REMUNOM
ISSN 2178-6925

Received: 13/01/2026 - **Accepted:** 18/01/2026
Vol: 01.02
DOI: 10.61164/wc3f6213
Pages: 1-26

ANÁLISE DO PERFIL DAS VÍTIMAS DE ACIDENTES FATAIS POR QUEDAS DE ALTURA DE 1996 A 2023

ANALYSIS OF THE PROFILE OF VICTIMS OF FATAL ACCIDENTS DUE TO FALLS FROM HEIGHT FROM 1996 TO 2023

ANÁLISIS DEL PERFIL DE LAS VÍCTIMAS DE ACCIDENTES FATALES POR CAÍDAS DE ALTURA DE 1996 A 2023

Fabricio Rodrigues de Souza
Mestrando em Computação Aplicada,
Universidade Federal do Mato Grosso do Sul, Brasil
E-mail: fabricio.r@ufms.br

Dionisio Machado Leite Filho
Doutor em Ciências da Computação e Matemática Computacional,
Universidade Federal do Mato Grosso do Sul, Brasil
E-mail: dionisio.leite@ufms.br

Resumo

As quedas em altura representam um grave problema de saúde pública no Brasil, extrapolando os limites do trabalho formal e exigindo estratégias baseadas em evidências. Este estudo caracterizou perfis de vulnerabilidade em óbitos por quedas de escada de mão (W11), andaimes (W12) e edifícios (W13), utilizando dados do DATASUS (1996-2023). Aplicou-se a mineração de dados (K-Means) para segmentação por sexo, idade e escolaridade, validada pelos métodos cotovelo e silhueta. Os resultados revelaram um núcleo de vulnerabilidade crítica composto predominantemente por homens (50-59 anos) em todas as categorias, destacando-se idosos (60-69 anos) nas quedas de escada de mão. Nos grupos de maior risco, predominou a escolaridade de 4 a 7 anos(fundamental incompleto). Conclui-se que o K-Means isolou eficazmente padrões de risco, evidenciando que a fatalidade atinge severamente trabalhadores em envelhecimento e baixa instrução, em contextos laborais e informais.

Palavras-chave: Quedas em altura; Mineração de dados; Algoritmo K-Means; Acidentes de trabalho.

REMUNOM
ISSN 2178-6925

Received: 13/01/2026 - **Accepted:** 18/01/2026
Vol: 01.02
DOI: 10.61164/wc3f6213
Pages: 1-26

Abstract

Falls from height constitute a serious public health issue in Brazil, extending beyond the boundaries of formal employment and requiring evidence-based strategies. This study characterized vulnerability profiles in fatalities caused by falls from ladders (W11), scaffolds (W12), and buildings (W13), utilizing data from DATASUS (1996–2023). Data mining (K-Means) was applied for segmentation by sex, age, and education level, validated by the Elbow and Silhouette methods. The results revealed a critical vulnerability core composed predominantly of men (50–59 years) across all categories, with older adults (60–69 years) standing out in falls from ladders. In the highest-risk groups, the predominant education level was 4 to 7 years of schooling (incomplete elementary education). It is concluded that K-Means effectively isolated risk patterns, demonstrating that fatalities severely affect aging workers with low educational attainment in both occupational and informal contexts.

Keywords: Falls from height; Data mining; K-Means algorithm; Occupational accidents.

Resumen

Las caídas de altura representan un grave problema de salud pública en Brasil, trascendiendo los límites del trabajo formal y requiriendo estrategias basadas en evidencia. Este estudio caracterizó perfiles de vulnerabilidad en defunciones por caídas desde escaleras de mano (W11), andamios (W12) y edificios (W13), utilizando datos del DATASUS (1996-2023). Se aplicó minería de datos (K-Means) para la segmentación por sexo, edad y escolaridad, validada por los métodos del codo y de la silueta. Los resultados revelaron un núcleo de vulnerabilidad crítica compuesto predominantemente por hombres (50-59 años) en todas las categorías, destacándose los adultos mayores (60-69 años) en las caídas desde escaleras de mano. En los grupos de mayor riesgo, predominó la escolaridad de 4 a 7 años (educación primaria incompleta). Se concluye que el K-Means aisló eficazmente patrones de riesgo, evidenciando que la fatalidad afecta severamente a trabajadores en proceso de envejecimiento y con baja instrucción, tanto en contextos laborales como informales.

Palabras clave: Caídas de altura; Minería de datos; Algoritmo K-Means; Accidentes de trabajo.

1. Introdução

As quedas em altura configuram-se como um grave problema de saúde pública e ocupacional. Segundo a Organização Mundial da Saúde (2024), elas representam a segunda principal causa de mortes por lesões não intencionais em todo o mundo. Essa magnitude evidencia que o problema não respeita as fronteiras burocráticas entre o trabalho formal e informal, gerando óbitos tanto em

REMUNOM
ISSN 2178-6925

Received: 13/01/2026 - **Accepted:** 18/01/2026
Vol: 01.02
DOI: 10.61164/wc3f6213
Pages: 1-26

canteiros de obras regulados quanto em atividades de manutenção predial autônoma ou doméstica.

No Brasil, as quedas em altura estão entre as causas mais frequentes de fatalidades no ambiente laboral, exigindo uma atenção rigorosa quanto às normas de segurança. O trabalho formal é regido pela Consolidação das Leis do Trabalho (CLT) e as práticas de prevenção são orientadas pela Norma Regulamentadora 35 (NR-35), que estabelece requisitos para qualquer atividade executada acima de 2,00 metros do nível inferior (BRASIL, 2018).

Contudo, a proteção legal restrita ao mercado formal não tem sido suficiente para conter a letalidade desses eventos. Como observa Saraiva (2023), apesar das regulamentações específicas, a persistência de altos índices de mortalidade evidencia lacunas na segurança que exigem análises mais profundas sobre a vulnerabilidade dos trabalhadores.

Essa realidade sugere que uma parcela significativa dos acidentes ocorre em contextos de informalidade ou manutenção precária, alinhando-se ao alerta da Organização Mundial da Saúde (2024) sobre as quedas constituírem um desafio de saúde pública global que extrapola o ambiente ocupacional regulado.

Para a compreensão desse cenário, o Brasil dispõe de bases de dados abrangentes, como o datasus. No entanto, analisar o vasto volume de registros acumulados ao longo de décadas cobrindo todas as regiões do país é uma tarefa complexa.

O gerenciamento de grandes volumes de dados através de métodos manuais é suscetível a erros e pode não revelar padrões ocultos essenciais para a prevenção. Nesse contexto, a utilização de técnicas estatísticas avançadas e ferramentas computacionais, como a mineração de dados, torna-se indispensável para transformar dados brutos em conhecimento aplicável à tomada de decisões (SÍCÂKYÜZ et al., 2024).

Diante dessa necessidade analítica, a técnica de agrupamento de dados apresenta-se como uma estratégia fundamental, destacando-se por sua capacidade de identificar similaridades e segmentar grandes conjuntos de dados em grupos com características homogêneas, sem a necessidade de

REMUNOM
ISSN 2178-6925

Received: 13/01/2026 - **Accepted:** 18/01/2026
Vol: 01.02
DOI: 10.61164/wc3f6213
Pages: 1-26

categorização prévia (Dinh et al., 2025).

