

# Análise de Padrões de Burnout e Episódios Depressivos no Ambiente de Trabalho Utilizando Métodos Estatísticos e Aprendizado de Máquina

Nicole V. L. da Costa<sup>1</sup>, Laiane M. de Souza<sup>1</sup>, Leno R. Martins<sup>1</sup>, Marta de O. Barreiros<sup>1</sup>

<sup>1</sup>Centro de Ciências Naturais e Tecnologias - Universidade do Estado do Pará (UEPA)  
Castanhal - PA - Brasil

{nicole.costa, laiane.md.souza}@aluno.uepa.br {marta,  
leno.martins}@uepa.br

**Abstract.** Depression and burnout represent significant public health challenges among Brazilian workers. This study applies statistical and clustering techniques to DATASUS data (2016-2019 and 2022-2025) to identify profiles associated with these disorders. In the overall dataset, a female predominance was observed (~75%), along with a higher number of cases in the Southeast region and a mean age of 40.7 years ( $SD=9.6$ ). As a result, the K-means model with FAMD achieved a silhouette score of 0.418, revealing three profiles that highlight educational, occupational, and territorial patterns in the country's mental illness landscape. Therefore, this study contributes to the understanding and mitigation of this national problem.

**Resumo.** A depressão e o burnout representam importantes desafios de saúde pública entre trabalhadores brasileiros. Este estudo aplica técnicas de estatística e clusterização em dados do DATASUS (2016-2019 e 2022-2025) para identificar perfis associados a esses transtornos. Na totalidade dos dados, observou-se predominância feminina (~75%), maior número de casos na região Sudeste e idade média de 40,7 anos ( $DP=9,6$ ). Como resultado, o modelo K-means com FAMD obteve silhouette score de 0,418, revelando três perfis que evidenciam padrões educacionais, ocupacionais e territoriais no adoecimento mental do país. O estudo contribui, portanto, para a compreensão e mitigação desse problema nacional.

## 1. Introdução

Episódios depressivos e a Síndrome de Burnout são transtornos mentais com impactos crescentes no ambiente de trabalho. Segundo o ministério da saúde, enquanto a depressão envolve tristeza e perda de interesse geral, muitas vezes agravada por frustrações profissionais, o burnout é uma resposta direta ao esgotamento crônico no trabalho, caracterizada por despersonalização e exaustão [Brasil 2001]. No Brasil, os afastamentos por essas condições têm crescido alarmantemente; em 2025, o ministério da Previdência Social registrou mais de 546 mil afastamentos por transtornos mentais, com os casos de burnout triplicando em relação a 2023 [Brasil 2026].

Compreender as particularidades socioambientais desses transtornos no cenário brasileiro exige a análise de bases governamentais de abrangência nacional, como o Sistema de Informações de Agravos de Notificações (SINAN) do DATASUS. Contudo, a complexidade estrutural e a inconsistência frequente nesses registros desafiam abordagens estatísticas tradicionais [Salvadori 2024].

Nesse contexto, técnicas de Machine Learning (ML), especificamente a clusterização, destacam-se por descobrir padrões ocultos e inter-relações em dados não rotulados, viabilizando intervenções mais precisas [De Melo Silva et al. 2025]. Diante disso, este estudo busca aplicar técnicas de clusterização em dados do DATASUS para identificar perfis e padrões subjacentes aos transtornos mentais relacionados ao trabalho no Brasil. Ferramentas de Inteligência Artificial generativa foram utilizadas neste trabalho para fins de correções linguísticas.

## 2. Trabalhos Correlatos

A literatura científica tem apresentado diversos estudos que exploram a análise estatística e a clusterização para compreender a dinâmica de agrupamento de dados. Assim, esta seção contextualiza o estado da arte sobre burnout e depressão ocupacional, além da aplicação de clusterização em bases oficiais, destacando as contribuições deste estudo.

