac{TP+FP}{|D|}\left(\frac{TP}{TP+FP} - \frac{|D^+|}{|D|}\right), \quad (2)$$

Essa metrica de avaliação equilibra entre a cobertura $\frac{TP+FP}{|D|}$ e a precisão relativa $\left(\frac{TP}{TP+FP} - \frac{|D^+|}{|D|}\right)$. O WRAcc varia de -0.25 a +0.25, o qual +0.25 representa um subgrupo puro (Lucas et al., 2018).

Lift e uma métrica de avaliação em mineração de dados que tem como objetivo medir a resposta de um modelo em relação a probabilidade de uma escolha aleatória da população como um todo. Nesse sentido, o modelo tem um bom desempenho se a resposta é muito melhor que a média da população em uma propriedade de interesse (Tuffery, 2011). O lift é definido na Equação (3):

$$lift\ (s) = \frac{\frac{TP}{TP+FP}}{\frac{|D^+|}{|D|}} \quad (3)$$

Outra metrica de avaliação bem conhecida é o Qg, definido na Equação (4), onde g é o parâmetro de generalização. Dessa forma, altos valores de g frequentemente retornam subgrupos mais genéricos e menos precisos, enquanto que baixos valores de g retornam subgrupos específicos e de maior precisão (Lavrač et al., 1999).

$$Qg(s) = \frac{TP}{FP+g} \quad (4)$$

Existem varias métricas de avaliação para subgrupos (Herrera et al., 2011). A escolha da melhor métrica de avaliação frequentemente depende do problema ou convicções de especialistas (Lucas et al., 2018). Logo, é importante que a análise dos resultados considere diferentes perspectivas sobre os subgrupos.

2.3 Simple Search Discriminative Patterns Plus

O Simple Search Discriminative Patterns Plus (SSDP+) é um algoritmo genético para mineração de subgrupos mono-objetivo com foco em data sets de alta dimensionalidade (Lucas et al., 2018). O algoritmo inicia com subgrupos de uma dimensão. A seleção é feita por torneio binário. Então, o SSDP+ utiliza os operadores genéticos para gerar novos candidatos. As taxas de mutação e crossover são alteradas dinamicamente de forma que os k melhores subgrupos sejam propagados para a geração seguinte. Quando há evolução entre gerações, é aumentado o crossover e diminuída a mutação. Quando não há evolução, o processo se inverte, diminuindo o crossover e aumentando a mutação.

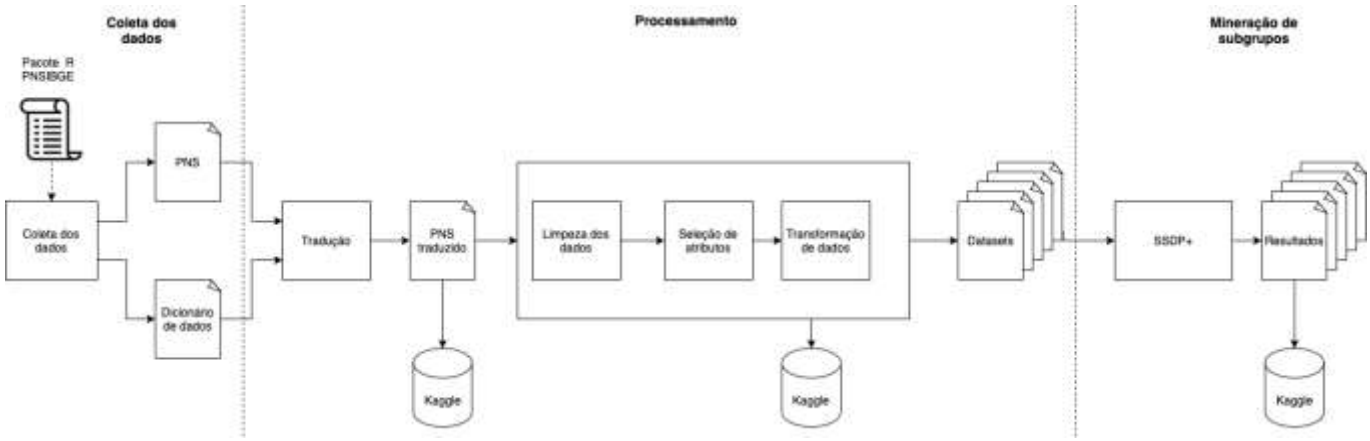
O critério de parada do SSDP+ é a estabilização dos melhores k subgrupos após a população ter sido reiniciada duas vezes. A população é reiniciada quando não há mudança nos melhores k subgrupos por três gerações consecutivas e a taxa de mutação é igual a um (Lucas et al., 2018). O SSDP+ está disponível em um repositório do Github na linguagem Java[1].