A motivação para este estudo surge da necessidade prática, no campo da engenharia de segurança do trabalho, de ir além da análise quantitativa simples e identificar padrões ocultos nos dados históricos do datasus. Compreender quais grupos de trabalhadores apresentam maior propensão a acidentes fatais é essencial para subsidiar a tomada de decisão técnica.

Diante disso, a questão que norteia esta pesquisa é: como a técnica de clusterização k-means pode ser aplicada aos dados históricos de acidentes fatais para identificar perfis de vulnerabilidade e subsidiar medidas preventivas?

O objetivo geral desta pesquisa é aplicar o algoritmo k-means para identificar os perfis demográficos de vítimas de acidentes fatais por quedas de altura (escadas, andaimes e edifícios e outras estruturas), com base em dados históricos do datasus (1996–2023), a fim de gerar informações que apoiem a formulação de estratégias de prevenção e melhoria na segurança contra acidentes.

Para alcançar esse propósito, definem-se os seguintes objetivos específicos:

- Estruturar e pré-processar datasets a partir da base histórica, focando nas variáveis de quedas (W11-Quedas em escada de mão, W12-Quedas em ou de um andaime e W13 - Quedas de edifícios e outras estruturas);
- Aplicar o algoritmo K-means, determinando o número ideal de grupos através de métodos de validação (Cotovelo e Silhueta);
- Avaliar e interpretar os clusters obtidos para caracterizar os perfis de risco (idade, escolaridade e gênero);
- Gerar informações referentes aos perfis sociodemográficos
- mais suscetíveis a quedas de altura, visando subsidiar a elaboração de medidas preventivas direcionadas.

2. Estrutura teórica

2.1 Quedas em altura

REMUNOM
ISSN 2178-6925

Received: 13/01/2026 - **Accepted:** 18/01/2026
Vol: 01.02
DOI: 10.61164/wc3f6213
Pages: 1-26

No Brasil, o trabalho formal é regido pela Consolidação das Leis do Trabalho (CLT) e amparado pela Lei nº 8.213/1991, que define acidente de trabalho como aquele ocorrido pelo exercício do trabalho a serviço de empresa (BRASIL, 1991).

As práticas de segurança são orientadas pelas Normas Regulamentadoras (NRs), sendo a NR-35, que trata do trabalho em altura, uma das mais relevantes. Essa norma estabelece que qualquer atividade executada acima de 2,00 metros do nível inferior, com risco de queda, é considerada trabalho em altura (BRASIL, 2018).

Para mitigar os riscos de quedas, a Norma Regulamentadora 35 (NR-35) estabelece requisitos mínimos e medidas de proteção para o trabalho em altura, envolvendo planejamento, organização e execução. Sob a ótica prevencionista, a aplicação rigorosa dessa norma é a barreira primária contra fatalidades, exigindo uso de EPIs, sistemas de ancoragem e capacitação técnica (BRASIL, 2018).

A gravidade das quedas vai além do cenário ocupacional regulado e configura-se como um desafio global de saúde pública. De acordo com as estimativas mais recentes da Organização Mundial da Saúde, as quedas representam a segunda principal causa de mortes por lesões não intencionais em todo o mundo.

Os dados globais indicam que, em 2021, as quedas foram responsáveis por aproximadamente 684 mil óbitos (ORGANIZAÇÃO MUNDIAL DA SAÚDE, 2024).

Apesar da regulamentação rigorosa da NR-35 para o mercado formal, é importante notar que as quedas em altura também atingem trabalhadores em contextos de informalidade, autoconstrução e ambientes domésticos. Nesses cenários, a ausência de fiscalização e proteção coletiva amplia a vulnerabilidade, gerando um perfil de acidentalidade misto que, conforme será explorado nesta pesquisa, é captado pelos registros universais de mortalidade pública datasus.

2.2 Análise de clusters e métodos de agrupamento

A análise de clusters é uma técnica de aprendizado de máquina não supervisionado, utilizada para segmentar dados em grupos que compartilham

REMUNOM
ISSN 2178-6925

Received: 13/01/2026 - **Accepted:** 18/01/2026
Vol: 01.02
DOI: 10.61164/wc3f6213
Pages: 1-26

características semelhantes, buscando maximizar a homogeneidade dentro dos grupos e a heterogeneidade entre eles (JAIN, 2010).

Devido a essa natureza exploratória, a técnica pode ser aplicada em estudos de segurança contra quedas para entender o comportamento dos acidentes a partir de características comuns dos envolvidos.

2.2.1 Algoritmo K-means

O K-Means é um dos algoritmos de clustering mais utilizados devido à sua simplicidade e eficiência. O algoritmo busca dividir um conjunto de dados em K clusters predefinidos, nos quais os pontos de dados em cada cluster são os mais semelhantes possível. O processo começa com a escolha de K centros iniciais (chamados de centroides) e, em seguida, os dados são atribuídos ao cluster mais próximo desses centros. Após essa etapa, os centroides são recalculados com base nas novas atribuições, e o processo é repetido até que as mudanças nos centroides sejam mínimas, indicando que o algoritmo convergiu (MACQUEEN, 1967).

O K-Means é amplamente usado por sua capacidade de processar grandes volumes de dados rapidamente e por ser relativamente fácil de implementar. Ele é eficaz na identificação de agrupamentos quando os dados possuem características claras e distintas entre os clusters (MACQUEEN, 1967).

2.2.2 Método do cotovelo

O Método do Cotovelo é uma técnica usada para identificar o número ideal de clusters (K) a ser utilizado no algoritmo K-Means. O método envolve calcular o valor da soma das distâncias quadradas entre os pontos de dados e seus respectivos centroides para diferentes valores de K. O gráfico resultante mostra uma diminuição rápida do erro à medida que o número de clusters aumenta, seguida por uma desaceleração significativa. O ponto de inflexão, onde a redução do erro começa a diminuir de forma mais gradual, é chamado de "cotovelo" e indica o número ótimo de clusters (THORNDIKE, 1953).

Esse método é essencial para evitar o sobreajuste ou o subajuste dos dados, garantindo que o número de clusters seja adequado à estrutura dos dados,

REMUNOM
ISSN 2178-6925

Received: 13/01/2026 - **Accepted:** 18/01/2026
Vol: 01.02
DOI: 10.61164/wc3f6213
Pages: 1-26

sem dividir excessivamente os grupos nem manter os clusters excessivamente grandes (THORNDIKE, 1953).

2.2.3 Método da silhueta

Após a definição do número de clusters, o método da silhueta é utilizado para validar a qualidade dos clusters formados. A silhueta é uma medida que avalia quão bem os pontos dentro de um cluster estão separados dos pontos de outros clusters. A pontuação da silhueta varia de -1 a 1, com valores próximos de 1 indicando que os pontos estão bem agrupados dentro de seus clusters, enquanto valores próximos de -1 sugerem que os pontos podem ter sido atribuídos ao cluster errado (ROUSSEEUW, 1987).

A fórmula da silhueta para um ponto (i) é dada por:

$$S(i) = \frac{b(i) - a(i)}{Max(a(i), b(i))}$$

Onde:

- a(i) é a média da distância entre o ponto (i) e todos os outros pontos dentro do mesmo cluster.
- b(i) é a média da distância entre o ponto (i) e todos os pontos do cluster mais próximo

Este método é fundamental para garantir que os clusters formados sejam bem definidos e coesos, ajudando a identificar se a segmentação realizada é válida e se os grupos encontrados refletem verdadeiramente os padrões presentes nos dados (ROUSSEEUW, 1987).