Parte significativa dos trabalhos apoia-se em estatística descritiva para delinear o cenário da saúde mental. No Brasil, [Bailo et al. 2024] utilizaram dados do DATASUS (2013-2023) para traçar o perfil demográfico de trabalhadores, evidenciando o crescimento de casos. No contexto estrangeiro, [Pachi et al. 2022] investigaram enfermeiros gregos durante a pandemia da COVID-19 e identificaram que quase metade da amostra apresentou valores patológicos para burnout e depressão, confirmando as mulheres como as mais vulneráveis e apontando que o Senso de Coerência (SOC) atua como um redutor desse impacto.

Para superar as limitações preditivas das abordagens descritivas, pesquisas recentes têm adotado algoritmos de clusterização para extrair perfis latentes de esgotamento. Modelos de mistura gaussiana e K-means foram aplicados com sucesso para segmentar profissionais em diferentes níveis de burnout. Nos Estados Unidos, [Yaghmour et al. 2024] identificaram quatro perfis de engajamento entre médicos residentes, revelando que a maioria (55,2%) não apresentava sinais de esgotamento. Na Polônia, [Makara-Studzinska et al. 2022] distinguiram nove perfis psicológicos em profissionais de alto risco, apontando os operadores de emergência como os mais expostos ao estresse crônico. Já na Arábia Saudita, [Baghdadi et al. 2022] mapearam o perfil de risco entre professoras mais jovens, com menos tempo de serviço e privação de sono. Nestes estudos, a clusterização revelou-se eficaz para identificar grupos vulneráveis e, no caso de [Baghdadi et al. 2022], elevar a precisão dos modelos preditivos para até 100%.

No Brasil, embora a clusterização venha sendo explorada em bases de dados governamentais volumosas, como na avaliação do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), onde [De Melo Silva et al. 2025] conseguiram definir os indicadores mais críticos para a melhoria de disparidades na educação pública, e na segurança, com [Barcelar et al. 2025] separando com eficácia seis padrões criminais recorrentes, ainda há uma lacuna metodológica. São poucos os estudos que unem algoritmos de agrupamento não supervisionado a bases oficiais de saúde para avaliar a saúde mental do trabalhador. Portanto, este trabalho contribui para o estado da arte ao aplicar técnicas de clusterização aos dados do SINAN, ampliando a compreensão epidemiológica da depressão e do burnout em escala nacional.

### 3. Metodologia

Para a realização deste estudo, utilizou-se uma abordagem quantitativa e descritiva para analisar a semelhança entre os padrões de síndrome de burnout e episódios depressivos no Brasil, bem como definir perfis para esses diagnósticos. Para tal, foram utilizados dados secundários de notificações oficiais do SINAN para transtornos mentais relacionados ao trabalho, nos anos de 2016 a 2019 e de 2022 a 2025, tendo em vista que os dados de 2020 e 2021 não estavam disponíveis no momento da obtenção. Além disso, a linguagem Python na versão 3.12.12 foi primordial para as etapas de pré-processamento, análise e aplicação do algoritmo de agrupamento K-means, por meio da biblioteca scikit-learn, utilizado para o agrupamento de perfis de burnout e episódios depressivos no Brasil.

#### 3.1. Base de Dados

A base de dados de notificações de agravos de transtornos mentais relacionados a trabalho foi obtida por meio da plataforma governamental DATASUS, que contém 62 variáveis. Dentre elas, 16 variáveis foram utilizadas para análise e transformação, conforme descritas na Tabela 1.

