Programa de Residência em TI

Desenvolvimento de Modelo de Previsão de Novos Casos - Área Cível

MANUEL LUCALA ZENGO

Outubro 2025

Resumo

Este relatório parcial apresenta as atividades realizadas até o momento no projeto de análise e previsão de novos casos da área cível. O objetivo central do trabalho é desenvolver modelos de previsão do volume de entrada de novos processos, em níveis agregados por área de ação, de modo a subsidiar decisões operacionais e planejamento. Nesta etapa foram realizadas a organização e inspeção da base de dados, limpeza e transformações iniciais, análises exploratórias descritivas e o desenho do pipeline de pré-processamento e modelagem. O relatório documenta as decisões metodológicas, resultados parciais, limitações encontradas e recomendações para as próximas etapas.

Objetivo

Desenvolver um pipeline reprodutível para previsão de novos casos cíveis, com as seguintes metas específicas:

1. Caracterizar a série temporal de entrada de novos processos (tendência e sazonalidade).
2. Construir modelos preditivos robustos (baselines estatísticos e modelos de machine learning) para horizontes de curto e médio prazo.
3. Avaliar a performance por métrica adequada (MAE, RMSE, MAPE) e propor implantação escalável por área de ação.

Introdução

A gestão eficiente do Poder Judiciário demanda uma compreensão quantitativa e preditiva do fluxo processual. A capacidade de antecipar a entrada de novos casos possui valor estratégico, permitindo o planejamento robusto da alocação de recursos, o dimensionamento de equipes de magistrados e servidores, e a otimização da infraestrutura jurisdicional. Prognósticos acurados sobre a demanda futura são, portanto, ferramentas essenciais para uma administração judiciária proativa e baseada em evidências.

Este projeto aborda este desafio por meio da análise e modelagem de uma extensa base de dados de processos cíveis, compreendendo **4.744.946 registros** individuais. O dataset é caracterizado por 11 variáveis primárias, que incluem identificadores únicos NUMERO, marcadores temporais DATA DE RECEBIMENTO, atributos processuais PRIORIDADE, CLASSE, ASSUNTOS, metadados jurisdicionais SERVENTIA, COMARCA e variáveis econômicas VALOR DA CAUSA. A riqueza e o volume destes dados oferecem uma oportunidade singular para a construção de modelos preditivos, mas também impõem desafios relacionados à qualidade, consistência e alta dimensionalidade, que foram sistematicamente tratados ao longo do projeto.

O objetivo central deste trabalho é o desenvolvimento de um pipeline metodológico para a previsão do volume de novos casos, utilizando técnicas de análise de séries temporais e aprendizado de máquina. As atividades descritas a seguir detalham o percurso técnico adotado, desde o saneamento dos dados brutos até a avaliação comparativa dos modelos de prognóstico.

Descrição da Atividade e Escopo do Projeto

As atividades conduzidas foram organizadas em um fluxo de trabalho estruturado, alinhado às melhores práticas em projetos de ciência de dados. As tarefas foram distribuídas nas seguintes fases:

1. **Levantamento e Carregamento dos Dados:** A primeira etapa consistiu na ingestão da base de dados por meio da biblioteca *pandas*. Uma inspeção inicial foi realizada para verificar a integridade estrutural, os tipos de dados e as estatísticas descritivas básicas (utilizando funções como `df.info()`, `head()` e `describe()`).

2. **Pré-processamento e Saneamento dos Dados:** Esta fase foi dedicada à limpeza e padronização dos dados, um pré-requisito para qualquer análise subsequente. As principais sub-tarefas incluíram:

    - A conversão da coluna `DATA_RECEBIMENTO` para o formato *datetime*, com tratamento de erros para registros com datas inválidas.
    - A normalização de campos textuais de alta cardinalidade (`COMARCA`, `SERVENTIA`, `ASSUNTOS`), por meio da remoção de acentos, conversão para minúsculas e eliminação de caracteres especiais, a fim de consolidar categorias semanticamente equivalentes.
    - O tratamento de valores ausentes (*missing values*) e a padronização de tipos em colunas como `CLASSE` e `AREA_ACAO`.