3. Metodologia

Este estudo tem como objetivo analisar os acidentes fatais causados por quedas em altura, como em andaimes, escadas, edifícios e outras estruturas, com foco em identificar os perfis mais vulneráveis a esses acidentes.

A coleta direta de dados foi realizada via interface web na plataforma pública TabNet (tabnet.datasus.gov.br), utilizando como fonte o Sistema de Informações sobre Mortalidade (SIM). A consulta abrangeu a série histórica de 1996 a 2023, englobando todas as grandes regiões do Brasil (Norte, Nordeste, Sudeste, Sul e Centro-Oeste).

REMUNOM
ISSN 2178-6925

Received: 13/01/2026 - **Accepted:** 18/01/2026
Vol: 01.02
DOI: 10.61164/wc3f6213
Pages: 1-26

O critério de inclusão baseou-se na causa básica do óbito, segundo a Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde (CID-10), filtrando-se estritamente as categorias: W11 (Queda de ou em escada de mão), W12 (Queda de ou em andaime) e W13 (Queda de ou em edifícios ou outras estruturas). Ressalta-se como limitação do estudo que a extração via TabNet/Datasus não possibilita filtrar separadamente cada subcategoria das categorias existentes nos acidentes citados, visto que os dados públicos disponibilizados abrangem o contexto geral de cada queda. Os dados, em sua totalidade, referem-se aos óbitos por essas causas, englobando tanto acidentes laborais quanto domésticos ou de lazer. Assume-se, contudo, que as categorias W12 (Andaimes) e W13 (Edifícios e outras estruturas) possuem forte correlação com atividades laborais, enquanto W11 (Escadas) apresenta natureza mista.

Foram criados três datasets, cada dataset foi organizado com base em perfis demográficos, onde cada linha representa uma combinação das variáveis sexo, escolaridade, idade e o número total de acidentes associados a esse perfil, e não um acidente isolado. Ou seja, para cada perfil, foi registrado o número de acidentes fatais que ocorreram para as diferentes combinações dessas variáveis.

As variáveis analisadas incluem sexo, escolaridade e faixa etária. A variável sexo foi categorizada entre masculino e feminino. A variável escolaridade foi dividida em várias categorias, 4 a 7 anos, 8 a 11 anos, 1 a 3 anos, 1 a 8 anos, 12 anos ou mais e nenhuma. Quanto à idade, foram criadas as faixas etárias de 15 a 19 anos, 20 a 29 anos, 30 a 39 anos, 40 a 49 anos, 50 a 59 anos, 60 a 69 anos, 70 a 79 anos e 80 anos ou mais.

Preliminarmente ao processamento, foi realizada uma etapa de limpeza dos dados que englobou a exclusão de perfis com escolaridade ignorada e idade ignorada. Essas variáveis indicam falhas no registro de informações do acidentado no momento da coleta de dados e sua remoção foi necessária para garantir a coesão dos grupos formados, evitando distorções no cálculo das distâncias pelo algoritmo.

REMUNOM
ISSN 2178-6925

Received: 13/01/2026 - **Accepted:** 18/01/2026
Vol: 01.02
DOI: 10.61164/wc3f6213
Pages: 1-26

Após a construção dos datasets, foi realizado o pré-processamento dos dados, incluindo as seguintes etapas:

- Normalização das Variáveis Numéricas: A variável acidentes, representando o número de acidentes fatais, foi normalizada utilizando o StandardScaler (Z-Score), por meio da ferramenta StandardScaler da biblioteca scikit-learn. Isso garantiu que a variável tivesse média 0 e desvio padrão 1, um passo essencial para que o K-Means pudesse processar os dados sem ser influenciado pela escala das variáveis. Para evitar que variáveis com maiores magnitudes dominassem o cálculo da distância Euclidiana, aplicou-se a normalização utilizando o StandardScaler (Z-Score), reescalando todos os atributos para uma distribuição com média 0 e desvio padrão 1.
- Codificação das Variáveis Categóricas: As variáveis sexo, faixa etária e escolaridade foram codificadas numericamente para permitir que o K-Means as utilizasse corretamente no cálculo da distância. O One-Hot Encoding foi aplicado para transformar essas variáveis em colunas binárias, representando as diferentes categorias de forma adequada para o algoritmo. Isso impede que o algoritmo atribua erroneamente uma hierarquia ou distância matemática inexistente entre as categorias (como interpretar que o valor 2 seria o dobro do valor 1), eliminando o viés numérico. Dessa forma, todas as variáveis categóricas e faixas etárias passaram a operar na mesma escala de magnitude (0 ou 1), dispensando a necessidade de normalização numérica adicional para estes atributos, uma vez que o viés de escala foi eliminado pela binarização.

Essas etapas de normalização e codificação foram implementadas utilizando as bibliotecas pandas e numpy para manipulação de dados e as ferramentas da biblioteca scikit-learn para a aplicação dos métodos de transformação.

REMUNOM
ISSN 2178-6925

Received: 13/01/2026 - **Accepted:** 18/01/2026
Vol: 01.02
DOI: 10.61164/wc3f6213
Pages: 1-26

Para identificar padrões nos dados, foi aplicada a técnica de análise de clusters, utilizando o algoritmo K-Means. O número ideal de clusters foi definido pelo método do cotovelo, que analisou a soma das distâncias dentro de cada cluster em função do número de clusters. O ponto de inflexão identificado no gráfico indicou o número ideal de clusters, que foi validado posteriormente através do método da silhueta, garantindo que os clusters fossem bem definidos e separáveis.

Com a definição do número de clusters, o algoritmo K-Means foi utilizado para agrupar os dados e identificar os perfis mais suscetíveis a acidentes fatais. Esse processo foi realizado para os três datasets, com cada um representando uma categoria de queda em altura. Isso permitiu identificar os perfis mais vulneráveis a acidentes fatais em cada categoria.

Com base nos resultados obtidos, essas informações serão essenciais para o planejamento e a implementação de medidas preventivas direcionadas, com o objetivo de reduzir os índices de fatalidades.

4. Resultados e discussões

4.1 Descrição do conjunto de dados

O Datasus é uma plataforma pública que oferece dados sobre mortes ocorridas no Brasil. Esses dados são categorizados segundo o sistema CID-10, um padrão internacional de codificação adotado pela Organização Mundial da Saúde. A CID-10 classifica doenças, condições médicas e acidentes, permitindo que essas informações sejam registradas de forma padronizada. No caso de lesões e acidentes, os códigos da CID-10 começam com a letra 'W', que abrange as lesões causadas por acidentes, incluindo especificamente as quedas, como as que ocorrem em escadas, andaimes e edifícios.

Foram criados 3 datasets referentes aos seguintes acidentes de quedas em altura: W11 - Quedas em escada de mão, W12 - Quedas em andaime e W13 - Quedas de edifícios e outras estruturas."