3. Metodologia

A metodologia utilizada neste trabalho pode ser dividida em três fases: (1) coleta dos dados; (2) processamento; e (3) mineração de subgrupos. Os passos que realizamos em cada fase estão ilustrados na Figura 1. Cada quadro na Figura 1 representa um script de processamento, um arquivo de base de dados ou repositório em que está armazenado. As setas na Figura 1 indicam que cada script recebe como entrada uma base de dados e o resultado está armazenado em um repositório.

[1] https://github.com/tarcisiodpl/ssdp plus

Figura 1. Processo de mineração de subgrupos.



(a)

Fonte: Autores (2021).

Na primeira fase, nós coletamos os dados da Pesquisa Nacional de Saúde (PNS). Considerando a edição de 2019, a PNS possui um enorme volume de dados, com uma amostragem de mais 279 mil pessoas entrevistadas e mais de 800 variáveis (dimensões). A pesquisa é dividida em quatro volumes, dos quais são subdivididos em módulos. Cada módulo tem foco em um tema específico. As referências para cada volume não seguem necessariamente uma ordem sequencial. A Tabela 1 possui o detalhamento dos volumes e cada módulo presente, assim como as respectivas descrições.

A PNS utiliza uma metodologia de amostra probabilística. Cada amostra está descrita a nível de grão pessoa e possui um atributo categórico que informa se a pessoa tem depressão ou não. A base de dados está disponível para download através do portal do IBGE[2], ou a partir do um pacote PNSIBGE[3], disponível na linguagem R.

Nós geramos um script utilizando a linguagem R para coleta da base de dados da PNS do ano de 2019 por meio do pacote PNSIBGE. Essa base de dados foi armazenada junto com o dicionário de dados, arquivo necessário para tradução dos valores de cada atributo e os respectivos códigos. Por exemplo, o atributo V0001 corresponde a "Unidade Federativa"e o valor 26 deste atributo significa "Pernambuco".

Na fase de processamento (Figura 1), com os arquivos da base de dados da PNS e o dicionario de dados, foi possível gerar um novo script na linguagem Python para realizar a tradução automática dos atributos e possíveis valores. Os scripts de coleta dos dados e tradução estão disponíveis em um repositório do Github[4]. Já a base de dados da PNS no formato traduzido pode ser consultado em um repositório do Kaggle[5].

[2] https://www.ibge.gov.br/estatisticas/sociais/saude/9160-pesquisa-nacional-de-saude.html
[3] https://cran.r-project.org/web/packages/PNSIBGE/readme/README.html
[4] https://github.com/filipecmedeiros/PNS
[5] https://www.kaggle.com/filipemedeiros/pns-2019

Tabela 1. PNS divida em volumes e módulos.

Volume	Módulo	Descrição	Decisão
1	A	Informações do domicílio	Remover
	B	Visitas domiciliares de Equipe de Saúde	Remover
	C	Características gerais dos moradores	Manter
	D	Características de educação	Manter
	E	Características de trabalho e rendimento	Manter
	F	Rendimentos	Manter
	I	Cobertura de Plano de Saúde	Remover
	J	Utilização de Serviços de Saúde	Remover
2	H	Atendimento médico e de saúde	Remover
	W	Informações antropométricas	Remover
4	N	Percepção do estado de saúde	Remover
	P	Estilos de Vida: habitos de alimentação, prática de atividade física, uso de bebidas alcóolicas e fumo	Manter
	Q	Doenças crônicas	Remover
	U	Saúde Bucal	Remover
5	O	Acidentes: trânsito e trabalho	Remover
	V	Violência	Manter
	T	Doenças Transmissíveis	Remover
	Y	Atividade Sexual	Remover
	M	Características do trabalho e apoio social do morador	Manter

Fonte: IBGE (2021).