3. **Análise Exploratória de Dados (AED):** Com os dados limpos, foram conduzidas análises para extrair padrões e insights. Foram investigadas as distribuições de processos por comarca e serventia, a frequência de priorização e segredo de justiça, e a distribuição da variável `VALOR_CAUSA`, que exigiu transformações logarítmicas para mitigar o efeito de outliers. Adicionalmente, foi realizada uma análise textual inicial sobre a coluna `ASSUNTOS` para identificar os temas mais recorrentes.

4. **Desenho do Pipeline de Modelagem Preditiva:** A fase final do escopo atual consistiu no delineamento da estratégia de modelagem. Foram definidas as seguintes diretrizes técnicas:

    - *Engenharia de Features Temporais:* Especificação de um conjunto de variáveis preditoras a serem criadas a partir da data, incluindo *lags* (valores da série em instantes de tempo anteriores, como t-1, t-7), médias móveis (estatísticas de janela,

como média dos últimos 7 ou 30 dias) e indicadores de calendário (dia da semana, mês, trimestre, feriados).

- *Estratégia de Validação:* Definição de uma metodologia de validação que respeita a natureza temporal dos dados, como a validação *out-of-time* (divisão treino/teste cronológica) ou a validação cruzada para séries temporais (*time-series cross-validation*) com janelas móveis.
- *Seleção de Modelos:* Escolha de um portfólio de modelos para comparação, incluindo baselines estatísticos (ex: Naïve, SARIMA), que servem como referência, e modelos de aprendizado de máquina supervisionado (ex: LightGBM, XGBoost), que utilizam as features de engenharia temporal.

Procedimentos e Desenvolvimento Metodológico

O desenvolvimento do projeto foi estruturado com base no *Cross-Industry Standard Process for Data Mining* (CRISP-DM), um framework robusto e iterativo que garante a correta aplicação de técnicas de ciência de dados para a solução de problemas de negócio. As fases foram adaptadas para as especificidades de um problema de previsão de séries temporais com grande volume de dados.

1. Compreensão do Negócio (Business Understanding)

A fase inicial concentrou-se na tradução do problema de gestão em um objetivo técnico de modelagem. O objetivo principal foi definido como a **previsão do volume de novos casos com granularidade mensal**, visando a um horizonte de até 12 meses. As previsões geradas devem suportar decisões estratégicas e operacionais, como o planejamento de alocação de pessoal (magistrados e servidores), a otimização de orçamentos e a gestão proativa de cargas de trabalho em diferentes comarcas e varas. O sucesso do projeto foi atrelado a métricas de acurácia preditiva, como o Erro Absoluto Médio (MAE), que quantifica o desvio médio das previsões em número de casos.

2. Compreensão dos Dados (Data Understanding)

Nesta fase, foi realizada uma Análise Exploratória de Dados (AED) para caracterizar as propriedades estatísticas do dataset, identificar potenciais problemas de qualidade e formular hipóteses. A inspeção inicial da base de dados (`df.info()`, estatísticas descritivas) revelou características críticas, como a presença de colunas com texto livre (`ASSUNTOS`), campos com inconsistências de nomenclatura (`COMARCA`, `SERVENTIA`), e a distribuição assimétrica da variável `VALOR_CAUSA`. A análise da série temporal agregada indicou a presença de **tendência e sazonalidade**, características fundamentais que direcionaram as escolhas na preparação dos dados e na modelagem.

3. Preparação dos Dados (Data Preparation)

Esta foi a etapa mais intensiva, focada em transformar os dados brutos em um formato estruturado e de alta qualidade para a modelagem. O pipeline de preparação incluiu:

- **Limpeza e Padronização:** A conversão da coluna de data (`DATA_RECEBIMENTO`) para o formato *datetime* foi o passo primordial. Valores nulos foram tratados por meio de imputação contextual (ex: `VALOR_CAUSA` preenchido com zero). A padronização textual (remoção de acentos, conversão para minúsculas) foi aplicada em campos categóricos para evitar a fragmentação de categorias por grafias distintas.
- **Engenharia de Features (Feature Engineering):** Para enriquecer o poder preditivo dos modelos, novas variáveis foram construídas. As principais foram:
  - *Features Temporais:* Criação de variáveis de calendário (mês, ano, trimestre), *lags* da série temporal (valores de meses anteriores, como t-1, t-12) e estatísticas de janela móvel (médias dos últimos 3, 6 e 12 meses). Estas features fornecem ao modelo uma ”memória” de padrões passados e cíclicos.
  - *Transformação de Variáveis:* A variável `VALOR_CAUSA` foi submetida a uma transformação logarítmica (`log1p`) para normalizar sua distribuição e reduzir o impacto de valores extremos.

