

CARACTERIZAÇÃO DE PERFIS MATERNO-INFANTIS A PARTIR DE DADOS DO SINASC UTILIZANDO TÉCNICAS DE CLUSTERIZAÇÃO

**CHARACTERIZATION OF MATERNAL AND CHILD PROFILES USING SINASC
DATA AND CLUSTERING TECHNIQUES**

Ciências Exatas e da Terra, Ciências da Saúde • 27/10/2026

REGISTRO DOI: [10.70773/revistatopicos/793075326](https://doi.org/10.70773/revistatopicos/793075326)

Cainan Linard de Albuquerque
Luiz Fernando Pinheiro Pereira Filho
Dadilton Bastos Melo

RESUMO

O Sistema de Informações sobre Nascidos Vivos (SINASC) reúne grandes volumes de dados sobre gestação, parto e recém-nascido, e a clusterização permite identificar perfis materno-infantis que apoiem políticas de saúde direcionadas. Como a base combina variáveis numéricas e categóricas, a escolha do algoritmo pode influenciar os agrupamentos obtidos. Comparar os algoritmos K-means e K-prototypes na formação de agrupamentos interpretáveis e com valor para a saúde pública a partir de dados do SINASC. Foram usados registros de 2022 a 2024, com 15 variáveis selecionadas; após a limpeza, restaram 240.791 dos 287.342 registros iniciais. As variáveis numéricas passaram por normalização Min-Max e, para o K-means, as categóricas foram convertidas por Label Encoding e One-Hot Encoding, enquanto o K-prototypes manteve a representação original. O número de agrupamentos foi definido pelos métodos do cotovelo e silhouette e pela análise em diferentes sementes, com implementação em Python 3.12.7. O K-means formou oito agrupamentos, com concentrações extremas em sexo e raça/cor da mãe, como clusters compostos integralmente por recém-nascidos de um mesmo sexo. O K-prototypes formou cinco agrupamentos, com distribuição mais equilibrada das variáveis categóricas e maior integração com os perfis numéricos. Ambos identificaram grupos neonatais contrastantes, diferenciados principalmente por peso ao nascer, Apgar 5 e semanas de gestação. O estudo mostra que o tratamento das variáveis categóricas altera a estrutura dos agrupamentos, apoiando a escolha de métodos adequados a dados mistos. Os agrupamentos devem ser interpretados como padrões exploratórios, e não como relações de causalidade.

Palavras-chave: Clusterização; K-means; K-prototypes; SINASC; Saúde materno-infantil.

ABSTRACT

The Brazilian Live Birth Information System (SINASC) gathers large volumes of data on pregnancy, delivery, and newborns, and clustering makes it possible to identify maternal-infant profiles that support targeted health policies. Because the database combines numerical and categorical variables, the choice of algorithm may influence the clusters obtained. This study compares the K-means and K-prototypes algorithms in forming interpretable clusters of value to public health, using SINASC data. Records from 2022 to 2024 were used, with 15 selected variables; after cleaning, 240,791 of the initial 287,342 records remained. Numerical variables underwent Min-Max normalization and, for K-means, categorical variables were converted using Label Encoding and One-Hot Encoding, while K-prototypes kept the original representation. The number of clusters was defined by the elbow and silhouette methods and by analysis across different seeds, with implementation in Python 3.12.7. K-means formed eight clusters, with extreme concentrations in newborn sex and maternal race/skin color, such as clusters composed entirely of newborns of the same sex. K-prototypes formed five clusters, with a more balanced distribution of categorical variables and greater integration with the numerical profiles. Both identified contrasting neonatal groups, differentiated mainly by birth weight, 5-minute Apgar score, and gestational weeks. The study shows that the treatment of categorical variables alters the structure of the clusters, supporting the choice of methods suited to mixed data. The clusters should be interpreted as exploratory patterns, not as causal relationships.

Keywords: Clustering; K-means; K-prototypes; SINASC; Maternal and child health.

1. INTRODUÇÃO

O volume de dados produzido hoje é enorme, e analisá-los, além de necessário, é um desafio (Han; Kamber; Pei, 2011). É nesse contexto que surge o aprendizado de máquina, responsável pelo processamento automatizado dessas informações. Segundo Han, Kamber e Pei (2011, p. 24), “Aprendizagem de máquina investiga como os computadores podem aprender (ou melhorar seu desempenho) com base em dados”. A necessidade aparece com clareza na saúde pública brasileira, que coleta quantidades massivas de dados. De acordo com Fernandes e Chiavegatto Filho (2019), o aprendizado de máquina (machine learning) é frequentemente usado na saúde para auxiliar no diagnóstico e na predição de riscos. Um exemplo dessa coleta é o Sistema de Informações sobre Nascidos Vivos (SINASC), que registrou 2.561.922 nascidos vivos em 2022 (Ministério da Saúde, 2022). Na saúde pública, o clustering já foi aplicado por Olson et al. (2016), que agruparam dados semelhantes de caminhoneiros para identificar quais comportamentos afetam o equilíbrio energético e testar diferenças em fatores psicossociais e de saúde.

Uma das necessidades da análise de dados é identificar perfis distintos para orientar políticas direcionadas. Nesse aspecto, destaca-se o aprendizado de máquina não supervisionado, em que, o algoritmo recebe um conjunto de dados sem rótulos e os agrupa por conta própria, com base em suas semelhanças. Essa abordagem auxilia a análise exploratória, a compreensão inicial dos dados e a tomada de decisões (Fernandes; Chiavegatto Filho, 2019). Dentro dela está o clustering, ou agrupamento, que consiste em dividir um conjunto de dados em subconjuntos de acordo com a similaridade, extraindo os padrões resultantes (Han; Kamber; Pei, 2011). K-means e K-prototypes são exemplos de algoritmos que fazem isso por abordagens diferentes.

O K-means agrupa os registros ao redor de centróides, definidos pelo operador, pela distância euclidiana e os atualiza pela média (Han; Kamber; Pei, 2011). Suas limitações incluem a definição prévia do número de agrupamentos, a sensibilidade a ruídos e outliers, a tendência a clusters esféricos e de tamanhos semelhantes e a restrição a dados numéricos, pois a média distorce os centróides com dados categóricos. O K-prototypes, projetado para dados de tipos mistos (Gan, 2026), supera essa última limitação usando a moda nos atributos categóricos e a média nos numéricos, mantendo as demais restrições. Por isso, é o mais indicado para o SINASC, que combina variáveis numéricas e categóricas.

A clusterização em saúde pública tem se mostrado uma ferramenta relevante para a identificação de perfis populacionais e o direcionamento de ações preventivas. Fernandes e Chiavegatto Filho (2019) apontam que aplicações de mineração de dados com algoritmos de aprendizado de máquina para análises em saúde pública apresentam melhor desempenho em comparação com análises tradicionais, destacando o clustering como alternativa eficaz para o agrupamento de perfis populacionais a partir de grandes bases de dados.

Este estudo aplica as duas técnicas, K-means e K-prototypes, à base pública do SINASC. O objetivo é analisar a capacidade de cada uma de formar agrupamentos com “valor” para a saúde pública, verificando se há padrões, se eles são interpretáveis e como se compara a eficiência das duas técnicas nesse contexto.

2. FUNDAMENTAÇÃO TEÓRICA

Olson et al. (2016) aplicaram agrupamento hierárquico a dados auto relatados de 452 caminhoneiros com obesidade, para identificar grupos com padrões semelhantes de sono, alimentação e exercício e testar diferenças em fatores de saúde e psicossociais. Os cinco clusters obtidos, após divisão dos comportamentos, diferiram em idade, tabagismo, diabetes, estresse e apoio social, mas não em índice de massa corporal. O estudo evidencia a utilidade da clusterização para derivar perfis em saúde, mas usou outro método e um recorte ocupacional, sem comparação entre algoritmos.

Huang (1998) propôs o k-modes e o k-prototypes como extensões do K-means para dados categóricos e mistos, o segundo por uma medida de dissimilaridade combinada. Testou-os em conjuntos clássicos (soybean disease e credit approval) e em duas bases reais com meio milhão de objetos cada, concluindo que são eficientes em grandes volumes. O trabalho fundamenta o uso do K-prototypes em dados mistos, mas ocorreu fora do contexto do SINASC e priorizou o desempenho do agrupamento, não a interpretação epidemiológica.

Nunes et al. (2021) analisaram dados do SINASC de 16.474 mães de São Luís, Maranhão, em estudo transversal com qui-quadrado de Pearson e odds ratio. Adolescentes tiveram maior chance de nascimento prematuro (OR = 1,37) e mulheres em idade avançada, maior risco de baixo peso ao nascer (OR = 1,22); em ambos os extremos etários, foi maior a chance de menos de seis consultas pré-natais. Embora use variáveis também presentes nesta pesquisa, parte de uma hipótese prévia sobre a idade materna e não busca perfis multivariados.