**Tabela 1. Descrição da base de dados**

| Variável  | Descrição                                                   | Variável   | Descrição                                         |
|-----------|-------------------------------------------------------------|------------|---------------------------------------------------|
| NU_ANO    | Ano de notificação                                          | ANO_NASC   | Ano de nascimento do paciente                     |
| DIAG_ESP  | Diagnóstico específico                                      | CS_SEXO    | Sexo do paciente                                  |
| ALCOOL    | Se o paciente possui hábito de ingerir álcool               | CS_RACA    | Raça do paciente                                  |
| FUMA      | Se o paciente possui hábito de fumar                        | CS_ESCOL_N | Escolaridade do paciente                          |
| DROGAS    | Se o paciente possui hábito de consumir drogas psicoativas  | ID_OCUPA_N | Ocupação do paciente                              |
| TRAB_DOE  | Se houve outros colegas de trabalho com o mesmo diagnóstico | ID_MN_RESI | Código IBGE do município em que o paciente reside |
| PSIC_FARM | Se o paciente possui hábito de consumir psicofármacos       | UF_EMP     | Unidade Federativa da empresa contratante         |
| SIT_TRAB  | Situação de trabalho                                        | MUN_EMP    | Código IBGE do município da empresa contratante   |

#### 3.2 Limpeza e processamento dos dados

Primeiramente, foi feita a filtragem pelo diagnóstico (DIAG\_ESP) para os dados corresponderem apenas àqueles referentes a pacientes com burnout e episódios depressivos. Em seguida, foram selecionadas as variáveis relevantes ao trabalho, conforme a Tabela 1. Assim, foi feito o tratamento de registros duplicados e de valores nulos, cujo procedimento resultou na exclusão deles, mantendo a pureza dos dados.

A coluna IDADE foi criada para representar a idade dos pacientes, subtraindo o ano de notificação do ano de nascimento. Por se tratar da única coluna numérica, realizou-se o tratamento de valores extremos utilizando o método do Intervalo Interquartil (IQR). Nesse método, o IQR é definido como  $Q3 - Q1$  e qualquer dado que estiver fora de  $Q3 + 1.5 * IQR$  ou  $Q1 - 1.5 * IQR$  é considerado um outlier, o qual, no presente estudo, foi substituído pela mediana. Além disso, também foram criadas a coluna REGIÃO, a partir da coluna de unidade federativa, e a coluna MORA\_MUNICIPIO, que consiste na comparação entre o município de residência e o município da empresa, para checar se o paciente mora no município em que trabalha.

Para a identificação dos municípios, cujos valores estão representados por códigos, foi essencial cruzar os dados com a Tabela de Municípios do Instituto Brasileiro de Geografia e Estatística (IBGE). Da mesma forma, a coluna referente à ocupação também está codificada, o que exigiu o acesso à Classificação Brasileira de Ocupações (CBO). Assim, foi possível criar a coluna AREA\_OCUPACAO, identificando as ocupações que são diretamente ligadas às áreas de Saúde, Atendimento ao Público, Banco e Educação. Os valores categóricos “Ignorado”, isto é, que não carregam nenhuma informação por falta de preenchimento adequado, foram removidos.

### 3.3 Análise estatística

Após a limpeza e processamento dos dados, foram realizados cálculos descritivos para as variáveis numéricas, como média e desvio padrão, e para as variáveis categóricas foram calculadas a proporção de cada categoria, além de agrupamentos por variáveis, para fins de análise geral da amostra.

Após o processo de agrupamento com K-means, foi avaliada a normalidade dos dados numéricos, foi aplicado o teste de Kolmogorov-Smirnov, que verifica a distribuição de uma variável específica. Em casos de distribuição não normal, utilizou-se o teste de Kruskal-Wallis para comparar os clusters formados e para as variáveis categóricas, foi aplicado o teste do Qui-quadrado. Em ambos os testes, os resultados foram considerados estatisticamente significativos quando o p-valor foi menor que 0,05, cujo valor indica a probabilidade de observar um resultado mais extremo que o obtido, permitindo rejeitar a hipótese nula. Além disso, foram realizados cálculos de proporção das variáveis dentro de cada cluster, a fim de identificar de forma observável os perfis de cada agrupamento.

### 3.4 Análise Fatorial para Dados Mistos(FAMD)

A FAMD é um método estatístico que une a Análise de Componentes Principais (PCA) e a Análise de Correspondências Múltiplas (MCA) para analisar simultaneamente dados quantitativos e qualitativos. Nesse método, as variáveis quantitativas são previamente padronizadas e as variáveis qualitativas são transformadas em variáveis binárias. Para misturar os dois tipos de dados, a FAMD assegura o equilíbrio no sentido da inércia axial máxima de cada variável, assim, os dois tipos de variáveis assumem um peso equilibrado na busca pelos eixos de inércia máxima [Pagès 2004].