4. Modelagem (Modeling)

A fase de modelagem foi conduzida como um processo experimental comparativo, no qual diferentes classes de algoritmos foram treinadas e avaliadas.

- **Seleção de Modelos:** Foi adotado um portfólio de modelos para garantir a cobertura de diferentes abordagens:
  - *Baselines Clássicos:* Modelos como o Naïve Sazonal e médias móveis foram implementados para estabelecer um limiar mínimo de performance.
  - *Modelos Estatísticos:* Foram explorados modelos da família ARIMA (como o **SARIMA**), que são robustos para séries com componentes de tendência e sazonalidade bem definidos.
  - *Modelos Supervisionados:* Algoritmos de *Gradient Boosting* (**XGBoost**, LightGBM) foram treinados utilizando as *features* de calendário e de janela móvel. Tais modelos são capazes de capturar relações complexas e não-lineares entre as variáveis preditoras e a contagem de casos.
- **Estratégia de Validação:** Diferentemente de problemas de classificação, a validação de séries temporais deve respeitar a ordem cronológica para evitar vazamento de dados do

futuro para o treino (*data leakage*). Foi utilizada uma abordagem de **validação out-of-time**, com uma janela de treino fixa e um período de teste subsequente. A performance foi medida em horizontes de previsão distintos, utilizando as métricas MAE, RMSE e MAPE.

5. Avaliação (Evaluation)

Nesta fase, os resultados dos modelos foram analisados não apenas por suas métricas técnicas, mas também por sua utilidade prática. O modelo com a melhor performance (SARIMA, conforme detalhado na seção de Resultados) foi selecionado. O erro de previsão foi contextualizado em termos operacionais (desvio médio em número de casos por mês) para que os gestores pudessem avaliar se a acurácia atingida era suficiente para suportar as decisões de negócio estabelecidas na primeira fase.

6. Implantação e Reprodutibilidade (Deployment)

Para garantir a reprodutibilidade e a futura operacionalização dos resultados, recomenda-se a adoção de boas práticas de engenharia de software. O pipeline de dados e modelagem deve ser modularizado em scripts executáveis (ex: `preprocessing.py`, `features.py`, `train_model.py`) e versionado com Git. Os artefatos gerados (modelos treinados, transformadores de dados, encoders) devem ser serializados com bibliotecas como `joblib` ou `pickle`, permitindo seu reuso em um ambiente de produção para gerar previsões de forma automatizada (seja via processos em lote ou por meio de uma API).

Resultados e Análise

Esta seção detalha as descobertas obtidas em cada fase do ciclo de vida do projeto, desde a compreensão dos dados brutos até a avaliação final dos modelos preditivos, seguindo a metodologia CRISP-DM.

II. Compreensão dos Dados (Data Understanding)

A análise exploratória inicial foi conduzida sobre um volume de **4.744.946 registros**, distribuídos em 11 variáveis que caracterizam os processos judiciais. Esta etapa, focada na inspeção e visualização, revelou padrões cruciais que nortearam as fases subsequentes de preparação e modelagem. As figuras a seguir consolidam as principais descobertas.

