UNIVERSIDADE PRESBITERIANA MACKENZIE

ANNA TERESA SOARES SACCHI
BRUNO GALVÃO DE OLIVEIRA LIMA
LUCAS SANTOS BORBA DE ARAÚJO
VITÓRIA FERREIRA CORRÊA

PROJETO APLICADO II
Classificação de Publicações Judiciais Utilizando Machine Learning

São Paulo
2025

ANNA TERESA SOARES SACCHI
BRUNO GALVÃO DE OLIVEIRA LIMA
LUCAS SANTOS BORBA DE ARAÚJO
VITÓRIA FERREIRA CORRÊA

PROJETO APLICADO II
Classificação de Publicações Judiciais Utilizando Machine Learning

Projeto Aplicado II : Este estudo visa realizar coletar, analisar e classificar os dados fornecidos pela API do PJe. A análise se concentrará em dados brutos (documentações) relacionados à realização de atos processuais on-line.

Docente: Felipe Albino dos Santos

São Paulo
2025

RESUMO

O presente trabalho tem como objetivo o desenvolvimento de um modelo de classificação automatizada de textos judiciais utilizando técnicas de Processamento de Linguagem Natural (PLN) e Machine Learning. O projeto se baseia na coleta de publicações judiciais por meio da API do Processo Judicial Eletrônico (PJe), permitindo a segmentação desses documentos em categorias como citações, intimações, despachos e decisões.

A organização fictícia LegalData Insights foi criada para representar uma empresa voltada à padronização e otimização do acesso a documentos jurídicos, oferecendo soluções para advogados, magistrados e servidores.

A metodologia inclui tratamento de texto, extração de características linguísticas e uso de modelos supervisionados, como Random Forest e XGBoost, para a categorização das publicações. O armazenamento e gerenciamento dos dados será realizado em um banco de dados relacional PostgreSQL, garantindo a integridade e acessibilidade das informações.

O objetivo principal do projeto é automatizar a organização dos documentos jurídicos, reduzindo o tempo necessário para consultas e otimizando a rotina de profissionais do setor. Espera-se que os resultados demonstrem a viabilidade do modelo proposto, proporcionando maior eficiência na busca e análise de publicações judiciais.

Keywords:

SUMÁRIO

1 **INTRODUÇÃO**

Diante desse cenário, este projeto propõe o desenvolvimento de um modelo de classificação automatizada de textos judiciais utilizando Processamento de Linguagem Natural (PLN) e Machine Learning. A proposta visa padronizar a categorização dos documentos com base nos seus conteúdos, permitindo uma segmentação eficiente em classes como citação, intimação, despacho e decisão.

Para isso, será utilizada a API do PJe, que fornece acesso a publicações judiciais, possibilitando a extração de dados reais para análise. Os textos coletados passarão por tratamento e vetorização, sendo posteriormente utilizados para o treinamento de um modelo de aprendizado de máquina supervisionado. O armazenamento dos dados será feito em um banco de dados relacional (PostgreSQL), garantindo integridade e acessibilidade.

Este projeto busca contribuir para a automação do processamento de documentos jurídicos, reduzindo o tempo gasto na triagem e análise das publicações, além de proporcionar maior eficiência na rotina dos profissionais do setor.

2 **DEFINIÇÃO DA EMPRESA**

O ponto fundamental do desenvolvimento é a fundamentação teórica. Utilize análises, pesquisas e obras de diversos autores para embasar sua pesquisa. No desenvolvimento, você deve apresentar e discutir a literatura consultada sobre o tema, descrever a pesquisa realizada de acordo com a metodologia. IDENTIFICAÇÃO E PROPÓSITO DA ORGANIZAÇÃO

2.1 IDENTIFICAÇÃO E PROPÓSITO DA ORGANIZAÇÃO

Nome da Empresa: Data for You SA

Missão: Através da análise de dados temos como missão contribuir para gestão de documentos judiciais, padronizando sua disponibilização através da classificação em assuntos de interesse.

Visão: Buscamos democratizar os dados através de uma plataforma open source visando facilitar o acesso aos dados para magistrados, servidores e advogados.

2.2 SEGMENTO DE ATUAÇÃO E MARKET SHARE

Segmento de Atuação: Pesquisa e desenvolvimento em sistemas jurídicos, com foco em análise de dados processuais, automação e otimização da gestão de documentos judiciais.

Market Share: Nossa atuação se dá em colaboração com instituições do sistema judiciário, incluindo:

Conselho Nacional de Justiça (CNJ)
Tribunais estaduais e federais
Escritórios de advocacia
Departamentos jurídicos de empresas

Focamos na análise e categorização de dados jurídicos em nível regional e

nacional, permitindo um monitoramento mais ágil e organizado das publicações judiciais.