REMUNOM
ISSN 2178-6925

Received: 13/01/2026 - **Accepted:** 18/01/2026
Vol: 01.02
DOI: 10.61164/wc3f6213
Pages: 1-26

Tabela 1. Quantitativo de registros e acidentes por categoria de queda

Tipo de queda	Nº de perfis	Nº de acidentes
W11 – Escada de mão	890	1134
W12 – Andaime	1258	2353
W13 – Edifícios / Estruturas	3796	13484

Fonte: Elaborado pelos autores com base em dados do DATASUS (2025).

Para a correta interpretação dos dados expostos, faz-se necessária uma distinção metodológica entre as métricas apresentadas. A coluna 'Nº de Acidentes' contabiliza a totalidade absoluta de óbitos registrados (vítimas fatais) no período. Por outro lado, a coluna 'Nº de Perfis' indica o tamanho da amostra efetivamente processada pelo algoritmo K-Means, correspondendo à quantidade de linhas únicas na base de dados. Essa diferença numérica ocorre devido à natureza agregada dos dados do Datasus, onde um único perfil (uma combinação idêntica de ano, região, sexo, idade e escolaridade) pode englobar múltiplos óbitos. Portanto, o agrupamento foi realizado sobre os padrões de ocorrência, ponderando-se a densidade de cada perfil.

4.2 Determinação e validação do número de clusters

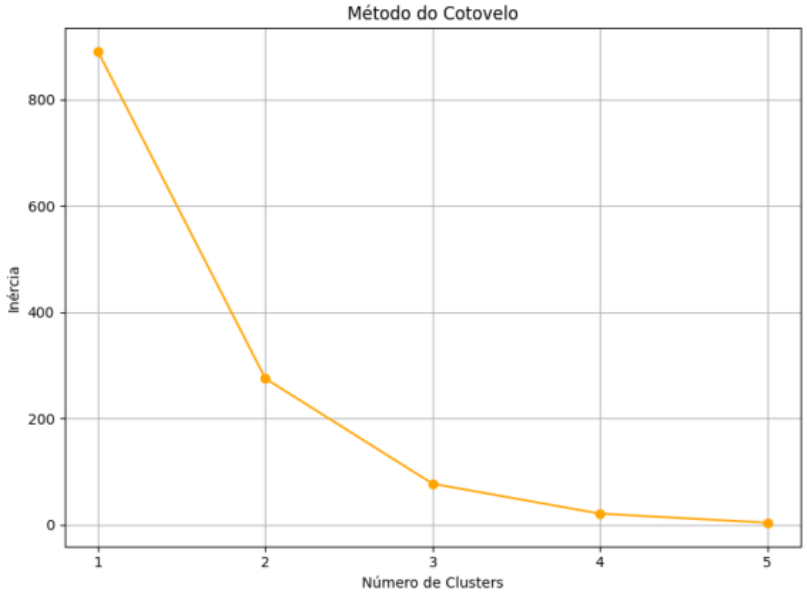
Para garantir que o agrupamento dos dados refletisse perfis reais de vulnerabilidade e não divisões aleatórias, a definição do número ideal de clusters (k) foi realizada através do método do cotovelo. A consistência desses agrupamentos foi validada pelo cálculo do coeficiente de silhueta.

Para o conjunto de dados referente a Quedas em escada de mão (W11), a análise da soma dos erros quadráticos em função do número de clusters indicou um ponto de inflexão ("cotovelo") em k=3, conforme apresentado na Figura 1.

REMUNOM
ISSN 2178-6925

Received: 13/01/2026 - **Accepted:** 18/01/2026
Vol: 01.02
DOI: 10.61164/wc3f6213
Pages: 1-26

Figura 1. Aplicação do Método do Cotovelo para determinação do número de clusters no dataset W11 (Quedas em escada de mão)



Fonte: Elaborado pelos autores (2025)

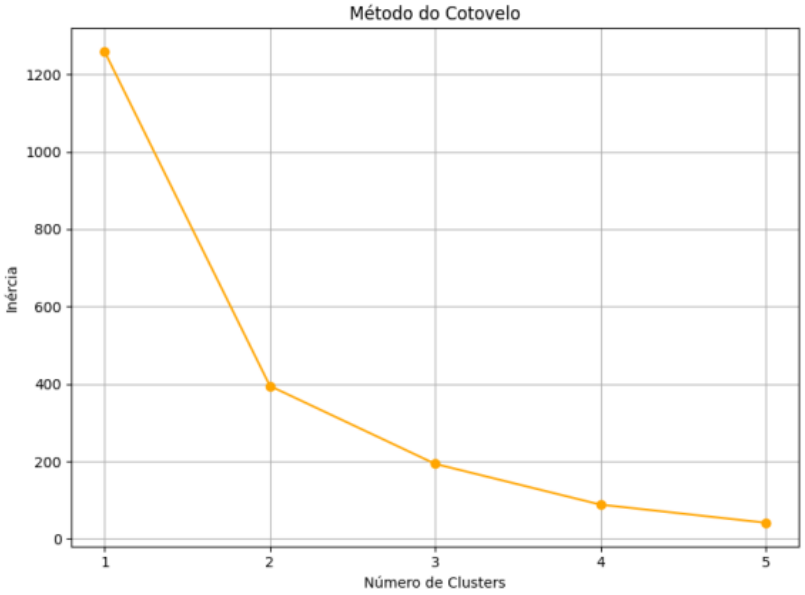
A escolha de 3 clusters foi confirmada pelo coeficiente de silhueta, que apresentou um valor médio de 0.9488, indicando uma coesão adequada dos grupos formados.

Para os cenários de Quedas em andaime (W12) e Quedas de edifícios (W13), a aplicação do método indicou um comportamento similar, onde a redução da variância se estabilizou significativamente a partir de k=2, sugerindo que a divisão em dois grupos principais é a que melhor representa a estrutura desses dados.

REMUNOM
ISSN 2178-6925

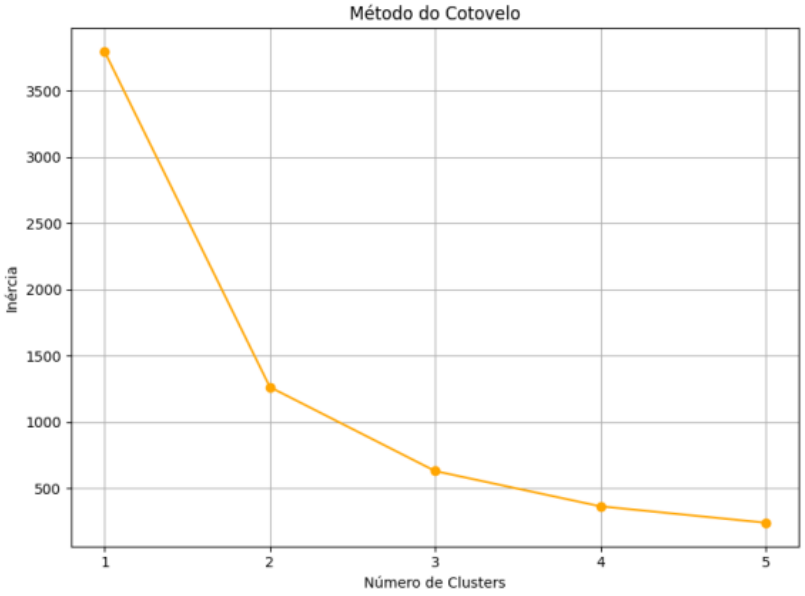
Received: 13/01/2026 - **Accepted:** 18/01/2026
Vol: 01.02
DOI: 10.61164/wc3f6213
Pages: 1-26

Figura 2. Aplicação do Método do Cotovelo para determinação do número de clusters no dataset W12 (Quedas em andaime)



Fonte: Elaborado pelos autores (2025)

Figura 3. Aplicação do Método do Cotovelo para determinação do número de clusters no dataset W13 (Quedas de edifícios e outras estruturas)



Fonte: Elaborado pelos autores (2025)

A validação por meio da silhueta para estes datasets confirmou a qualidade da segmentação, apresentando índices de 0.7864 para andaimes e 0.8049 para

REMUNOM
ISSN 2178-6925

Received: 13/01/2026 - **Accepted:** 18/01/2026
Vol: 01.02
DOI: 10.61164/wc3f6213
Pages: 1-26

edifícios. A tabela 2 resume os parâmetros de validação adotados para cada agrupamento.

