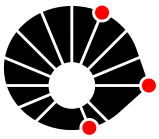
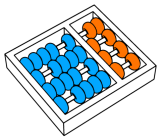




Modelo preditivo para a Justiça do Trabalho: Um estudo sobre viés algorítmico e explicabilidade

Luiza Coelho de Souza *Emanuel Felipe Duarte*

Relatório Técnico - IC-PFG-26-16
Projeto Final de Graduação
2026 - Junho

UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE COMPUTAÇÃO

Modelo preditivo para a Justiça do Trabalho: Um estudo sobre viés algorítmico e explicabilidade

Luiza Coelho de Souza* Emanuel Felipe Duarte†

Resumo

Este trabalho apresenta o desenvolvimento de um pipeline de aprendizado de máquina para prever desfechos processuais no Tribunal Regional do Trabalho da 15ª Região (TRT-15), usando dados públicos do DataJud. Foram implementados dois modelos supervisionados: um classificador binário com LightGBM para prever a ocorrência de acordos judiciais, e um classificador multiclasse com CatBoost para distinguir três desfechos possíveis: decisões favoráveis ao trabalhador, improcedências e extinções sem resolução de mérito. A interpretação dos modelos foi feita com o *framework* SHAP, que permite identificar o impacto de cada variável nas predições e avaliar assimetrias na base histórica. Os resultados mostram que variáveis geográficas e institucionais, como o município e o código da vara trabalhista, são predominantes para o comportamento preditivo, o que indica que o modelo aprende mais sobre o contexto de tramitação do que sobre o mérito individual de cada ação. A partir disso, discutem-se as implicações éticas dessas dependências e propõe-se uma análise crítica sobre o desenvolvimento responsável de sistemas preditivos na esfera jurídica, mostrando que é possível conciliar desempenho técnico com princípios de proteção de dados e explicabilidade.

Palavras-chave: Aprendizado de Máquina, Predição de Desfechos Judiciais, Explicabilidade, SHAP, Viés Algorítmico, Justiça do Trabalho, DataJud, LightGBM, CatBoost, Ética em Inteligência Artificial.

*Instituto de Computação, UNICAMP, 13083-852 Campinas, SP. l247257@dac.unicamp.br
†Instituto de Computação, UNICAMP, 13083-852 Campinas, SP. emanuel@ic.unicamp.br

Sumário

1 Introdução

O uso de algoritmos de aprendizado de máquina para apoiar decisões em domínios de alto impacto passou por uma expansão acelerada na última década, especialmente nos últimos 5 anos, alcançando áreas como saúde, financeiro, segurança pública e o sistema de justiça. No Brasil, essa expansão foi impulsionada pelo aumento da disponibildiade de dados processuais públicos, especialmente após a criação do DataJud pelo Conselho Nacional de Justiça, que passou a centralizar e disponibilizar informações estruturadas sobre a tramitação processual em diferentes tribunais do país.

A Justiça do Trabalho, responsável por um dos maiores volumes de litígios do Judiciário brasileiro, concentra grande parte desse potencial de análise quantitativa. Modelos preditivos capazes de estimar a probabilidade de acordo judicial ou o desfecho de mérito de uma ação têm valor prático cada vez mais evidênciado, tanto para partes e advogados que buscam estimar riscos processuais, quanto para o próprio Judiciário. No entanto, a aplicação de algoritmos complexos sobre dados gerados por um sistema social e institucionalmente desigual levanta questões que vão além do desempenho estatístico dos modelos.

Modelos de aprendizado de máquina aprendem os padrões presentes nos dados de treino sem distinguir, por si só, entre correlações que refletem o mérito jurídico de um caso e correlações que refletem assimetrias estruturais do próprio sistema, como diferenças de prática entre varas ou regiões. Caso essas assimetrias não sejam identificadas e discutidas, um modelo com boa performance métrica pode, ainda assim, reproduzir e potencialmente amplificar desigualdades já existentes na tramitação processual, o que torna a explicabilidade algorítmica uma ferramenta metodológica essencial.

Este trabalho apresenta o desenvolvimento e a avaliação de um pipeline de aprendizado de máquina para predição de desfechos processuais no Tribunal Regional do Trabalho da 15ª Região, construído exclusivamente a partir de dados públicos do DataJud. Foram implementados dois classificadores supervisionados, um modelo binário para estimar a ocorrência de acordo judicial e um modelo multiclasse para distinguir entre desfechos favoráveis ao trabalhador, improcedências e extinções sem resolução de mérito, ambos submetidos a uma auditoria de explicabilidade baseada no framework SHAP.

O restante deste texto está organizado da seguinte forma: a Seção 2 apresenta a justificativa do trabalho sob a ótica da mitigação de vieses e da governança de dados judiciários; a Seção 3 traz a fundamentação teórica, cobrindo o funcionamento da Justiça do Trabalho, os algoritmos de aprendizado usados e os princípios de explicabilidade e regulação de dados; a Seção 4 detalha os objetivos técnico e analítico do trabalho; a Seção 5 descreve a metodologia empregada, da coleta de dados ao treinamento dos modelos; a Seção 6 apresenta os resultados obtidos, incluindo o desempenho dos classificadores e as análises de explicabilidade; a Seção 7 discute as implicações éticas e os limites de generalização dos resultados; e a Seção 8 conclui o trabalho, sintetizando suas contribuições e apontando direções futuras.

2 Justificativa

2.1 Mitigação do efeito caixa-preta e replicação de vieses

Modelos de aprendizado de máquina funcionam por otimização estatística, com isso, dado um conjunto de dados e uma função de perda, o algoritmo ajusta seus parâmetros para minimizar erros de predição. Esse processo é matemático e não distingue correlações legítimas e que possivelmente auxiliem em análises de dados de distorções históricas ou desigualdades socioeconômicas presentes nos dados. Caso a base de treinamento reflita assimetrias do sistema judiciário, o modelo tende a absorvê-las e repeti-las nas predições futuras.

Na Justiça do Trabalho, isso é um problema a ser considerado. O histórico de decisões de cada vara é influenciado não só pelas características dos casos, mas também por restrições institucionais, variações regionais e inconsistências que afastam a prática do ideal de uniformidade jurídica. Um modelo treinado sobre esses dados tende a tratar as variações como padrões normais de inferência, o que aplica os viéses também nos resultados. Com isso, o viés algorítmico gerado pelo modelo, nesse cenário, não é considerado um defeito de implementação, mas sim uma consequência de se aplicar algoritmos complexos sobre dados gerados por dinâmicas sociais desiguais.

Dessa forma, considera-se a explicabilidade como ferramenta essencial para o desenvolvimento do modelo. Sem métodos concretos para auditar e governar a contribuição de cada variável nos resultados do modelo, não há como identificar se a acurácia do modelo é consequência de conexões causais de fato ou da reprodução de padrões históricos. Essa preocupação guiou a estrutura metodológica deste projeto, que prioriza a avaliação dos critérios de aprendizado dos modelos antes de qualquer conclusão sobre utilidade prática.

2.2 Governança ativa no uso de dados judiciários

Plataformas como o DataJud disponibilizam um volume expressivo de dados processuais públicos e anonimizados, mas ainda não existem regulamentações específicas sobre os limites de uso dessas informações para desenvolvimento de modelos preditivos. Esse fato não retira a responsabilidade dos desenvolvedores de projetar um sistema justo e ético, e aumenta diretamente o poder das escolhas de engenharia do sistema.

Essa responsabilidade tem duas dimensões principais. A técnica exige rastrear o ciclo de vida dos dados, garantindo consistência e compreendendo as transformações aplicadas por meio da engenharia e os efeitos adversos possíveis. Já a ética exige que o uso dessas bases não prejudique a finalidade de transparência que motivou sua abertura pública, evitando expor os envolvidos a riscos processuais ou econômicos desproporcionais, mesmo que tais análises sejam tecnicamente viáveis.

Com base nisso, as escolhas de projeto buscaram tornar transparentes os impactos de cada decisão de modelagem, da seleção das fontes de dado à seleção das métricas de performance pelas quais o modelo seria avaliado. Dessa forma, o objetivo não foi somente construir um modelo funcional, como também documentar e justificar cada etapa para que o processo de manufatura do modelo possa ser auditado e sua governança e aplicação ética avaliada.

3 Fundamentação Teórica

Esta seção apresenta os conceitos que fundamentam o trabalho: a dinâmica operacional da Justiça do Trabalho brasileira, os algoritmos de aprendizado de máquina usados nas tarefas de classificação, e os princípios de equidade, explicabilidade e governança de dados que orientam o estudo.

3.1 O judiciário trabalhista brasileiro

3.1.1 Estrutura do sistema judiciário trabalhista

O Poder Judiciário brasileiro é organizado em ramos especializados por tipo de demanda. A Justiça do Trabalho cuida das relações laborais, intermediando conflitos entre trabalhadores e empregadores regidos pela CLT e pela Constituição, como questões de contratos, verbas rescisórias e condições de trabalho.

A organização interna segue três níveis hierárquicos. As Varas do Trabalho representam a primeira instância e estão distribuídas municipalmente, concentrando o maior volume decisório do sistema. Os Tribunais Regionais do Trabalho (TRTs) julgam recursos contra as sentenças de primeiro grau, nesse caso, o país é dividido em 24 regiões, cada uma com seu TRT. No topo está o Tribunal Superior do Trabalho (TST), em Brasília, responsável pela uniformização da jurisprudência trabalhista em âmbito nacional.

Os dados deste estudo vêm do TRT da 15ª Região, com sede em Campinas, que abrange o interior e o litoral de São Paulo. O expressivo volume processual desse tribunal o torna uma das bases mais representativas para análises quantitativas de dados jurídicos no país, além de ser o tribunal correspondente a localização da UNICAMP.

3.1.2 O processo trabalhista: da abertura ao desfecho

O processo trabalhista começa com a petição inicial do reclamante (trabalhador) contra o reclamado (empregador), expondo as alegações, os fundamentos legais e os pedidos. A partir desse momento, a ação segue as etapas seguintes até seu fim definitivo.

A fase postulatória tem como objetivo a audiência de conciliação, em que o juiz tenta promover um acordo entre as partes. Se houver concordância, ele é homologado e o processo encerra. Sem acordo, inicia a fase de instrução, com produção de provas, depoimentos e laudos periciais.

Após a instrução, o juiz profere sentença, que pode ser procedente (pedidos do trabalhador acolhidos total ou parcialmente), improcedente (reivindicações rejeitadas) ou resultar em extinção sem resolução de mérito (ação interrompida por falhas formais, defeitos processuais ou desistência, sem avaliação do mérito). Este trabalho se restringe aos desfechos de primeiro grau, conforme registrados nos sistemas de informação judiciária.

3.1.3 A Reforma Trabalhista de 2017 e seu impacto nos dados

A Lei nº 13.467, vigente desde novembro de 2017, alterou configurações da CLT, tanto em aspectos contratuais quanto em regras processuais. Duas mudanças afetam diretamente os

dados analisados nesta pesquisa.

A primeira foi a criação de novas modalidades contratuais, como o trabalho intermitente e a regulamentação do regime autônomo exclusivo. A segunda, com impacto direto nos dados estatísticos utilizados durante a etapa de análise exploratória de dados, foi o estabelecimento dos honorários de sucumbência recíproca: a parte vencida passa a responder pelos custos advocatícios da vencedora, inclusive em casos de beneficiários da justiça gratuita. Essa mudança desincentivou o ajuizamento de ações com teses jurídicas mais frágeis, reduzindo o volume de distribuições e alterando o perfil dos pedidos nos anos seguintes.

Do ponto de vista da modelagem preditiva, a reforma criou uma quebra estrutural na série temporal. Processos anteriores e posteriores a 2017 pertencem a regimes de risco econômico e distribuições estatísticas distintos. Para garantir a homogeneidade dos dados, o intervalo amostral foi delimitado entre 2018 e 2023.

3.1.4 O DataJud como fonte de dados pública

O DataJud é a base de dados unificada do Poder Judiciário, criada pela Resolução CNJ nº 331/2020 [3] e gerida pelo Conselho Nacional de Justiça. A plataforma centraliza informações estruturadas sobre o andamento processual de diferentes tribunais e as disponibiliza publicamente via APIs de dados abertos.