No presente estudo, a FAMD foi empregada com o objetivo de identificar os componentes que melhor sintetizam a variabilidade do conjunto de dados. Foi utilizada a variância acumulada para avaliar a proporção da variabilidade total explicada por cada

conjunto de componentes. As variáveis utilizadas para essa etapa foram: idade, sexo, raça, escolaridade, ocupação, situação de trabalho, hábito de álcool, hábito de psicofármacos, hábito de drogas psicoativas, hábito de fumar, região, se há outros colegas com mesmo diagnóstico e se mora no mesmo município do trabalho. As colunas referentes à geolocalização, exceto região, foram excluídas para evitar enviesamento e redundância estrutural no FAMD, assim como a coluna relativa ao diagnóstico.

### 3.5 K-Means

O K-means é um algoritmo de aprendizado não supervisionado amplamente utilizado em tarefas de clusterização, cujo objetivo é agrupar observações em conjuntos (clusters) de forma que elementos pertencentes ao mesmo grupo sejam mais semelhantes entre si do que em relação aos elementos de outros grupos. Nesse método, o número de clusters  $k$  deve ser definido previamente, e cada grupo é representado por um centróide, que corresponde à média dos pontos pertencentes ao cluster. A qualidade do agrupamento é avaliada através de uma função de custo, que mede o erro quadrático das amostras em relação aos centros de seus respectivos grupos [Jain, A. K. 2010].

Na prática, o funcionamento do algoritmo ocorre de maneira iterativa. Inicialmente,  $k$  centroides são selecionados aleatoriamente. Em seguida, cada observação do conjunto de dados é atribuída ao cluster cujo centróide esteja mais próximo, de acordo com a distância euclidiana, normalmente. Após essa atribuição, os centróides são recalculados com base na média dos elementos pertencentes a cada cluster. Esse processo de atribuição e atualização dos centróides é repetido até que não ocorram mais mudanças significativas nos agrupamentos [Jain, A. K. 2010]. Dessa forma, por apresentar uma solução aproximada eficiente e ser amplamente empregado em análises exploratórias de dados, este algoritmo foi adotado neste trabalho.

## 4. Resultados e Discussão

A filtragem dos dados resultou em 4.803 notificações para Síndrome de Burnout e Episódios Depressivos e, após a etapa de limpeza e processamento de dados, a base de dados obteve 1.553 registros e 18 colunas. Do total, 789 correspondiam aos casos de burnout e 764, aos episódios depressivos.

### 4.1 Panorama Geral e Regional

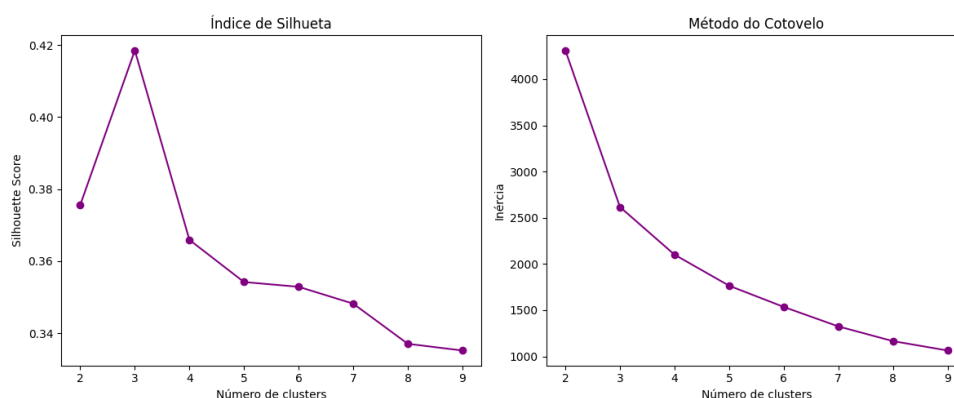
A média de idade de pessoas diagnosticadas com a síndrome de burnout ou episódios depressivos ocupacionais foi de 40,7 e desvio padrão de 9,6, o que confirma que esses transtornos afetam pessoas em idade ativa, ou seja, entre 30 e 50 anos. A análise descritiva dos dados demonstrou que proporção de notificações referidas a mulheres foi ainda maior que a percebida no estudo de [Bailo et al 2024] para transtornos ocupacionais gerais, indo de 65,8% para 74,5%, demonstrando que as mulheres são as mais afetadas pela burnout e episódios depressivos. Mais da metade dos dados (53,5%) se referiam a pessoas brancas, 35,6% a pessoas autodeclaradas pardas e menos de 12% a pessoas preta, indígenas ou amarelas, o que não reflete os dados do [IBGE, 2023], que mostra que a maioria da população é parda (45,3%). Outro ponto a ser destacado também o fato dos transtornos serem associados a mão de obra especializada, uma vez que a maioria das notificações são de pessoas com ensino médio ou superior completo, sendo que a burnout possui um número maior de graduados, com 58,5%.