O índice de Apgar, criado pela anestesilogista Virginia Apgar em 1953, avalia a vitalidade do recém-nascido no primeiro e no quinto

minuto de vida por meio de cinco critérios clínicos pontuados de zero a dois, totalizando até dez pontos (Apgar, 1953). O escore do quinto minuto, utilizado neste estudo, é o mais relevante clinicamente: valores inferiores a sete estão fortemente associados à mortalidade neonatal, à morbidade neurológica grave e a desfechos adversos a longo prazo (Simon; Hashmi; Bragg, 2023). Dondé, Soncini e Nunes (2019) identificaram que fatores como número insuficiente de consultas pré-natais e intercorrências gestacionais estão associados ao baixo índice de Apgar no quinto minuto, evidenciando sua relação com variáveis também presentes na base do SINASC utilizada neste trabalho.

Os estudos agregam-se ao analisar perfis em saúde, mas diferem no propósito: explorar perfis comportamentais (Olson et al., 2016), desenvolver um algoritmo (Huang, 1998) e testar associações no SINASC (Nunes et al., 2021). O índice de Apgar, por sua vez, fundamenta a escolha do APGAR5 como variável de análise, dado seu reconhecido valor prognóstico neonatal. Nesse conjunto, o presente estudo aplica K-means e K-prototypes à base do SINASC de 2022 a 2024 (240.791 registros após a limpeza) e compara como o tratamento das variáveis categóricas altera os agrupamentos, em análise exploratória que não indica relações causais.

3. METODOLOGIA

Os dados vêm da base de dados abertos do SINASC, disponibilizada pelo Ministério da Saúde, e correspondem aos registros de 2022, 2023 e 2024. O SINASC é uma importante fonte nacional de informações sobre nascimentos. Reúne variáveis sociodemográficas maternas, características da gestação e do parto e dados do recém-nascido, o que permite construir indicadores para estudos

epidemiológicos e para o monitoramento da saúde materno-infantil (Romero; Cunha, 2007; Szwarcwald et al., 2019).

As variáveis foram escolhidas com base no dicionário de dados do SINASC e na relevância epidemiológica das características maternas, gestacionais e neonatais. Estudos brasileiros apontam como relevantes para caracterizar os nascimentos e investigar fatores associados a desfechos perinatais a idade materna, a escolaridade, a situação conjugal, o número de consultas pré-natais, a história reprodutiva, a duração da gestação, o peso ao nascer, o sexo, o tipo de parto e outras informações da Declaração de Nascido Vivo (Romero; Cunha, 2007; Barbas et al., 2009).

Foram selecionadas 15 variáveis, numéricas e categóricas. As numéricas passaram por normalização Min-Max, e as categóricas foram codificadas conforme sua natureza, nominal ou ordinal. Uma escala comum é especialmente importante em algoritmos baseados em distância, porque variáveis em unidades ou amplitudes diferentes influenciam o agrupamento de forma desigual (Han; Kamber; Pei, 2011). O Quadro 1 apresenta as variáveis, seu tipo e a técnica aplicada.

Quadro 1. Variáveis utilizadas

Nome da variável	Tipo de dado	Descrição	Técnica aplicada
PESO	Numérica contínua	Peso do recém-nascido ao nascer (gramas)	Min-Max Scaling
APGAR5	Numérica contínua	Índice Apgar no 5º minuto de vida	Min-Max Scaling

IDADEMAE	Numérica contínua	Idade da mãe em anos concluídos	Min-Max Scaling
SEMAGESTAC	Numérica contínua	Número de semanas de gestação	Min-Max Scaling
QTDGESTANT	Numérica contínua	Número de gestações anteriores	Min-Max Scaling
QTDFILMORT	Numérica contínua	Número de perdas fetais/abortos anteriores	Min-Max Scaling
ESMAE2010	Categórica ordinal	Nível de escolaridade da mãe	Label Encoding + Min-Max Scaling no K-means; apenas Label Encoding no K-prototypes
CONSULTAS	Categórica ordinal	Faixa de consultas pré-natais realizadas	Label Encoding + Min-Max Scaling no K-means; apenas Label Encoding no K-prototypes
SEXO	Categórica nominal	Sexo do recém-nascido	One-Hot Encoding
RACACORMAE	Categórica nominal	Raça/cor da mãe	One-Hot Encoding
ESTCIVMAE	Categórica nominal	Estado civil da mãe	One-Hot Encoding
LOCNASC	Categórica nominal	Local do nascimento (hospital, domicílio etc.)	One-Hot Encoding

GRAVIDEZ	Categórica nominal	Tipo de gravidez (única, dupla, tripla ou mais)	One-Hot Encoding
PARTO	Categórica nominal	Tipo de parto (vaginal, cesáreo)	One-Hot Encoding
IDANOMAL	Categórica nominal	Presença de anomalia congênita identificada	One-Hot Encoding

Fonte: Elaborada pelos autores (2026).

Essas variáveis têm respaldo na literatura sobre a qualidade e o potencial epidemiológico do SINASC. Romero e Cunha (2007) destacam que a base reúne informações sociodemográficas, reprodutivas, assistenciais e das condições do nascimento. Avaliações posteriores mostraram elevada cobertura e boa confiabilidade para várias variáveis, embora algumas apresentem mais inconsistências ou menor concordância (Szwarcwald et al., 2019; Silva et al., 2014).

O índice de Apgar é uma medida padronizada da condição do recém-nascido logo após o nascimento e de sua resposta à reanimação. Ele não deve, porém, ser interpretado isoladamente como indicador de asfixia nem como preditor individual de mortalidade ou de desfecho neurológico (Apgar, 1953; American Academy of Pediatrics; American College of Obstetricians and Gynecologists, 2015).

A implementação foi feita em Python 3.12.7, usando o Visual Studio Code como ambiente de desenvolvimento. As bibliotecas

empregadas foram Matplotlib (1.1), NumPy (2.4.2), pandas (3.0.1) e scikit-learn (1.9.0), para visualização, operações numéricas, manipulação de dados e procedimentos de análise, respectivamente.

3.1. Coleta e Limpeza dos Dados

Na coleta, as bases de 2022, 2023 e 2024 foram concatenadas e, entre as 61 variáveis originais de cada arquivo, foram selecionadas 15 de interesse. O resultado foi uma base com 287.342 registros (Quadro 2).

Quadro 2. Bases brutas

Base	Quantidade de registros	Número de colunas (variáveis)
Base do ano de 2022	97964	61
Base do ano de 2023	97205	61
Base do ano de 2024	92173	61
Base concatenada	287342	15

Fonte: Elaborada pelos autores (2026).

Na limpeza, foram identificados os códigos que o SINASC usa para registros ignorados ou sem informação. Em SEXO, os códigos “I” e “0” foram tratados como ausentes. Nas variáveis ESCMAE2010, CONSULTAS, ESTCIVMAE, LOCNASC, GRAVIDEZ, PARTO e

IDANOMAL, o código 9 foi tratado como informação ignorada e convertido em valor ausente.

Esse tratamento é relevante porque estudos de qualidade da informação indicam boa completude para a maioria das variáveis do SINASC, mas apontam problemas específicos em história reprodutiva, escolaridade, raça/cor, consultas pré-natais e outras, dependendo do contexto (Romero; Cunha, 2007; Silva et al., 2014). Identificar explicitamente os códigos de ausência evita que valores administrativos sejam lidos como observações válidas.

Por fim, foram eliminados os registros com valor ausente em pelo menos uma das variáveis selecionadas. Restaram 240.791 observações, o que corresponde à remoção de 46.551 registros, cerca de 16,2% da base inicial (Quadro 3).

Quadro 3. Registros antes e depois da limpeza

Base	Quantidade de registros	Quantidade de registros removidos
Base concatenada	287342	0
Base pós-limpeza	240791	46.551

Fonte: Elaborada pelos autores (2026).

Adotou-se a análise de casos completos. A exclusão de dados ausentes exige cautela, pois reduz a amostra e pode introduzir viés se não for aleatória (Little; Rubin, 2019). Por isso, a proporção de registros excluídos deve ser sempre avaliada para garantir a validade externa dos agrupamentos.

3.2. Transformação dos Dados

Cada algoritmo exigiu um tratamento diferente. O K-means precisa de dados numéricos para calcular distâncias e atualizar os centróides pela média (Macqueen, 1967). Nele, as variáveis numéricas passaram por Min-Max para ficarem na mesma escala. As ordinais (ESCMAE2010 e CONSULTAS) receberam Label Encoding seguido de normalização. Essa técnica preserva a ordem das categorias, mas exige cautela, pois não garante que a distância matemática entre elas tenha o mesmo significado substantivo. As nominais foram convertidas em variáveis binárias por One-Hot Encoding, que permite a representação numérica, mas pode aumentar a dimensionalidade e afetar a geometria do espaço em métodos baseados em distância (Cerdeira; Varoquaux, 2022).

