Relatório – Trabalho Final

Disciplina: Ciência de Dados
Especialização: Residência em Tecnologia da Informação
Equipe: Júlio César Prado Souza Rodrigues
Base de dados utilizada: Arquivos CSV contendo registros de processos judiciais do Tribunal de Justiça do Estado de Goiás (TJGO), organizados por ano e nomeados no padrão: processos_<ano>.csv

1. Business Understanding

1.1 Negócio por trás dos dados

O dataset representa o setor judiciário, mais especificamente os processos judiciais do Tribunal de Justiça, com foco no acompanhamento de advogados (OAB) e a proporção de processos sob segredo de justiça.

1.2 Objetivos estratégicos da organização

- Aumentar a transparência e eficiência no acompanhamento da atuação de advogados em processos sigilosos.
- Identificar padrões de comportamento que possam indicar concentração, especialização ou irregularidades no uso de sigilo em determinados escritórios ou advogados.

1.3 Problema de negócio

A organização busca entender se há crescimento ou concentração atípica de processos sigilosos em determinados advogados, podendo afetar a eficiência, a equidade e a credibilidade do sistema judicial.

1.4 Tradução para Ciência de Dados

O problema de negócio é transformado em:

- Classificação e segmentação de advogados conforme seu perfil de atuação em casos sigilosos (ex.: expansão, novos focos, transição, fora do foco).
- Análise de séries temporais e proporções, verificando variações entre anos (2022–2024).

1.5 Hipóteses iniciais

- H1: A proporção de casos sigilosos está crescendo para um grupo restrito de advogados, indicando especialização.
- H2: A maioria dos advogados mantém um equilíbrio entre casos sigilosos e não sigilosos, sem variações expressivas ao longo dos anos.

1.6 Restrições

- Qualidade dos dados: presença de OABs inválidas e inconsistências.
- Tempo de processamento: volume considerável de dados históricos (milhares de processos).
- Questões legais/éticas: tratamento de dados sensíveis exige cuidado com anonimização e privacidade.

1.7 Critério de sucesso

O projeto será bem-sucedido se conseguir classificar corretamente os perfis de advogados quanto à proporção de casos sigilosos, fornecendo subsídios para relatórios estratégicos e tomada de decisão institucional.

1.8 Métricas de avaliação

- Taxa de acerto na classificação (comparando critérios heurísticos e validações manuais).
- Métricas estatísticas: variação percentual de sigilosos por advogado, média e desvio padrão.
- Visualizações interpretáveis (funnel plot, dispersão, tabelas comparativas).

2. Data Understanding

2.1 Coleta

- Origem dos dados: bases internas do Tribunal (arquivos CSV por ano, ex.: processos_2022.csv, processos_2023.csv, processos_2024.csv).

- Volume: dezenas de milhares de registros, com colunas como processo, data_distribuicao, is_segredo_justica, oab, comarca, classe, assunto.

- Granularidade: nível de processo individual, com possibilidade de expansão por advogado (OAB).

- Limitações: dados contêm OABs inválidas, duplicações e valores ausentes. Há também implicações éticas/legais no uso de informações sobre sigilo.

2.2 Exploração

- Valores faltantes/inconsistentes: algumas OABs nulas ou fora do padrão; datas inconsistentes.
- Outliers: advogados com volume anormalmente alto de processos sigilosos.
- Balanceamento: a variável-alvo (is_segredo_justica) não é perfeitamente balanceada, havendo predominância de não sigilosos.
- Correlação: forte relação entre ano_distribuicao e variação da proporção de sigilosos.
- Sugestões preliminares: há indícios de que determinados advogados concentram a maior parte dos sigilosos.
- Necessidade de enriquecimento: incluir eventualmente dados externos (perfil do advogado, tipo de cliente, área de atuação) poderia ampliar a análise.
- Viés: risco de sobrerrepresentação de advogados com registros inválidos ou escritórios grandes que concentram processos.
- Limpeza e transformação: validação e padronização de OAB, conversão de datas e exclusão de registros inválidos são passos necessários.
- Variável-alvo: claramente definida (is_segredo_justica: True/False).

3. Data Preparation

3.1 Seleção de atributos

Foram selecionadas as variáveis mais relevantes para o problema:

- processo (identificação única)
- data_distribuicao (para cálculo do ano de entrada)
- is_segredo_justica (variável-alvo binária)
- oab (identificação do advogado)

- comarca, serventia, codg_classe, codg_assunto (variáveis contextuais para análises futuras).

3.2 Tratamento de registros

- Remoção de OABs inválidas ou nulas, validadas via regex.
- Conversão de colunas de datas para o tipo datetime.
- Exclusão de registros totalmente inconsistentes (datas nulas e OAB inválida simultaneamente).