No que diz respeito aos hábitos individuais, 8,9% afirma ter o hábito de ingerir bebidas alcoólicas, 10% diz fumar ou ter sido fumante, 8,4% diz fazer uso de drogas psicoativas e 62,7% fazem uso recorrente de psicofármacos, o que indica que a maioria dos diagnosticados já fazem acompanhamento médico para tratamento da saúde mental. 71,2% dos casos de Burnout são de pessoas que têm colegas de trabalho com o mesmo diagnóstico, ante 56% das pessoas com episódios depressivos, reafirmando que se trata de transtornos cujo ambiente de trabalho exerce forte influência. Dois estados também se destacam em número de notificações: São Paulo (26,4%) e Minas Gerais (21,7%). Isso reflete o fato de que o Sudeste é a região com o maior número de casos, com 51,5%, mas não reflete a proporção geral do Centro-Oeste, que é de 7,1%.

Outro ponto que difere entre as regiões é a proporção dos diagnósticos. Mesmo que a amostra total apresente uma distribuição equilibrada entre os casos de burnout e de episódios depressivos, pode-se observar que essa prevalência varia conforme a localidade. Por exemplo, a região Centro-Oeste teve o maior índice de episódios depressivos (75,2%) e o Sudeste teve o maior de síndrome de burnout (58,9%). Ademais, só a cidade de São Paulo aglomera 14,6% de todas as notificações do Brasil e 27,6% da região Sudeste.

#### 4.2 Agrupamento com FAMD e K-means

A aplicação da FAMD permitiu que os dois primeiros componentes principais retivessem 70% da variância total do conjunto de dados. A análise das contribuições das variáveis indicou que o Componente 1 foi formado principalmente por ocupação (36,7%), situação de trabalho (20,8%) e região (15,9%); já o componente 2 foi influenciado principalmente por ocupação (39,1%), escolaridade (24,5%) e situação de trabalho (14,2%). Com base nessa redução de dimensionalidade, aplicaram-se a métrica do Coeficiente de Silhueta e o método do Cotovelo (Elbow Method) para determinar o número ideal de clusters (k). Ante a divergência entre os resultados de ambos os métodos, optou-se pelo valor de k que apresentou o maior Índice de Silhueta, ou seja,  $k=3$  com 0,418 como pode ser visto na figura 2.



**Figura 2.** Gráficos com métodos para k ideal. (a) Índice de Silhueta; (b) Método do Cotovelo

A clusterização via K-means ( $k=3$ ) gerou grupos com volumetrias próximas: 706(Cluster 0), 430 (Cluster 1) e 417 (Cluster 2). Essa configuração reflete uma

distribuição de dados relativamente equilibrada entre os agrupamentos. A coluna Idade apresentou distribuição não normal e p-valor menor que 0,00001. Também foi identificado que todas as variáveis categóricas utilizadas tiveram diferença estatística entre os clusters, como pode ser observado na Tabela 2.

