

ANÁLISE DE PERFIS SOCIAIS E COMPORTAMENTAIS ASSOCIADOS A SINTOMAS DEPRESSIVOS (PHQ-9) NA PNS 2019

Autores: Mateus Grandel, Giovanni Urbani, Eduardo Camargo

Data: 04 Dezembro de 2025

RESUMO

O presente trabalho aplica técnicas de aprendizado de máquina não supervisionado para investigar a alta prevalência de depressão na população brasileira. Utilizando dados da Pesquisa Nacional de Saúde (PNS) de 2019, que inclui o questionário PHQ-9 (Patient Health Questionnaire-9), o projeto teve como objetivo identificar perfis sociais, econômicos e comportamentais estatisticamente associados a diferentes níveis de sintomas depressivos.

Foram aplicados o algoritmo de **Agrupamento K-Means** para segmentar a população em perfis homogêneos e o algoritmo **Apriori** de **Regras de Associação** para descobrir padrões de co-ocorrência entre variáveis sociodemográficas, de saúde e as categorias de gravidade do PHQ-9. Os resultados da mineração de regras apontaram fortes associações ($Lift > 2.0$) entre o perfil de **mulheres de baixa renda e sedentárias** e a presença de **depressão moderada** e **consumo frequente de álcool**, oferecendo *insights* diretos para a formulação de políticas públicas focadas em grupos vulneráveis.

INTRODUÇÃO

A depressão é reconhecida como um dos maiores problemas de saúde pública no Brasil, afetando milhões de pessoas e impondo custos sociais, emocionais e econômicos expressivos. Frequentemente, a depressão acompanha a desigualdade social, a baixa escolaridade e o acesso limitado a serviços de saúde.

A Pesquisa Nacional de Saúde (PNS) de 2019, com seu grande volume de dados comportamentais, socioeconômicos e emocionais, incluindo as respostas ao PHQ-9, viabiliza a aplicação de métodos avançados de Mineração de Dados. A análise tradicional de dados de saúde costuma focar em estatísticas descritivas, deixando lacunas na identificação de padrões complexos que emergem de forma mais eficiente com o uso de técnicas não supervisionadas, como o K-Means e o Apriori.

O propósito deste estudo é preencher essa lacuna, aplicando o Apriori para revelar conjuntos de variáveis (antecedentes) que predizem com alta probabilidade a ocorrência de categorias de gravidade do PHQ-9 (consequentes), como Depressão Moderada, Moderadamente Grave ou Grave.

METODOLOGIA

Base de Dados e Pré-processamento

Os dados foram extraídos da PNS 2019. Após a etapa de pré-processamento, a base de dados finalizou com 47.346 linhas (registros) e 52 colunas (variáveis).

O pré-processamento envolveu:

- Remoção de *missing values*.
- Criação do Score_PHQ9 (soma dos pontos do questionário).
- Categorização da Gravidade do PHQ-9 em: Mínima (0-4), Leve (5-9), Moderada (10-14), Moderadamente Grave (15-19) e Grave (20-27).
- One-Hot Encoding (OHE) para as variáveis categóricas (e.g., Tipo de área, Sexo, Estado Civil).
- Ordenação e Padronização para variáveis ordinais (e.g., Escolaridade, Renda) e numéricas (e.g., Idade).

As variáveis utilizadas no projeto abrangem:

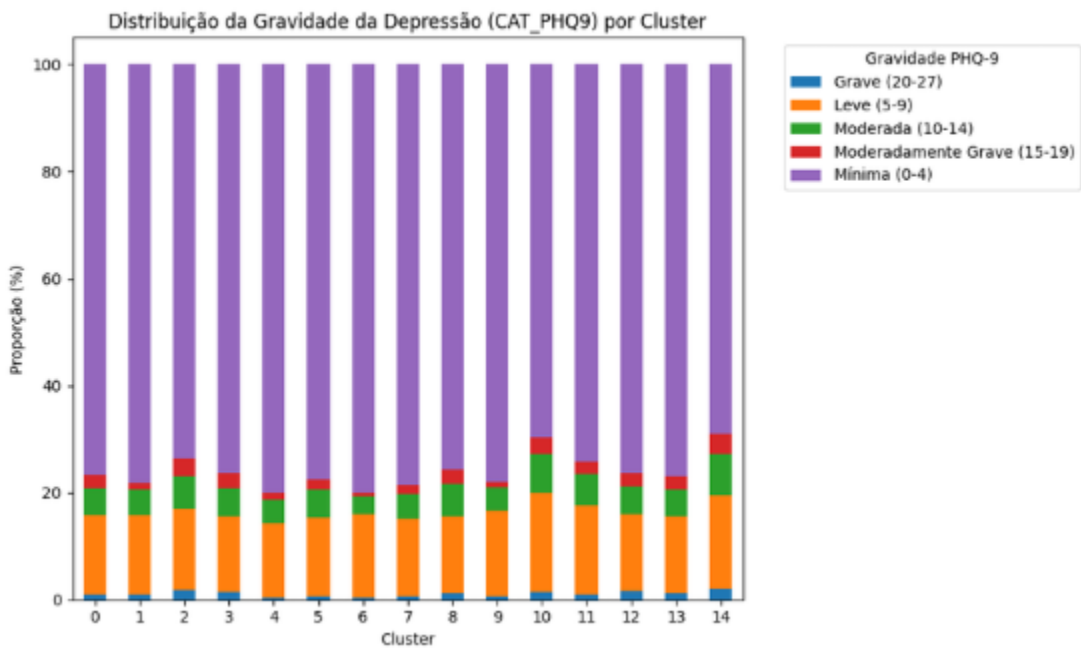
- Demográficas e Geográficas (Tipo de área, Sexo, Idade, Etnia, Estado Civil).
- Saúde e Comportamento (Autoavaliação de saúde mental, Consumo de álcool/tabaco, Hipertensão, Diabetes, Problemas na coluna, Prática de exercício, Rede de apoio).
- Socioeconômicas (Nível de instrução, Renda domiciliar per capita, Renda individual, Acesso à internet).

Algoritmos e Parâmetros

O projeto empregou uma abordagem dupla:

No Agrupamento (K-Means) segmentamos a base em perfis sociais. O melhor resultado foi alcançado com K=15 *Clusters* e PCA com 8 componentes, otimizando a proporcionalidade entre os registros e a diferença nas categorias do PHQ-9 entre os

grupos. Abaixo segue o gráfico obtido dessas proporções e a tabela com as porcentagens descritas:



Cluster	Mínima (0-4) (%)	Leve (5-9) (%)	Moderada (10-14) (%)	Moderadamente Grave (15-19) (%)	Grave (20-27) (%)	% PHQ9 ≥ 10 (Soma Risco) (%)
6	79,99	15,6	3,46	0,66	0,29	4,41%
9	78	16,04	4,28	1,05	0,64	5,97%
4	80,12	13,88	4,39	1,19	0,42	6,00%
7	78,65	14,6	4,47	1,72	0,56	6,75%
1	78,17	14,74	4,74	1,34	1,01	7,09%
5	77,5	14,69	5,26	1,89	0,66	7,81%
0	76,68	14,75	4,93	2,56	1,08	8,57%
13	77,01	14,36	4,95	2,54	1,15	8,64%
11	74,25	16,72	5,86	2,23	0,94	9,03%
12	76,27	14,37	5,07	2,65	1,64	9,36%
3	76,37	14,11	5,32	2,84	1,36	9,52%
8	75,75	14,29	5,98	2,73	1,26	9,97%
2	73,57	15,18	6	3,43	1,82	11,25%
10	69,75	18,53	7,45	2,95	1,32	11,72%
14	69,01	17,45	7,68	3,77	2,1	13,55%