Os registros incluem metadados como o número único do processo, a identificação da vara, o assunto principal conforme as Tabelas Processuais Unificadas (TPU), a classe da ação, o município de origem, os marcos temporais de movimentação e os códigos de andamento. Os dados passam por mascaramento, sendo retirados nomes e números de CPF e CNPJ de todas as partes.

Enquanto essa anonimização protege a privacidade, também limita o aprendizado dos algoritmos. Sem variáveis diretas relacionadas ao porte econômico das empresas ou o perfil socioeconômico dos trabalhadores, o modelo falha em explorar hipóteses baseadas na assimetria de poder entre as partes ou em reincidências litigiosas específicas, que era uma tentativa inicial para o desenvolvimento do trabalho. Essas restrições e seus reflexos na capacidade dos classificadores são detalhados nas seções seguintes.

3.2 Aprendizado de máquina para classificação

3.2.1 Gradient Boosting: a família de modelos

Os modelos implementados nesta pesquisa usam *gradient boosting*, uma técnica de aprendizado por conjunto (*ensemble*) que combina sequencialmente preditores fracos, em geral árvores de decisão rasas. O conceito principal é treinar cada nova árvore para aproximar os resíduos gerados pelo conjunto acumulado nas iterações anteriores, de forma que a soma ponderada de modelos simples resulte em um estimador robusto.

Formalmente, dado um conjunto de treinamento $\{(x_i, y_i)\}_{i=1}^{n}$, busca-se uma função $F(x)$ que minimize uma função de perda empírica $\mathcal{L}(y, F(x))$. O algoritmo expande $F(x)$ de forma aditiva a cada iteração m:

$$F_m(x) = F_{m-1}(x) + \eta \cdot h_m(x) \tag{1}$$

onde $h_m(x)$ é a árvore fraca ajustada sobre os gradientes negativos da função de perda avaliada em $F_{m-1}(x)$, e η é a taxa de aprendizado (*learning rate*), que escala o passo de atualização para evitar sobreajuste.

3.2.2 XGBoost

O XGBoost (*Extreme Gradient Boosting*) [2] se consolidou como uma das principais referências em conjuntos de árvores pela sua eficiência e robustez. O diferencial do algoritmo está na inclusão de termos de regularização explícitos na função objetivo, que penalizam a complexidade estrutural das árvores para reduzir o risco de *overfitting*. A função de custo na iteração m incorpora a penalidade $\Omega(h_m)$, que monitora o número de folhas e a magnitude dos pesos do modelo:

$$\mathcal{L}^{(m)} = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i^{(m-1)} + h_m(x_i)) + \Omega(h_m) \tag{2}$$

Apesar da popularidade entre cientistas, o XGBoost tem limitações de desempenho em volumes de dados muito grandes, pois o crescimento das árvores nível por nível (*level-wise*) exige alto custo de varredura e ordenação de atributos.

3.2.3 LightGBM

O LightGBM (*Light Gradient Boosting Machine*) [6], desenvolvido pela Microsoft, foi projetado para escalar em matrizes de alta dimensionalidade. Duas modificações do XGBoost mudam sua aplicação de forma considerável. A primeira é o crescimento de árvore por folha (*leaf-wise*), no qual, ao invés de expandir todos os nós de um nível, o LightGBM avalia todas as folhas disponíveis e expande apenas a que gera a maior redução na função de perda. Isso produz árvores assimétricas e mais profundas, que convergem com menos partições, mas exigem controle maior e mais rigoroso dos hiperparâmetros para evitar sobreajuste em amostras menores.

A segunda inovação é o GOSS (*Gradient-based One-Side Sampling*), que filtra instâncias de treinamento pela magnitude dos gradientes. Instâncias com gradientes pequenos indicam erros já baixos e com isso, o GOSS preserva todas as de gradiente elevado e amostra aleatoriamente as demais, ajustando os pesos para manter a consistência da distribuição:

$$\tilde{g}_j = \frac{1}{n} \left(\sum_{x_i \in A} g_i + \frac{1-a}{b} \sum_{x_i \in B} g_i \right) \tag{3}$$

onde A é a fração a de instâncias com gradientes altos, B é a amostra de fração b dos gradientes baixos, e g_i é o gradiente do elemento i. O LightGBM também usa o EFB (*Exclusive Feature Bundling*) para agrupar atributos esparsos e mutuamente exclusivos, economizando memória sem perda de informação.

3.2.4 CatBoost

O CatBoost (*Categorical Boosting*) [9] foi desenvolvido para lidar de forma nativa com atributos categóricos de alta e baixa cardinalidade, sem precisar de codificações extras como *one-hot encoding*. O fator central do algoritmo é o *ordered boosting*, que busca evitar o vazamento de dados típico de codificações de alvo convencionais.

Em codificações de alvo comuns, o valor numérico de uma categoria é calculado incluindo o próprio registro avaliado, o que introduz um viés chamado *target leakage*. O CatBoost resolve esse problema aplicando permutações aleatórias nos dados e estimando o valor de um registro com base apenas nos elementos que vêm antes dele na sequência:

$$\hat{x}_i^k = \frac{\sum_{j=1}^{p-1} \mathbf{1}[x_{\sigma_j}^k = x_{\sigma_p}^k] \cdot y_{\sigma_j} + a \cdot P}{\sum_{j=1}^{p-1} \mathbf{1}[x_{\sigma_j}^k = x_{\sigma_p}^k] + a} \tag{4}$$

onde σ é a permutação estocástica, P é o valor *a priori* da classe, e a é o parâmetro de suavização. O modelo usa ainda árvores simétricas (*oblivious trees*), com critérios de divisão iguais em todos os nós de um mesmo nível, o que acelera a inferência.

3.2.5 Engenharia de features

Engenharia de features é o processo de mapear e transformar dados brutos em representações numéricas que exponham correlações relevantes para os classificadores. A qualidade dos atributos construídos influencia amplamente os limites de aprendizado dos modelos, mesmo em arquiteturas baseadas em árvores, que já lidam razoavelmente bem com atributos brutos e não normalizados.

Em problemas de predição sobre dados administrativos e jurídicos, é comum organizar os atributos derivados em categorias que capturam diferentes dimensões do fenômeno modelado: características temporais do evento, propriedades intrínsecas do processo ou registro, variáveis exógenas de contexto socioeconômico, e métricas de comportamento histórico associadas às entidades envolvidas, como a frequência e a recência de eventos similares em uma mesma unidade administrativa. Essa última categoria costuma se apoiar em métricas do tipo RFM (*recency, frequency, monetary*), adaptadas da modelagem de comportamento transacional para o contexto de séries de eventos institucionais.

Um cuidado metodológico central na construção de atributos históricos é evitar o vazamento de dados temporais (*data leakage*). Em séries temporais, métricas agregadas calculadas sem restrição cronológica podem incorporar informações do futuro nos dados de treino, inflando artificialmente o desempenho do modelo em validação sem que essa performance se sustente em produção. Por isso, variáveis derivadas de histórico devem ser calculadas com travas temporais, de forma que o cálculo de um registro inclua apenas eventos anteriores ao momento avaliado.

3.2.6 Lidar com classes desbalanceadas

O desbalanceamento na distribuição de classes é comum em diversos tipos de problema de classificação, incluindo problemas jurídicos, em que certos desfechos processuais tendem a

ser bem mais raros que outros. Quando esse desbalanceamento não é tratado de alguma forma, os modelos classificadores tendem a maximizar a acurácia global priorizando as classes majoritárias, o que não reflete sua performance real sobre as classes minoritárias, muitas vezes as mais relevantes para o problema em questão.

Duas frentes complementares são comumente usadas para mitigar esse problema. A primeira é a seleção de métricas de validação que penalizam o desempenho assimétrico entre classes, como o F1-macro ou métricas calculadas por classe, ao invés de métricas agregadas que escondem o mau desempenho sobre categorias raras. A segunda é a inclusão de pesos de classe no treinamento, aumentando o custo dos erros cometidos sobre os registros minoritários e forçando o otimizador a considerá-los de forma proporcional à sua relevância, e não apenas à sua frequência na base.

3.2.7 Métricas de avaliação

A escolha das métricas de desempenho envolve critérios metodológicos e de governança. A acurácia global não representa o cenário real em cenários desbalanceados, pois uma regra que classifica todas as observações na classe mais frequente já geraria índices artificialmente altos.

A precisão indica a proporção de verdadeiros positivos entre os elementos classificados como positivos; o *recall* mede a sensibilidade do modelo para cobrir a classe alvo real. O F1-score combina os dois via média harmônica:

$$F1 = 2 \cdot \frac{\text{Precisão} \times \text{Recall}}{\text{Precisão} + \text{Recall}} \tag{5}$$

Para o Modelo 2 (multiclasse), usou-se o F1-macro, calculado pela média simples dos F1 de cada categoria, o que garante que a classe menos frequente tenha o mesmo peso na avaliação.

Para o Modelo 1 (binário), a métrica principal foi a Área sob a Curva ROC (AUC-ROC), que relaciona taxa de verdadeiros positivos e falsos positivos em diferentes limiares de decisão. Para o Modelo 2, usou-se a estratégia AUC-OvR (*One-vs-Rest*), calculando a média das curvas ROC de cada classe isolada contra as demais.

3.3 Equidade, explicabilidade e aspectos legais

3.3.1 Ciclos de retroalimentação em sistemas preditivos

Quando algoritmos preditivos passam a orientar decisões de alto impacto, existe o risco de criar ciclos de retroalimentação (*feedback loops*): as predições do modelo influenciam o comportamento dos seres humanos responsáveis pelo uso do modelo, o que gera novos dados que retroalimentam o próprio modelo, reafirmando padrões preexistentes.

No contexto jurídico, isso pode ocorrer se um modelo indicar baixa probabilidade de procedência para ações em certas empresas. Advogados influenciados por essa estimativa podem deixar de propor ações nessas localidades, reduzindo a amostragem de dados e privando o sistema de novos casos que contradiriam a tendência observada. O retreinamento sobre dados já considerando as ações reafirmaria o viés inicial. Por isso, a avaliação de um

sistema preditivo não pode se limitar à metricas e resultados mas também é necessário analisar seus efeitos colaterais ao longo do tempo.

3.3.2 Explicabilidade com SHAP

Da teoria dos jogos ao aprendizado de máquina. O framework SHAP (*SHapley Additive exPlanations*) [8] é um método de explicabilidade baseado na teoria dos jogos cooperativos. Ele adapta o conceito de valor de Shapley [10], originalmente proposto para distribuir recompensas de forma equitativa entre participantes de uma coalizão.

No seu uso para aprendizado de máquina, os atributos assumem o papel dos jogadores, e a recompensa a distribuir é o desvio da predição de um exemplo específico em relação à média global. O valor de Shapley de uma variável quantifica qual é a sua contribuição marginal média para o resultado, avaliando todas as combinações de atributos possíveis.

Definição formal. Sendo $N = \{1, 2, \ldots, p\}$ o conjunto total de atributos e $v(S)$ a função que mapeia um subconjunto $S \subseteq N$ à predição estimada pelo modelo, o valor de Shapley para o atributo i é:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} \left[v(S \cup \{i\}) - v(S) \right] \tag{6}$$

O termo combinatório pondera a probabilidade de cada coalizão S em permutações aleatórias, e a diferença $[v(S \cup \{i\}) - v(S)]$ isola o ganho marginal trazido pela variável examinada.

Propriedades e garantias. Os valores de Shapley têm base matemática garantida por quatro pilares: eficiência (a soma das contribuições individuais equivale à diferença entre a predição local e o valor esperado global), simetria (variáveis com contribuições marginais equivalentes recebem atribuições iguais), nulidade (atributos sem impacto no resultado têm valor zero) e aditividade (explicações de modelos compostos podem ser combinadas linearmente).

Aplicação prática: TreeSHAP. O cálculo exato da Equação 6 é computacionalmente inviável para bases complexas, pois envolve 2^p combinações. Para isso, usou-se o TreeSHAP [7], implementação especializada para modelos baseados em árvores que calcula os valores exatos de Shapley em tempo polinomial, explorando a estrutura das ramificações das árvores.