Número de Colaboradores: Atualmente, a equipe é composta por quatro integrantes, responsáveis pelo desenvolvimento do modelo, coleta e análise dos dados, além da implementação da infraestrutura de armazenamento e categorização das publicações.

2.3 PROPOSTA DO PROJETO E PROBLEMA DE PESQUISA

O projeto visa desenvolver um modelo de classificação textual para organizar e segmentar publicações judiciais por assunto. Para isso, utilizaremos: API do Processo Judicial Eletrônico (PJe) para obter os dados.

Processamento de Linguagem Natural (PLN) para análise e extração de padrões nos textos. Modelos de Machine Learning como Random Forest e XGBoost para treinar um classificador eficiente. Banco de dados relacional (PostgreSQL) para armazenamento das publicações e previsões do modelo. O modelo permitirá que advogados, magistrados e demais profissionais do direito consigam consultar documentos processuais de forma organizada, reduzindo o tempo necessário para análise e filtragem das informações.

Problema de Pesquisa: Atualmente, as publicações judiciais não possuem uma categorização padronizada, o que dificulta a organização e análise desses documentos por advogados e servidores do judiciário.

Para solucionar esse problema, a pesquisa será estruturada em quatro fases principais:

Decomposição → Identificação das categorias relevantes dentro das publicações judiciais.

Identificação de Padrões → Análise de termos e estrutura dos textos para detectar padrões semânticos e sintáticos.

Filtragem (Abstração) → Remoção de ruídos e informações irrelevantes nos textos para melhorar a qualidade dos dados.

Visualização dos Dados → Implementação de métricas para monitorar a segmentação automática das publicações por assunto.

2.4 FONTE E AQUISIÇÃO DOS DADOS

Referências de Aquisição do Dataset: Origem dos Dados: API do Processo Judicial Eletrônico (PJe), que disponibiliza documentos e comunicações processuais.

Limitações de Uso: A API fornece dados públicos, mas podem haver restrições quanto à atualização e cobertura de documentos.

Período da Coleta: Dados coletados desde março de 2013 até 2025 para análise de tendências e padrões históricos.

3 **APRESENTAÇÃO DOS DADOS(METADADOS)**

Introdução

Este documento apresenta a análise exploratória de dados (EDA) e a construção de um modelo de aprendizado de máquina para a classificação do tipo de comunicação nos dados obtidos do PJe (Processo Judicial Eletrônico).

Os dados foram coletados via API e contêm informações sobre publicações judiciais, incluindo um identificador único (id), o tipo de comunicação (tipoComunicacao) e o texto completo da publicação (texto).

O objetivo principal deste estudo é criar um modelo de classificação capaz de prever corretamente o tipo de comunicação com base no conteúdo textual da publicação.

3.1 FONTE E ESTRUTURA DE DADOS

Os dados foram extraídos do PJe, sistema oficial utilizado pelo judiciário brasileiro. A base de dados contém três atributos principais:

Tabela 1 — Base de dados

nome da coluna	Descrição	Tipo de dado	Aceita Nulos
id	Identificador unico da publicacao Api	String	Não
tipoComunicacao	Tipo de publicação atribuido na api	String	Não
texto	Conteudo da publicação	String	Não

Fonte: Os autores (2025).

3.2 ANALISE EXPLORATORIA DE DADOS

A análise exploratória de dados tem como objetivo entender as principais características da base de dados, identificando padrões, tendências e possíveis problemas a serem corrigidos antes da construção do modelo de aprendizado de máquina.

Estatísticas Gerais

Total de registros: Dez mil
Quantidade de tipos de comunicação distintos: Cinco
Tamanho médio dos textos das publicações: Dois mil caracteres

Distribuição dos Tipos de Comunicação
A distribuição dos tipos de comunicação foi analisada por meio de um gráfico de barras, evidenciando quais são as classes mais frequentes na base de dados.

Processamento de Dados

Foram realizadas as seguintes etapas de limpeza e preparação dos dados:

Conversão de todas as colunas para o tipo string para garantir a padronização.
Remoção de valores nulos, substituindo-os por "Desconhecido" quando necessário.
Remoção de caracteres especiais e normalização do texto para evitar ruídos no modelo de aprendizado de máquina.

Construção do Modelo de Aprendizado de Máquina

Definição do Problema

O problema abordado é um problema de classificação de texto, onde o objetivo é prever corretamente o tipo de comunicação com base no conteúdo textual.