Ainda na fase de processamento, foi desenvolvido um script na linguagem Python que realizou três tarefas na base de dados da PNS traduzida: (1) limpeza dos dados, isto é, a remoção de amostras sem respostas para depressão; (2) seleção de atributos relacionados à depressão; e (3) transformação de atributos numéricos em categóricos por intervalos inter-quartis. Os dados foram limitados para pessoas entre 15 e 29 anos.

Para a etapa de seleção de atributos, foi necessário analisar cada módulo para filtrar os temas que fugiam da proposta deste trabalho, ou seja, investigar grupos de características socioeconômicas e de estilo de vida relacionadas a depressão entre jovens brasileiros. Baseado nesses critérios, a decisão de filtrar cada módulo está descrita na Tabela 1.

Ao final da etapa de processamento, a base de dados gerada era formada por 17.709 pessoas entre 15 e 29 anos, sendo 926 delas diagnosticadas com depressão, representando uma taxa de 5,2%. Cada pessoa na base foi descrita por 95 atributos.
Por fim, na fase de mineração de subgrupos (Figura 1), os dados gerados pela fase de processamento foram submetidos ao algoritmo SSDP+ com k=10 e parâmetro de similaridade máxima entre os subgrupos configurado para 10%. O SSDP+ foi executado com foco primeiramente na métrica WRAcc e em seguida para a métrica Qg. Os itens com valores NULL (dados faltante) foram ignorados pelo algoritmo. Os resultados foram divididos em diferentes execuções: primeiramente na população geral jovem, entre 15 e 29 anos e posteriormente essa mesma base foi subdividida entre gênero (masculino e feminino) e idade, sendo população mais jovem entre 15 e 20 anos e a população menos jovem entre 21 e 29 anos.

Todas as tabelas de resultados, tanto para as execuções com foco no WRAcc quanto para o Qg, estão disponíveis integralmente em um repositório do Kaggle[6].

4. Resultados e Discussão

Nesta sessão serão apresentados os resultados obtidos a partir da execução da metodologia proposta na base de dados. Cada tabela possui os melhores 10 subgrupos de cada execução, com o ID de referência, a descrição do subgrupo, o número de pessoas com depressão (Sim), o número de pessoas sem depressão (Não) e o lift associado, respectivamente.

4.1 Geral

A partir da Tabela 2 é possível identificar os principais subgrupos da execução com foco no WRAcc na base de dados, onde a taxa de pessoas com depressão foi de 5,2%. Estes subgrupos são responsáveis por descrever 92,98% de todos os casos de depressão. É possível destacar que o subgrupo 1 envolveu o maior número de pessoas e está relacionado à mulher, com 689 casos de depressão, enquanto que atributos relacionados à violência emergem com maior lift, como nos subgrupos 2 e 3. Por exemplo: o subgrupo 3 descreve pessoas que sofreram abuso sexual e tem o lift de 3,68. Em outras palavras, esse subgrupo tem uma taxa de depressão de 268% maior que o comportamento padrão. Além destes, metade dos subgrupos possuem atributos relacionados à vida urbana (2, 5, 8, 9 e 10), mas com lift menor. Por exemplo: o subgrupo 10 apresenta um lift de 1,80 e descreve pessoas que moram na capital e a maior escolaridade alcançada foi o ensino superior completo.

Tabela 2. Resultado Geral – WRAcc.

ID	Descrição	Sim	Não	Lift
1	Mulher	689	8562	1.42
2	Vive na cidade, Sofreu humilhação recentemente	237	1472	2.65
3	Sofreu abuso sexual	153	643	3.68
4	Viu propagandas contra cigarros, Rendimento domiciliar per capita superior a R$1200	172	1824	1.65
5	Vive na cidade, Come doces diariamente	144	1498	1.68
6	Faz faxina pesada, Tem acesso a local publico para praticar exercícios físicos	133	1314	1.76
7	Nao costuma comer peixe, Viu propagandas de cigarros	145	1598	1.59
8	Vive na cidade, Fez trabalho voluntário algumas vezes recentemente	105	970	1.87
9	Vive na cidade, Fuma diariamente	106	1014	1.81
10	Mora na capital, Maior escolaridade alcançada foi o ensino superior completo	98	942	1.80

Fonte: Autores (2021).