Visualizações SHAP. As explicações geradas incluem análises globais e locais. O gráfico de importância global mostra a média dos valores absolutos de SHAP por atributo no conjunto de teste, estabelecendo um ranking de relevância. O gráfico *beeswarm* complementa essa visão, mostrando a distribuição dos impactos locais de cada registro e relacionando os valores originais do atributo à direção do impacto preditivo.

3.3.3 Direito à explicação: LGPD e GDPR

O desenvolvimento de soluções de machine learning em âmbitos sensíveis está cada vez mais sujeito a governança e regulamentação jurídica. A exigência de explicabilidade não é somente focada em boas práticas de engenharia mas também é uma prerrogativa legal estabelecida por marcos regulatórios nacionais e internacionais.

LGPD: Lei nº 13.709/2018. A Lei Geral de Proteção de Dados, em vigor desde 2020, regula o tratamento de informações pessoais com base nos princípios de finalidade específica, adequação e transparência.

Para sistemas de decisão automatizada, o Artigo 20 garante ao titular o direito de solicitar revisão de decisões tomadas exclusivamente por tratamentos automatizados que afetem seus interesses civis ou profissionais. O parágrafo primeiro impõe ao controlador o dever de informar os critérios e procedimentos usados pelo algoritmo, respeitados os limites de segredo comercial. Embora este trabalho use dados descaracterizados na origem do seu desenvolvimento, o que os coloca fora do escopo direto de dados pessoais, os princípios de transparência funcionaram como requisitos arquiteturais no design do projeto.

GDPR: Regulamento (UE) 2016/679. O GDPR europeu é a principal referência global para governança de dados digitais e influenciou diretamente a legislação brasileira.

O Artigo 22 restringe decisões baseadas exclusivamente em processamento automatizado quando geram efeitos jurídicos ou impactos significativos em indivíduos, incluindo criação de perfis. Nas exceções previstas pelo artigo, o regulamento exige salvaguardas como o direito de contestação e a garantia de intervenção humana.

Além disso, o Considerando 71 amplia essa interpretação, sugerindo que os indivíduos afetados devem receber explicações sobre a lógica do sistema após a avaliação automatizada. Esse entendimento consolidou na literatura o conceito de "direito à explicação" (*right to explanation*), estimulando o desenvolvimento de metodologias de IA explicável na interseção entre computação e direito [4].

4 Objetivos

O objetivo geral deste trabalho é desenvolver e avaliar, sob uma perspectiva sociotécnica, um pipeline de aprendizado de máquina para predição de desfechos processuais no TRT-15, usando exclusivamente dados públicos do DataJud. A pesquisa se organiza em dois eixos complementares. O eixo técnico foca na construção de um ambiente preditivo estruturado, com engenharia de features, modelagem supervisionada e validação cronológica. O eixo analítico e ético usa métodos de explicabilidade algorítmica como ferramentas de auditoria para mapear padrões decisórios e possíveis vieses sistêmicos, promovendo uma reflexão sobre o uso responsável de inteligência artificial em ambientes governamentais sensíveis.

4.1 Objetivo técnico

O eixo técnico visa projetar e implementar o fluxo completo de engenharia de dados e modelagem, compreendendo extração, saneamento, transformação, treinamento e validação dos classificadores. Para isso, foram desenvolvidos dois modelos com escopos distintos: um classificador binário para estimar a probabilidade de acordos judiciais, e um classificador multiclasse para discriminar entre três desfechos de mérito (decisões favoráveis ao trabalhador, improcedências e extinções sem julgamento). Operar os dois modelos em paralelo permite explorar abordagens específicas de engenharia de atributos, estratégias para distribuições desbalanceadas e métricas customizadas de avaliação, fornecendo uma base metodológica sólida para as análises interpretativas seguintes.

4.2 Objetivo analítico

O eixo analítico parte da premissa de que modelos treinados sobre dados históricos tendem a absorver e propagar as distorções estruturais presentes nesses registros. Usando os valores de contribuição marginal do SHAP, o trabalho busca auditar os critérios de inferência dos modelos para identificar quais variáveis mais influenciam as predições e o que essa dependência revela sobre a dinâmica decisória do TRT-15. Com isso, pretende-se fundamentar uma discussão crítica sobre os riscos de viés algorítmico em plataformas jurídicas, a importância da auditabilidade em sistemas de suporte à decisão e os desafios de governança no uso de dados processuais públicos.

5 Metodologia

O desenvolvimento do trabalho seguiu um pipeline estruturado em seis etapas sequenciais. Além do aspecto técnico em cada fase, o processo envolveu escolhas deliberadas orientadas por princípios de responsabilidade de dados, com atenção especial à seleção de features, definição jurídica dos alvos e escolha de métricas sem viés.

5.1 Contexto da coleta: o TRT-15 e o DataJud

A escolha do TRT-15 como escopo amostral se justifica por volume e consistência estatística. Sediado em Campinas e cobrindo o interior e o litoral paulista, o tribunal tem uma das maiores movimentações processuais do país, oferecendo dados suficientes para o ajuste estável dos modelos, inclusive os cortes para treino, teste e validação. Além disso, restringir o trabalho a um único tribunal elimina possíveis inconsistências silenciosas que são geradas por diferenças entre regiões administrativas distintas.

A extração dos dados foi feita via API pública do DataJud, mantida pelo CNJ. Para cada processo, foram coletados: número único identificador, código da vara trabalhista, classe da ação, assunto segundo as TPUs, município de origem, marcos temporais de protocolo e baixa, e o histórico de andamentos codificados.

É importante ressaltar que a API do DataJud já funciona com anonimização dos dados, omitindo nomes, CPF/CNPJ e o valor nominal da causa. Essa restrição foi mantida ao longo

do projeto, sem cruzamento de dados ou técnicas de reidentificação. Os efeitos dessa escolha no desempenho do classificador e na análise ética são discutidos na seção de Discussão.

5.2 Coleta de dados

A coleta principal foi feita por meio de requisições à API do DataJud, filtrando processos do TRT-15 com autuação entre 2018 e 2023. O recorte de 2018 em diante garante consistência dos dados após a Reforma Trabalhista de 2017, que alterou o perfil dos processos, conforme discutido na Fundamentação Teórica.

Para cada processo, foram extraídos: identificador único, código do tribunal e da vara, assunto principal e classe processual pelas tabelas unificadas do CNJ, município de origem (com o respectivo código IBGE), data de autuação, data de encerramento e histórico de movimentos processuais com seus códigos e datas.

Os dados socioeconômicos foram obtidos diretamente do IBGE [5] e cruzados com a base processual pelo código de município. Para cada município presente nos dados do DataJud, foram coletados: PIB *per capita* (Pesquisa do PIB dos Municípios), estimativa populacional (Estimativas da População), taxa de desemprego e indicadores de mercado de trabalho (PNAD Contínua) e classificação microrregional geográfica. Como os dados do IBGE têm granularidade anual, cada processo recebeu os indicadores do município de origem referentes ao ano de sua autuação, de modo que variações econômicas ao longo do período amostral ficassem refletidas nas features.

5.3 Pré-processamento e definição dos desfechos

O pré-processamento envolveu a limpeza da base de dados e a estruturação dos alvos. Foram descartados registros com ausência de campos obrigatórios, como identificação da vara, assunto principal ou data de autuação. Os códigos categóricos de assunto e classe foram normalizados conforme as tabelas unificadas do CNJ, permitindo comparações padronizadas entre comarcas e períodos.

A rotulação dos alvos exigiu mapear o histórico de andamentos processuais. A partir dos códigos padronizados de movimentação do CNJ, foram identificados os atos que determinavam o encerramento em primeiro grau, classificados em categorias mutuamente exclusivas conforme seu significado jurídico.

Para o Modelo 1 (binário), os desfechos foram consolidados em duas classes: "acordo" (processos encerrados por conciliação ou homologação de transação judicial) e "não acordo" (todos os demais encerramentos).

Para o Modelo 2 (multiclasse), foram definidas três categorias: "favorável ao trabalhador" (sentenças que acolheram total ou parcialmente os pedidos), "improcedente" (rejeição integral dos pleitos) e "extinção sem resolução de mérito" (encerramento por vícios formais, ilegitimidade ou abandono). Registros com andamentos contraditórios ou códigos ambíguos foram removidos para garantir a qualidade da rotulação.

A distribuição amostral resultante revelou desbalanceamento entre as classes de desfecho, sobretudo na categoria de extinção sem resolução de mérito, que representa menos de 8% da base geral. Conforme os princípios discutidos na Fundamentação Teórica, esse

desbalanceamento exigiu tratamento específico nas etapas subsequentes de treinamento e avaliação dos modelos.

5.4 Engenharia de features

O processo resultou em 42 variáveis preditivas, organizadas em seis blocos:

- **Features Temporais:** Codificam características do momento de autuação da ação (mês, dia da semana, trimestre e indicadores de volume sazonal da vara) para capturar flutuações de carga de trabalho.
- **Features Processuais:** Mapeiam propriedades técnicas da ação, incluindo o código do assunto principal, a classe processual TPU, o grau de jurisdição e o rito adotado.
- **Features Macroeconômicas:** Integram dados socioeconômicos do município de origem do IBGE, como PIB *per capita* (em escala logarítmica), estimativa populacional, taxas de desemprego e divisão microrregional.
- **Features de Recência RFM:** Calculam o intervalo temporal desde o último processo protocolado na mesma vara com o mesmo assunto, adaptando métricas de comportamento transacional para o fluxo de litígios.
- **Features de Frequência RFM:** Quantificam o volume acumulado de ações por vara, assunto e comarca em diferentes janelas de tempo, mapeando a densidade de litigância de cada unidade jurídica.
- **Taxas Históricas:** Calculam proporções passadas de desfechos (acordos, procedências e improcedências) vinculadas a cada vara, assunto ou comarca, fornecendo ao modelo um indicador prévio do comportamento decisório de cada contexto.

O controle contra vazamento de dados temporais foi implementado nessa etapa. Todas as métricas baseadas em histórico foram geradas com restrição cronológica, incluindo apenas processos ajuizados antes do registro em processamento. Isso garante que o classificador dê suas previsões com informações que de fato estariam disponíveis em um cenário de produção real.

5.5 Estratégia de validação

A separação dos conjuntos de dados seguiu critério estritamente cronológico, evitando divisões aleatórias que quebrariam a dependência temporal dos registros. Processos de 2018 a 2021 foram separados para treino, enquanto os de 2022 para validação e os de 2023 para o teste de performance.

Essa estratégia de separação dos dados atende a dois objetivos. Do ponto de vista técnico, reproduz o cenário real de operação de sistemas inteligentes, em que o modelo é treinado com dados históricos para fazer inferências sobre eventos futuros. Do ponto de vista da validação, expõe a estabilidade temporal da solução, indicando se os padrões aprendidos se mantêm diante de mudanças graduais no comportamento do sistema judiciário.

5.6 Treinamento e otimização dos modelos

O pipeline incluiu o ajuste de duas arquiteturas independentes.

O Modelo 1 é um classificador binário baseado em LightGBM, focado em estimar a ocorrência de acordo judicial. Esse alvo foi escolhido pela utilidade prática do indicador para as partes, já que a decisão de conciliação representa um dos momentos de maior impacto econômico no litígio.

O Modelo 2 é um classificador multiclasse baseado em CatBoost, para prever os três desfechos de mérito. A escolha pelo CatBoost se justifica pela sua capacidade de processar variáveis categóricas de alta cardinalidade, como os códigos TPU de assunto e vara, sem precisar de transformações que expandam a dimensionalidade.

Para lidar com o desbalanceamento identificado na distribuição das classes de desfecho, foram adotados pesos de classe no treinamento do Modelo 2, penalizando com maior intensidade os erros cometidos sobre a classe de extinção sem resolução de mérito, minoritária na base. Essa escolha foi acompanhada da seleção de métricas de validação sensíveis a esse desbalanceamento, como o F1-macro e a AUC-OvR, apresentadas na Fundamentação Teórica.

