Tarefa de agrupamento com K-Means na base de dados de violência contra mulheres - Secretaria de estado de Minas Gerais (SES)

Matheus Souza Azevedo

Sistemas de Apoio à Decisão

icea

UFOP

DECSI
DEPARTAMENTO DE COMPUTAÇÃO E SISTEMAS

Base de dados

Os dados foram retirados do **Portal de Dados Abertos do Estado de Minas Gerais**

Arquivos .csv únicos de 2010 à 2025

- Intervalo de data 01/2010 à 12/2025

Os registros contemplam casos suspeitos ou confirmados de violência **interpessoal** ou **autoprovocada**, inseridos no **Sistema de Informação de Agravos de Notificação (SINAN)** que é alimentado pelas unidades de saúde

Base de dados

- **dt_notific:** data de notificação da violência
- **dt_nasc:** data de nascimento da vítima
- **nu_idade_n:** idade da vítima
- **cs_sexo:** sexo da vítima
- **cs_raca:** raça da vítima
- **id_mn_resi:** município de residência da vítima
- **local_ocor:** local de ocorrência da violência
- **out_vezes:** quantas vezes a violência ocorreu
- **les_autop:** a lesão foi autoprovocada?
- **viol_fisic:** é violência física?
- **viol_psico:** é violência psicológica?
- **viol_sexu:** é violência sexual?
- **num_envolv:** quantidade de envolvidos
- **autor_sexo:** sexo do autor da violência
- **orient_sex:** orientação sexual da vítima
- **ident_gen:** identidade de gênero da vítima

 # de linhas: 476.019
 # colunas: 16

Modelagem

Como os dados são coletados através de uma ficha de múltipla escolha, é natural que as variáveis sejam categóricas, totalizando 94% das features. Portanto, para lidar com múltiplas categorias, a modelagem e o pré-processamento são fundamentais

Por exemplo, a variável **dt_notificacao** gerará outras variáveis:

- dia_util
- fim_semana
- 1º, 2º, 3º e 4º trimestre

Modelagem

É comum que em fichas haja campos como Não se aplica, Ignorado, ou que o responsável simplesmente ignore o campo. Portanto, variáveis com frequentes categorias desse tipo, foram mapeadas manualmente, **excluindo** essas ocorrências

raca_vitima	# de registros
Parda	213837
Branca	160611
Preta	52409
Ignorado	35679
Não preenchido	5983
Amarela	3633
Indígena	1755

 # de colunas após a modelagem: 46

Conclusão

O modelo conseguiu identificar com precisão perfis críticos, como a violência autoprovocada entre **adultos/jovens mulheres (Cluster 1)** e a diferença entre as ocorrências de violência no **fim de semana (Cluster 0)** e em **dias úteis da semana (Cluster 2)**.

Devido à natureza categórica de muitas variáveis (0 e 1), os clusters ainda apresentam certa proximidade visual, evidenciada pelos gráfcos, o que explica o **score médio de silhueta não ser tão elevado** (próximo de 0.1).

Referências

MINAS GERAIS. Secretaria de Estado de Saúde (SES). **Violência**. Portal de Dados Abertos do Estado de Minas Gerais. Disponível em: https://dados.mg.gov.br/dataset/violencia_ses. Acesso em: 01 dez. 2025.

Obrigado!