3.3 Subconjuntos de dados

O conjunto poderá ser dividido em:

- **Treino (80%) –** desenvolvimento e ajuste das análises/modelos.
- **Validação (10%) –** ajuste de parâmetros e avaliação intermediária.
- **Teste (10%) –** avaliação final, para medir generalização.

3.4 Outliers

- Identificação de advogados com volumes desproporcionais de processos sigilosos.
- Mantidos inicialmente, pois podem representar comportamentos estratégicos reais (especialistas em determinadas áreas).

3.5 Criação de novas variáveis (feature engineering)

- ano_distribuicao – derivada da coluna data_distribuicao.
- Proporções calculadas: proporcao_sigilosos, proporcao_nao_sigilosos.
- Variáveis de agregação por OAB e ano (ex.: total de processos, variação percentual).

3.6 Codificação de variáveis categóricas

- oab é mantida como identificador, mas poderá ser transformada em variável categórica (Label Encoding) caso usada em modelos.

- comarca, serventia e classe podem ser submetidas a One-Hot Encoding caso sejam utilizadas em aprendizado supervisionado.

3.7 Normalização/padronização de variáveis numéricas

- Proporções (% de sigilosos) já estão normalizadas entre 0–100.

- Para modelos de Machine Learning mais sensíveis à escala (ex.: KNN, SVM), poderá ser aplicada Min-Max Scaling ou Z-score.

3.8 Formatação dos dados

- Todos os datasets foram unificados em um único DataFrame Pandas.

- Datas convertidas para datetime.

- OABs padronizadas em maiúsculas e validadas.

- Variáveis derivadas já calculadas para facilitar visualização e modelagem.

3.9 Balanceamento da variável-alvo

- A variável is_segredo_justica apresenta leve desbalanceamento (predomínio de não sigilosos).

- Poderão ser aplicadas técnicas como SMOTE (oversampling) ou undersampling para equilibrar os conjuntos em tarefas preditivas.

3.10 Conjunto final

O dataset preparado contém apenas OABs válidas, colunas de datas tratadas, variáveis derivadas e proporções calculadas. Ele reflete fielmente o problema de negócio e está pronto para análises estatísticas e aplicação de modelos de Machine Learning.

4. Modeling

4.1 Escolha da tarefa

Indique qual abordagem foi escolhida:

(x) **Classificação** (ex.: prever se um processo terá duração de 'Curto Prazo', 'Médio Prazo' ou 'Longo Prazo');

( ) **Regressão** (ex.: prever valor de vendas, preços ou notas);

( ) **Clusterização** (ex.: segmentar clientes ou agrupar documentos).

4.2 Algoritmos utilizados

Foram escolhidos e comparados três algoritmos de classificação:

- **Random Forest (Floresta Aleatória):** Um modelo de ensemble robusto, com boa performance geral e capacidade de interpretar a importância das features.
- **XGBoost (Extreme Gradient Boosting):** Um algoritmo de gradient boosting de alta performance, conhecido por sua velocidade e precisão em competições de dados.
- **LightGBM (Light Gradient Boosting Machine):** Outro algoritmo de gradient boosting, otimizado para ser ainda mais rápido e eficiente em memória com grandes volumes de dados.

Justificativa: O Random Forest foi escolhido como um baseline sólido. XGBoost e LightGBM foram adicionados para comparar o desempenho com algoritmos de ponta (state-of-the-art), que frequentemente superam o Random Forest em performance preditiva.

4.3 Preparação para modelagem

Foram necessários diversos ajustes antes da modelagem:

- **Amostragem:** Para evitar erros de memória e agilizar o treinamento, foi utilizada uma amostra aleatória de 300.000 registros do conjunto de dados.
- **Engenharia de Features:** A variável-alvo (categoria_duracao) foi criada a partir do cálculo da duração dos processos. Features de perfil do advogado (total_processos, percentual_sigilo_advogado) foram criadas com base nas descobertas da análise estatística.
- **Encoding e Limpeza:** Features categóricas foram transformadas via One-Hot Encoding (pd.get_dummies). Os nomes das colunas resultantes foram limpos para garantir a compatibilidade com todos os modelos.

- **Label Encoding:** A variável-alvo de texto foi convertida para formato numérico (0, 1, 2) para garantir a compatibilidade com o XGBoost.
- **Balanceamento de Classes:** Os modelos foram treinados com o parâmetro class_weight='balanced' para mitigar o forte desbalanceamento da classe "Longo Prazo" (<1% dos dados).

A divisão dos dados foi de 80% para treino e 20% para teste, com estratificação para manter a proporção original das classes em ambos os conjuntos.