A busca pelos hiperparâmetros ótimos foi feita com o framework Optuna [1], usando otimização bayesiana com validação temporal. Para o Modelo 1, a função objetivo maximizou a AUC-ROC e para o Modelo 2, o F1-macro, garantindo peso equivalente entre as classes na calibração.

5.7 Análise de explicabilidade

A interpretação dos modelos foi feita com o SHAP, aplicando o `TreeExplainer` sobre os conjuntos de teste de ambas as arquiteturas. O mapeamento das contribuições marginais foi estruturado em duas perspectivas.

A abordagem macro avalia a importância global das variáveis pela média dos valores absolutos de SHAP em toda a base de teste, identificando quais atributos têm maior peso no comportamento preditivo geral. A abordagem micro investiga a orientação dos impactos locais via gráficos *beeswarm*, relacionando os valores originais das variáveis à flutuação positiva ou negativa dos escores de decisão.

No Modelo 2, essa auditoria foi feita separadamente para cada uma das três classes. Essa decomposição é fundamental para o objetivo analítico do trabalho pois permite rastrear se o modelo baseia suas inferências em atributos associados ao mérito jurídico ou em indicadores circunstanciais de contexto (como comarca ou código de vara), o que evidenciaria assimetrias sistêmicas nas decisões.

6 Resultados

Esta seção detalha as evidências empíricas obtidas nas fases de experimentação: análise exploratória, desempenho do classificador binário (Modelo 1), performance do classificador multiclasse (Modelo 2) e auditoria de explicabilidade via SHAP para ambas as configurações.

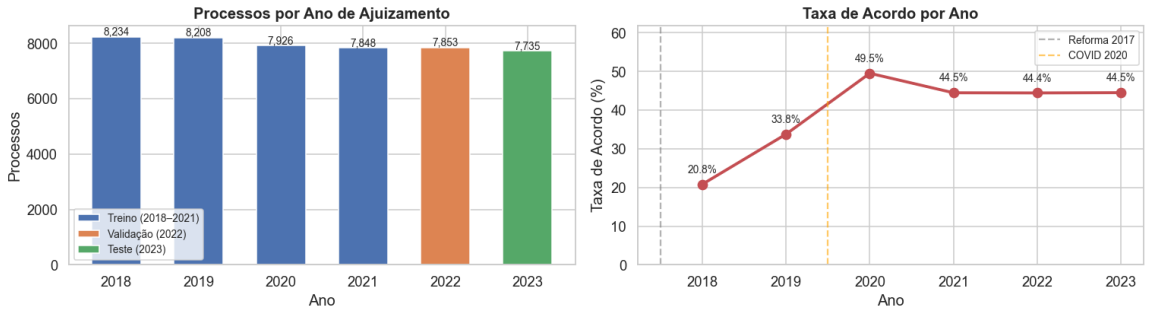


Figura 1: Volume de processos por ano de ajuizamento e taxa de acordo anual. As linhas tracejadas indicam a Reforma Trabalhista (2017) e a pandemia de COVID-19 (2020).

6.1 Análise exploratória dos dados

A amostra final incluiu ações do TRT-15 de 2018 a 2023, divididas entre treino (2018–2021), validação (2022) e teste (2023). O volume anual de processos se manteve estável ao longo do período, oscilando entre 7.700 e 8.200 processos por ano, conforme ilustrado na Figura 1.

A taxa histórica de conciliação, por outro lado, variou bastante no período. Em 2018, os acordos representavam 20,8% dos desfechos. Esse índice cresceu progressivamente até atingir 49,5% em 2020, reflexo das políticas de incentivo à conciliação virtual adotadas durante a pandemia de COVID-19. A partir de 2021, a taxa se estabilizou próximo a 44% nos anos de 2022 (44,4%) e 2023 (44,5%). Essa oscilação histórica reforça a necessidade da divisão temporal de validação para monitorar a aderência dos modelos a quebras de tendência.

A análise dos 25 assuntos mais recorrentes, apresentada na Figura 2, mostra variação expressiva nas taxas de acordo conforme o tema jurídico da demanda. Processos de "Reconhecimento de Relação de Emprego" chegaram a médias de conciliação de 61%, enquanto demandas de "Indenização/Dobra/Terço Constitucional" ficaram em torno de 4%. Essa dispersão indica que a distribuição dos temas pelas TPUs carrega poder preditivo.

A análise de correlação linear das features numéricas (Figura 3) indicou baixa colinearidade geral entre os grupos de features. As exceções ficaram em pareamentos esperados, como o número de assuntos e sua transformação logarítmica ($r = 0{,}91$), e indicadores macroeconômicos colineares, como taxas percentuais de desemprego e o indicador binário de desemprego elevado ($r = 0{,}94$). A independência entre as demais variáveis valida o desenho do conjunto de features.

6.2 Modelo 1: classificação binária

O Modelo 1 foi desenvolvido para estimar a probabilidade de encerramento de um processo via acordo judicial. Após a otimização com Optuna, a arquitetura LightGBM superou o CatBoost nos conjuntos de validação e teste, sendo definida como a configuração final do pipeline binário.

Os indicadores de performance estão na Tabela 1.

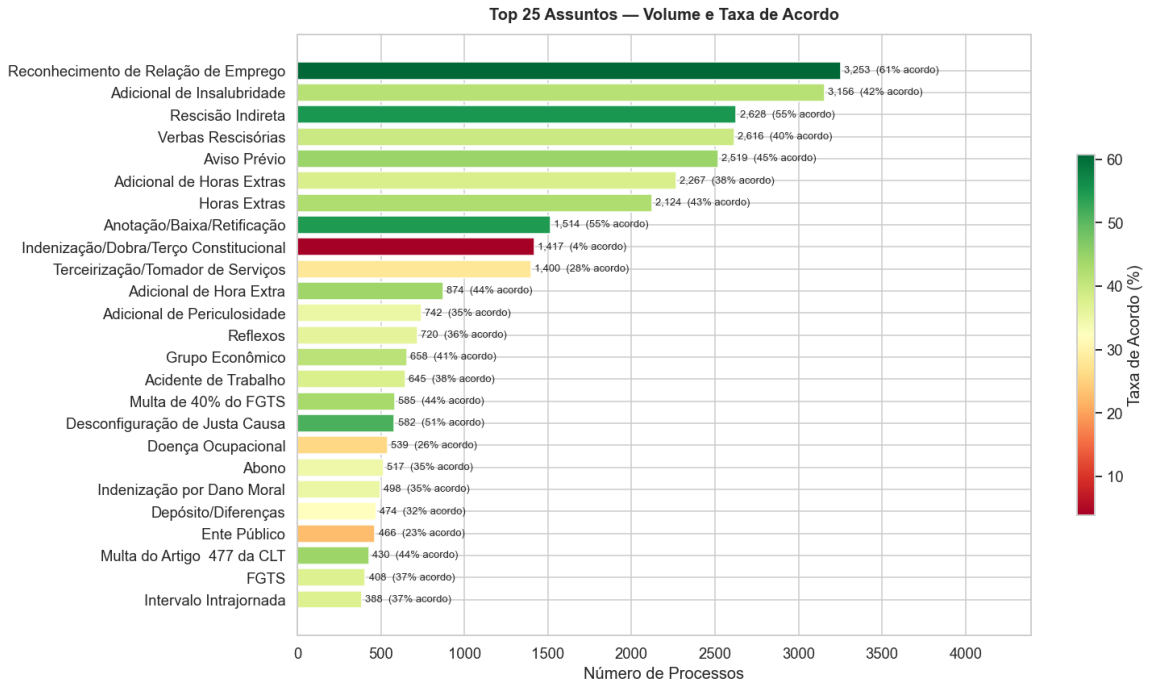


Figura 2: Top 25 assuntos por volume de processos. A acentuação cromática indica a taxa de acordo histórica do assunto.

Tabela 1: Desempenho do Modelo 1 (LightGBM) nos conjuntos de validação e teste.

Métrica	**Validação 2022**	**Teste 2023**
AUC-ROC	0,7561	0,7378
F1-score (macro)	0,6750	0,6660

O classificador registrou AUC-ROC de 0,7561 na validação e 0,7378 no teste, com a diminuição de 1,8 ponto percentual entre os períodos. Essa diferença pequena atesta a estabilidade temporal do modelo, dado que os padrões aprendidos entre 2018 e 2021 mantêm capacidade preditiva sobre dados de 2023. A comparação das curvas ROC (Figura 4) mostra desempenhos próximos entre LightGBM e CatBoost.

As matrizes de confusão (Figura 5) mostram equilíbrio na distribuição de erros. No conjunto de teste, o modelo classificou corretamente 7.684 processos que resultaram em conciliação e 9.294 da classe contrária. Os erros foram simétricos: 3.164 acordos classificados como não-acordo e 5.299 não-acordos classificados como acordo.

6.3 Modelo 2: classificação multiclasse

O Modelo 2 discrimina o desfecho processual entre três classes. Nessa formulação, o CatBoost superou o LightGBM no F1-macro em ambas as partições temporais.

A distribuição amostral revelou desbalanceamento considerável, exigindo técnicas de

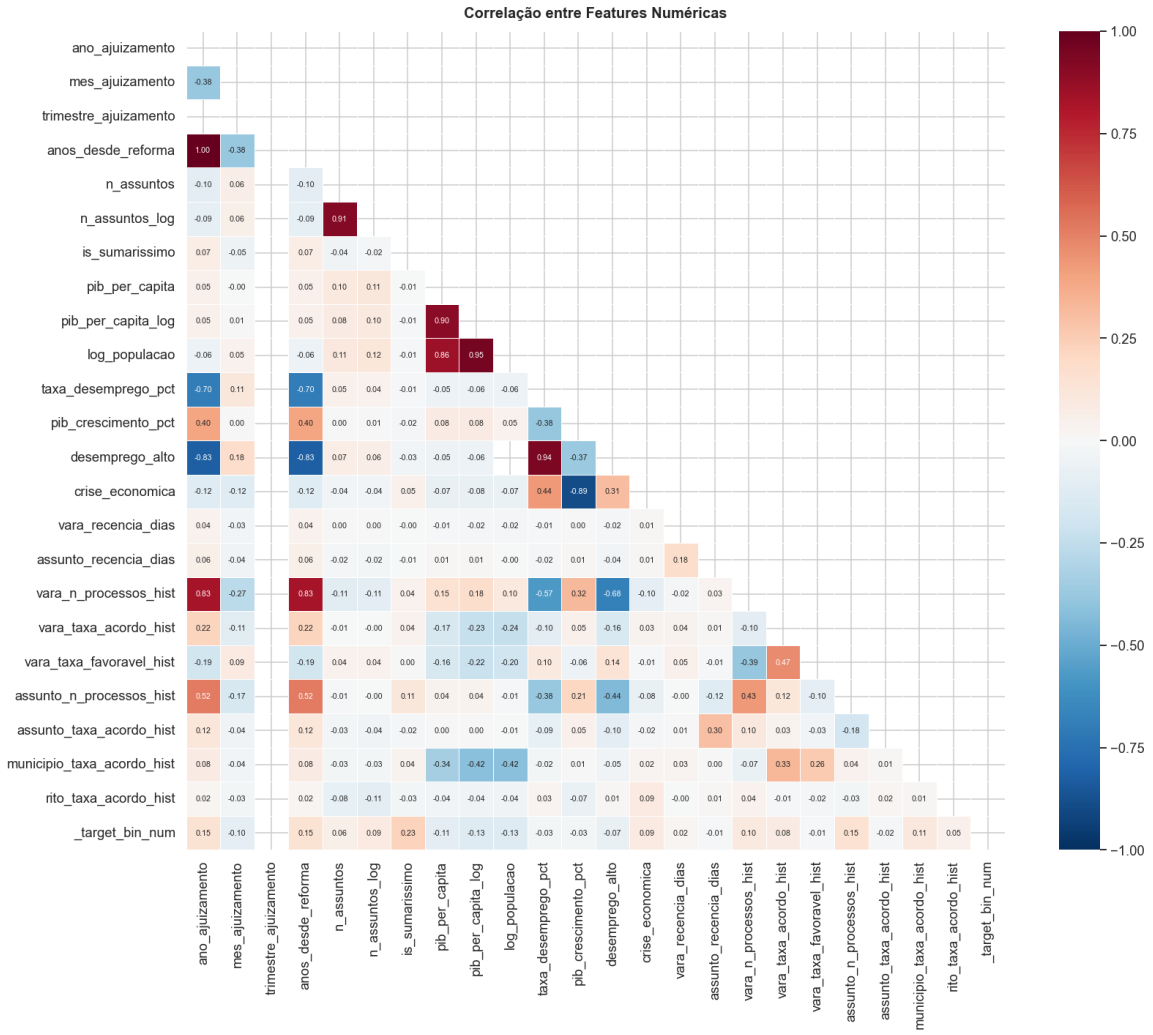


Figura 3: Matriz de correlação entre as features numéricas do conjunto de dados. Apenas correlações com $|r| \geq 0{,}1$ são destacadas.

balanceamento dos dados no treino. A classe de extinção sem resolução de mérito recebeu peso multiplicativo de 13,4 em relação à classe favorável ao trabalhador, por representar menos de 8% do banco geral. Os resultados estão na Tabela 2.