Quando considerado os resultados da mesma base em relação à métrica Qg, o lift varia entre 7,7 e 16,73. Entretanto, estes subgrupos são muito específicos e com baixa significância estatística, tendo o tamanho da amostra variando entre 7 e 29 ocorrências de depressão. A cobertura de casos de depressão para todos os subgrupos do Qg é de 8,32%. Ainda assim, a

[6] https://www.kaggle.com/filipemedeiros/resultados-minerao-de-subgrupos-pns

ocorrência da depressão em associação com atributos de violência se torna evidente, ocorrendo em todos os 10 melhores subgrupos do Qg. Por exemplo, apesar de cobrir apenas 8 pessoas, o subgrupo que descreve pessoas brancas, que frequentam escola, sofreram xingamentos, sofreram abuso sexuais e possui rendimento domiciliar maior que R$3.700, possui uma taxa de depressão quase 17 vezes maior em relação ao comportamento padrão.

4.2 Depressão e os gêneros

Apesar do gênero estar relativamente equilibrado na base, sendo 47,76% a população masculina e 52,24% a população feminina, o total de casos de depressão masculina representa apenas 25,59% dos casos totais.

Tabela 3. Resultado por genero masculino – WRAcc.

ID	Descrição	Sim	Não	Lift
1	Vive na cidade, Pessoa branca, Solteiro(a), Nao fuma	99	1717	1.95
2	Vive na cidade, Sofreu humilhação recentemente	65	656	3.22
3	Pessoa branca	115	2512	1.56
4	Não auxiliou crianças, idosos ou enfermos, Tem acesso a local público para praticar exercícios físicos, Viu informações pela internet	48	700	2.29
5	Nao realizou tarefas domésticas em domicílio de parentes, Fuma diariamente	54	931	1.96
6	Nao mora com cônjuge, Fumou diariamente	38	578	2.20
7	Come doces diariamente, Tem acesso a local publico para praticar´ exercícios físicos	34	488	2.32
8	Nao assiste TV, Tem acesso a local público para praticar exercícios físicos	37	606	2.05
9	Mora com até 2 pessoas, Rendimento domiciliar superior a R$3994	30	397	2.51
10	Nao costuma comer peixe, Alto consumo de sal	36	681	1.79

Fonte: Autores (2021).

A Tabela 3 mostra os resultados do WRAcc obtidos para a população masculina, que tem uma taxa de depressão de 2,8% na base de dados. A porcentagem de cobertura nestes subgrupos representa 86,50% dos casos de depressão. Assim como nos resultados gerais, atributos relacionados à violência, têm o maior lift, como no subgrupo 2. Este subgrupo representa homens que vivem na cidade e sofreram humilhação recentemente e tem lift de 3.22, isto é, um aumento de 222% em relação a taxa padrão. Alguns hábitos relacionados à alimentação surgem nos subgrupos 7 e 10, o qual retrata a população masculina que não costuma comer peixe e tem alto consumo de sal.

Com referência à métrica Qg, os resultados da população masculina possuem lift variando entre 9.99 e 35.69. A cobertura é de 15,19% dos casos de depressão. Os subgrupos mantém alguns aspectos dos resultados gerais, como a baixa abrangência de casos e a ocorrência de atributos relacionados à violência. Por exemplo: o subgrupo de maior lift caracteriza homens que sofreram abuso sexual, não moram com o cônjuge, podem contar com até duas pessoas em momentos ruins, consomem sal de maneira adequada e não viram propaganda de cigarros, todavia este subgrupo representa apenas 5 pessoas. Também vale ressaltar na execução da base de dados masculina com foco em Qg a presença de atributos associados aos hábitos de fumar e de alimentação. Por exemplo: o subgrupo da população masculina que mora com até uma pessoa, substitui almoço por lanches em cinco dias da semana e fuma diariamente possui lift de 13,38 e cobertura de 8 pessoas.