Figura 1 — Série Temporal de Novos Casos

Interpretação e Implicações para Modelagem:

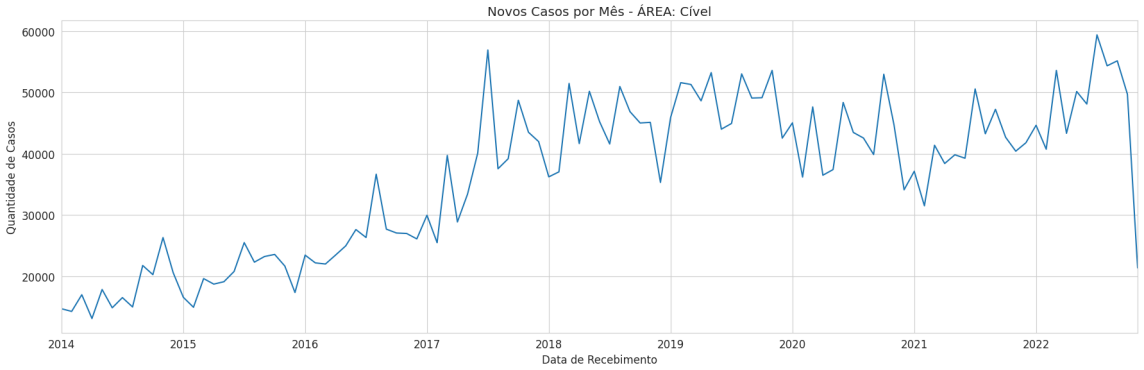

Figure 1: Série temporal mensal de novos casos, de 2014 a 2022.

- **Não-estacionariedade:** A série exibe uma clara tendência de crescimento no volume de casos entre 2014 e 2018, seguida por um patamar de maior volume. Esta característica de não-estacionariedade exige o uso de técnicas de diferenciação em modelos estatísticos para estabilizar a média ao longo do tempo.
- **Heterocedasticidade:** O aumento da amplitude das oscilações ao longo do tempo sugere que a variância não é constante. Para mitigar este efeito em modelos sensíveis à escala, transformações como a logarítmica (`log1p`) são recomendadas na preparação dos dados.
- **Sazonalidade Anual:** Observa-se um padrão cíclico que se repete anualmente, indicando a presença de sazonalidade, possivelmente influenciada por períodos de recesso e variações na atividade econômica.

Figura 2 e 6 — Análise da Distribuição de Variáveis Categóricas

A análise da distribuição por `CLASSE`, `ASSUNTOS` e `COMARCA` revelou uma forte concentração de casos em poucas categorias, um padrão conhecido como longa cauda. Goiânia, por exemplo, concentra a maior parte da demanda processual (Figura 6), enquanto poucas classes processuais respondem pela maioria dos registros (Figura 2). Esta heterogeneidade sugere que modelos preditivos podem se beneficiar de abordagens segmentadas, tratando as categorias de maior volume de forma distinta.

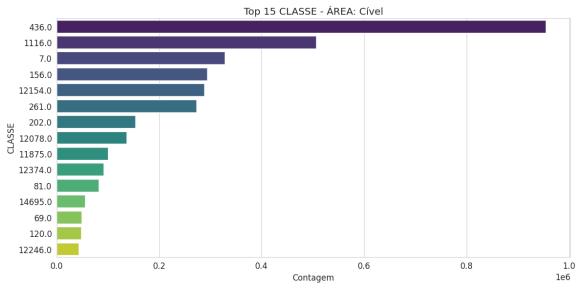

Figure 2: Ranking das 15 classes processuais mais frequentes.

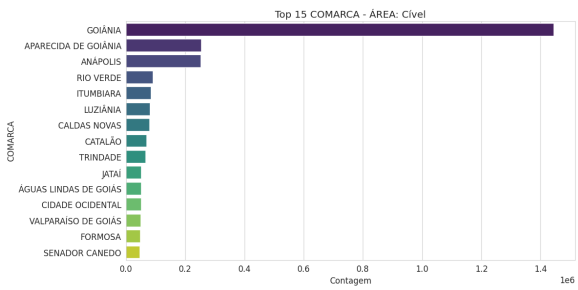

Figure 3: Ranking das 15 comarcas com maior número de casos.