Tabela 2: Desempenho do Modelo 2 (CatBoost) nos conjuntos de validação e teste.

Métrica	**Validação 2022**	**Teste 2023**
F1-macro	0,4279	0,4089
F1-weighted	0,5806	0,5829
AUC-OvR	0,7063	0,7033

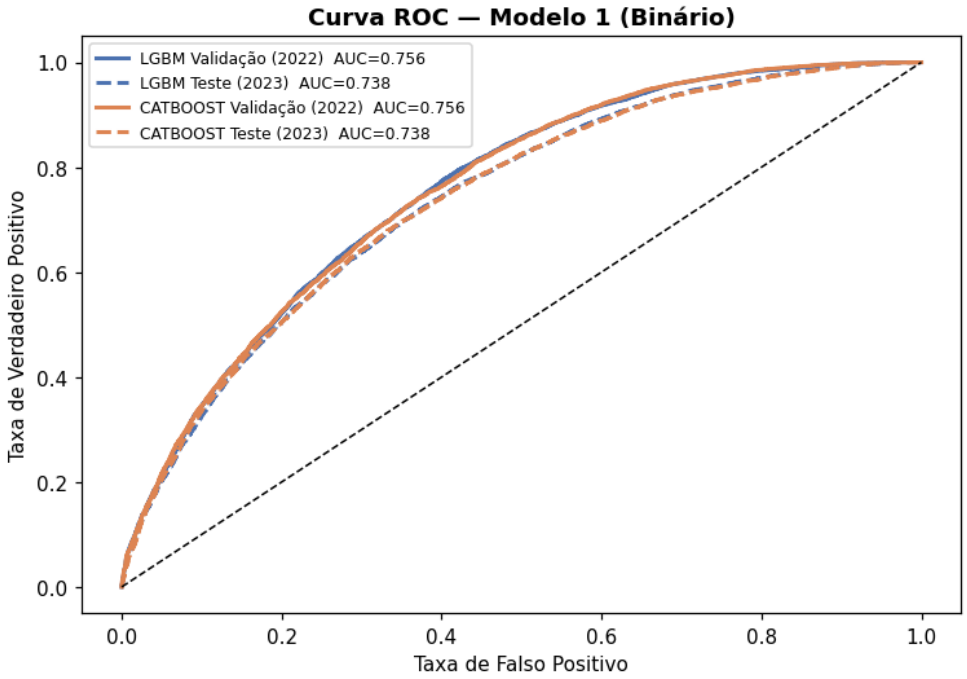


Figura 4: Curva ROC do Modelo 1 para LightGBM e CatBoost nos conjuntos de validação (2022) e teste (2023).

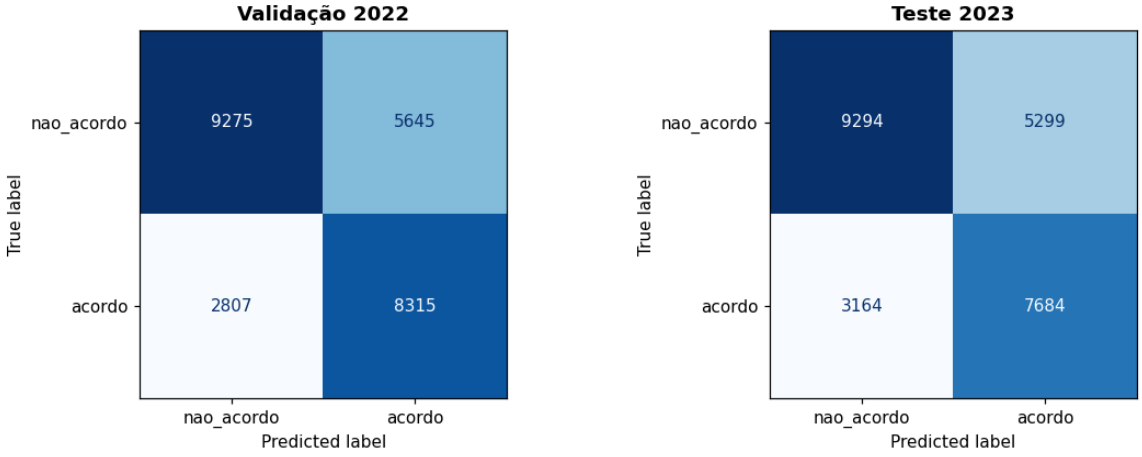


Figura 5: Matrizes de confusão do Modelo 1 para os conjuntos de validação (2022) e teste (2023).

O F1-macro de 0,4089 no teste reflete a dificuldade de modelar a classe minoritária de extinção. A análise por categoria (Figura 6) demonstra que o classificador atinge F1 de 0,634 para decisões favoráveis e 0,461 para improcedências, mas apenas 0,132 para extinções no teste. A queda no F1 de extinção entre validação (0,176) e teste (0,132) indica maior

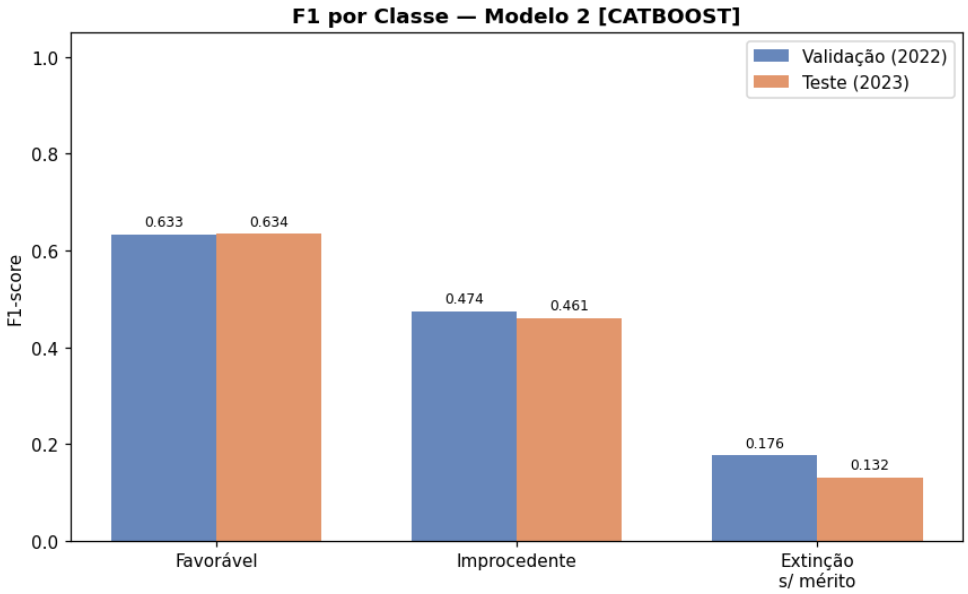


Figura 6: F1-score por classe do Modelo 2 (CatBoost) nos conjuntos de validação (2022) e teste (2023).

instabilidade preditiva dessa categoria ao longo do tempo.

As matrizes de confusão do Modelo 2 (Figura 7) evidenciam a sobreposição estatística entre extinção e as demais classes. No teste, de 276 ocorrências reais de extinção, o modelo identificou corretamente 157 casos, classificando erroneamente 79 como favoráveis e 40 como improcedentes. Há também confusão mútua esperada entre as classes favorável e improcedente, decorrente do compartilhamento de atributos estruturais comuns na petição inicial.

6.4 Explicabilidade SHAP: Modelo 1

A análise SHAP do Modelo 1 foi feita sobre o conjunto de teste. A Figura 8 mostra quais variáveis tiveram mais peso nas predições, medido pela média dos valores absolutos de SHAP de cada feature.

As três variáveis mais importantes foram o Município (IBGE), o Código da vara e o Assunto principal, com valores médios de 0,41, 0,29 e 0,27. O ponto mais relevante desse resultado é o que essas variáveis têm em comum dado que nenhuma delas descreve o conteúdo jurídico do processo. Todas descrevem onde e sobre qual tema o processo tramita, não o que foi pedido ou quais provas foram produzidas.

O gráfico *beeswarm* (Figura 9) mostra como cada variável impacta as predições individualmente. Para o Município, os pontos aparecem espalhados nos dois sentidos do eixo, algumas comarcas aumentam a probabilidade de acordo, outras a reduzem, e não há uma tendência linear clara. Para o Assunto principal, os processos com valores baixos da variável se concentram no lado negativo, o que indica que certos tipos de demanda estão sistemati-

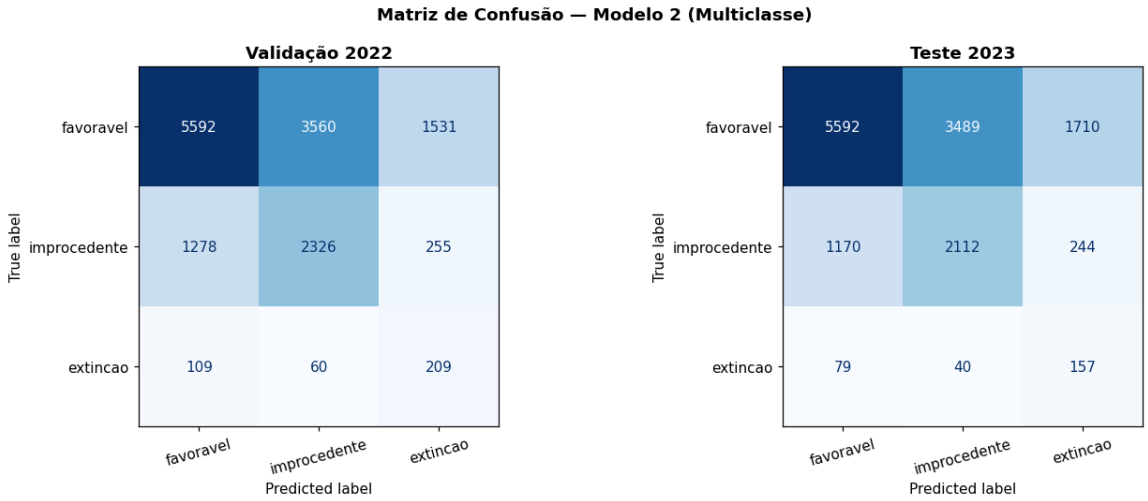


Figura 7: Matrizes de confusão do Modelo 2 para os conjuntos de validação (2022) e teste (2023).

camente associados a menores taxas de conciliação.

Para ver como o modelo raciocina em casos concretos, a Figura 10 mostra os gráficos *waterfall* de dois processos do conjunto de teste: um com probabilidade de acordo de 86,24% (Figura 10a) e outro com apenas 4,66% (Figura 10b).

No processo com alta probabilidade de acordo, o Município é a variável que mais empurra a predição para cima (+0,79), com contribuições adicionais do Assunto principal (+0,32), do Código da classe (+0,18) e da Taxa histórica do rito (+0,18). No processo oposto, o Assunto principal derruba o escore sozinho em −1,61, com o Código da vara (−0,33), o Número de assuntos logarítmico (−0,32) e o Município (−0,28) reforçando a direção negativa.