O K-prototypes, por sua vez, permitiu tratar as variáveis em conjunto, sem converter artificialmente as categorias nominais em coordenadas numéricas. Proposto por Huang (1997, 1998) para atributos mistos, o método combina uma medida de distância numérica com uma medida de dissimilaridade para os categóricos. Neste estudo, apenas as variáveis numéricas passaram por Min-Max, e as categóricas mantiveram sua representação original (Quadro 4).

Quadro 4. Bases transformadas

Base	Número de colunas (variáveis)
Base pós-limpeza	15
Base transformada para o K-means	24
Base transformada para o K-prototypes	15

Fonte: Elaborada pelos autores (2026).

3.3. Influência dos Categóricos no K-Means

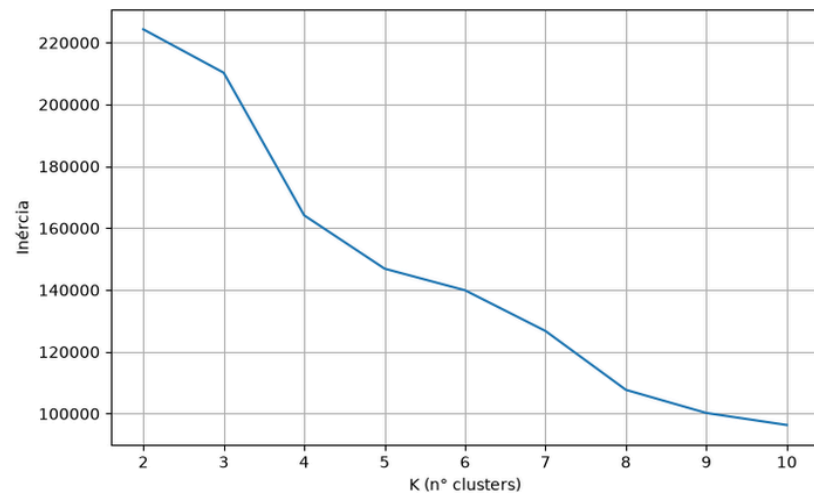
Os dados categóricos influenciam fortemente os agrupamentos do K-means, já que o método atualiza os centróides pela média. Variáveis codificadas por Label Encoding ou One-Hot Encoding geram uma “gravidade” artificial nas distâncias euclidianas, criando concentrações em pontos específicos do espaço (os valores escalonados de uma categoria ou os extremos 0 e 1 do One-Hot). Isso afeta a formação dos centróides, em contraste com a distribuição contínua das variáveis numéricas. O K-prototypes contorna o problema ao separar os tipos de dados na atualização dos centróides, com média para os numéricos e moda para os categóricos. Avaliar essa diferença metodológica, entre um algoritmo para dados numéricos aplicado a uma base transformada e outro concebido para dados mistos, é um dos fundamentos centrais deste estudo.

3.4. Geração dos Gráficos e Escolha do Número de Agrupamentos

Definir o número de agrupamentos é uma das etapas mais importantes da análise exploratória por clustering, pois valores diferentes de k podem gerar estruturas bem diferentes. Jain (2010) ressalta que a análise de agrupamentos é essencialmente exploratória e que não existe um algoritmo universal capaz de determinar a estrutura correta para qualquer conjunto de dados.

Para o K-means, foram usados os gráficos de cotovelo e de silhouette com random state 26 (o escolhido para executar o algoritmo), além da análise da inércia em outras sementes.

Figura 1. Gráfico do cotovelo k-means

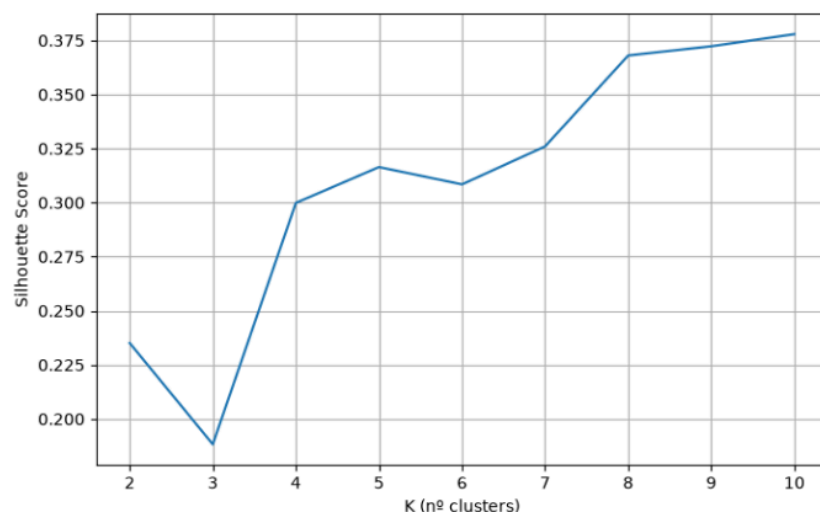


Fonte: Elaborada pelos autores (2026).

A inércia do K-means é a soma dos erros quadráticos intracluster e diminui à medida que se acrescentam agrupamentos. A busca pelo ponto de inflexão procura, portanto, um número de clusters que reduza bem a dispersão sem aumentar a complexidade desnecessariamente.

O gráfico de silhouette reforçou os pontos com picos em 4, 5 e 8.

Figura 2. Gráfico do silhouette k-means



Fonte: Elaborada pelos autores (2026).

Assim, os gráficos de cotovelo e silhouette indicaram pontos de interesse em $k = 4$, $k = 5$ e $k = 8$. A inércia em diferentes sementes

(Quadro 5) foi a evidência adicional para escolher $k = 8$, que apresentou maior estabilidade relativa entre as sementes consideradas.

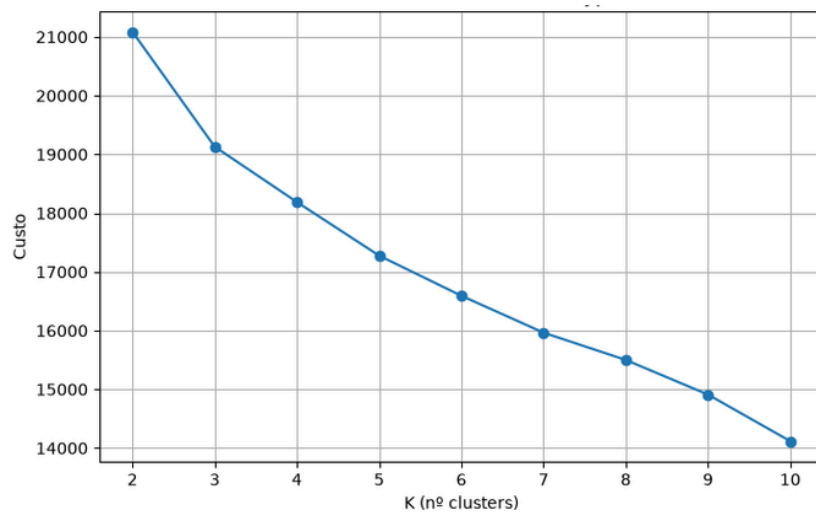
Quadro 5. Inércias k-means

Sementes	Inércia para 4	Inércia para 5	Inércia para 8
Semente 26	164156.32	146896.93	107698.66
Semente 42	187576.38	163079.40	106088.71
Semente 7	164156.32	146759.68	108072.77
Semente 123	202940.18	164058.53	119300.07
Semente 0	172991.69	155304.80	115159.79
Semente 99	182664.35	155089.23	113521.31

Fonte: Elaborada pelos autores (2026).

Para o K-prototypes, foram usados o método do cotovelo com *random state* 26 (também o escolhido para a execução), a função de custo em outras sementes e o *silhouette* com distância de Gower, calculado em uma amostra de 10.000 registros. As sementes são as mesmas do K-means, por padronização. A distância de Gower é adequada a conjuntos com variáveis numéricas e categóricas, pois integra atributos de tipos diferentes em uma única medida de similaridade/dissimilaridade (Gower, 1971; Kaufman; Rousseeuw, 1990).

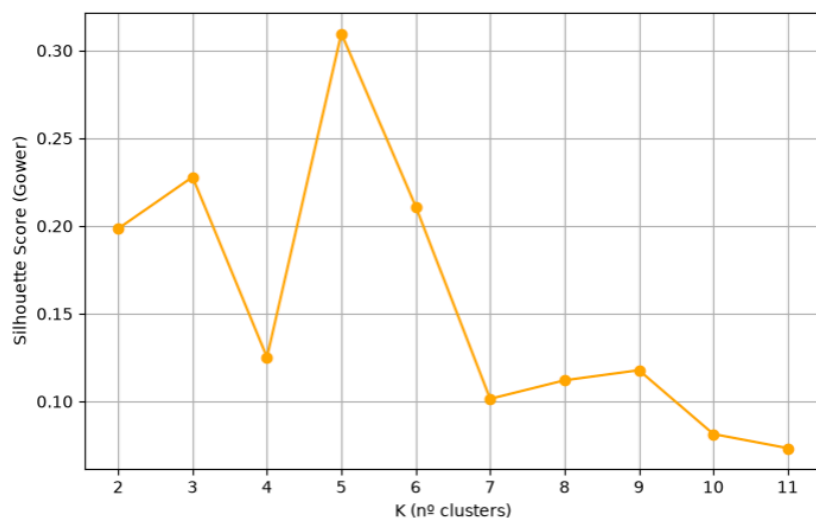
Figura 3. Gráfico do cotovelo k-prototypes



Fonte: Elaborada pelos autores (2026).

