

Exploração de Dados (TJMA) — Preparação para Classe Usucapião

Objetivo: Documentar o notebook de preparação/exploração que identificou a classe processual USUCAPIÃO a partir do dataset geral (tjma-geral.csv), gerando recortes por municípios e amostras para análises subsequentes.

1. Carregamento de Base e Visão Geral

- Leitura do arquivo 'tjma-geral.csv' com encoding 'cp1252' e separador ';' (skip de linhas problemáticas).
- Contagem das classes em 'Nome_da_Ultima_classe' para identificar as mais frequentes e localizar 'USUCAPIÃO'.

2. Filtro da Classe USUCAPIÃO e Exportação

- Criação do subconjunto 'df_usucapiao' filtrando 'Nome_da_Ultima_classe' == 'USUCAPIÃO'.
- Resultado do filtro: ~1114 casos (linhas) com 29 colunas na amostra exibida.
- Exportação do recorte para 'tjma-usucapiao.csv' (encoding 'cp1252', separador ';').

3. Exploração por Municípios (Usucapião)

- Leitura de 'tjma-usucapiao-municipios.csv' para análise de distribuição por município.
- Contagem de casos por 'Municipio' e seleção de cidades com contagens > 10 e < 120 (faixa balanceada).

Top municípios (exemplos)	IMPERATRIZ (118), BALSAS (66), CAXIAS (64), ANAJATUBA (56), CAROLINA (56)
Faixa de corte	> 10 e < 120

4. Amostra Filtrada e Exportação

- Construção de 'municipios_lista' com as cidades selecionadas e filtragem de 'df_usucap_municip'.
- Amostra resultante ('df_filtrado') com ~572 linhas e 29 colunas (exibida).
- Exportação da amostra para 'tjma-usucapiao-amostra.csv' (encoding 'cp1252', separador ';').

5. Observações e Boas Práticas

- Padronize encodings (cp1252/UTF-8) e separadores ao consolidar bases; registre 'on_bad_lines' para auditoria de perdas.
- Documente critérios de seleção (faixas e filtros) e salve os recortes gerados com carimbo de data/hora.
- Integre estes recortes ao pipeline de rotulagem e fairness, assegurando consistência de chaves (nº do processo, município, vara/comarca).