6.5 Explicabilidade SHAP: Modelo 2

A análise SHAP do Modelo 2 foi feita separadamente para cada uma das três classes. A Figura 11 mostra a importância global de cada feature, com as barras divididas pela contribuição de cada classe.

O Município se manteve como a variável mais importante no modelo multiclasse, com índice cumulativo acima de 0,6, seguido pelo Assunto principal e pelo Código da vara. O que muda em relação ao Modelo 1 é a distribuição por classe dado que o Município contribui de forma parecida para os três desfechos, o Assunto principal pesa mais para improcedências e extinções e o Código da vara é especialmente relevante para identificar extinções sem resolução de mérito.

Os gráficos *beeswarm* por classe (Figuras 12, 13 e 14) mostram como essas variáveis atuam em cada desfecho.

Para a classe de extinção (Figura 12), o Município apresenta pontos espalhados nos dois lados do eixo, sem um padrão claro entre o valor da variável e a direção do impacto. Isso sugere que o encerramento formal de processos não segue uma lógica uniforme, mas reflete

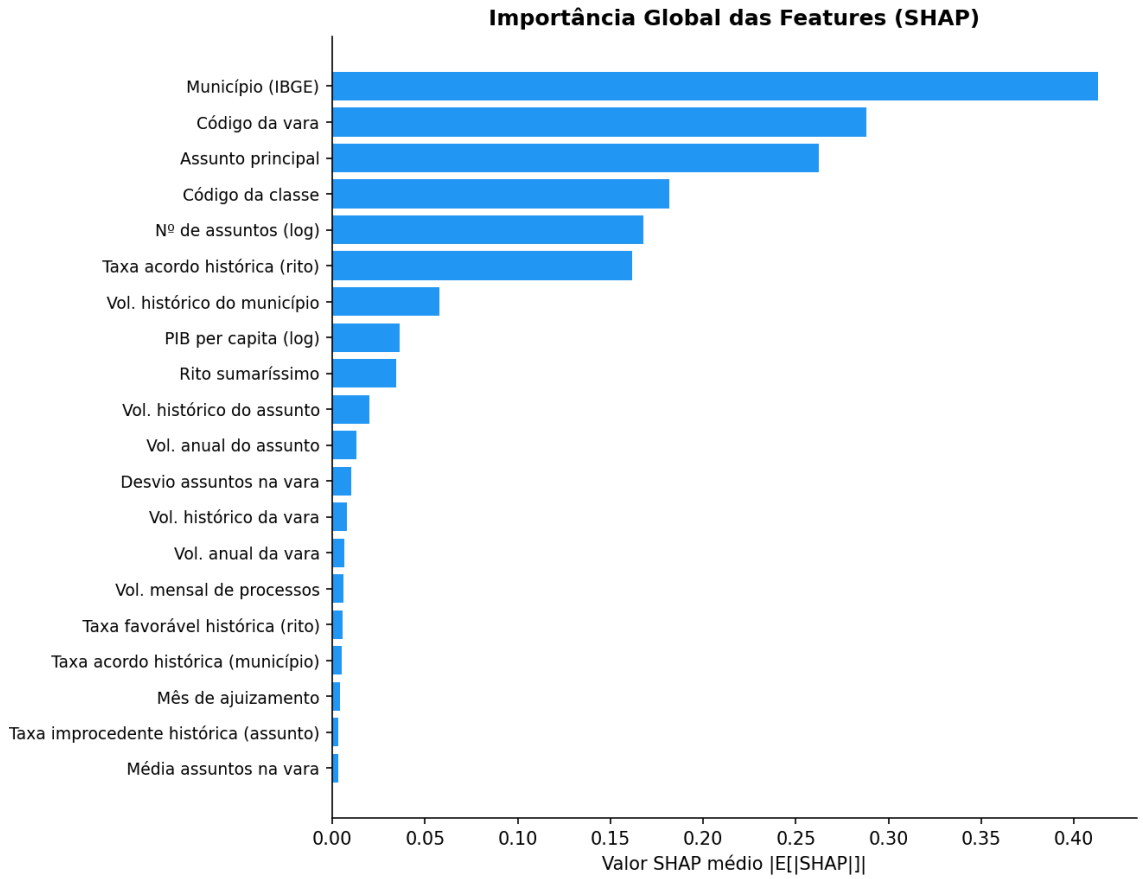


Figura 8: Importância global das features do Modelo 1, medida pela média dos valores absolutos de SHAP.

práticas específicas de cada vara ou microrregião.

Para desfechos favoráveis ao trabalhador (Figura 13), Município e Código da vara continuam dominando, ambos com pontos bem espalhados nos dois sentidos. O Assunto principal aparece com coloração mais neutra, o que é esperado para uma variável categórica com muitos valores distintos, os temas jurídicos dessa classe se distribuem de forma equilibrada, sem que nenhum grupo concentre o impacto em uma direção só.

Para a improcedência (Figura 14), o Assunto principal e o Município têm os maiores espalhamentos, com assimetria clara, onde certas comarcas e certos tipos de demanda aparecem com força em um dos lados, indicando vínculos consolidados com a rejeição dos pedidos.

Em síntese, as análises de ambos os modelos apontam para o mesmo padrão no qual as predições são dominadas por variáveis que descrevem o contexto da tramitação, como localização, vara e tipo de demanda, e não por informações sobre o mérito individual de cada caso.

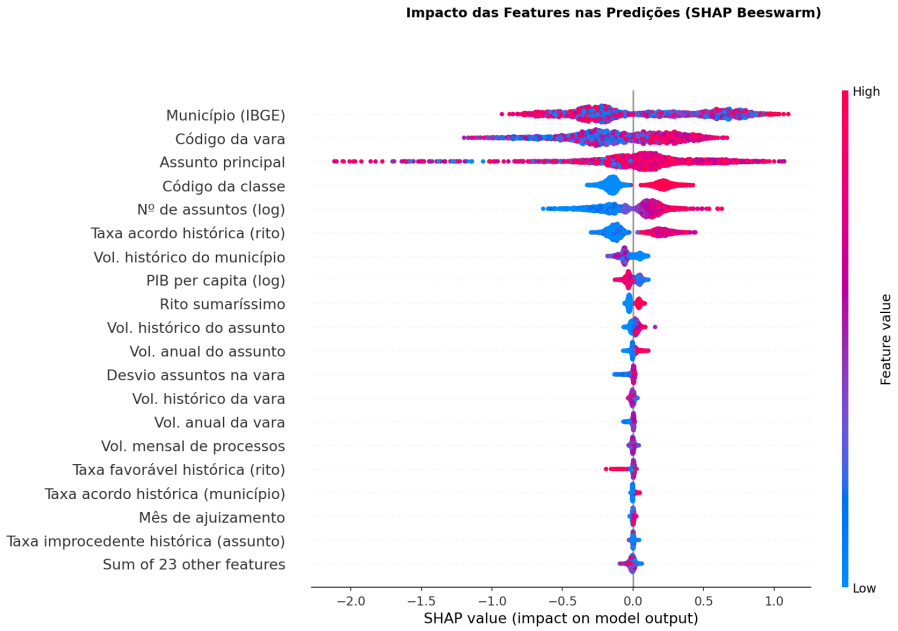
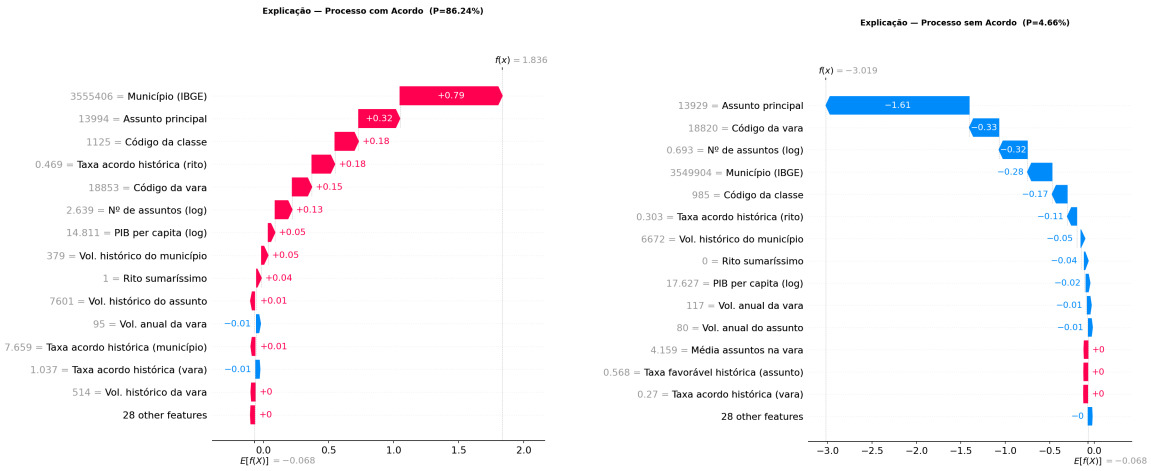


Figura 9: Gráfico *beeswarm* dos valores SHAP do Modelo 1. Cada ponto representa um processo do conjunto de teste. A cor indica o valor original da feature (vermelho: alto; azul: baixo).



(a) Processo com acordo (P=86,24%).

(b) Processo sem acordo (P=4,66%).

Figura 10: Gráficos *waterfall* SHAP para dois processos individuais do conjunto de teste. Os valores à esquerda indicam o valor original de cada feature.

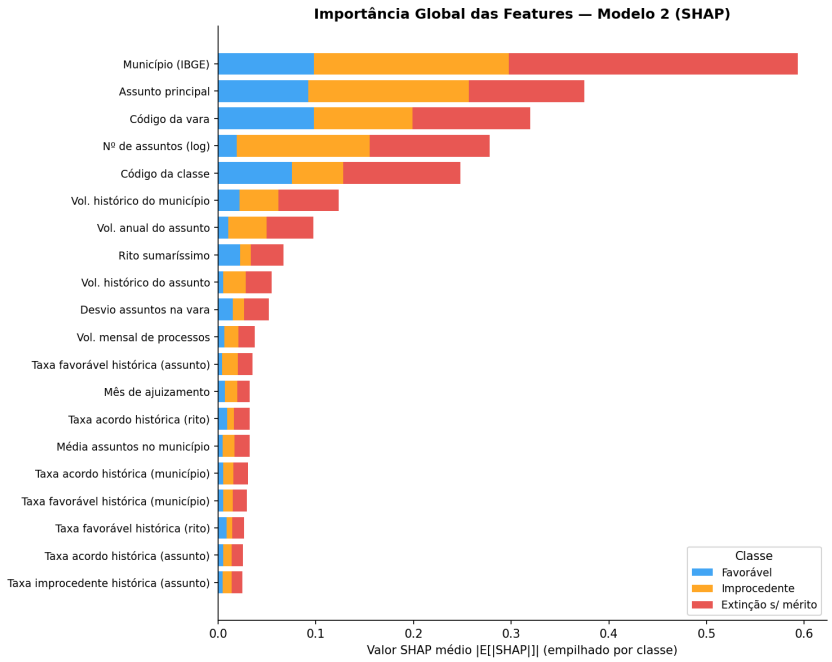


Figura 11: Importância global das features do Modelo 2, decomposta por classe. Cada segmento da barra representa a contribuição média absoluta de SHAP para aquela classe.