O cotovelo funciona como no K-means, mudando apenas o cálculo: nos dados numéricos usa o mesmo SSE e, nos categóricos, a distância de Hamming, combinados em uma função de custo. O silhouette, calculado com a distância de Gower, é o que permite gerar esse gráfico em dados de tipos mistos.

Figura 4. Gráfico do silhouette k-prototypes



Fonte: Elaborada pelos autores (2026).

O silhouette foi aplicado a uma amostra por causa do alto custo computacional no banco completo. Por isso, ele tem menos precisão que a versão usada no K-means.

No K-prototypes, o cotovelo apresentou maior acentuação em $k = 3$, e os valores $k = 4$ e $k = 5$ também mostraram comportamento relevante. O silhouette teve picos em $k = 3$ e $k = 5$. A função de custo em diferentes sementes (Quadro 6) mostrou maior estabilidade relativa com cinco agrupamentos.

Quadro 6. Funções de custo k-prototypes

Sementes	Função de custo para 3	Função de custo para 4	Função de custo para 5
Semente 26	19130.80	18187.82	17275.44
Semente 42	19512.82	17923.70	17296.69
Semente 7	19068.97	18258.08	17291.16
Semente 123	19272.12	18141.97	17291.16
Semente 0	19149.78	17923.70	17291.16
Semente 99	19271.12	17923.70	17380.24

Fonte: Elaborada pelos autores (2026).

3.5. Execução dos Algoritmos

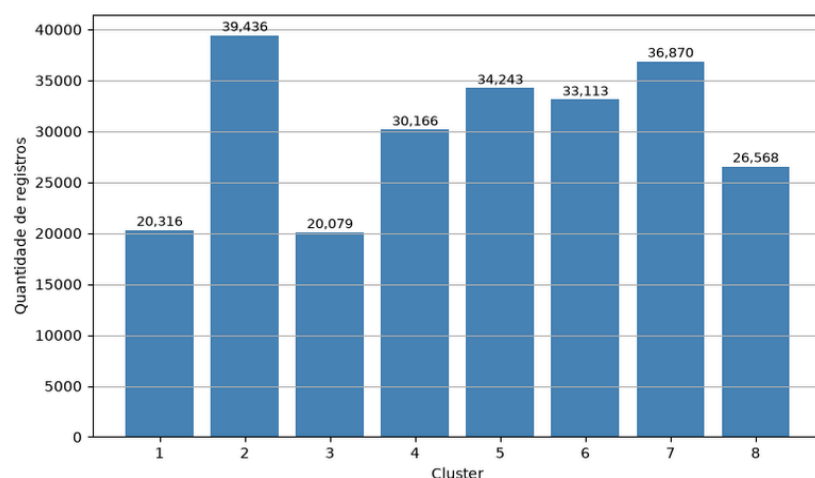
No K-means, foram criados 8 agrupamentos, escolha baseada nos gráficos de cotovelo e silhouette, que apresentaram acentuação nesse ponto, e nos testes de inércia com outras sementes. O random state foi 26. Vale lembrar a influência dos categóricos, que pode distorcer cada um desses critérios.

No K-prototypes, foram criados 5 agrupamentos, por causa da proximidade desse valor com o único ponto de acentuação do cotovelo ($k = 3$), do pico no silhouette e da maior estabilidade da função de custo entre as sementes. O random state também foi 26.

4. RESULTADOS E DISCUSSÃO

O K-means gerou oito agrupamentos a partir dos 240.791 registros analisados. Cada cluster foi interpretado combinando as médias das variáveis numéricas (mapa de calor) com a distribuição das categóricas. A análise não se limitou a identificar valores altos ou baixos isolados, mas buscou entender quais combinações de características distinguem cada agrupamento.

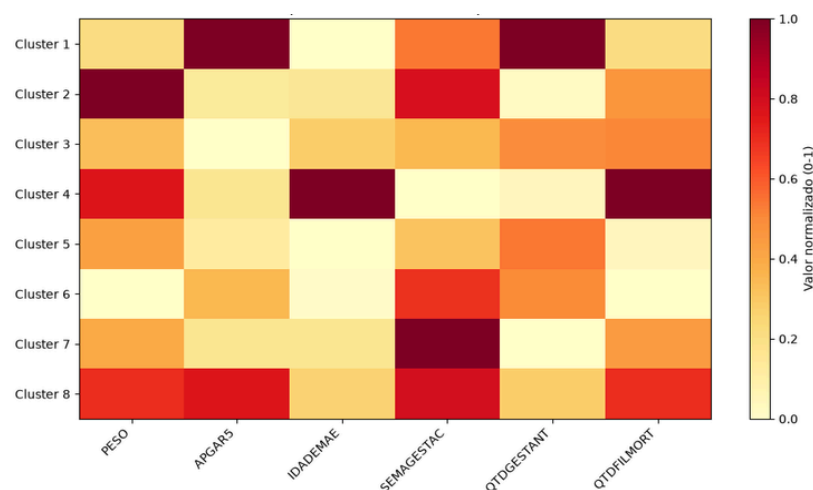
Figura 5. Tamanho dos clusters k-means



Fonte: Elaborada pelos autores (2026).

O mapa de calor das médias numéricas mostra diferenças relevantes entre os oito agrupamentos.

Figura 6. Mapa de calor das médias numéricas k-means



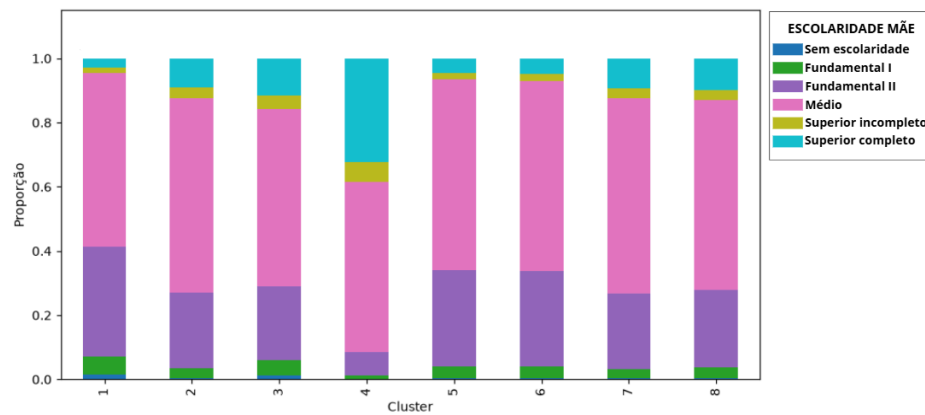
Fonte: Elaborada pelos autores (2026).

O cluster 1 apresenta valores elevados de Apgar 5 e de gestações anteriores, associados a menor idade materna. O cluster 2 se destaca pelo elevado peso ao nascer e pelo menor número de gestações prévias. O cluster 3 tem o menor Apgar 5, com as demais variáveis em níveis intermediários. O cluster 4 concentra a maior idade materna e o maior número de filhos mortos, com menos semanas de gestação e menos gestações anteriores. O cluster 5 tem perfil predominantemente intermediário, com menor idade materna e menor número de filhos mortos. O cluster 6 combina semanas de gestação relativamente elevadas com menores valores de peso, idade materna e filhos mortos. O cluster 7 se caracteriza por longa duração gestacional e menos gestações anteriores. O cluster 8 tem valores intermediários a elevados de peso, Apgar 5, semanas de gestação e filhos mortos.

A diferenciação entre os grupos não vem de uma única variável, mas da combinação de atributos maternos, gestacionais e neonatais, com destaque para os contrastes em idade materna, semanas de gestação e peso ao nascer. Esses atributos têm relevância reconhecida na literatura: Nunes et al. (2021) associaram extremos de idade materna a prematuridade e baixo peso ao nascer, o que dá aos agrupamentos forte coerência epidemiológica, embora não

devam ser lidos como relações causais diretas. Nos gráficos de barras das variáveis categóricas, foram selecionados os que mais chamaram a atenção em relação aos agrupamentos.