Utilizamos as Regras de Associação (algoritmo Apriori) para identificar as combinações de fatores que se associam às categorias de depressão. Abaixo segue as regras e métricas obtidas:

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
frozenset({'Exercicio=Não', 'Sexo=Mulher', 'Renda=Baixa'})	frozenset({'Alcool=Frequente', 'PHQ9=Moderada (10-14)'})	0.134798953	0.035647913	0.010136847	0.0751599749	2.109513347
frozenset({'Sexo=Mulher', 'Tabaco=NaoFuma', 'Renda=Baixa', 'Idade=Adulto'})	frozenset({'Alcool=Frequente', 'PHQ9=Moderada (10-14)'})	0.152960804	0.035647913	0.011108295	0.072621842	2.0317197543
frozenset({'Sexo=Mulher', 'Internet=Sim', 'Renda=Baixa', 'Idade=Adulto'})	frozenset({'Alcool=Frequente', 'PHQ9=Moderada (10-14)'})	0.142887312	0.035647913	0.010326913	0.07227313	2.027415444
frozenset({'Sexo=Mulher', 'Renda=Baixa', 'Idade=Adulto'})	frozenset({'Alcool=Frequente', 'PHQ9=Moderada (10-14)'})	0.169327589	0.035647913	0.012206454	0.072087802	2.022122166
frozenset({'Sexo=Mulher', 'Tabaco=NaoFuma', 'Renda=Baixa'})	frozenset({'Alcool=Frequente', 'PHQ9=Moderada (10-14)'})	0.185757729	0.035647913	0.013346849	0.071850841	2.015569335
frozenset({'Sexo=Mulher', 'Tabaco=NaoFuma', 'Internet=Sim', 'Renda=Baixa'})	frozenset({'Alcool=Frequente', 'PHQ9=Moderada (10-14)'})	0.157026662	0.035647913	0.011256124	0.071678322	2.010727992
frozenset({'Sexo=Mulher', 'Renda=Baixa'})	frozenset({'Alcool=Frequente', 'PHQ9=Moderada (10-14)'})	0.204447542	0.035647913	0.014592837	0.071376924	2.00227494
frozenset({'Sexo=Mulher', 'Internet=Sim', 'Renda=Baixa'})	frozenset({'Alcool=Frequente', 'PHQ9=Moderada (10-14)'})	0.171523906	0.035647913	0.012185335	0.071041615	1.992868822
frozenset({'Internet=Sim', 'Exercicio=Não', 'Tabaco=NaoFuma', 'Idade=Adulto', 'Sexo=Mulher'})	frozenset({'Alcool=Frequente', 'PHQ9=Moderada (10-14)'})	0.159148505	0.035647913	0.01150532	0.070063694	1.965436049
frozenset({'Exercicio=Não', 'Sexo=Mulher', 'Internet=Sim', 'Idade=Adulto'})	frozenset({'Alcool=Frequente', 'PHQ9=Moderada (10-14)'})	0.176993358	0.035647913	0.01239652	0.070039375	1.964753836
frozenset({'Sexo=Mulher', 'Internet=Sim', 'Renda=Baixa'})	frozenset({'PHQ9=Moderada (10-14)', 'Tabaco=NaoFuma', 'Idade=Adulto'})	0.171523906	0.037020612	0.012016388	0.070056636	1.892368421
frozenset({'Exercicio=Não', 'Sexo=Mulher', 'Renda=Baixa'})	frozenset({'PHQ9=Moderada (10-14)', 'Idade=Adulto'})	0.134798953	0.043081602	0.010897111	0.080839731	1.876432804
frozenset({'Exercicio=Não', 'Sexo=Mulher', 'Internet=Sim', 'Renda=Baixa'})	frozenset({'PHQ9=Moderada (10-14)'})	0.110407163	0.053788647	0.010939348	0.0909081867	1.842059113
frozenset({'Sexo=Mulher', 'Internet=Sim', 'Renda=Baixa'})	frozenset({'PHQ9=Moderada (10-14)', 'Idade=Adulto'})	0.171523906	0.043081602	0.01347356	0.078552081	1.823324168
frozenset({'Exercicio=Não', 'Local=Capital', 'Sexo=Mulher'})	frozenset({'PHQ9=Moderada (10-14)'})	0.102572225	0.053788647	0.010052374	0.098002382	1.821999407
frozenset({'Sexo=Mulher', 'Internet=Sim', 'Renda=Baixa', 'Idade=Adulto'})	frozenset({'PHQ9=Moderada (10-14)', 'Tabaco=NaoFuma'})	0.142887312	0.046376077	0.012016388	0.084096955	1.813369322