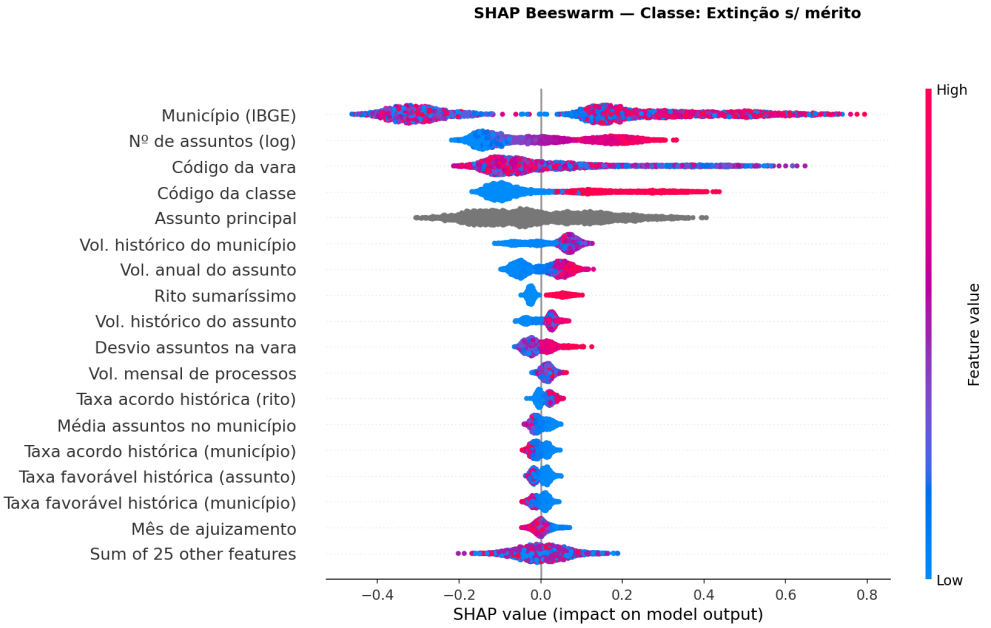


Figura 12: Gráfico *beeswarm* SHAP para a classe extinção sem mérito (Modelo 2).

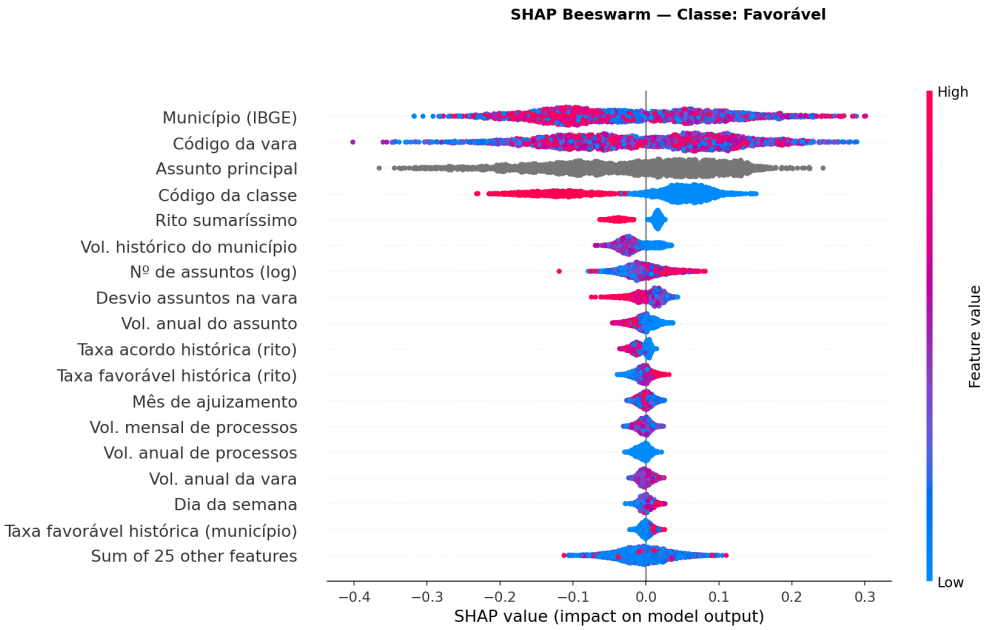


Figura 13: Gráfico *beeswarm* SHAP para a classe favorável ao trabalhador (Modelo 2).

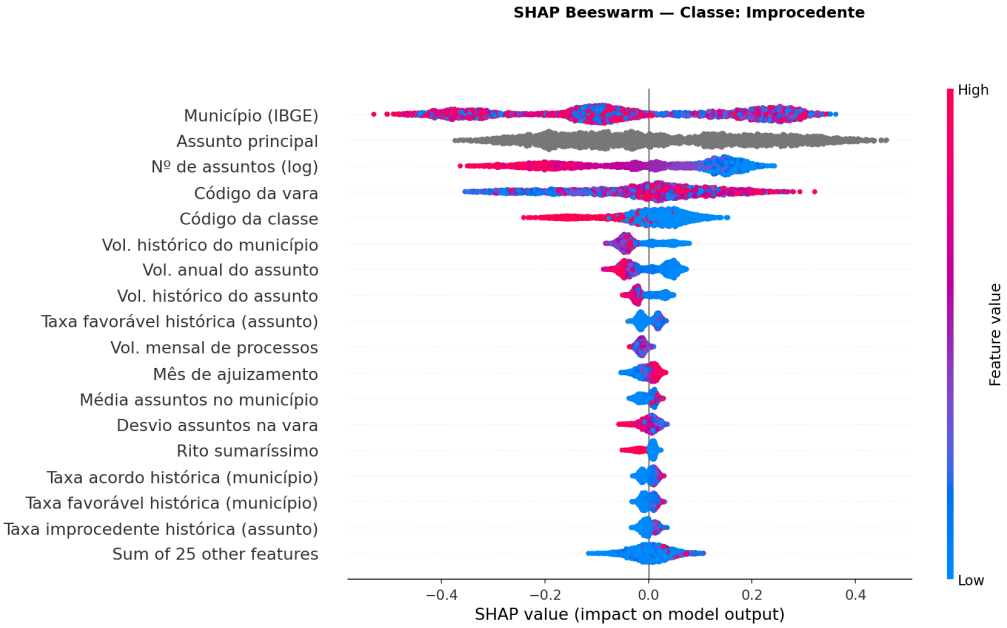


Figura 14: Gráfico *beeswarm* SHAP para a classe improcedente (Modelo 2).

7 Discussão

Os padrões revelados pelas análises de explicabilidade exigem uma avaliação que vai além das métricas de performance algorítmica. Esta seção discute a natureza do viés geográfico mapeado pelo SHAP, os limites de generalização do escopo amostral, as implicações éticas da ausência de dados das partes sob a ótica de privacidade por design, riscos de retroalimentação de dados e os critérios para implantação responsável de ferramentas inteligentes em sistemas de justiça.

7.1 Viés geográfico e estrutural

Os resultados relacionados à análise de explicabilidade deste trabalho têm origem nas features que coordenam o comportamento preditivo dos modelos. Em ambas as arquiteturas, o maior peso de decisão se concentra no município de origem, no código da vara trabalhista e no assunto temático. Nenhum desses componentes descreve fundamentação jurídica individual ou a qualidade probatória do caso, nesse caso, mapeiam o ambiente geográfico e institucional ao qual o processo está inserido.

Nesse sentido, o modelo funciona como um estimador das tendências históricas das unidades judiciárias. A comarca e a vara atuam como *proxies* das práticas de gestão e perfis decisórios acumulados pelas secretarias ao longo do tempo. O algoritmo não extrai os fundamentos que determinam a determinação de um acordo ou a procedência de um pedido, entretanto, mapeiam os contextos nos quais esses resultados aparecem com maior frequência.

Esse comportamento reflete as limitações das variáveis disponíveis na API pública para capturar as características do mérito das ações, o que força o modelo a otimizar com base em dados puramente estruturais. Do ponto de vista substantivo, a dominância das variáveis espaciais pode indicar assimetrias reais na operação da Justiça do Trabalho, sugerindo que o desfecho de uma ação sofre influência de fatores exógenos ao direito, associados ao local de distribuição e à unidade responsável pelo julgamento.

Os dois cenários impõem restrições ao uso prático de um modelo de machine learning. No primeiro, o modelo tem utilidade preditiva estatística, mas não é utilizável para uma operação que exija avaliação individualizada do direito de um litigante. No segundo, a ferramenta passa a propagar distorções sistêmicas ao guiar decisões com base em assimetrias históricas locais. Nesse caso, o sistema preditivo perdura os vieses já presentes na base. Esses fatores reforçam a relevância dos métodos de explicabilidade, cujo papel é expor as dependências, viabilizando a auditoria crítica do sistema.

7.2 Limites de generalização

O ajuste dos classificadores com dados restritos ao TRT-15 entre 2018 e 2023 estabelece limites claros para a aplicação dos resultados.

A restrição espacial é o fator principal. O TRT-15 cobre o interior e o litoral de São Paulo, região com dinâmicas econômicas e agroindustriais particulares. Os padrões de conciliação e julgamento de comarcas como Campinas ou Ribeirão Preto não se repetem

necessariamente em outros tribunais, como o TRT-2 (Grande São Paulo) ou o TRT-4 (Rio Grande do Sul). Os parâmetros estimados por estes modelos não devem ser transferidos para outros tribunais sem retreinamento local completo.

A delimitação temporal impõe restrições semelhantes. A base de treino (2018–2021) captura os impactos iniciais de adaptação à Reforma Trabalhista e as distorções volumétricas da pandemia de COVID-19. Os padrões aprendidos refletem essa conjuntura histórica específica. À medida que os efeitos da pandemia e a jurisprudência da reforma se alteram, os critérios do algoritmo podem perder representatividade. Além disso, o escopo limitado ao primeiro grau impede que os modelos avaliem reversões de mérito ocorridas em sede de recurso nas instâncias superiores.

Por fim, o escopo está limitado ao primeiro grau. Desfechos revertidos em instâncias superiores não entram na definição de resultado usada aqui. Esse comportamento é uma consequência do que o DataJud disponibiliza, mas é uma limitação que precisa ser considerada ao interpretar a relevância prática das predições.

7.3 Ausência de dados das partes e privacidade por design

O DataJud público não disponibiliza dados de identificação das partes, nomes, CPF, CNPJ ou valores de causa. Essa restrição foi mantida ao longo do projeto como uma decisão deliberada de arquitetura.

Durante o desenvolvimento, foi avaliada a possibilidade de enriquecer a base cruzando os dados do DataJud com plataformas como o Jusbrasil, que disponibilizam dados das partes por *scraping* ou consultas pagas. Tecnicamente, seria possível reidentificar empresas rés e seus setores de atividade.

Essa alternativa foi descartada principalmente por uma razão ética, considerando que os dados do DataJud foram abertos pelo CNJ com finalidade de transparência pública sobre o funcionamento do sistema judiciário como um todo. Usá-los como ponto de partida para reidentificar partes individuais extrapola essa finalidade, contrariando o princípio da finalidade do Art. 6º da LGPD.

Além disso, a ausência de dados sensíveis aponta para algo a ser considerado além da LGPD: mais dados não são necessariamente melhores. A decisão de incluir uma variável precisa considerar não só o ganho preditivo que ela traz, mas o que o modelo vai aprender com ela e quais comportamentos esse aprendizado pode gerar. Dados sobre a identidade das partes aumentariam a acurácia histórica do modelo, mas ao custo de produzir predições que retroalimentam e reforçam desvantagens acumuladas. A governança dos dados começa antes do treinamento, nas escolhas sobre como as features serão construídas e quais dados devem ser coletados.

A restrição aos metadados do DataJud define, portanto, um modelo de *privacidade por design* dado que o escopo de dados foi delimitado pelo que era metodologicamente justificável e eticamente adequado utilizar. Existe um custo de performance associado a isso, e os modelos ficaram sem acesso a informações que poderiam ser relevantes para o mérito dos casos. Mas esse custo foi assumido de forma consciente.

7.4 Risco de ciclos de retroalimentação com dados das partes

A decisão de não incorporar dados identificáveis das partes, discutida acima, se sustenta não apenas por razões de finalidade e conformidade legal, mas também por um risco de que esse tipo de variável crie exatamente as condições para um ciclo de retroalimentação (*feedback loop*), mecanismo apresentado na Fundamentação Teórica (Seção 3.3.1). É importante destacar como esse risco se manifestaria caso a decisão de arquitetura tomada neste trabalho fosse revertida.

Variáveis capazes de identificar ou perfilar empresas rés, como reincidência em litígios, setor de atividade ou histórico de acordos anteriores, tendem a ter forte poder preditivo, dado que capturam padrões de comportamento dessas empresas perante a Justiça do Trabalho. Entretanto, se um sistema baseado nesse tipo de variável fosse consultado por advogados antes do ajuizamento de uma ação, a decisão de litigar ou de aceitar um acordo passaria a ser influenciada pela reputação estatística da empresa ré, e não apenas pelo mérito do caso individual. Empresas identificadas como "boas pagadoras" tenderiam a atrair mais acordos consensuais, enquanto empresas com histórico de resistência processual poderiam ser evitadas por escritórios que preferem retornos mais previsíveis.