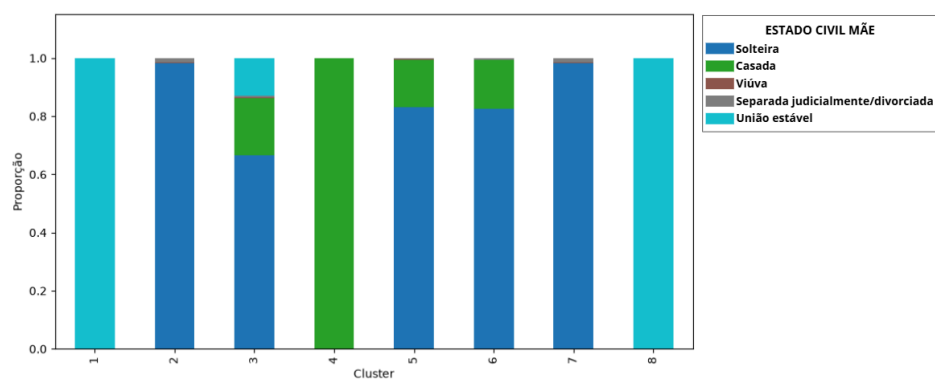
Figura 7. Proporção de escolaridade da mãe por cluster k-means



Fonte: Elaborada pelos autores (2026).

A escolaridade materna teve distribuição relativamente semelhante entre os agrupamentos, com predomínio do ensino médio completo. O cluster 4, porém, teve maior participação relativa de mães com ensino superior incompleto e completo. Essa diferença não parece suficiente, sozinha, para explicar a formação dos clusters, mas ajuda a caracterizar a heterogeneidade sociodemográfica da base. A escolaridade é reconhecida como importante marcador social ligado às condições de nascimento: em análise nacional com dados do SINASC, Silvestrin et al. (2020) observaram diferenças no peso ao nascer entre níveis de escolaridade materna.

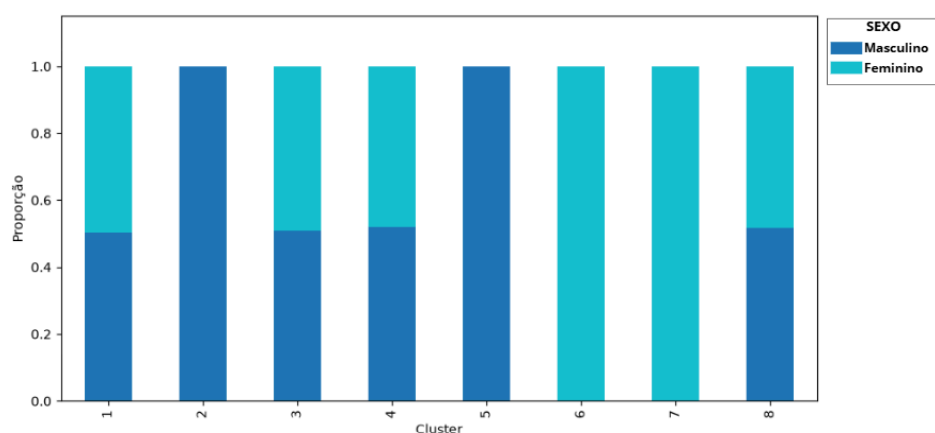
Figura 8. Proporção do estado civil da mãe por cluster k-means



Fonte: Elaborada pelos autores (2026).

Quanto ao estado civil, os clusters 1 e 8 predominam em união estável, o cluster 4 concentra a maior proporção de mães casadas, e os clusters 2 e 7 são formados majoritariamente por mães solteiras. Isso reforça que os agrupamentos não se diferenciam apenas por características neonatais, mas também por características sociodemográficas maternas. Essa composição, no entanto, é descritiva da população agrupada e não indica efeito causal sobre os demais atributos.

Figura 9. Proporção do sexo por cluster

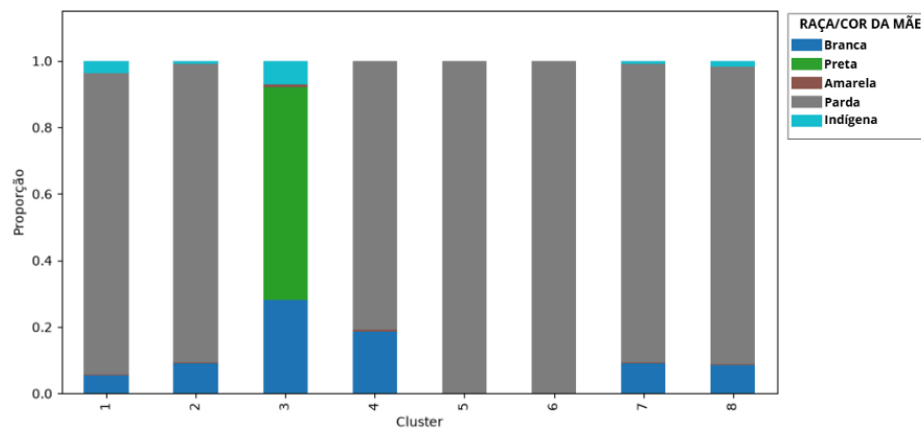


Fonte: Elaborada pelos autores (2026).

O sexo apresentou concentração particularmente alta em alguns agrupamentos. Os clusters 2 e 5 foram formados integralmente por recém-nascidos do sexo masculino, e os clusters 6 e 7, integralmente por recém-nascidos do sexo feminino. Esse comportamento interessa sobretudo do ponto de vista metodológico, pois a variável

foi convertida para representação numérica por One-Hot Encoding antes da aplicação do K-means.

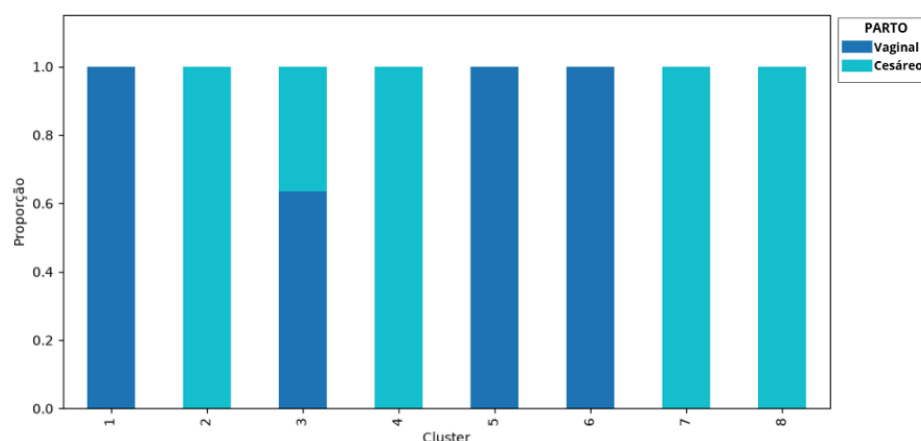
Figura 10. Proporção da Raça/cor da mãe por cluster



Fonte: Elaborada pelos autores (2026).

Na raça/cor materna, a concentração é ainda mais marcante. O cluster 3 reúne todas as mães classificadas como pretas, tem a maior participação de mães indígenas e não possui mães pardas. Já os clusters 5 e 6 são compostos integralmente por mães pardas. Isso é importante para interpretar a solução, mas não é evidência de que raça/cor determine as demais características do agrupamento.

Figura 11. Proporção do tipo de parto por cluster



Fonte: Elaborada pelos autores (2026).

No tipo de parto, os clusters 1, 5 e 6 predominam em partos vaginais, e os clusters 2, 4, 7 e 8, em cesáreas. Analisado junto com as variáveis

numéricas, o tipo de parto mostra que grupos com o mesmo predomínio podem ter perfis bem distintos. A cesárea aparece tanto no cluster 2, caracterizado por maior peso e maior duração gestacional, quanto no cluster 4, com menor duração gestacional e menor Apgar 5. Portanto, o tipo de parto, isoladamente, não explica a formação dos agrupamentos.

Integrando as variáveis analisadas, podem-se estabelecer os seguintes perfis exploratórios:

- Grupo 1: mães jovens e pardas, em união estável e com ensino médio completo, com elevado histórico de gestações anteriores e predomínio do parto vaginal. Os bebês tiveram peso de baixo a intermediário e o maior Apgar 5 entre os agrupamentos.
- Grupo 2: mães jovens, pardas e solteiras, com ensino médio e menor histórico de gestações, que optaram quase integralmente pela cesariana. Todos os recém-nascidos são do sexo masculino, com peso e duração gestacional elevados, mas Apgar 5 de baixo a intermediário.
- Grupo 3: composição racial distinta (todas as mães pretas, maior proporção de indígenas e nenhuma parda), com idade e escolaridade intermediárias e predomínio de solteiras. Predominou o parto vaginal, e os bebês nasceram com peso intermediário e o menor Apgar 5 da análise.
- Grupo 4: mães pardas, mais velhas e casadas, com a maior escolaridade (ensino superior), maior número de filhos mortos e menor tempo de gestação. Predominou a cesariana, e os bebês (majoritariamente do sexo masculino) nasceram com peso elevado, mas Apgar 5 baixo.