Figura 3 — Análise da Variável Econômica

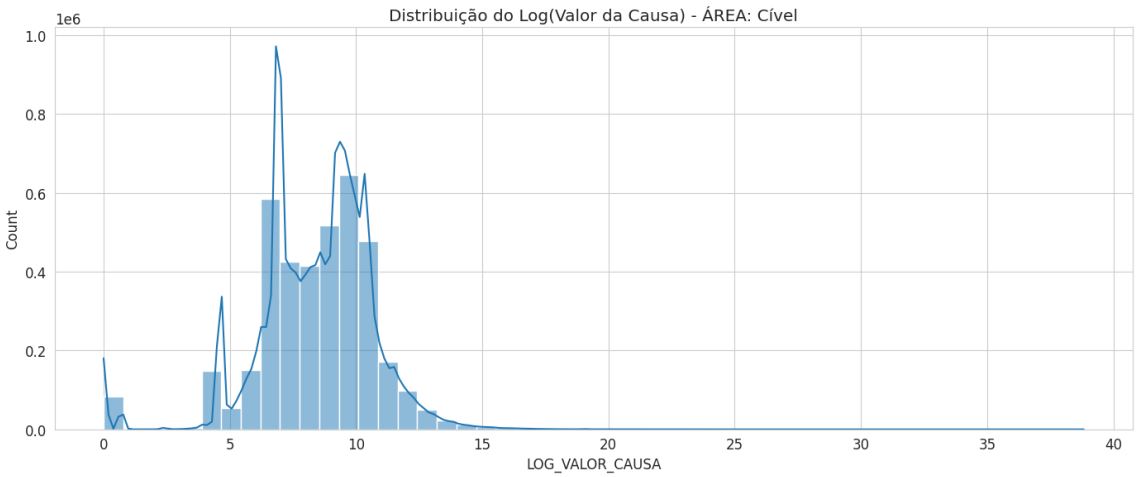

Figure 4: Histograma da transformação logarítmica de `VALOR_CAUSA`.

Interpretação e Implicações:

- A distribuição do `VALOR_CAUSA` apresenta forte assimetria à direita, mesmo após a transformação logarítmica, o que indica a presença de valores extremos (*outliers*) que podem influenciar desproporcionalmente os modelos.
- **Ações Recomendadas:** Manter a transformação `log1p` para modelos sensíveis à escala e considerar técnicas de tratamento de outliers, como a winsorização.

Figura 4 — Decomposição da Série Temporal

Interpretação e Implicações: A decomposição formal da série (Figura 4) isola seus componentes principais, confirmando a tendência de crescimento e a sazonalidade anual bem definida. Os resíduos, que representam a variabilidade não explicada pela tendência e sazonalidade, não são puramente aleatórios, sugerindo a influência de eventos externos que podem ser explorados com variáveis exógenas.

Figura 7 — Análise de Padrões Intra-anuais e Semanais

Interpretação e Implicações: O gráfico por dia da semana (Figura 7, direita) evidencia um claro padrão operacional, com maior volume de entrada em dias úteis e uma redução drástica nos finais de semana. Esta informação é crítica para a engenharia de features em modelos de previsão diária, justificando a inclusão de variáveis como dia da semana e indicadores de feriados.