Esse comportamento alteraria o próprio processo gerador dos dados de treino futuros. Como os casos que chegam a julgamento deixariam de ser uma amostra representativa de todos os litígios potenciais contra aquela empresa, e passariam a refletir apenas os casos que advogados julgaram valer a pena levar adiante dado o perfil sinalizado pelo modelo, o retreinamento sobre essa base filtrada reforçaria o padrão original em vez de corrigi-lo. Diferentemente do viés geográfico e institucional discutido nas seções anteriores, que reflete práticas difusas de varas e comarcas, esse ciclo operaria sobre identidades específicas, seja de empresas, seja em contextos com menor anonimização, de trabalhadores, aproximando o sistema de uma forma de filtro individual que o Artigo 20 da LGPD e o Artigo 22 do GDPR, discutidos na Fundamentação Teórica, buscam restringir.

Esse risco é um dos argumentos para a decisão de manter o modelo restrito aos metadados anonimizados do DataJud. Um sistema que aprende sobre o contexto processual de tramitação, como o discutido nas seções anteriores, já impõe limites ao seu uso individualizado, entretanto, um sistema que aprendesse sobre a identidade das partes teria o potencial de institucionalizar, de forma automatizada e autorreforçada, vantagens e desvantagens específicas de empresas e trabalhadores.

7.5 O uso responsável de IA em domínios sensíveis

Nos últimos anos, projetos de aprendizado de máquina passaram a ser aplicados com rapidez crescente em domínios de alta consequência como saúde, crédito, segurança pública, justiça. Mas esse ritmo criou um problema estrutural, dado que a capacidade de construir e implantar sistemas cresceu muito mais rápido do que a capacidade de governá-los. Modelos são lançados, escalam e passam a influenciar decisões antes que alguém tenha verificado o que eles aprenderam de fato.

Os resultados deste trabalho mostram como isso pode acontecer de forma não intencional. Os modelos têm desempenho estatístico e se fossem avaliados apenas por essas métricas,

seriam implantados. Mas quando se abre o modelo e se examina o que dirige as predições, o que aparece é que o sistema está operando com base em padrões geográficos e institucionais, não em critérios jurídicos. Um sistema implantado sem essa análise seria uma caixa preta funcional pois acertaria com frequência razoável, mas por razões que ninguém teria verificado e que poderiam reproduzir exatamente as assimetrias que o sistema de justiça deveria corrigir.

Esse é o argumento central para tratar explicabilidade como requisito de projeto e não como etapa opcional. Em domínios sensíveis, é necessário compreender de forma extensa porque o modelo acerta e quais dados contribuem para suas decisões. Um modelo que discrimina bem entre acordos e não-acordos com base no município de origem pode ser estatisticamente útil mas pode ter inúmeros problemas de governança e ética. A diferença entre os dois não aparece nas métricas de performance, aparece apenas com análise de explicabilidade e consciência dos dados que estão sendo adicionados ao modelo.

A governança também precisa acompanhar o ciclo de vida do sistema depois da implantação. Modelos mudam de comportamento à medida que os dados mudam, e sem monitoramento periódico esse desvio passa despercebido. No contexto judiciário, isso tem consequências práticas, dado que padrões aprendidos em um período de pandemia ou de transição legislativa podem não refletir mais o comportamento atual do sistema, e um modelo desatualizado continuando a gerar predições é mais perigoso do que nenhum modelo.

8 Conclusão

Este trabalho desenvolveu e avaliou um pipeline de aprendizado de máquina para predição de desfechos processuais no TRT-15, construído inteiramente sobre dados públicos e anonimizados do DataJud. Os dois classificadores implementados alcançaram desempenho estatístico consistente entre os conjuntos de validação e teste: o Modelo 1 (LightGBM) obteve AUC-ROC de 0,7378 na predição de acordos judiciais, e o Modelo 2 (CatBoost) atingiu F1-macro de 0,4089 na discriminação entre os três desfechos de mérito, com desempenho particularmente limitado sobre a classe minoritária de extinção sem resolução de mérito. A estabilidade entre validação e teste, obtida por meio de uma separação estritamente cronológica dos dados, indica que os padrões aprendidos generalizam, ao menos dentro do próprio tribunal e período analisado.

Assim como os resultados obtidos, os resultados da análise da explicabilidade sobre a origem desse desempenho também merecem relevância. Em ambos os modelos, as variáveis de maior peso preditivo, o município de origem, o código da vara trabalhista e o assunto processual, descrevem o contexto geográfico e institucional que estão inseridos, e não o mérito individual de cada ação. Isso mostra que os modelos aprendem mais sobre as práticas históricas de cada unidade decisória do que sobre os fundamentos jurídicos dos casos individuais, o que impõe limites claros ao seu uso para avaliação de forma individual das partes do processo.

Esses resultados reforçam o argumento de que desempenho estatístico não necessariamente acompanha adequação ética, e que a explicabilidade não deveria ser tratada como uma etapa posterior ao desenvolvimento de um modelo, mas sim como parte integrante

de seu projeto desde a escolha das variáveis até a definição das métricas de avaliação. As decisões tomadas ao longo deste trabalho, como a restrição a dados anonimizados publicamente disponíveis e a rejeição deliberada de fontes que permitiriam reidentificar as partes, refletem essa postura, ainda que ao custo de alguma perda de poder preditivo.

As limitações discutidas ao longo do texto, sobretudo a restrição espacial ao TRT-15, a delimitação temporal entre 2018 e 2023 e o escopo circunscrito ao primeiro grau de jurisdição, indicam caminhos naturais para trabalhos futuros. Entre eles, destacam-se a extensão da metodologia para outros tribunais regionais, permitindo avaliar se o padrão de dominância geográfica observado se repete em contextos distintos; a incorporação de mecanismos de monitoramento contínuo do modelo após uma eventual implantação, capazes de detectar desvios de comportamento ao longo do tempo; e a investigação de técnicas de mitigação de viés que atuem diretamente sobre as variáveis geográficas e institucionais identificadas como dominantes, sem comprometer a utilidade estatística dos modelos.

Em síntese, este trabalho mostra que é possível construir modelos preditivos tecnicamente competentes para o domínio judiciário trabalhista, mas que essa competência técnica, isoladamente, não garante um sistema adequado para uso em produção. A combinação entre modelagem supervisionada e auditoria de explicabilidade permitiu não apenas avaliar a performance dos classificadores, mas também compreender os mecanismos por trás de suas predições, revelando um padrão de dependência estrutural que, sem essa análise, permaneceria invisível.

Agradecimentos

O desenvolvimento deste trabalho, assim como toda a minha jornada na universidade, não seria possível sem pessoas-chave que estiveram ao meu lado ao longo desse caminho.

Agradeço primeiramente aos meus pais, Solange e Edson, que são exemplos diários para mim de pessoas e profissionais e que me inspiraram a seguir o caminho da computação. Ao longo de toda a minha vida, me proporcionaram um ambiente extremamente propício para o meu desenvolvimento, me dando todas as ferramentas necessárias para que eu pudesse perseguir os meus sonhos. Essa conquista passa diretamente pelas mãos de vocês.

À minha irmã, Julia, agradeço por ter estado ao meu lado em cada fase da minha jornada na vida e na universidade, me apoiando em cada desafio e me transmitindo confiança e inspiração, acreditando em mim quando muitas vezes eu não acreditava. Aos olhos dela, nada era impossível para mim.

Agradeço também à minha tia e madrinha, Cilene, que me proporcionou muito acolhimento e suporte, e que, durante minha jornada universitária, me deu estabilidade, segurança e apoio em todos os momentos, me auxiliando de todas as formas possíveis.

Estendo esse agradecimento a toda a minha família que permaneceu ao meu lado ao longo desses anos e me apoiou nos momentos difíceis, especialmente meus avós, Valmiria e Francisco, minha prima, Sofia, e meu tio, Denis.

Aos amigos que fiz durante o período de faculdade, sem vocês essa conquista não seria possível. Raphael, Pedro, Luigi, Clara, Isabella, Luisa e Tiago foram pessoas extremamente importantes para minha vivência acadêmica: nas madrugadas de estudo, nas provas e nos

trabalhos, vocês estiveram ao meu lado, me apoiando em todos os momentos. Tenho certeza de que fiz amigos para a vida. Agradeço também às minhas amigas, Luiza Ferreira, Luiza Giffoni, Luana, Letícia, Marina e Giulia, que, mesmo fora da computação, me apoiaram em todos os momentos.

Agradeço aos meus colegas de trabalho e aos meus gestores, que nutriram em mim a paixão pela área de ciência de dados e que me deram suporte em cada momento da conciliação entre a universidade e o trabalho. Agradeço especialmente ao Josival e ao Jarbas, que me apoiaram ao longo de toda a minha jornada e nunca me fizeram despriorizar meus estudos e minha educação.

Ao meu orientador, Emanuel, agradeço por ter acreditado na ideia deste trabalho e por ter me direcionado e apoiado ao longo de todo o seu desenvolvimento. Sua orientação foi essencial. Agradeço também pelas ótimas aulas de Banco de Dados, extremamente importantes para minha formação e vida profissional.

Por fim, agradeço ao Instituto de Computação e à UNICAMP por terem sido um lugar de crescimento, acolhimento e descobertas ao longo desses cinco anos, que certamente formaram a pessoa que sou hoje.

Declaração de uso de Ferramentas de Inteligência Artificial

Os autores declaram que ferramentas de inteligência artificial generativa, especificamente os modelos de linguagem Claude e Gemini, foram utilizadas como apoio ao longo do desenvolvimento deste trabalho. Essas ferramentas serviram de apoio em pesquisas, auxiliando na consulta e organização de referências e conceitos técnicos relacionados aos algoritmos de aprendizado de máquina utilizados, à legislação aplicável e ao funcionamento do sistema judiciário trabalhista, sempre posteriormente verificados em sua integridade nas fontes primárias. Também foram usadas como apoio no desenvolvimento de código, principalmente para coleta e integração de dados públicos, como na obtenção de indicadores socioeconômicos do IBGE e seu cruzamento com a base processual do DataJud. E, por fim, serviram de apoio na revisão da escrita, principalmente quanto à adequação às normas de formatação acadêmica, clareza da redação e consistência terminológica ao longo do texto.

Os autores reforçam que todo o conteúdo presente neste trabalho são de sua inteira responsabilidade.

Referências

[1] Takuya Akiba et al. "Optuna: A Next-generation Hyperparameter Optimization Framework". Em: *Proceedings of the 25th ACM SIGKDD*. 2019.

[2] Tianqi Chen e Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". Em: *Proceedings of the 22nd ACM SIGKDD*. 2016.

[3] Conselho Nacional de Justiça. *DataJud: Base Nacional de Dados do Poder Judiciário*. `https://www.cnj.jus.br/sistemas/datajud/`. Acesso em: 10 03 2026. 2020.

[4] Bryce Goodman e Seth Flaxman. "European Union regulations on algorithmic decision-making and a "right to explanation"". Em: *AI Magazine* 38.3 (2017).

[5] Instituto Brasileiro de Geografia e Estatística. *IBGE: Instituto Brasileiro de Geografia e Estatística*. `https://www.ibge.gov.br`. Acesso em: 02 04 2026. 2024.

[6] Guolin Ke et al. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree". Em: *Advances in Neural Information Processing Systems*. 2017.

[7] Scott M Lundberg et al. "Consistent Individualized Feature Attribution for Tree Ensembles". Em: *arXiv preprint arXiv:1802.03888* (2018).

[8] Scott M Lundberg e Su-In Lee. "A Unified Approach to Interpreting Model Predictions". Em: *Advances in Neural Information Processing Systems*. 2017.

[9] Liudmila Prokhorenkova et al. "CatBoost: unbiased boosting with categorical features". Em: *Advances in Neural Information Processing Systems*. 2018.

[10] Lloyd S Shapley. "A Value for N-Person Games". Em: *Contributions to the Theory of Games* 2 (1953).