**Tabela 2. Resultados dos testes paramétricos para variáveis categóricas entre clusters**

| Variável       | Categorias                                                                           | P-Valor   |
|----------------|--------------------------------------------------------------------------------------|-----------|
| Sexo           | Masculino, Feminino                                                                  | < 0,00001 |
| Raça           | Branca, Preta, Parda, Amarela, Indígena                                              | < 0,0001  |
| Escolaridade   | EM completo, EF completo, EF incompleto, Ensino Superior Completo, 2 outros          | < 0,00001 |
| Situação_trab  | Empregado de carteira assinada, Empregado não registrado, Servidor público, 8 outros | < 0,00001 |
| Região         | Sudeste, Sul, Nordeste, Norte, Centro-Oeste                                          | < 0,00001 |
| Álcool         | Sim, Não                                                                             | < 0,001   |
| Drogas         | Sim, Não                                                                             | < 0,00001 |
| Psico_fármacos | Sim, Não                                                                             | < 0,00001 |
| Fuma           | Sim, Não, Ex-fumante                                                                 | < 0,00001 |
| Outros_colegas | Sim, Não                                                                             | < 0,00001 |
| Mora_municipio | Sim, Não                                                                             | < 0,00001 |
| Ocupação       | 355 ocupações específicas                                                            | < 0,00001 |

#### 4.2.1. Perfis sociodemográficos identificados pelo K-means

Os clusters formados revelaram padrões predominantemente sociodemográficos e comportamentais, o que será melhor investigado posteriormente. Não houve grande variação de raça entre os clusters, assim como não houve a predominância expressiva de nenhum diagnóstico, tendo em vista que nenhum superou 66% em nenhum cluster. A coluna referente à idade apresentou distinções suaves. A média de idade do cluster 0 foi de 39,5 ( $\pm 8,8$ ), do cluster 1 foi de 39,5 anos ( $\pm 10,5$ ) e o cluster 2 foi o mais velho com 44,1 anos ( $\pm 9,2$ ). Também foi percebido a predominância da região Sudeste e Nordeste para os três clusters. A Figura 3 contém um mapa com a distribuição espacial dos clusters pelo território brasileiro.

##### Cluster 0

O cluster 0 foi o que concentrou mais pessoas nas regiões Sudeste (60%) e Nordeste (36%), sobretudo para o estado de São Paulo e Minas Gerais, que contém 55,5% das ocorrências. Trata-se do agrupamento com o maior número de pessoas empregadas de carteira assinada (84,7%) e a maioria possui ensino superior completo (61%), trabalhando diretamente com a área bancária (31%). Além disso, 87,2% dos diagnosticados têm colegas na mesma situação, o maior índice entre os clusters e o

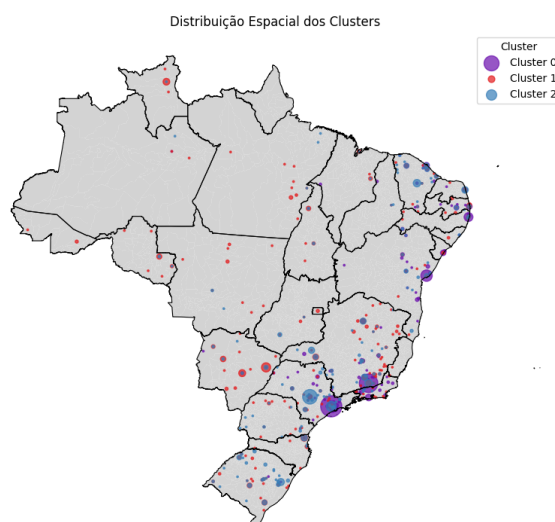
maior número de usuários de psicofármacos (74,3%), apesar de ter sido o agrupamento com o menor índice para os outros hábitos. Também foi o cluster com maior proporção de pessoas que não moram no mesmo município em que trabalham (33,4%), pelo menos 20% mais que os outros clusters. Esse dado indica uma associação entre o deslocamento para o trabalho e este perfil epidemiológico, também marcado pelo maior consumo de substâncias psicoativas.