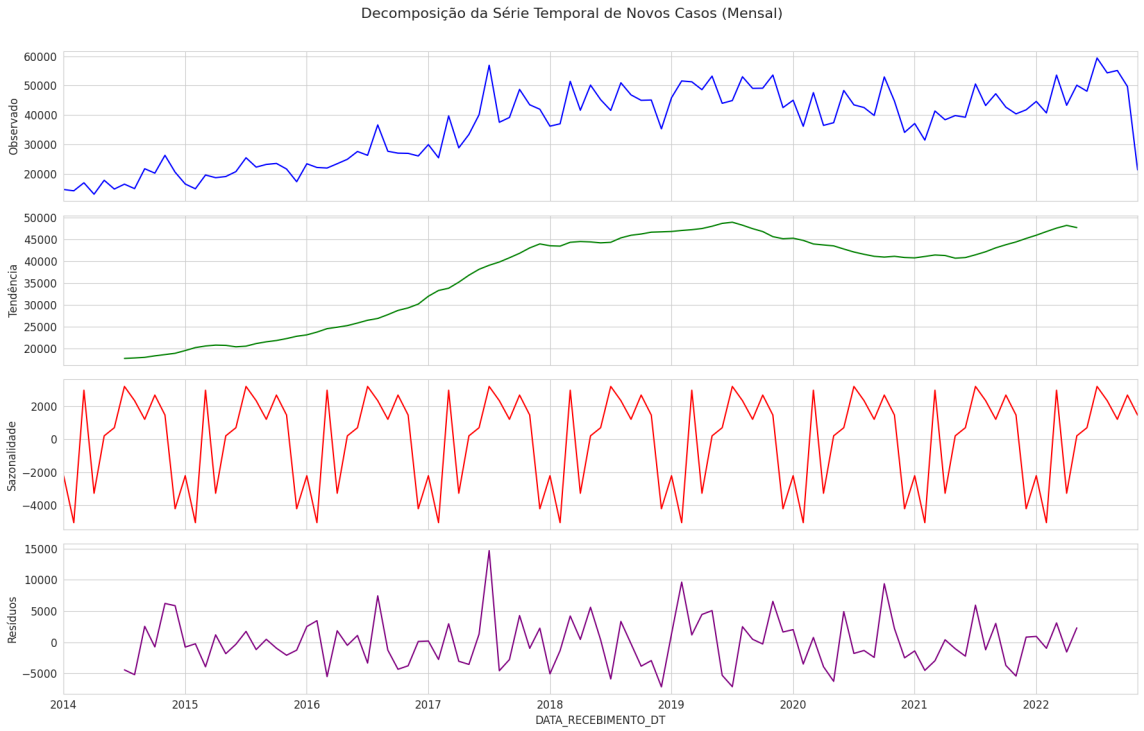

Figure 5: Decomposição aditiva da série mensal em tendência, sazonalidade e resíduos.

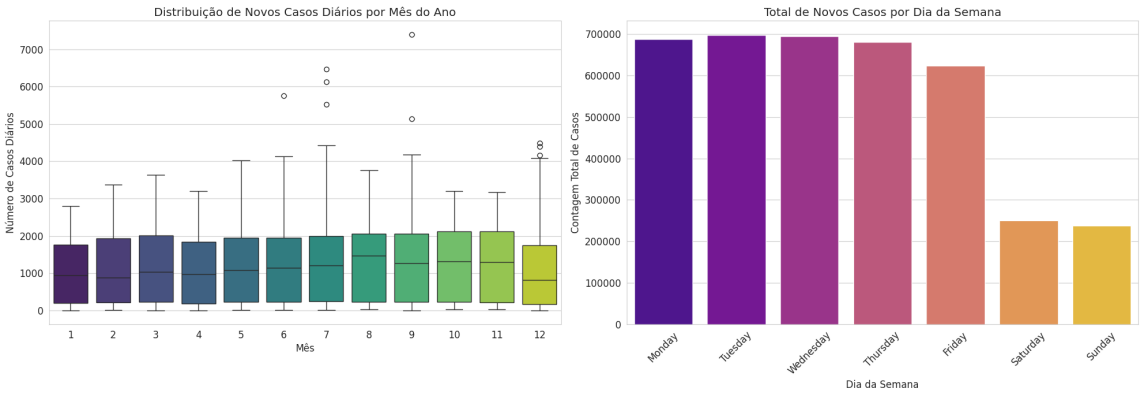

Figure 6: (Esquerda) Boxplots da distribuição diária por mês; (Direita) Contagem total por dia da semana.

III. Preparação de Dados (Data Preparation)

Com base nos insights da fase anterior, o dataset foi submetido a um processo de preparação para adequá-lo à modelagem. As principais ações foram:

- **Agregação Temporal:** O registro de processos individuais foi agregado em uma série temporal com frequência **mensal**, focando na análise de tendências de longo prazo e padrões anuais.
- **Limpeza e Tratamento:** Foram corrigidos os tipos de dados da coluna `DATA_RECEBIMENTO`. Valores ausentes, identificados como mínimos, foram tratados por meio de imputação.
- **Engenharia de Features:** Para os modelos de aprendizado de máquina, foram criadas variáveis preditoras baseadas em calendário, como `ano` e `mês`.

O resultado desta etapa foi um dataset limpo e estruturado, contendo a série temporal da contagem de novos casos mensais, pronto para a modelagem.

IV. Modelagem (Modeling)

A etapa de modelagem focou na seleção, treinamento e comparação de diferentes técnicas de previsão para identificar a abordagem mais acurada.