- Grupo 5: mães jovens, pardas e solteiras, com ensino médio e elevado histórico gestacional (mas poucas perdas), com 100% de partos vaginais. Todos os bebês são do sexo masculino, com semanas de gestação, peso e Apgar 5 entre níveis baixos e intermediários.
- Grupo 6: perfil materno muito semelhante ao do Grupo 5 (jovens, pardas, solteiras e 100% parto vaginal), mas com 100% de recém-nascidos do sexo feminino. Diferencia-se pela maior duração gestacional e pelo Apgar 5 mais elevado, embora com baixo peso ao nascer.
- Grupo 7: mães pardas e solteiras, de idade baixa a intermediária, com menor histórico de gestações (porém mais perdas anteriores). Predominou a cesariana, e 100% dos bebês são do sexo feminino, com gestação prolongada, peso intermediário e Apgar 5 de baixo a intermediário.
- Grupo 8: mães pardas, de idade intermediária e em união estável, com histórico elevado de filhos mortos. Predominou a cesariana, e os bebês nasceram com valores elevados de semanas de gestação, peso e Apgar 5.

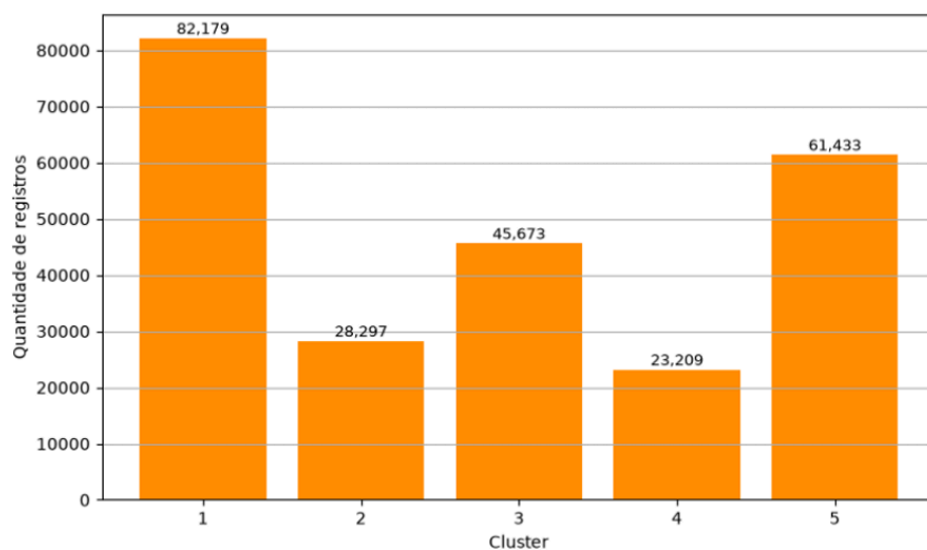
A análise dos oito agrupamentos mostra, portanto, que o K-means produziu perfis heterogêneos a partir da base transformada. Porém, a elevada concentração de certas categorias em alguns clusters indica que a representação das variáveis categóricas pesou bastante na solução encontrada. Por isso, os agrupamentos devem ser lidos como estruturas exploratórias de similaridade, e não como grupos naturais, diagnósticos ou relações causais. Jain (2010) destaca que a análise de clusters é essencialmente exploratória e que a solução

depende fortemente da representação dos dados, da medida de similaridade e do algoritmo empregado.

4.1. Resultados do K-Prototypes

O K-prototypes resultou em cinco agrupamentos. Diferentemente do K-means, esse algoritmo foi desenvolvido para conjuntos que combinam variáveis numéricas e categóricas, tratando cada tipo de atributo de acordo com sua natureza, o que torna a solução especialmente importante para a comparação proposta.

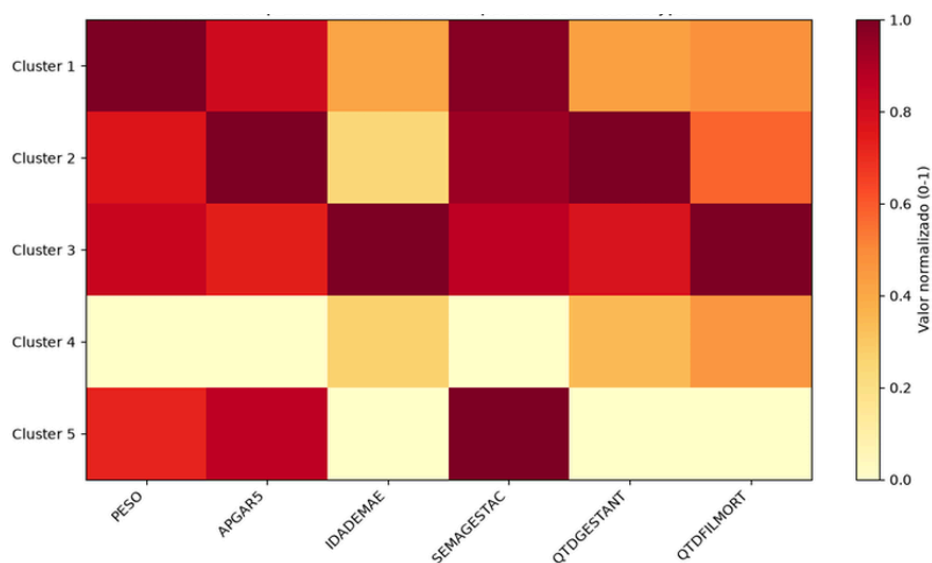
Figura 12. Tamanho dos clusters k-prototypes



Fonte: Elaborada pelos autores (2026).

O mapa de calor das variáveis numéricas mostra uma diferenciação mais concentrada entre os cinco agrupamentos.

Figura 13. Mapa de calor das médias numéricas k-prototypes



Fonte: Elaborada pelos autores (2026).

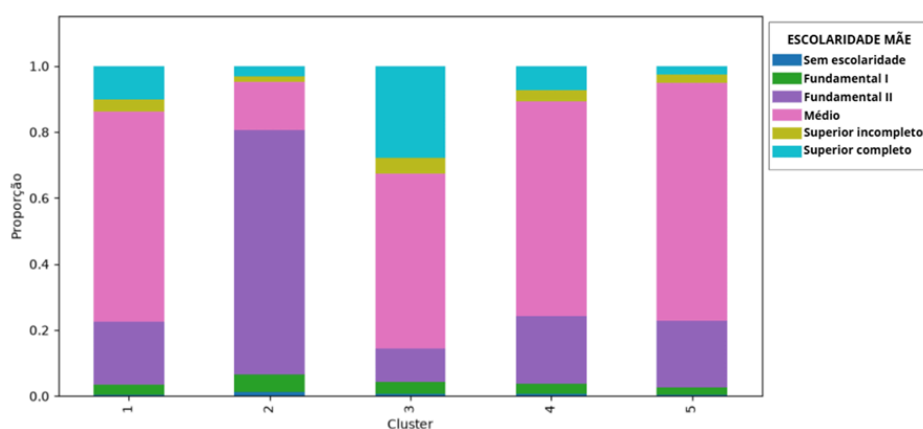
O cluster 1 tem médias elevadas de peso, Apgar 5 e semanas de gestação, com as demais variáveis em níveis predominantemente intermediários. O cluster 2 tem médias elevadas em quase todas as variáveis, exceto idade materna e quantidade de filhos mortos, que ficam em níveis intermediários. O cluster 3 concentra valores elevados em todas as variáveis numéricas. O cluster 4 tem níveis de baixos a intermediários em praticamente todas as variáveis. O cluster 5 tem valores elevados de peso, Apgar 5 e semanas de gestação, com valores baixos nas demais.

A estrutura numérica mostra dois padrões contrastantes. O primeiro é o do cluster 4, que concentra os menores níveis de peso ao nascer, Apgar 5 e semanas de gestação. O segundo é o dos clusters 3 e 5, com valores elevados nessas variáveis. A combinação entre duração gestacional e peso ao nascer é especialmente útil para distinguir perfis neonatais bem diferentes.

O cluster 4 chama atenção por reunir baixo peso, baixo Apgar 5 e menor duração gestacional. Esse conjunto é compatível com um perfil de maior vulnerabilidade neonatal, mas não basta para classificar todos os indivíduos como prematuros ou de baixo peso,

pois isso exigiria analisar os valores individuais segundo pontos de corte clínicos. O resultado deve, portanto, ser tratado como um perfil estatístico identificado pelo algoritmo.

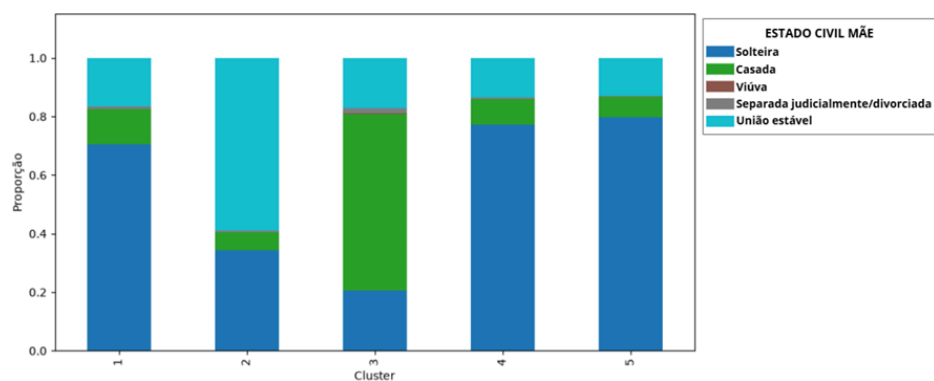
Figura 14. Proporção de escolaridade da mãe por cluster k-prototypes



Fonte: Elaborada pelos autores (2026).