Tabela 2. Número de clusters e índice de silhueta para diferentes tipos de quedas

Tipo de queda (Dataset)	Número de clusters (K)	Índice silhueta
W11 – Escada de mão	3	0,9488
W12 – Andaime	2	0,7864
W13–Edifícios/ Estruturas	2	0,8049

Fonte: Elaborado pelos autores com base em dados do DATASUS (2025).

4.3. Análise dos clusters - Quedas em escada de mão

A identificação do perfil de risco predominante baseou-se na análise dos centroides dos agrupamentos.

O Cluster 2 foi definido como o de maior vulnerabilidade por apresentar a maior densidade de acidentes por perfil (maior magnitude do centroide da variável: acidentes).

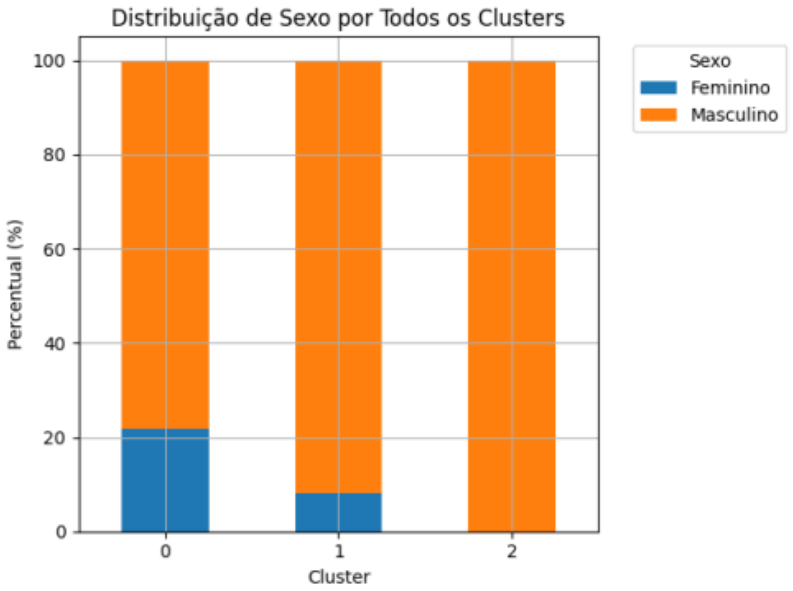
Essa métrica foi priorizada em detrimento da frequência absoluta total, pois isola as combinações demográficas (homens, 50-59 anos, escolaridade 4-7 anos) que geram fatalidades de forma sistemática e recorrente, distinguindo o risco ocupacional da dispersão observada nos demais grupos.

As Figuras 4, 5 e 6 apresentam a distribuição das variáveis demográficas (sexo, idade e escolaridade) entre os três agrupamentos formados.

REMUNOM
ISSN 2178-6925

Received: 13/01/2026 - **Accepted:** 18/01/2026
Vol: 01.02
DOI: 10.61164/wc3f6213
Pages: 1-26

Figura 4. Proporção de sexo por cluster (k-3)



Fonte: Elaborado pelos autores (2025)

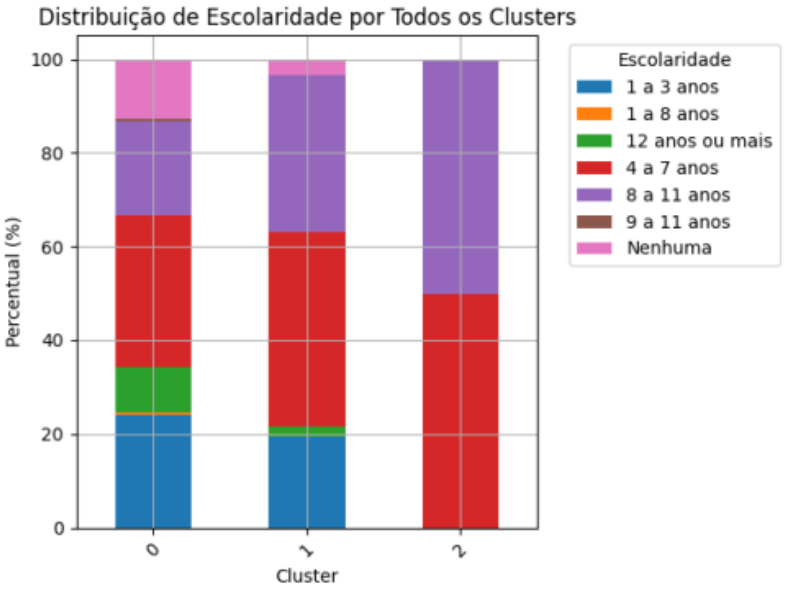
Figura 5. Proporção de idade por cluster (k-3)

Fonte: Elaborado pelos autores (2025)

REMUNOM
ISSN 2178-6925

Received: 13/01/2026 - **Accepted:** 18/01/2026
Vol: 01.02
DOI: 10.61164/wc3f6213
Pages: 1-26

Figura 6. Proporção de escolaridade por cluster (k-3)



Fonte: Elaborado pelos autores (2025)

A validação estatística de K=3 para o cenário de quedas em escadas demonstrou alta sensibilidade na identificação de subpopulações distintas, corroborando a hipótese de heterogeneidade deste tipo de acidente.

Tabela 3. Caracterização do Cluster Predominante (W11 – Quedas de Escada de Mão)

Cluster	Óbitos	%	Sexo	Idade	Escolaridade
0	625	55,11%	Masculino (78,8%)	50-59 (20,5%)	1-3 anos (35,4%)
1	173	15,26%	Masculino (97,9%)	60-69 (27,1%)	8-11 anos (45,8%)
2	336	29,63%	Masculino (82,5%)	50-59 (20,7%)	4-7 anos (100%)

Fonte: Elaborado pelos autores com base em dados do DATASUS (2025).