A população feminina, o qual possui uma taxa de depressão de 7,4%, tem os resultados descritos na Tabela 4. Os subgrupos descritos possuem uma porcentagem de cobertura de 75,04%. Novamente, subgrupos que possuem atributos associados à algum tipo de violência possuem lift elevado, a exemplo dos subgrupos 1, 2 e 8. No subgrupo 8, a pessoa mora com até 3 pessoas e teve algo destruído de propósito, representando um aumento de 201% em relação à taxa de depressão normal. É também notório que apesar de ter uma cobertura percentual menor em relação aos subgrupos da população masculina, os subgrupos da Tabela 4 descrevem uma parcela maior da população geral com depressão.

Tabela 4. Resultado por genero feminino – WRAcc.

ID	Descrição	Sim	Não	Lift
1	Sofreu xingamentos	250	1385	2.05
2	Nao recebe pensão do governo, Sofreu abuso sexual	99	275	3.55
3	Pessoa branca, Nao recebe pensão do governo, Pratica exercício físico	156	1181	1.57
4	Faz faxina pesada, Mora com até 3 pessoas	122	847	1.69
5	Assiste TV por menos de uma hora, Viu propagandas contra cigarros	104	703	1.73
6	Fez trabalho voluntário algumas vezes recentemente	88	612	1.69
7	Fuma diariamente, Viu propagandas contra cigarros	59	298	2.22
8	Teve algo destruído de propósito, Mora com até 3 pessoas	41	142	3.01
9	Alto consumo de sal, Viu propagandas contra cigarros	68	492	1.63
10	Come verduras ou legumes diariamente, Viu propagandas de cigarros	60	393	1.78

Fonte: Autores (2021).

Os resultados com foco em Qg na população feminina descrevem 11,61% dos casos de depressão, dos quais variam o lift entre 5.81 e 12.39. Atributos associados à violência voltam a ocorrer em todos os subgrupos, frequência ainda maior em comparação aos resultados da métrica Qg para a base de dados masculina. Como exemplo, o subgrupo de maior lift também tem baixa cobertura (13 pessoas). Este subgrupo descreve as mulheres que moram com até 3 pessoas, não moram com a mãe, tem acesso a local público para praticar exercícios físicos, sofreram violência física e sofreram abuso sexual.

4.3 Mais e menos jovens

A população mais jovem, entre 15 e 20 anos, representa 30,74% do total da população jovem na base de dados. Essa parcela da população corresponde a 24,51% dos casos de depressão.

Tabela 5. Resultado mais jovens – WRAcc.

ID	Descrição	Sim	Não	Lift
1	Vive na cidade, Mulher, Não recebe aluguel	138	1944	1.59
2	Solteiro(a), Sofreu humilhação recentemente	51	389	2.78
3	Viu propaganda de cigarros, Viu informações pela internet	44	379	2.49
4	Come verduras ou legumes diariamente, Mora com até 3 pessoas	41	444	2.03

5	Solteiro(a), Sofreu abuso sexual	21	54	6.71
6	Não pratica atividades religiosas, Passa 6 horas ou mais de lazer no computador ou celular	42	559	1.68
7	Come doces diariamente, Viu propagandas contra cigarros	33	372	1.95
8	Faz tarefas domesticas, Viu propaganda de cigarros, Rendimento domiciliar entre R$1100 e R$3598	29	302	2.10
9	Vive na cidade, Não recebe pensão do governo, Maior escolaridade alcançada foi o ensino superior incompleto	30	325	2.03
10	Vive na cidade, Faz faxina pesada	33	397	1.84

Fonte: Autores (2021).

A Tabela 5 descreve os subgrupos da execução com foco no WRAcc para a população mais jovem, que tem uma taxa de depressão de 4,2%. Os subgrupos contemplam 90,31% dos casos de depressão. É possível notar uma semelhança com os resultados gerais, com tendências da vida urbana nos subgrupos 1, 9 e 10 e lift elevado em atributos de violência, nos subgrupos 2 e 5. O subgrupo 5, que caracteriza o grupo de pessoas solteiras e que sofreram abuso sexual, representa um crescimento de 571% da taxa de depressão entre a população mais jovem.