Na escolaridade materna, o cluster 2 se destaca pela maior concentração relativa de mães com ensino fundamental II, e o cluster 3 reúne a maior participação dos níveis mais elevados. Essa diferença complementa a separação observada nas variáveis numéricas e sugere que o algoritmo identificou perfis em que características sociais e neonatais aparecem juntas, o que é compatível com a evidência de que a escolaridade materna é um importante marcador social relacionado ao peso ao nascer (Silvestrin et al., 2020).

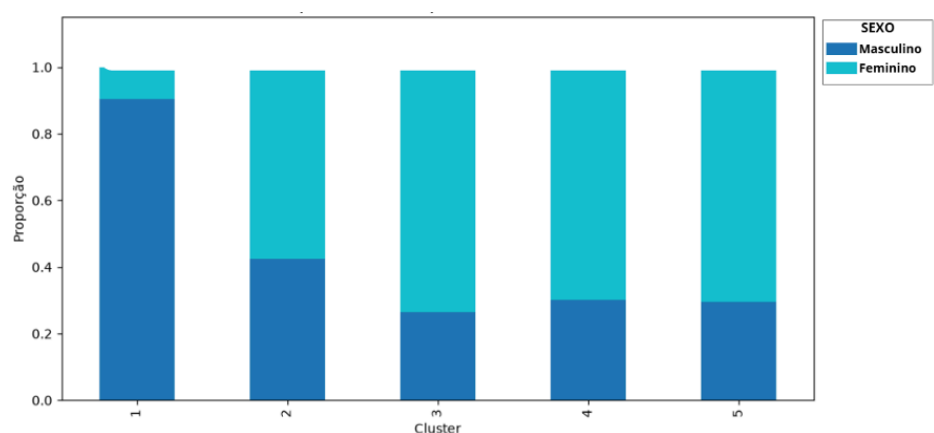
Figura 15. Proporção do estado civil da mãe por cluster k-prototypes



Fonte: Elaborada pelos autores (2026).

No estado civil, o cluster 2 tem maior predomínio de união estável e o cluster 3 concentra a maior proporção de mães casadas. Essas características não explicam sozinhas a formação dos grupos, mas ajudam a construir perfis multidimensionais e a diferenciar agrupamentos que seriam parecidos se apenas as variáveis numéricas fossem consideradas.

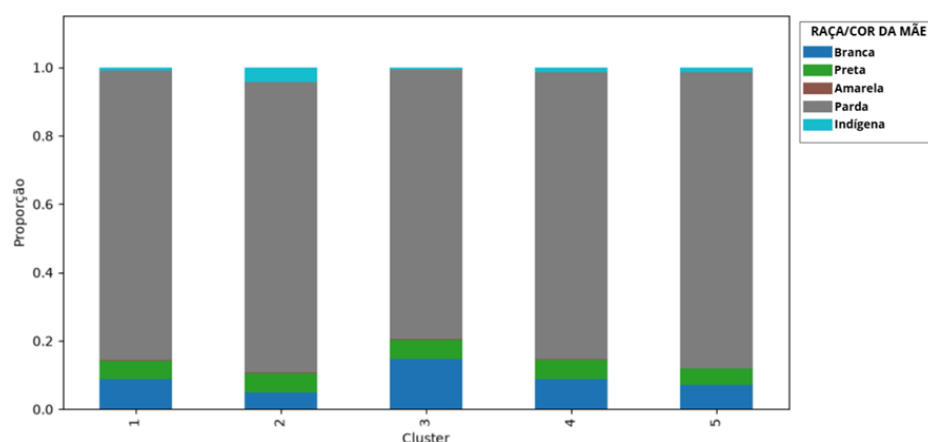
Figura 16. Proporção do Sexo por cluster k-prototypes



Fonte: Elaborada pelos autores (2026).

No sexo, o cluster 1 tem predominância masculina e os demais têm maioria feminina. Ao contrário do K-means, não houve concentrações extremas de 100% de uma categoria em vários agrupamentos. Isso é metodologicamente relevante para a comparação, pois sugere que o tratamento específico dos atributos categóricos pode gerar uma estrutura menos dependente da codificação binária.

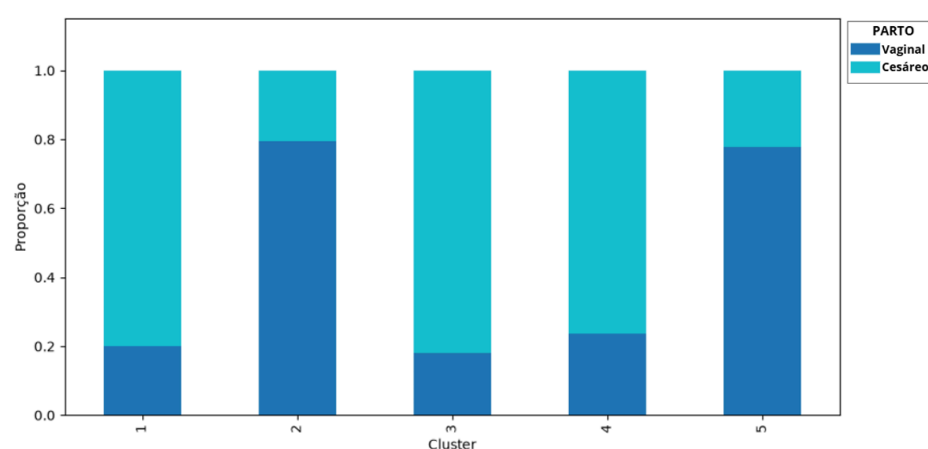
Figura 17. Proporção da raça/cor da mãe por cluster k-prototypes



Fonte: Elaborada pelos autores (2026).

A raça/cor materna varia pouco entre os cinco agrupamentos, em contraste direto com o K-means, em que a variável se concentrou fortemente em determinados clusters. A mudança sugere que parte da separação extrema da primeira solução pode estar ligada à forma como os atributos categóricos foram incorporados à distância usada pelo algoritmo. Essa variável é, portanto, uma das evidências mais importantes para comparar os dois métodos.

Figura 18. Proporção do tipo de parto por cluster k-prototypes



Fonte: Elaborada pelos autores (2026).

No tipo de parto, os clusters 2 e 5 predominam em partos vaginais e os demais em cesáreas. Como no K-means, a variável aparece combinada com perfis numéricos diferentes, o que reforça a

necessidade de interpretá-la junto com as demais características maternas e neonatais.

- Grupo 1: mães pardas e solteiras, com ensino médio e idade levemente acima da mediana, com tempo de gestação alto e parto cesáreo. Os bebês (ampla maioria do sexo masculino) nasceram com peso e Apgar 5 elevados.
- Grupo 2: mães pardas, em união estável e com ensino fundamental II, com o maior número de gestações anteriores e parto vaginal. Os bebês nasceram com peso elevado e Apgar 5 máximo.
- Grupo 3: mães pardas, casadas e mais velhas, com a maior proporção de nível superior e a maior incidência de filhos mortos, com parto cesáreo. Os bebês nasceram com peso e Apgar 5 elevados.
- Grupo 4: mães pardas e solteiras, com ensino médio e parto cesáreo. Os bebês nasceram com os menores índices de peso, Apgar 5 e semanas de gestação, o que caracteriza um claro perfil de prematuridade.
- Grupo 5: mães pardas, mais jovens e em sua grande maioria solteiras (indicativo de primíparas), que completaram o tempo máximo de gestação e tiveram parto vaginal. Os bebês nasceram com peso e Apgar 5 elevados.

4.2. Comparação Entre K-Means e K-Prototype

A comparação mostra diferenças estruturais importantes. O K-means (oito clusters) apresentou elevada concentração nas variáveis

categóricas, com extremos em sexo e raça/cor, enquanto o K-prototypes (cinco clusters) teve distribuição mais equilibrada e maior integração com os perfis numéricos. Isso reflete a concepção de cada algoritmo: o K-means depende da distância euclidiana sobre representações numéricas, o que força o peso de certas categorias na separação dos grupos, enquanto o K-prototypes foi concebido para tratar atributos numéricos e categóricos em conjunto (Huang, 1998).

Apesar da sensibilidade das variáveis sociodemográficas à metodologia, os dois algoritmos identificaram com clareza grupos com características neonatais contrastantes. Semanas de gestação, peso e Apgar 5 mantiveram, portanto, sua relevância na diferenciação dos perfis, qualquer que fosse a técnica. Em suma, a forma de tratar os dados categóricos altera substancialmente a estrutura de similaridade e a composição final dos clusters. A escolha do método de agrupamento não é só uma etapa técnica, mas uma decisão analítica central, que deve considerar a natureza mista da base.