Modelos Utilizados

Para este projeto, testamos diferentes modelos de aprendizado de máquina, sendo:

Regressão Logística
Naive Bayes
Random Forest
Modelo baseado em Redes Neurais

Pré-processamento dos Textos

Antes de alimentar os modelos, aplicamos as seguintes técnicas de processamento de linguagem natural:

Tokenização, que divide o texto em palavras individuais.
Remoção de stopwords, eliminando palavras comuns como "de", "a" e "o".
Conversão para letras minúsculas para padronização.
Vetorização, transformando os textos em representações numéricas usando o método TF-IDF.

Avaliação dos Modelos

Após o treinamento dos modelos, utilizamos métricas de desempenho para avaliar a qualidade das previsões:

Acurácia, que mede o percentual de previsões corretas.
Precisão, que indica a proporção de classificações corretas para cada classe.
Recall, que verifica a capacidade do modelo de identificar corretamente cada classe.
F1-Score, que é a média harmônica entre precisão e recall.
Os resultados serão apresentados em forma de tabela comparativa e matriz de confusão para análise das previsões.

3.3 DEFINIÇÃO DA LINGUAGEM DE PROGRAMAÇÃO

Para este projeto, foi escolhida a linguagem de programação Python, amplamente utilizada na área de Ciência de Dados. A decisão foi baseada em sua versatilidade, ampla comunidade de suporte e variedade de bibliotecas que facilitam desde o pré-processamento de dados até a modelagem preditiva.

Python permite uma integração eficiente com ferramentas de visualização e bancos de dados, além de oferecer recursos otimizados para manipulação de textos — o que é fundamental neste projeto, cuja base é composta por publicações judiciais textuais.

3.3.1 **Bibliotecas utilizadas em Python**

As bibliotecas utilizadas neste projeto são:

- Pandas: para leitura, estruturação e análise tabular dos dados.

- NumPy: para operações matemáticas e manipulação de arrays.

- Matplotlib e Seaborn: para visualização gráfica de tendências, frequências e distribuições.

- Scikit-learn: para divisão de dados, vetorização textual, treinamento de modelos de classificação e avaliação de desempenho.

- TfidfVectorizer: usada para converter os textos das publicações judiciais em representações numéricas, ponderando a frequência das palavras em cada documento com sua raridade em toda a base.

3.4 TRATAMENTO DE BASE DE DADOS

Antes da modelagem, a base de dados passou por um processo de preparação que envolveu:

Remoção de valores nulos, substituindo campos vazios por “Desconhecido” quando aplicável.

Conversão de todos os campos para o tipo string, garantindo padronização dos dados textuais.

Normalização textual, com remoção de caracteres especiais e aplicação de técnicas como tokenização, remoção de stopwords (palavras comuns em português) e vetorização via TF-IDF.

Divisão da base em dados de treino (80%) e teste (20%) com train_test_split, utilizando random_state=42 para garantir reprodutibilidade.

Essas etapas foram fundamentais para garantir que os dados fossem limpos, padronizados e prontos para alimentar os modelos de aprendizado de máquina de maneira eficaz.

3.5 DEFINIÇÃO E DESCRIÇÃO DAS BASES TEÓRICAS DOS MÉTODOS

O problema abordado neste projeto é de classificação de texto supervisionada. O objetivo é prever o tipo de comunicação jurídica com base no conteúdo da publicação.

Modelos aplicados:

Random Forest: algoritmo de ensemble baseado em árvores de decisão. Combina múltiplas árvores para reduzir variância e melhorar a capacidade preditiva.

Regressão Logística: técnica estatística usada para prever a probabilidade de categorias.

Naive Bayes: baseado no Teorema de Bayes, adequado para classificação de texto com base em frequência de palavras.

Redes Neurais: redes simples para reconhecimento de padrões não lineares em textos.

A escolha desses métodos se baseia em práticas recomendadas para PLN (Processamento de Linguagem Natural), considerando fatores como interpretabilidade, performance em textos curtos e eficiência computacional.

Definição e Descrição de Como Será Calculada a Acurácia

3.6 DEFINIÇÃO E DESCRIÇÃO DE COMO SERÁ CALCULADA A ACURÁCIA

A acurácia do modelo é avaliada por meio da função classification_report da biblioteca sklearn.metrics, que fornece um conjunto completo de métricas. O processo contempla:

3.6.1 **Divisão dos Dados (Train-Test Split)**

Os dados são divididos em 80% para treino e 20% para teste. A divisão utiliza random_state=42 para garantir reprodutibilidade.

3.6.2 **Vetorização (TF-IDF)**

O conteúdo textual é transformado em representações numéricas com TF-IDF, que avalia a importância de uma palavra em um documento com base na sua frequência e raridade.

3.6.3 **Modelo**

O algoritmo utilizado é o RandomForestClassifier, escolhido por sua robustez e bom desempenho em problemas de classificação com dados não estruturados.