Atributos associados à violência ocorrem em todos os principais subgrupos dos resultados com foco na métrica Qg. Geralmente combinados com hábitos de alimentação, tais subgrupos variam o lift entre 9.59 e 21.31. Por exemplo: o subgrupo de maior lift refere ao grupo de pessoas que vivem na cidade, solteiras, fazem tarefas domésticas, consomem sal de forma adequada, sofreram ameaças e sofreram abuso sexual. Os subgrupos do Qg representam 18,50% dos casos totais de depressão.

Tabela 6. Resultado menos jovens – WRAcc.

ID	Descrição	Sim	Não	Lift
1	Mulher	524	5945	1.42
2	Vive na cidade, Sofreu humilhação recentemente, Sofreu xingamentos	164	919	2.66
3	Não fuma, Sofreu abuso sexual	128	523	3.45
4	Vive na cidade, Pessoa branca, Rendimento domiciliar per capita superior a R$1300	144	1297	1.75
5	Mora com até 2 pessoas, Não mora com cônjuge, Não costuma comer peixe	108	1061	1.62
6	Mulher, Fez trabalho voluntário algumas vezes recentemente	67	421	2.41
7	Vive na cidade, Convive com fumante	100	983	1.62
8	Sofreu xingamentos, Maior escolaridade alcançada foi o ensino superior completo	55	252	3.14
9	Mora com até 2 pessoas, Solteiro(a), Cuida de pet	103	1089	1.52
10	Mulher, Não mora com cônjuge, Faz faxina pesada	62	461	2.08

Fonte: Autores (2021).

Por fim, a Tabela 6 descreve os resultados com foco no WRAcc para a população menos jovem da base de dados, que tem uma taxa de depressão de 5,7%. Estes subgrupos descrevem 90,56% dos casos de depressão. É possível verificar a

continuidade de padrões nos resultados em relação às execuções anteriores, com atributos relacionados à vida urbana ocorrendo nos subgrupos 2, 4 e 7 e atributos relacionados à algum tipo de violência apresentando alto lift, nos subgrupos 2, 3 e 8.

Os subgrupos revelados pela execução com foco em Qg tem cobertura de 10,01%. O menor lift entre estes subgrupos é de 8.57, em contrapartida do maior, com 14.26. Todos os subgrupos possuem associação com algum tipo de violência, tal como o subgrupo que descreve mulheres que moram com até 2 pessoas, não praticam atividades religiosas e sofreram abuso sexual.

5. Conclusão e Trabalhos Futuros

Este trabalho tem como principal objetivo investigar os grupos de características socioeconômicas e de estilo de vida relacionadas a depressão entre jovens brasileiros. Para isso, foi utilizado a base de dados PNS em conjunto com o algoritmo SSDP+ para aplicar a mineração de dados com diferentes métricas e encontrar possíveis regras de associação entre os dados disponíveis.

Conclui-se ao aplicar o método proposto que grupos de características conseguem contemplar os casos de depressão tanto em abrangência, como em precisão. Estes grupos de características emergem principalmente em pessoas que sofreram algum tipo de violência e nas mulheres. Outros grupos menores também se destacam, mas com menor lift e abrangência, como por exemplo as características da vida urbana.

Como trabalhos futuros, nós consideramos comparar os resultados obtidos neste trabalho com a execução da mesma metodologia na edição da PNS de 2013. Além disso, consideramos utilizar outras técnicas de inteligência artificial nas bases de dados. Dessa forma, é possível obter outras perspectivas e consequentemente um melhor conhecimento sobre quais fatores ambientais estão mais associados com a depressão nos jovens brasileiros.

Referências

Atzmueller, M. (2015). Subgroup discovery. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 5(1), 35-49.

Azuelo, N. C. S., de Souza Filho, Z. A., das Neves, A. L. M., de Oliveira Ferreira, B., de Lima Oliveira, D., & Tavares, N. K. C. (2020). Prevalência de depressão em pessoas que vivenciaram violência por parceiro íntimo: revisão sistemática com meta-análise. Research, Society and Development, 9(8), e84985094-e84985094.