A análise dos clusters permitiu isolar dinâmicas sociodemográficas que seriam mascaradas em uma análise binária simples:

REMUNOM
ISSN 2178-6925

Received: 13/01/2026 - **Accepted:** 18/01/2026
Vol: 01.02
DOI: 10.61164/wc3f6213
Pages: 1-26

- Cluster 1 (Perfil Idoso/Escolarizado): Embora represente a menor fatia quantitativa (15,26%), este grupo é qualitativamente o mais distinto. Caracteriza-se pela maior faixa etária predominante (60 a 69 anos) e, paradoxalmente, pela maior escolaridade relativa (8 a 11 anos). Este achado sugere uma vitimização associada não à precarização laboral clássica, mas possivelmente a idosos ativos em tarefas de manutenção ou atividades domésticas, onde a perda de reflexos motores e equilíbrio associada à idade se torna o fator crítico de risco.
- Cluster 2 (Perfil Ocupacional Padrão): Com 29,63% dos casos, este grupo apresenta uma homogeneidade absoluta na escolaridade (100% com 4 a 7 anos de estudo), alinhando-se perfeitamente ao perfil sociodemográfico médio da força de trabalho da construção civil e manutenção no Brasil.
- Cluster 0 (Perfil de Alta Vulnerabilidade Social): O grupo majoritário (55,11%) reúne as vítimas com os menores indicadores de instrução formal (1 a 3 anos), indicando um cenário de informalidade severa ou atividades domésticas em populações de baixa renda.

A adoção de três clusters foi, portanto, indispensável para distinguir o risco ocupacional (Cluster 2) do risco associado à vulnerabilidade social (Cluster 0) e do risco fisiológico do envelhecimento em população instruída (Cluster 1).

4.4. Análise dos clusters - Quedas de ou em andaimes

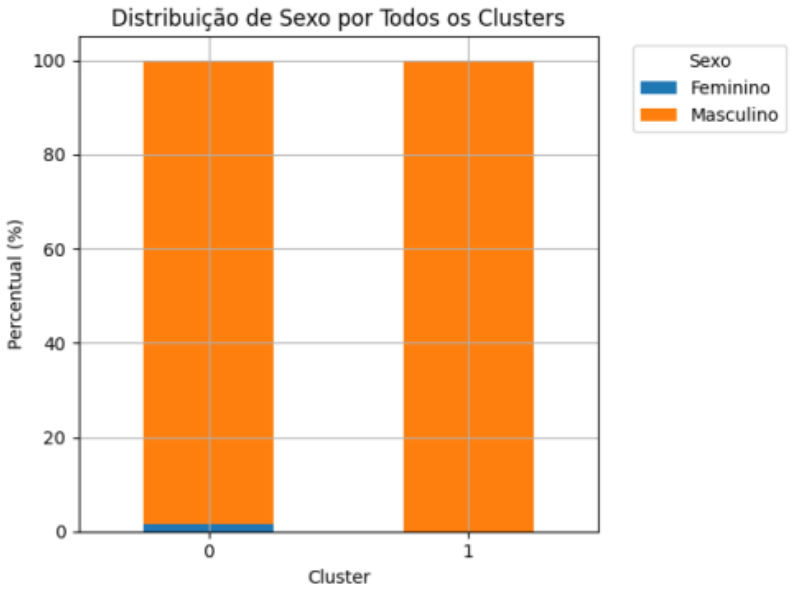
Para a categoria de andaimes, a determinação do grupo de risco seguiu o critério de intensidade. O Cluster 1 foi identificado como o perfil crítico, pois exibe uma densidade de acidentes por registro significativamente superior à do grupo geral. Isso indica que a combinação de suas variáveis (homogeneidade masculina, faixa etária de 50-59 anos e baixa escolaridade) possui uma correlação mais forte com a fatalidade do que os perfis diluídos no restante da amostra, validando-o como o foco prioritário para análise prevencionista.

As Figuras 8, 9 e 10 apresentam a distribuição das variáveis demográficas (sexo, idade e escolaridade) entre os dois agrupamentos formados.

REMUNOM
ISSN 2178-6925

Received: 13/01/2026 - **Accepted:** 18/01/2026
Vol: 01.02
DOI: 10.61164/wc3f6213
Pages: 1-26

Figura 7. Proporção de sexo por cluster (k-2)



Fonte: Elaborado pelos autores (2025)

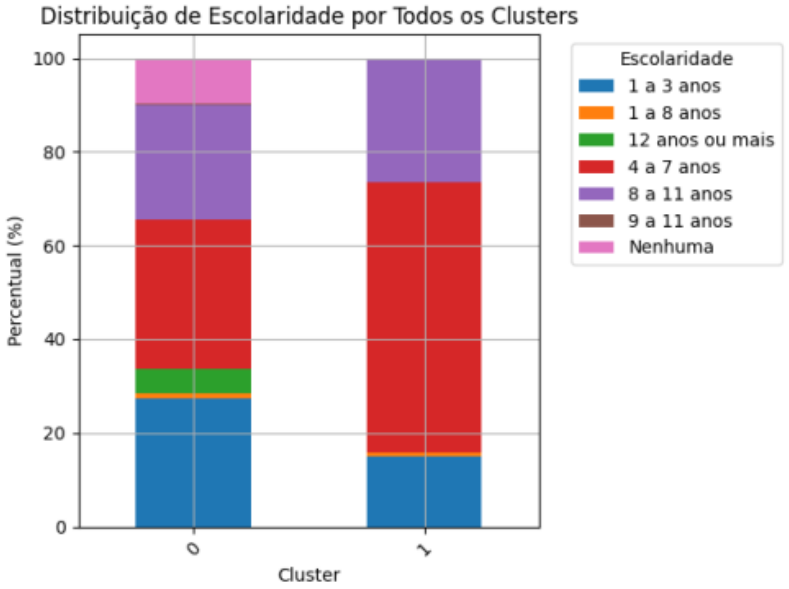
Figura 8. Proporção de idade por cluster (k-2)

Fonte: Elaborado pelos autores (2025)

REMUNOM
ISSN 2178-6925

Received: 13/01/2026 - **Accepted:** 18/01/2026
Vol: 01.02
DOI: 10.61164/wc3f6213
Pages: 1-26

Figura 9. Proporção de escolaridade por cluster (k-2)



Fonte: Elaborado pelos autores (2025)

A aplicação do método de agrupamento K-Means (K=2) demonstrou uma capacidade robusta de isolar o núcleo demográfico de maior vulnerabilidade, conforme detalhado na tabela 4.

Tabela 4. Caracterização do Cluster Predominante (W12 – Quedas de Andaime)

Cluster	Óbitos	%	Sexo	Idade	Escolaridade
0	1613	68,55%	Masculino (98,5%)	50-59 anos (19,4%)	4-7 anos (31,7%)
1	740	31,45%	Masculino (100%)	50-59 anos (35,3%)	4-7 anos (57,6%)

Fonte: Elaborado pelos autores com base em dados do DATASUS (2025).

A análise comparativa entre os clusters revela uma distinção clara baseada na intensidade das variáveis de risco:

- Cluster 1 (Núcleo de Vulnerabilidade Acentuada): Compreendendo 31,45% da amostra, este grupo atua como um marcador estatístico do perfil de risco severo. Observa-se uma concentração significativamente maior de vítimas na faixa etária de 50 a 59 anos

REMUNOM
ISSN 2178-6925

Received: 13/01/2026 - **Accepted:** 18/01/2026
Vol: 01.02
DOI: 10.61164/wc3f6213
Pages: 1-26

(35,3\%) e com escolaridade de 4 a 7 anos (57,6%), quando comparado ao restante da amostra. Este agrupamento sugere um cenário onde a combinação de idade avançada para o trabalho braçal e baixa qualificação formal potencializa drasticamente a ocorrência fatal.