frozenset({'Sexo=Mulher', 'Internet=Sim', 'Renda=Baixa'})	frozenset({'PHQ9=Moderada (10-14)', 'Tabaco=NaoFuma'})	0.171523906	0.046376077	0.014402771	0.083969466	1.810620281
frozenset({'Exercicio=Não', 'Sexo=Mulher', 'Renda=Baixa', 'Idade=Adulto'})	frozenset({'PHQ9=Moderada (10-14)'})	0.112012164	0.053788647	0.010897111	0.097285068	1.808654312
frozenset({'Alcool=Frequente', 'Sexo=Mulher', 'Renda=Baixa'})	frozenset({'PHQ9=Moderada (10-14)', 'Tabaco=NaoFuma'})	0.134270992	0.046376077	0.011256124	0.083831394	1.807643054
frozenset({'Exercicio=Não', 'Sexo=Mulher', 'Renda=Baixa'})	frozenset({'PHQ9=Moderada (10-14)', 'Tabaco=NaoFuma'})	0.134798953	0.046376077	0.011213887	0.083189723	1.793806808
frozenset({'Exercicio=Não', 'Sexo=Mulher', 'Renda=Baixa'})	frozenset({'PHQ9=Moderada (10-14)'})	0.134798953	0.053788647	0.012966717	0.096193013	1.788351605
frozenset({'Alcool=Frequente', 'Sexo=Mulher', 'Renda=Baixa', 'Idade=Adulto'})	frozenset({'PHQ9=Moderada (10-14)', 'Tabaco=NaoFuma'})	0.1341654	0.046376077	0.011108295	0.08279553	1.785308886
frozenset({'Alcool=Frequente', 'Sexo=Mulher', 'Internet=Sim', 'Renda=Baixa'})	frozenset({'PHQ9=Moderada (10-14)', 'Idade=Adulto'})	0.134270992	0.043081602	0.010326913	0.076910978	1.785239531
frozenset({'Sexo=Mulher', 'Renda=Baixa'})	frozenset({'PHQ9=Moderada (10-14)', 'Idade=Adulto'})	0.204447542	0.043081602	0.015690995	0.07674827	1.78146278
frozenset({'Sexo=Mulher', 'Tabaco=NaoFuma', 'Internet=Sim', 'Renda=Baixa'})	frozenset({'PHQ9=Moderada (10-14)', 'Idade=Adulto'})	0.157036662	0.043081602	0.012016388	0.076519634	1.776155745
frozenset({'Alcool=Frequente', 'Sexo=Mulher', 'Renda=Baixa'})	frozenset({'PHQ9=Moderada (10-14)', 'Tabaco=NaoFuma'})	0.162590809	0.046376077	0.013346849	0.082088583	1.770063105
frozenset({'Sexo=Mulher', 'Renda=Baixa'})	frozenset({'PHQ9=Moderada (10-14)', 'Tabaco=NaoFuma'})	0.204447542	0.046376077	0.016725798	0.08180973	1.764050252
frozenset({'Sexo=Mulher', 'Renda=Baixa', 'Idade=Adulto'})	frozenset({'PHQ9=Moderada (10-14)', 'Idade=Adulto'})	0.169327589	0.046376077	0.013790336	0.081441756	1.75611568
frozenset({'Sexo=Mulher', 'Internet=Sim', 'Renda=Baixa', 'Idade=Adulto'})	frozenset({'PHQ9=Moderada (10-14)'})	0.142887312	0.053788647	0.01347356	0.094295004	1.735005194
frozenset({'Internet=Sim', 'Alcool=Frequente', 'Exercicio=Não', 'Idade=Adulto', 'Sexo=Mulher'})	frozenset({'PHQ9=Moderada (10-14)', 'Tabaco=NaoFuma'})	0.137586586	0.046376077	0.011150532	0.081043745	1.747533435
frozenset({'Sexo=Mulher', 'Internet=Sim', 'Renda=Baixa'})	frozenset({'PHQ9=Moderada (10-14)'})	0.171523906	0.053788647	0.016092245	0.093819256	1.744220426
frozenset({'Alcool=Frequente', 'Sexo=Mulher', 'Renda=Baixa'})	frozenset({'PHQ9=Moderada (10-14)', 'Idade=Adulto'})	0.162590809	0.043081602	0.012206454	0.075074685	1.742615924
frozenset({'Exercicio=Não', 'Sexo=Mulher', 'Internet=Sim'})	frozenset({'PHQ9=Moderada (10-14)', 'Idade=Adulto'})	0.217878865	0.043081602	0.016345666	0.075021809	1.741388571
frozenset({'Exercicio=Não', 'Sexo=Mulher', 'Renda=Baixa'})	frozenset({'PHQ9=Moderada (10-14)', 'Internet=Sim'})	0.134798953	0.046671735	0.010939348	0.081153063	1.738805353

