**Detecção de Fraudes**

**Giovani Cancherini e Nicolas Pietro**

**1. Tecnologias Utilizadas**

* **Pandas**: Manipulação e análise de dados, leitura e escrita de arquivos CSV, limpeza e transformação de dados.
* **Matplotlib & Seaborn**: Visualização de dados, criação de gráficos e análise exploratória.
* **Scikit-Learn**: Ferramentas de machine learning, incluindo KMeans para clustering, LabelEncoder para codificação categórica e StandardScaler para normalização.
* **KMeans Clustering**: Algoritmo de aprendizado não supervisionado para agrupar dados em clusters baseados em similaridades.

**2. Código Aproveitado**

* A estrutura básica de pré-processamento e limpeza de dados (funções utilitárias como load\_and\_sample\_files e clean\_and\_preprocess\_data) foi aproveitada de repositórios públicos.\

**3. Implementação Própria**

* Implementação das análises de clusters com gráficos customizados.
* Procedimentos adicionais de normalização e análise de características de cada cluster.
* Integração de dados de múltiplos arquivos e manipulação específica de campos relacionados a fraudes.

**4. Coleta de Dados**

* Arquivos de boletins de ocorrência entre 2007-2016 foram carregados e amostrados para reduzir a dimensionalidade.

**5. Pré-processamento**

* **Eliminação de colunas irrelevantes**: Remoção de colunas como Unnamed.
* **Limpeza de dados**: Extração e padronização de idades, conversão de datas.
* **Imputação**: Tratamento de valores nulos com SimpleImputer.
* **Codificação**: Conversão de variáveis categóricas para numéricas com LabelEncoder.
* **Escalonamento**: Normalização de variáveis numéricas.

**6. Treinamento e Avaliação**

* **Clustering KMeans**:
  + Aplicado para agrupar os dados em 3 clusters.
  + Avaliação exploratória dos clusters por meio de gráficos.
* **Análise Visual**:
  + Distribuição de crimes por ano, mês e categorias.
  + Pairplots para inspeção de agrupamentos.

**7. Resultados**

* Os clusters mostram distinções significativas em variáveis como DESCR\_TIPO\_PESSOA e IDADE\_PESSOA.
* Gráficos destacaram tendências temporais e categorias de crime mais prevalentes.
* Clustering revelou padrões relevantes para suporte à análise de dados policiais.

**8. Melhorias Potenciais**

* **Aprimoramento dos Dados**: Uso de outras técnicas de imputação, como KNNImputer, para tratar valores ausentes.
* **Modelo de Clustering**: Testar algoritmos como DBSCAN ou Hierarchical Clustering para melhorar a identificação de padrões.
* **Análise Temporal**: Aplicar métodos de séries temporais para prever incidentes futuros.
* **Dimensionalidade**: Implementar PCA para reduzir a dimensionalidade e melhorar a eficiência computacional.

**9. Referências**

1. Pedregosa, F., et al., *Scikit-learn: Machine Learning in Python*, JMLR 12, 2011.
2. Wes McKinney, *Python for Data Analysis*, 2nd Edition.
3. Hunter, J. D., *Matplotlib: A 2D Graphics Environment*, Computing in Science & Engineering.