Figura 8 — Divisão da Série em Treino e Teste

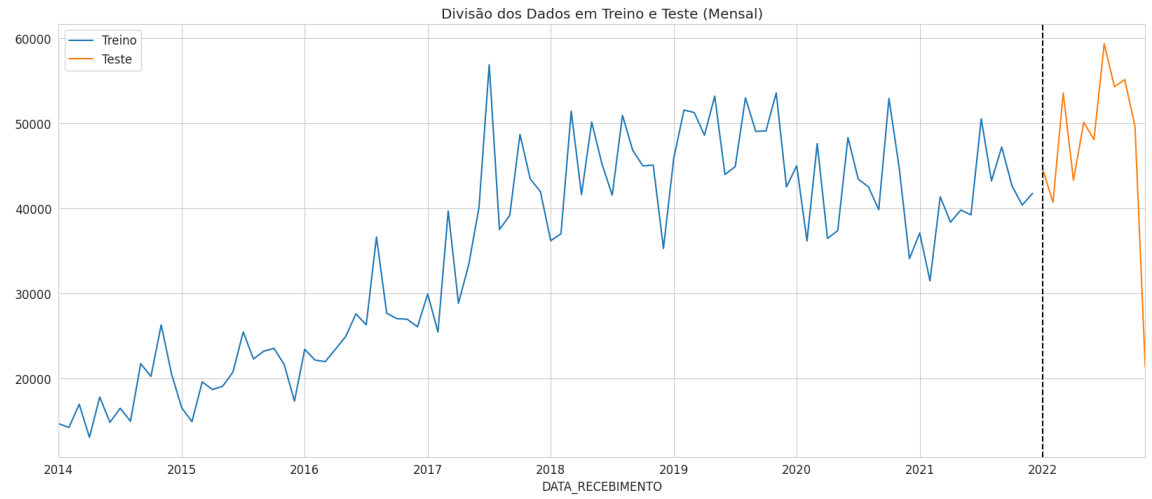

Figure 7: Divisão da série temporal mensal em dados de treino (2014-2021) e teste (2022).

Metodologia de Validação: A série foi dividida cronologicamente: dados de **2014 a 2021** foram usados para treino, e os de **2022** para teste (Figura 8). Esta abordagem simula um cenário real, no qual o modelo, treinado com o passado, é usado para prever o futuro.

Abordagens de Modelagem e Resultados

Foram implementadas e comparadas três abordagens: um modelo de **Baseline** (ingênuo sazonal), um modelo estatístico clássico (**SARIMA**) e um modelo de aprendizado de máquina (**XGBoost**). Os modelos foram avaliados utilizando as métricas de Erro Absoluto Médio (MAE) e Raiz do Erro Quadrático Médio (RMSE), com os resultados consolidados na Tabela 1.

Table 1: Comparativo de Performance dos Modelos de Previsão Mensal.

Modelo	MAE (Erro Absoluto Médio)	RMSE (Raiz do Erro Quadrático Médio)
SARIMA	5.890,24	9.849,78
Baseline	9.720,09	10.334,75
XGBoost	9.750,99	10.386,49

Análise de Performance: O modelo **SARIMA apresentou performance superior**, com os menores valores de MAE e RMSE. O erro absoluto médio de 5.890 indica que, em média, a previsão mensal do modelo desviou do valor real por aproximadamente 5.890 casos. O modelo XGBoost, com as features atuais, não conseguiu superar o desempenho do Baseline.

V. Avaliação (Evaluation)

A fase de avaliação finaliza o ciclo, analisando a performance do modelo sob uma ótica qualitativa e verificando se os resultados atendem aos objetivos práticos.