Barger, S. D., Messerli-Bürgy, N., & Barth, J. (2014). Social relationship correlates of major depressive disorder and depressive symptoms in Switzerland: nationally representative cross sectional study. BMC public health, 14(1), 1-10.

Brito, I. (2011). Ansiedade e depressão na adolescência. Revista Portuguesa de Medicina Geral e Familiar, 27(2), 208-14.

Carmona, C. J., González, P., del Jesus, M. J., Navío-Acosta, M., & Jiménez-Trevino, L. (2011). Evolutionary fuzzy rule extraction for subgroup discovery in a psychiatric emergency department. Soft Computing, 15(12), 2435-2448.

Carvalho, V. P. D. S. (2016). Análise da relação entre o estilo de vida da população economicamente ativa e a prevalência da depressão (Master's thesis, Universidade Federal de Pernambuco).

Daimi, K., & Banitaan, S. (2014). Using data mining to predict possible future depression cases. International Journal of Public Health Science (IJPHS), 3(4), 231-240.

da Rocha, A. C. B., Myva, L. M. M., & de Almeida, S. G. (2020). O papel da alimentação no tratamento do transtorno de ansiedade e depressão. Research, Society and Development, 9(9), e724997890-e724997890.

Gonçalves, C., Ferreira, D., Neto, C., Abelha, A., & Machado, J. (2020). Prediction of Mental Illness Associated with Unemployment Using Data Mining. Procedia Computer Science, 177, 556-561.

Helal, S., Li, J., Liu, L., Ebrahimie, E., Dawson, S., & Murray, D. J. (2019). Identifying key factors of student academic performance by subgroup discovery. International Journal of Data Science and Analytics, 7(3), 227-245.

Herrera, F., Carmona, C. J., González, P., & Del Jesus, M. J. (2011). An overview on subgroup discovery: foundations and applications. Knowledge and information systems, 29(3), 495-525.

Hullam, G., Antal, P., Petschner, P., Gonda, X., Bagdy, G., Deakin, B., & Juhasz, G. (2019). The UKB envirome of depression: From interactions to synergistic effects. Scientific reports, 9(1), 1-19.

IBGE (2021) Pesquisa Nacional da Saúde (PNS) O que é. Recuperado em novembro, 26, 2021, em https://www.ibge.gov.br/estatisticas/sociais/saude/9160-pesquisa-nacional-de-saude.html?=$t=o-que-e

Islam, M. R., Kabir, M. A., Ahmed, A., Kamal, A. R. M., Wang, H., & Ulhaq, A. (2018). Depression detection from social network data using machine learning techniques. Health information science and systems, 6(1), 1-12.

Justo, L. P., & Calil, H. M. (2006). Depressão: o mesmo acometimento para homens e mulheres?. Archives of Clinical Psychiatry (São Paulo), 33, 74-79.

Lavrač, N., Flach, P., & Zupan, B. (1999, June). Rule evaluation measures: A unifying view. In International Conference on Inductive Logic Programming (pp. 174-185). Springer, Berlin, Heidelberg.

Lucas, T., Vimieiro, R., & Ludermir, T. (2018, July). SSDP+: A diverse and more informative subgroup discovery approach for high dimensional data. In 2018 IEEE Congress on Evolutionary Computation (CEC) (pp. 1-8). IEEE.

OMS (2017) Depression and Other Common Mental Disorders. Recuperado em novembro, 30, 2021 em https://www.who.int/publications/i/item/depression-global-health-estimates

Pasini, A. L. W., da Silveira, F. L., da Silveira, G. B., Busatto, J. H., Pinheiro, J. M., Leal, T. G., ... & Carlesso, J. P. P. (2020). Suicídio e depressão na adolescência: fatores de risco e estratégias de prevenção. Research, Society and Development, 9(4), e36942767-e36942767.

Tufféry, S. (2011). Data mining and statistics for decision making. John Wiley & Sons.

Wrobel, S. (1997, June). An algorithm for multi-relational discovery of subgroups. In European symposium on principles of data mining and knowledge discovery (pp. 78-87). Springer, Berlin, Heidelberg.