- Cluster 0 (Perfil de Risco Disperso): Embora represente a maioria quantitativa dos casos (68,55%), este cluster apresenta uma diluição das características predominantes. A concentração na faixa etária de 50-59 anos cai para 19,4% e a escolaridade predominante para 31,7%. Isso indica que este grupo engloba uma variedade maior de perfis (incluindo trabalhadores mais jovens ou com níveis de instrução variados), representando o risco inerente à atividade de trabalho em altura que independe de fatores de vulnerabilidade social específicos.

A segmentação binária, portanto, foi essencial para evidenciar que o risco em andaimes não é uniforme, validando a necessidade de políticas de segurança focadas no envelhecimento e na capacitação de trabalhadores com menor escolaridade formal.

4.5 Análise dos clusters - Quedas de edifícios ou outras estruturas

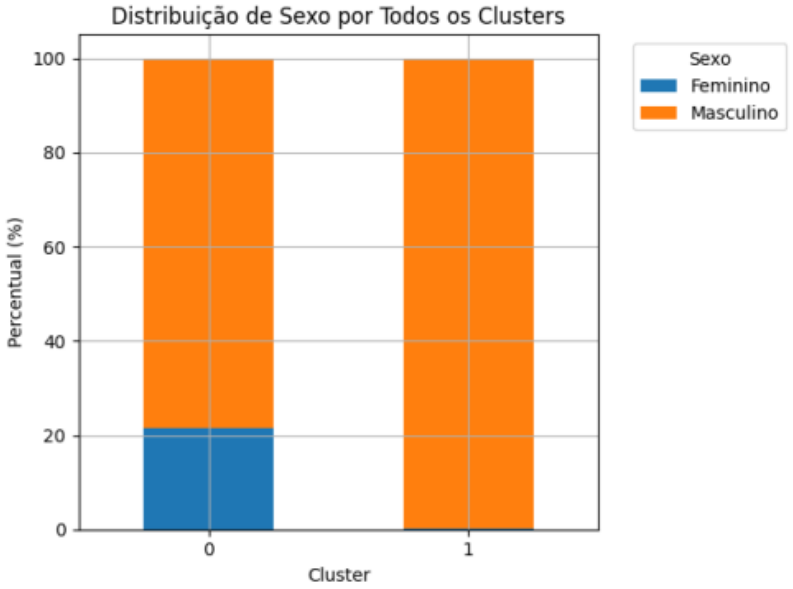
Na análise de quedas de edifícios, a polarização dos dados permitiu uma distinção clara baseada na severidade. O Cluster 1 foi classificado como o de maior risco com base na magnitude do centroide da variável 'acidentes'. Este indicador demonstra que este grupo específico concentra a maior taxa de fatalidade por perfil demográfico individual, validando estatisticamente que o risco neste cenário não é aleatório, mas fortemente associado ao perfil de trabalhadores em idade de pré-aposentadoria e com baixa instrução formal.

As Figuras 10, 11 e 12 apresentam a distribuição das variáveis demográficas (sexo, idade e escolaridade) entre os agrupamentos.

REMUNOM
ISSN 2178-6925

Received: 13/01/2026 - **Accepted:** 18/01/2026
Vol: 01.02
DOI: 10.61164/wc3f6213
Pages: 1-26

Figura 10. Proporção de sexo por cluster (k-2)



Fonte: Elaborado pelos autores (2025)

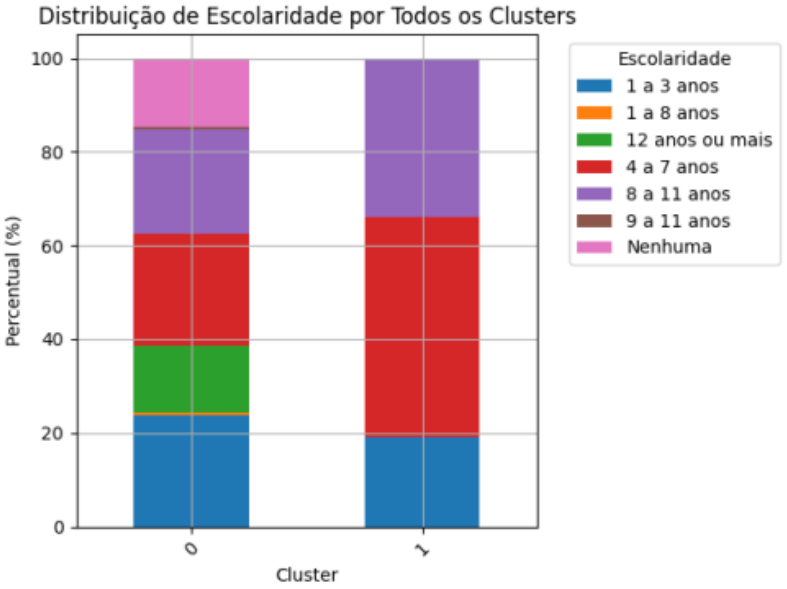
Figura 11. Proporção de idade por cluster (k-2)

Fonte: Elaborado pelos autores (2025)

REMUNOM
ISSN 2178-6925

Received: 13/01/2026 - **Accepted:** 18/01/2026
Vol: 01.02
DOI: 10.61164/wc3f6213
Pages: 1-26

Figura 12. Proporção de escolaridade por cluster (k-2)



Fonte: Elaborado pelos autores (2025)

A aplicação do algoritmo com K=2 revelou uma polarização demográfica distinta, validando a eficácia da segmentação binária para este tipo de acidente. A tabela 5 sintetiza essa caracterização.

Tabela 5. Caracterização do Cluster Predominante (W13 – Edifícios e outras estruturas)

Cluster	Óbitos	%	Sexo	Idade	Escolaridade
0	8078	59,9%	Masculino (78,4%)	50-59 anos (15,3%)	4-7 anos (23,8%)
1	5406	40,1%	Masculino (100%)	50-59 anos (24,5%)	4-7 anos (47,1%)

Fonte: Elaborado pelos autores com base em dados do DATASUS (2025).

A separação dos dados em dois grupos permitiu interpretar os clusters sob a ótica da exposição ao risco:

- Cluster 1 (Perfil de Vulnerabilidade Crítica): Representando 40,1% dos óbitos, este grupo isolou o perfil de maior risco ocupacional. Caracteriza-se por uma homogeneidade absoluta de gênero (100% masculino) e uma concentração acentuada de fatores de risco:

REMUNOM
ISSN 2178-6925

Received: 13/01/2026 - **Accepted:** 18/01/2026
Vol: 01.02
DOI: 10.61164/wc3f6213
Pages: 1-26

quase metade das vítimas (47,1%) possui baixa escolaridade e 24,5% encontram-se na faixa de 50 a 59 anos. Estatisticamente, este cluster atua como o núcleo representativo dos acidentes laborais típicos da construção civil, onde a interação entre a exigência física da tarefa e o envelhecimento da força de trabalho se mostra letal.

- Cluster 0 (Perfil Difuso): Com 59,9% dos casos, apresentou um perfil heterogêneo. A presença significativa de mulheres (21,6%) e a dispersão das variáveis de idade e escolaridade indicam que este grupo captura eventos de natureza diversa, possivelmente associados a contextos domésticos, urbanos ou de autoconstrução informal.