5. CONCLUSÃO

Este estudo aplicou K-means e K-prototypes à base do SINASC de 2022 a 2024 com o objetivo de verificar a capacidade de cada algoritmo de formar agrupamentos interpretáveis e comparar como o tratamento das variáveis categóricas afeta os resultados. A comparação mostrou que a escolha do algoritmo influencia diretamente a configuração dos clusters. O K-means produziu oito agrupamentos com alta concentração em categorias como sexo e raça/cor, indicando que a codificação por One-Hot Encoding pesou de forma desproporcional na formação dos grupos, o que limita a

interpretação epidemiológica desses agrupamentos. O K-prototypes gerou cinco clusters com distribuição mais equilibrada, integrando variáveis numéricas e categóricas de acordo com sua natureza. Em particular, identificou um agrupamento com baixo peso ao nascer, menor duração gestacional e menor Apgar 5, compatível com um perfil de vulnerabilidade neonatal, o que ilustra o potencial da técnica para apontar subgrupos de risco em bases populacionais. Apesar das diferenças estruturais, peso ao nascer, Apgar 5, idade materna e semanas de gestação continuaram relevantes nas duas abordagens, o que mostra que a representação e o tratamento dos dados afetam os resultados e exigem um método alinhado à natureza da base.

O estudo tem limitações metodológicas importantes, como a possível distorção no cálculo de distâncias ao transformar atributos categóricos em numéricos para o K-means. Soma-se a isso a natureza dos registros do SINASC, sujeitos a variações de preenchimento, completude e qualidade. Por isso, os agrupamentos devem ser interpretados como padrões exploratórios de similaridade, capazes de apontar associações mas não relações de causalidade, e não como uma classificação definitiva dos registros.

Como trabalhos futuros, recomenda-se validar a estabilidade dos agrupamentos em diferentes configurações, amostras e recortes temporais ou espaciais, e compará-los com outros métodos adequados a dados mistos. Por fim, os resultados mostram que as técnicas de clustering são úteis para explorar padrões complexos em grandes volumes de dados do SINASC, desde que a escolha do algoritmo e a interpretação dos perfis estejam vinculadas às características, limitações e ao tratamento aplicado aos dados.”

REFERÊNCIAS BIBLIOGRÁFICAS

AMERICAN ACADEMY OF PEDIATRICS. Committee on Fetus and Newborn; AMERICAN COLLEGE OF OBSTETRICIANS AND GYNECOLOGISTS. Committee on Obstetric Practice. The Apgar score. **Pediatrics**, v. 136, n. 4, p. 819–822, 2015.

APGAR, Virginia. A proposal for a new method of evaluation of the newborn infant. **Current Researches in Anesthesia and Analgesia**, v. 32, n. 4, p. 250–259, 1953.

BARBAS, Danielle da Silva; COSTA, Antonio José Leal; LUIZ, Ronir Raggio; KALE, Pauline Lorena. Determinantes do peso insuficiente e do baixo peso ao nascer na cidade do Rio de Janeiro, Brasil, 2001. **Epidemiologia e Serviços de Saúde**, Brasília, v. 18, n. 2, p. 161–170, 2009.

BRASIL. Ministério da Saúde. **Sistema de Informação sobre Nascidos Vivos (SINASC)**. Brasília, DF: Ministério da Saúde, 2022. Disponível em: <https://www.gov.br/saude/pt-br/composicao/svsa/sistemas-de-informacao/sinasc>. Acesso em: 26 set. 2026.

CERDA, Patricio; VAROQUAUX, Gaël. Encoding high-cardinality string categorical variables. **IEEE Transactions on Knowledge and Data Engineering**, v. 34, n. 3, p. 1164–1176, 2022.

DONDÉ, Juliana Perotoni; SONCINI, Thaise Cristina Brancher; NUNES, Rodrigo Dias. Fatores associados ao baixo índice de Apgar no quinto minuto de vida em recém-nascidos. **Arquivos Catarinenses de Medicina**, v. 49, n. 3, p. 69–80, 2020.

FERNANDES, Fernando Timoteo; CHIAVEGATTO FILHO, Alexandre Dias Porto. Perspectivas do uso de mineração de dados e aprendizado de máquina em saúde e segurança no trabalho. **Revista Brasileira de Saúde Ocupacional**, v. 44, e13, 2019.

GAN, Guojun. **Data clustering with Python: from theory to implementation**. Boca Raton: Chapman & Hall/CRC, 2026. 260 p.

GOWER, John C. A general coefficient of similarity and some of its properties. **Biometrics**, v. 27, n. 4, p. 857–871, 1971.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. **Data mining: concepts and techniques**. 3. ed. Waltham: Morgan Kaufmann, 2011. 744 p.

HUANG, Zhexue. Clustering large data sets with mixed numeric and categorical values. In: PROCEEDINGS OF THE FIRST PACIFIC ASIA KNOWLEDGE DISCOVERY AND DATA MINING CONFERENCE, 1997, Singapore. **Anais...** Singapore: World Scientific, 1997. p. 21–34.

HUANG, Zhexue. Extensions to the k-means algorithm for clustering large data sets with categorical values. **Data Mining and Knowledge Discovery**, v. 2, n. 3, p. 283–304, 1998.

JAIN, Anil K. Data clustering: 50 years beyond K-means. **Pattern Recognition Letters**, v. 31, n. 8, p. 651–666, 2010.

KAUFMAN, Leonard; ROUSSEEUW, Peter J. **Finding groups in data: an introduction to cluster analysis**. New York: Wiley, 1990. 342 p.

LITTLE, Roderick J. A.; RUBIN, Donald B. **Statistical analysis with missing data**. 3. ed. Hoboken: Wiley, 2019. 462 p.

MACQUEEN, James. Some methods for classification and analysis of multivariate observations. In: BERKELEY SYMPOSIUM ON MATHEMATICAL STATISTICS AND PROBABILITY, 5., 1967, Berkeley. **Proceedings...** Berkeley: University of California Press, 1967. v. 1, p. 281–297.

NUNES, Flávia Baluz Bezerra de Farias; SILVA, Polyana Cabral da; BARBOSA, Thays Luanny Santos Machado; LOPES, Maria Lucia Holanda; SILVA, Elza Lima da. Influence of maternal age in perinatal conditions in live births of São Luís, Maranhão. **Revista de Pesquisa: Cuidado é Fundamental Online**, v. 12, p. 292–299, 2021.

OLSON, Ryan; THOMPSON, Sharon V.; WIPFLI, Brad; HANSON, Ginger; ELLIOT, Diane L.; ANGER, W. Kent; BODNER, Todd; HAMMER, Leslie B.; HOHN, Elliot; PERRIN, Nancy A. Sleep, dietary, and exercise behavioral clusters among truck drivers with obesity: implications for interventions. **Journal of Occupational and Environmental Medicine**, v. 58, n. 3, p. 314–321, 2016.

ROMERO, Dalia E.; CUNHA, Cynthia Braga da. Avaliação da qualidade das variáveis epidemiológicas e demográficas do Sistema de Informações sobre Nascidos Vivos, 2002. **Cadernos de Saúde Pública**, v. 23, n. 3, p. 701–714, 2007.

SILVA, Laura Pedroza da; MOREIRA, Claudia Maria Marques; AMORIM, Maria Helena Costa; CASTRO, Denise Silveira de; ZANDONADE, Eliana. Avaliação da qualidade dos dados do Sistema de Informações sobre Nascidos Vivos e do Sistema de Informações sobre Mortalidade no período neonatal, Espírito Santo, Brasil, de 2007 a 2009. **Ciência & Saúde Coletiva**, v. 19, n. 7, p. 2011–2020, 2014.

SILVESTRIN, Sonia et al. Inequalities in birth weight and maternal education: a time-series study from 1996 to 2013 in Brazil. **Scientific Reports**, v. 10, art. 8707, 2020.

SIMON, Leslie V.; SHAH, Manan; BRAGG, Bradley N. APGAR Score. In: STATPEARLS. Treasure Island: StatPearls Publishing, 2026. Disponível em: <https://www.ncbi.nlm.nih.gov/books/NBK470569/>. Acesso em: 26 set. 2026.

SZWARCWALD, Célia Landmann; LEAL, Maria do Carmo; ESTEVES-PEREIRA, Ana Paula; ALMEIDA, Wanessa da Silva de; FRIAS, Paulo Germano de; DAMACENA, Giseli Nogueira; SOUZA JÚNIOR, Paulo Roberto Borges de; ROCHA, Narayani Martins; MULLACHERY, Priscila Melissa Honorato. Avaliação das informações do Sistema de Informações sobre Nascidos Vivos (SINASC), Brasil. **Cadernos de Saúde Pública**, v. 35, n. 10, e00214918, 2019.