4.4 Construção dos modelos

Os três modelos foram treinados em um loop comparativo sobre o mesmo conjunto de dados de treino. Foram utilizados parâmetros otimizados para menor consumo de memória, como n_jobs=1 e tree_method='hist' (para XGBoost). Não foi realizada uma etapa de ajuste fino de hiperparâmetros.

5. Evaluation

5.1 Avaliação do desempenho do Modelo Random Forest

As métricas utilizadas foram: Acurácia, Precisão, Recall, F1-Score e a Matriz de Confusão, justificadas por serem o padrão para avaliar problemas de classificação, especialmente em cenários com dados desbalanceados.

Os valores obtidos no conjunto de teste foram:

XGBoost:

- Acurácia: 74.53%
- Performance notável em "Curto Prazo" (F1-Score 0.82), mas falhou completamente em identificar a classe "Longo Prazo" (F1-Score 0.00).

Random Forest:

- Acurácia: 72.84%
- Apresentou o desempenho mais equilibrado, com performance razoável em "Curto" (F1 0.81) e "Médio Prazo" (F1 0.60), e sendo o único a identificar alguns casos de "Longo Prazo" (F1 0.08).

LightGBM:

- Acurácia: 61.82%
- Obteve o pior desempenho geral, com um comportamento anômalo para a classe "Longo Prazo" (Recall alto de 0.62, mas Precisão baixíssima de 0.03).

5.3 Comparação de modelos

A tabela comparativa final, ordenada pelo F1-Score Ponderado, foi:

Modelo	Acurácia	F1-Score Ponderado
XGBoost	74.53%	73.39%
Random Forest	72.84%	72.75%
LightGBM	61.82%	66.47%

O XGBoost apresentou o melhor desempenho quantitativo. No entanto, sua alta pontuação é enganosa, pois foi alcançada ao ignorar completamente a classe minoritária "Longo Prazo". Qualitativamente, o Random Forest se mostrou o modelo mais promissor como ponto de partida, por não ter descartado nenhuma classe durante o aprendizado.

5.4 Alinhamento com o problema de negócio

Os resultados atendem parcialmente ao critério de sucesso. O insight mais valioso não foi a performance preditiva, mas a confirmação, via importância de features, de que o perfil do advogado é o fator mais relevante. A alta precisão para "Curto Prazo" tem valor de negócio para planejamento, mas a falha na previsão de "Longo Prazo" é um ponto crítico que impede o uso do modelo para essa finalidade.

6. Conclusão e Encaminhamentos

6.1 Principais descobertas

- **Insight Principal:** O perfil do advogado (total_processos e percentual_sigilo_advogado) é o fator mais preditivo para determinar a duração de um processo.

- **Descoberta dos Modelos:** O modelo com a maior acurácia (XGBoost) não é necessariamente o mais útil, pois pode ignorar classes minoritárias críticas. Isso destaca a importância de analisar métricas detalhadas.
- **Descoberta da Análise:** Foi confirmado que a distribuição de processos segue um padrão de Lei de Potência (Pareto/Log-Normal) e que há uma tendência de crescimento de casos sigilosos.

6.2 Limitações do estudo

- **Desbalanceamento de Classes:** A principal limitação é o baixo desempenho na classe minoritária ("Longo Prazo"), o que torna o modelo menos útil para prever os casos mais demorados.
- **Generalização:** O modelo foi treinado com dados de um contexto específico e pode não ter o mesmo desempenho em outras localidades ou sistemas.
- **Amostragem:** Para viabilizar a execução, foi utilizada uma amostra dos dados, o que pode não refletir perfeitamente a distribuição completa.

6.3 Recomendações futuras

Melhorar o Modelo:

- Aplicar técnicas de oversampling como SMOTE para criar mais exemplos sintéticos da classe "Longo Prazo" e re-treinar os modelos.
- Realizar o ajuste fino de hiperparâmetros (tuning) para os modelos de melhor desempenho (XGBoost e Random Forest).

Aprofundar a Análise:

- Simplificar o problema para uma classificação binária ("Curto Prazo" vs. "Não Curto Prazo") para criar um primeiro modelo de alta performance.

6.4 Conclusão final

Nesse projeto foi realizada a análise e a extração de insights valiosos da base de dados. Foram comparados três modelos, estabelecendo um baseline de performance (Acurácia de ~75% com XGBoost) e, mais importante, foi gerado o insight estratégico de que o perfil do advogado é o principal fator preditivo da duração de um processo.

O estudo identificou claramente o desafio do desbalanceamento de classes e definiu os próximos passos para a evolução do modelo.