A separação destes dois universos confirma que a fragmentação em mais grupos (K=3) apenas diluiria a clareza deste perfil crítico isolado no Cluster 1, justificando a escolha metodológica.

5. Conclusão

A aplicação de técnicas de mineração de dados, especificamente o algoritmo K-Means, revelou-se uma estratégia eficaz para explorar a complexidade dos registros de óbitos por quedas no Brasil (1996-2023). Diante da natureza híbrida dos dados do Datasus, que não distinguem explicitamente a circunstância do óbito se laboral, doméstico ou urbano, a clusterização permitiu identificar padrões latentes que sugerem contextos distintos de ocorrência, superando as limitações das estatísticas descritivas tradicionais.

A principal contribuição científica deste estudo foi a capacidade de segmentar a massa de dados em perfis de risco bem definidos. Observou-se a existência consistente de um núcleo de vulnerabilidade crítica, identificado nos agrupamentos de maior densidade em todas as categorias, caracterizado predominantemente por homens, com escolaridade de nível fundamental incompleto 4 a 7 anos e idade entre 50 e 69 anos. A homogeneidade deste perfil sugere uma forte correlação com atividades de construção civil e manutenção

REMUNOM
ISSN 2178-6925

Received: 13/01/2026 - **Accepted:** 18/01/2026
Vol: 01.02
DOI: 10.61164/wc3f6213
Pages: 1-26

predial, seja formal, informal ou autoconstrução, onde o envelhecimento fisiológico do trabalhador colide com a exigência física e o risco de altura.

Paralelamente, a identificação de clusters mais difusos como o cluster 0 em W11 e W13, que apresentam uma participação feminina expressiva acima de 20\% e uma maior dispersão nas faixas etárias abrangendo desde jovens até idosos fora da faixa produtiva central, aponta para a relevância das quedas em contextos domésticos e urbanos. Essas características demográficas divergem do perfil típico da construção civil, sugerindo um cenário de risco híbrido que configura um grave problema de Saúde Pública.

Portanto, as implicações deste estudo vão além da segurança do trabalho tradicional e apontam para frentes de ação integradas:

No âmbito ocupacional a necessidade urgente de adaptar as normas de segurança como a NR-35 para a realidade de uma força de trabalho em envelhecimento, focando em ergonomia, exames médicos mais rigorosos para maiores de 50 anos e capacitação adaptada para trabalhadores com menor instrução formal.

No âmbito da saúde coletiva a demanda por políticas de habitação e conscientização sobre os riscos da autoconstrução e manutenção doméstica precária, que vitimam o mesmo perfil demográfico fora do ambiente regulado de trabalho.

No âmbito da gestão de dados o estudo evidencia a necessidade de aprimoramento no preenchimento das notificações de acidentes e óbitos, especialmente na variável escolaridade, cuja incompletude ainda representa um desafio para modelos preditivos mais precisos.

Conclui-se que a integração entre ciência de dados e segurança do trabalho oferece um novo paradigma para a prevenção, transformando registros administrativos em inteligência aplicada capaz de orientar tanto gestores de segurança corporativa quanto formuladores de políticas públicas na preservação da vida.

REMUNOM
ISSN 2178-6925

Received: 13/01/2026 - **Accepted:** 18/01/2026
Vol: 01.02
DOI: 10.61164/wc3f6213
Pages: 1-26

Referências

BRASIL. **Portaria n. 3.214, de 08 de junho de 1978**. Aprova as normas regulamentadoras NRs do Capítulo V, Título II, da Consolidação das Leis do Trabalho, relativas à segurança e medicina do trabalho. Diário Oficial da União. Disponível em: https://www.gov.br/trabalho-e-emprego/pt-br/assuntos/inspecao-do-trabalho/seguranca-e-saude-no-trabalho/sst-portarias/1978/portaria_3-214_aprova_as_nrs.pdf. Acesso em: 19 nov. 2025.

BRASIL. **Lei n. 8.213, de 24 de julho de 1991**. Dispõe sobre os planos de benefícios da previdência social e dá outras providências. Diário Oficial da União. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/l8213cons.htm. Acesso em: 15 nov. 2025.

BRASIL. **Norma Regulamentadora n. 35: Trabalho em altura**. 2018. Disponível em: https://www.gov.br/trabalho-e-previdencia/pt-br/composicao/orgaos-especificos/secretaria-de-trabalho/inspecao/seguranca-e-saude-no-trabalho/sst-portarias/2012/portaria_313_aprova_a_nr_35.pdf/view. Acesso em: 20 nov. 2025.

CAÑAVERAS PEREA, R. M.; TEJADA PONCE, A.; SÁNCHEZ GONZÁLEZ, M. P. How to prevent 3 million deaths worldwide: a systematic review of occupational accident research - a factor- and cost-based approach. **European Journal of Public Health**, v. 35, n. 1, p. 91-100, 2025. Disponível em: https://doi.org/10.1093/eurpub/ckae197. Acesso em: 16 nov. 2025.

DINH, T. et al. Data clustering: an essential technique in data science. arXiv, 2412(18760v2):1-17, 2025. Disponível em: https://arxiv.org/abs/2412.18760v2. Acesso em: 19 nov. 2025.

REMUNOM
ISSN 2178-6925

Received: 13/01/2026 - **Accepted:** 18/01/2026
Vol: 01.02
DOI: 10.61164/wc3f6213
Pages: 1-26

JAIN, A. K. Data clustering: 50 years beyond K-means. **Pattern Recognition Letters**, v. 31, n. 8, p. 651-666, 2010. Disponível em: https://doi.org/10.1016/j.patrec.2009.09.011. Acesso em: 20 nov. 2025.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: **Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability**, v. 1, p. 281-297. University of California Press, 1967. Acesso em: 25 nov. 2025.

ORGANIZAÇÃO MUNDIAL DA SAÚDE. **Global Health Estimates 2021: Global Deaths by Cause**, Age and Sex, 2000-2021. Geneva: World Health Organization, 2024.

ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Graphical Statistics**, v. 1, n. 2, p. 53-65, 1987. Acesso em: 25 nov. 2025.

SARAIVA, O. L. C. Segurança do trabalho quanto a queda de altura: particularidades, impactos e prevenção. **GETEC**, n. 13, p. 80-100, 2023. Acesso em: 17 nov. 2025.

SICAKYUZ, C.; EDALATPANAH, S. A.; PAMUCAR, D. Data mining applications in risk research: A systematic literature review. **International Journal of Knowledge-Based and Intelligent Engineering Systems**, v. 0, n. 0, p. 1-40, 2024. Disponível em: https://doi.org/10.1177/13272314241296866. Acesso em: 18 nov. 2025.

SILVA, R. F.; FERREIRA, R. V.; PEREIRA, V. A. Sustentabilidade em sistemas de segurança do trabalho na construção civil. **Brazilian Journal of Development**, v. 8, n. 8, p. 56951-56969, 2022. Disponível em: https://doi.org/10.34117/bjdv8n8-139. Acesso em: 18 nov. 2025.

REMUNOM
ISSN 2178-6925

Received: 13/01/2026 - **Accepted:** 18/01/2026
Vol: 01.02
DOI: 10.61164/wc3f6213
Pages: 1-26

THORNDIKE, R. L. Who belongs in the family? **Psychometrika**, v. 18, n. 4, p. 267-276, 1953. Acesso em: 22 nov. 2025.