Parâmetros do Apriori

Para mitigar os desafios de volume de dados e complexidade computacional, o algoritmo Apriori foi aplicado com os seguintes parâmetros definidos na metodologia:

Métrica	Parâmetro Desejado	Uso no Projeto
Suporte Mínimo	~2.0%	Define a frequência mínima de ocorrência do conjunto de itens.
Confiança Mínima	~7.0%	Define a probabilidade do consequente ocorrer, dada o antecedente.
Lift	> 1.0	Utilizado como filtro primário para garantir que a regra tem associação real (não aleatória).

RESULTADOS E DISCUSSÃO

Resultados da Clusterização (K-Means)

A análise de agrupamento com K=15 revelou perfis distintos, como, por exemplo:

- **Cluster 14 (RISCO):** Grupo com hábito sedentário e alto consumo de cigarro e bebida alcóolica, representando 7.68%de Moderada, 3.77%de Moderadamente Grave e 2,10% de Grave (Maior taxa de todas).

- **Cluster 10 (RISCO):** Grupo com baixa escolaridade, baixa renda, com vulnerabilidade estrutural e etnia parda, representando 7,45% de moderada e 2,95% de Moderadamente grave.
- **Cluster 2 (RISCO):** Grupo com baixa escolarização e exercícios físicos, representando 6,00% de moderada e 3,43% de Moderadamente grave.
- **Cluster 8 (RISCO):** Grupo de etnia branca, classe média e pouco exercício, representando 7,45% de moderada e 2,95% de Moderadamente grave.
- **Cluster 6 (PROTETIVA):** Grupo de etnia branca, com alto poder econômico, casado, formação alta e praticante de exercícios físicos, tendo as taxas mais baixas de sintomas depressivos.
- **Cluster 4 (PROTETIVA):** Grupo de etnia branca, classe média-alta, praticante de exercícios físicos.

Regras de Associação (Apriori)

As regras de associação foram aplicadas focando nos alvos de PHQ-9 com maior impacto (Score ≥ 10), ou seja, categorias Moderada, Moderadamente Grave e Grave.