### Cluster 1

Os registros do cluster 1 estão distribuídos, principalmente, entre as regiões Sudeste, Nordeste e Sul, mas com grande parte do estado de São Paulo (30,4%) e Minas Gerais (21,8%). 80% são empregados de carteira assinada, sendo que a maioria (60%) tem ensino médio completo e é o cluster com o maior número de pessoas com ensino fundamental completo e incompleto, se comparado com os outros clusters. Aqui, é possível identificar uma maior variabilidade de ocupações, mas dando destaque para as que lidam diretamente com o público (20,2%). O cluster 1 também foi o que obteve uma menor proporção de pacientes com colegas com mesmo diagnóstico (34,1%), mas também foi o agrupamento com a maior concentração de pessoas que usam drogas psicoativas (20,2%), que fumam ou foram fumantes(16%) e com o hábito de ingerir álcool(13,5%). A formação deste cluster identifica um grupos de pessoas mais comum pelo Brasil, mas com tendência a hábitos prejudiciais à saúde.

### Cluster 2

O cluster 2, diferentemente dos outros dois, foi o que obteve a maior concentração de servidores públicos (83,6%). Apesar dos outros clusters terem a composição majoritariamente feminina, a proporção do cluster 1 foi ainda maior, com 92,3%, pelo menos 25% a mais que os outros clusters. A distribuição entre as regiões foi bem balanceada, porém quase todos os dados do Norte e Centro-Oeste concentram-se neste grupo e é o único que o estado de São Paulo não ganha destaque de incidência. A maioria dos pacientes(54,6%), assim como no cluster 0, possui colegas com o mesmo diagnóstico e 70% dos registros possuem ensino superior, sendo que 49,8% trabalham diretamente com a área da saúde e 18,4% na educação. Quanto aos hábitos, foi o grupo que menos apresentou qualquer destaque, evidenciando uma formação pautada, principalmente, no modo e no tipo de trabalho.



### **Figura 3. Distribuição espacial dos clusters pelo Brasil**

Como mencionado, a Figura 3 demonstra, de forma visual, como os clusters estão distribuídos pelo Brasil, revelando uma concentração nas regiões Sudeste e Nordeste, principalmente dos clusters 0 e 1. O cluster 0, portanto, foi caracterizado pela predominância de trabalhadores formais com maior nível de escolaridade, especialmente na área bancária, e com alta incidência de colegas de trabalho com o mesmo diagnóstico e da necessidade de locomoção intermunicipal. O cluster 1 apresentou maior diversidade ocupacional e de nível de instrução, com destaque para atividades de atendimento ao público e maior presença de comportamentos de risco, como uso de álcool, fumo e drogas psicoativas. Já o cluster 2 foi marcado pela forte presença de servidores públicos, especialmente profissionais da área da saúde e educação, além de apresentar a maior proporção de mulheres entre os grupos analisados e de representar o grupo sutilmente mais velho. Esses resultados evidenciam que perfis epidemiológicos específicos estão fortemente associados a fatores ocupacionais, organizacionais e sociodemográficos. Além disso, a presença de múltiplos casos dentro de um mesmo ambiente profissional reforça a correlação entre as condições laborais e o adoecimento mental dos trabalhadores, caracterizando um importante padrão observacional.

## **5. Conclusão**

Os resultados do estudo evidenciaram que os transtornos mentais ocupacionais afetam predominantemente mulheres e profissionais com escolaridade de nível médio a superior. Observou-se uma concentração de casos no Sudeste (principalmente em São Paulo) e no Nordeste, um provável reflexo da densidade demográfica e de disparidades regionais na capacidade de notificação. A clusterização revelou três perfis epidemiológicos principais, dois muito presentes nas regiões Sudeste e Nordeste e um comum em todo território brasileiro, pautado exclusivamente no tipo de trabalho.