Figura 9 — Análise Visual da Acurácia do Modelo

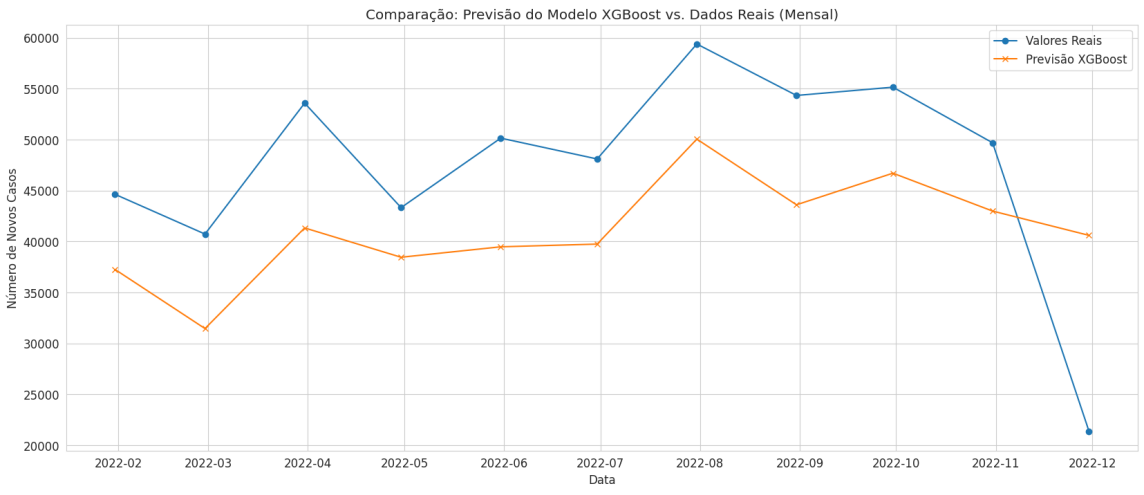

Figure 8: Comparativo visual: valores reais (azul) vs. previsões de um dos modelos testados (laranja).

Interpretação detalhada: A análise visual (Figura 9) ilustra a performance de um dos modelos menos acurados (XGBoost). Embora a direção geral da série seja capturada, o modelo

demonstra dificuldade em prever a magnitude dos picos e vales. Esta observação corrobora as métricas quantitativas e reforça a superioridade do modelo **SARIMA**, que se mostrou mais robusto para capturar os componentes da série.

Conclusão e Próximos Passos:

- Com base na análise comparativa, o modelo **SARIMA foi selecionado** como a solução preditiva final para este projeto, devido à sua acurácia significativamente superior.
- **Recomendações:** Sugere-se a implantação deste modelo para gerar previsões operacionais com horizonte de 6 a 12 meses, auxiliando no planejamento de recursos. Adicionalmente, recomenda-se a implementação de um ciclo de monitoramento contínuo para reavaliar e, se necessário, re-treinar o modelo com dados mais recentes.

Limitações e Próximos Passos

A principal limitação evidenciada até o momento refere-se à qualidade e à disponibilidade de variáveis exógenas que poderiam melhorar as previsões (variáveis socioeconômicas, indicadores locais, séries administrativas complementares). Alguns conjuntos consultados (IBGE e outras fontes) apresentaram períodos muito curtos (menos de 3 anos), o que reduz seu valor para modelagem temporal robusta. Ademais, a heterogeneidade na nomenclatura de `COMARCA` e `SERVENTIA` exige limpeza adicional para evitar dispersão de categorias.

Próximos passos prioritários:

1. Extrair e incorporar indicadores exógenos viáveis e documentar a qualidade e o período de cobertura.
2. Construir baselines estatísticos (Prophet / SARIMA) e um primeiro modelo de árvore (LightGBM) com validação temporal para comparação de performance.
3. Gerar relatórios com métricas por comarca e preparar roteiro de implantação (salvamento de modelos por unidade quando pertinente).
4. Estender a abordagem para a área criminal (metodologia análoga) assim que consolidada a pipeline para a área cível.

Conclusão

O trabalho avançou de forma consistente nas etapas iniciais de entendimento, limpeza e exploração dos dados. Foi estabelecida uma base técnica sólida para avançar à etapa de modelagem e validação temporal. Para que as previsões tenham robustez prática, é essencial: (i) padronizar e documentar o pipeline de pré-processamento; (ii) formalizar a estratégia de

validação temporal; e (iii) incorporar variáveis exógenas quando disponíveis e de qualidade adequada. O próximo ciclo de trabalho deve priorizar a criação de baselines, avaliação comparativa e documentação reprodutível do pipeline.