As regras mais relevantes encontradas, que demonstram uma alta associação $Lift > 2.0$ com categorias de sintomas depressivos, são apresentadas abaixo:

Regra	Suporte	Confiança	Lift	Interpretação da Regra
'Exercicio=Não', 'Sexo=Mulher', 'Renda=Baixa' > 'Alcool=Freqüente', 'PHQ9=Moderada (10-14)'	0.0101	0.075	2.1	Mulheres de baixa renda e sedentárias apresentam mais que o dobro da probabilidade de ter depressão moderada e consumir álcool frequentemente.
'Sexo=Mulher', 'Tabaco=NaoFuma', 'Renda=Baixa', 'Idade=Adulto'}	0.0111	0.072	2.03	Mulheres adultas de baixa renda apresentam mais que o dobro da probabilidade de ter

> 'Alcool=Frequente', 'PHQ9=Moderada (10-14)'				depressão moderada e consumir álcool frequentemente.
'Sexo=Mulher', 'Renda=Baixa', 'Internet=Sim', 'Idade=Adulto' > 'Alcool=Frequente', 'PHQ9=Moderada (10-14)'	0.0103	0.0722	2.02	Mulheres de renda baixa, com internet disponível e de idade adulta apresentam o dobro de probabilidade de apresentar sintomas depressivos moderados

O valor de Lift acima de 2.0 nas regras é um achado significativo, indicando que a co-ocorrência dos itens no antecedente e no consequente é pelo menos duas vezes maior do que seria esperado por acaso.

As regras reforçam a evidência de que o perfil de Mulher, Baixa Renda, Adulta e a presença de Sedentarismo estão altamente associados ao risco de sintomas depressivos moderados.

CONCLUSÃO E CONTRIBUIÇÕES

O uso combinado de Agrupamento K-Means e Regras de Associação Apriori na base da PNS 2019 permitiu identificar padrões complexos e estatisticamente relevantes associados à sintomas depressivos.

Os resultados das Regras de Associação confirmam que fatores socioeconômicos (Baixa Renda) e comportamentais (Sedentarismo, Consumo de Álcool) possuem uma forte ligação com a presença de sintomas depressivos moderados (PHQ-9 ≥ 10), especialmente em mulheres.

Esses achados podem ser úteis para as seguintes contribuições, conforme a justificativa do projeto:

- **Identificação de Grupos Vulneráveis:** Priorização de políticas de saúde mental focadas em mulheres de baixa renda e seus fatores comportamentais associados (e.g., sedentarismo e consumo de álcool).
- **Formulações de Políticas Públicas:** Criação de estratégias de prevenção que abordem a saúde mental em conjunto com a promoção de exercícios físicos e a moderação do consumo de álcool para o público-alvo identificado.

As técnicas de mineração de dados trouxeram à luz padrões populacionais complexos que reforçam a necessidade de intervenções sociais e de saúde pública direcionadas e interseccionais.

REFERÊNCIAS BIBLIOGRÁFICAS

- KIRKBRIDE, J. B. et al. The social determinants of mental health and disorder: Evidence, prevention and recommendations. *World Psychiatry : Official Journal of the World Psychiatric Association (WPA)*, v. 23, n. 1, p. 58–90, 2024.
- SCHUCH, F. B. et al. Physical activity and incident depression: A Meta-Analysis of prospective cohort studies. *The American Journal of Psychiatry*, v. 175, n. 7, p. 631–648, 2018.
- MREJEN, M.; HONE, T.; ROCHA, R. Socioeconomic and racial/ethnic inequalities in depression prevalence and the treatment gap in Brazil: A decomposition analysis. *SSM - Population Health*, p. 101266, out. 2022.
- ALBERT, P. Why is depression more prevalent in women? *Journal of Psychiatry & Neuroscience*, v. 40, n. 4, p. 219–221, 1 jul. 2015.
- BODEN, J. M.; FERGUSON, D. M. Alcohol and depression. *Addiction*, v. 106, n. 5, p. 906–914, 7 mar. 2011.