Metodologicamente, o estudo atesta a eficácia do uso conjunto de análise estatística e algoritmos de Machine Learning na exploração de grandes bases públicas, suprimindo a escassez de pesquisas de agrupamento. Embora a dependência de dados secundários traga limitações inerentes, como subnotificações e excesso de valores nulos, além da falta de dados no período da pandemia da COVID-19 (2020 e 2021), a abordagem demonstrou robustez na extração de padrões ocultos. Em síntese, este trabalho reforça a importância de políticas públicas e estratégias organizacionais voltadas à promoção da saúde mental no ambiente de trabalho, especialmente em setores e perfis profissionais mais vulneráveis, apoiando a formulação de intervenções mais direcionadas e baseadas em evidências. Como trabalhos futuros, pode-se ter a integração com outras bases de dados governamentais e a avaliação de algoritmos de clusterização alternativos, visando aprimorar e validar as descobertas deste estudo, assim como a inclusão da Inteligência Artificial Explicável(XAI).

## Referências

- Bailo D. W. et al. (2024). Impacto dos Transtornos Mentais Relacionados ao Trabalho no Brasil: Uma Revisão de Dados e Perspectivas Epidemiológicas. *Brazilian Journal of Implantology and Health Sciences*, 6(8): 4723-4733.
- Baghdadi, N. A. et al. (2022). An Analysis of Burnout among Female Nurse Educators in Saudi Arabia Using K-Means Clustering. *European Journal of Investigation in Health, Psychology and Education*, 13(1):33-55.
- Barcelar, R. R. et al. (2025). Analysis of Criminal Patterns in Police Report Narratives using Spectral Clustering with K-means. In *XII Symposium on Knowledge Discovery, Mining and Learning(KDMILE).SBC*
- Brasil. Ministério da Previdência Social (2026). Previdência Social concede 546.254 benefícios por incapacidade temporária por transtornos mentais e comportamentais, <https://www.gov.br/previdencia/pt-br/noticias/2026/janeiro/previdencia-social-concede-546-254-beneficios-por-incapacidade-temporaria-por-transtornos-mentais-e-comportamentais>
- Brasil. Ministério da Saúde do Brasil (2001). Doenças relacionadas ao trabalho manual e procedimentos para os serviços de saúde. Ministério Da Saúde Do Brasil, Brasília.
- Brasil. Ministério do Trabalho e Emprego (2023). Classificação Brasileira de Ocupações (CBO), <https://www.gov.br/trabalho-e-emprego/pt-br/assuntos/cbo>
- De Melo Silva, M. L. et al. (2025). Desigualdades Educacionais no Brasil: Uma Análise por Clusterização de Indicadores Educacionais e Desempenho Escolar. In *XL Simpósio Brasileiro de Banco de Dados(SBBD).SBC*.
- Instituto Brasileiro de Geografia e Estatística - IBGE (2023). Censo Demográfico 2022, <https://www.ibge.gov.br/estatisticas/sociais/trabalho/22827-censo-demografico-2022.html>.
- Instituto Brasileiro de Geografia e Estatística - IBGE (2026). Códigos dos municípios. <https://www.ibge.gov.br/explica/codigos-dos-municipios.php>.
- Jain, A. K. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, v. 31, n. 8, p. 651–666, 1 jun. 2010.
- Makara-Studzinska, M. et al. (2022). Profiles of Occupational Burnout in the Group of Representatives High-Risk Professions in Poland. *International Journal of Environmental Research and Public Health*, 19(10): 6297.
- Pachi, A. et al. (2022). Burnout, Depression and Sense of Coherence in Nurses during the Pandemic Crisis. *Healthcare*, 10(1): 134.
- Page's, J. (2004). Analyse Factorielle de Données Mixtes. *Revue de Statistique Appliquée*, 52(4): 93-111.
- Salvadori, I. L. (2024). Análise e Publicação de Dados de Processos Eletrônicos em Organizações Públicas. In *XXXIX Simpósio Brasileiro de Bancos de Dados (SBBD). SBC*.