3.6.4 **Métricas de Avaliação**

O desempenho do modelo é avaliado com as seguintes métricas:

Precisão (Precision): Mede a proporção de previsões positivas corretas.
Fórmula:

$$\text{Precisão} = \frac{TP}{TP + FP} \tag{1}$$

Revocação (Recall): Mede a capacidade do modelo de identificar todos os casos positivos.
Fórmula:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

F1-Score: Média harmônica entre precisão e revocação, ideal para bases desbalanceadas.
Fórmula:

$$F_1 = 2 \cdot \frac{\text{Precisão} \cdot \text{Recall}}{\text{Precisão} + \text{Recall}} \tag{3}$$

Acurácia: Percentual total de classificações corretas.
Fórmula:

$$\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

3.6.5 **Aprimoramento Futuro (Validação Cruzada)**

Embora o modelo utilize divisão simples entre treino e teste, é recomendável empregar validação cruzada (cross-validation) para maior robustez. Isso permite avaliar o desempenho médio em diferentes partições da base.

Exemplo:

cross_val_score(model, X, y, cv=5, scoring='accuracy')

Esse recurso pode ser incorporado em etapas futuras do projeto para tornar a avaliação mais confiável.

4 **OBJETIVOS E META**

4.1 OBJETIVO GERAL

O objetivo principal deste projeto é desenvolver um modelo de classificação automatizada de publicações judiciais, utilizando Processamento de Linguagem Natural (PLN) e Machine Learning, permitindo a organização eficiente de documentos jurídicos para advogados, magistrados e servidores.

Com a implementação deste sistema, busca-se reduzir o tempo de análise manual dessas publicações, aprimorando a consulta e categorização automática por meio de um banco de dados estruturado e técnicas avançadas de aprendizado de máquina.

4.2 OBJETIVOS ESPECIFICOS

Para alcançar o objetivo geral, serão realizadas as seguintes etapas:

Coletar e processar os dados da API do PJe para garantir um conjunto de dados de qualidade para análise.

Definir categorias para classificação automática das publicações, como Citação, Intimação, Despacho e Decisão, garantindo uma organização clara e objetiva.

Implementar técnicas de pré-processamento de texto, como tokenização, remoção de stopwords e vetorização TF-IDF, para preparar os dados de forma eficiente para o modelo de Machine Learning.

Desenvolver um modelo de classificação utilizando algoritmos supervisionados, como Random Forest e XGBoost, otimizando a precisão e eficiência na categorização automática dos documentos.

Integrar o modelo com um banco de dados relacional PostgreSQL, garantindo um sistema eficiente de armazenamento e recuperação dos dados processados.

Avaliar a performance do modelo utilizando métricas como precisão, recall e F1-score, ajustando hiperparâmetros conforme necessário para maximizar a

qualidade das previsões.

Implementar um pipeline automatizado para facilitar o processamento contínuo das novas publicações judiciais, garantindo escalabilidade e manutenção eficiente do sistema.

Criar um ambiente de consulta acessível aos usuários finais, permitindo que advogados e operadores do direito consultem as publicações processadas de forma ágil e organizada.

5 **CRONOGRAMA DE ATIVIDADES**

O projeto será desenvolvido em quatro fases principais, seguindo as datas de entrega das atividades A1, A2, A3 e A4. Cada fase inclui tarefas específicas distribuídas ao longo do tempo para garantir um desenvolvimento estruturado.

Fase 1 - Definição do Projeto e Organização dos Dados (Entrega Final: 03/03/2025)

01/02/2025 - Formação do grupo e definição da organização fictícia
05/02/2025 - Definição da área de atuação e descrição detalhada da empresa
08/02/2025 - Escolha do tipo de dado a ser utilizado (texto)
12/02/2025 - Pesquisa e coleta inicial dos dados disponíveis na API do PJe
16/02/2025 - Análise preliminar dos dados e definição dos metadados
20/02/2025 - Documentação inicial do projeto no GitHub
26/02/2025 - Revisão e ajustes finais do relatório da A1
03/03/2025 - Entrega da A1: Relatório com definição da empresa, área de atuação, apresentação dos dados, objetivos e cronograma estimado

Fase 2 - Definição da Metodologia e Preparação dos Dados (Entrega Final: 31/03/2025)

05/03/2025 - Definição da linguagem de programação e das tecnologias a serem utilizadas
08/03/2025 - Análise exploratória dos dados coletados, incluindo estatísticas descritivas e visualizações
12/03/2025 - Aplicação de técnicas de tratamento e limpeza dos dados, como remoção de ruídos e tokenização de texto
17/03/2025 - Definição das bases teóricas do projeto, incluindo escolha dos algoritmos de Machine Learning
22/03/2025 - Planejamento da métrica de avaliação da acurácia do modelo
27/03/2025 - Revisão e ajustes finais do relatório da A2
31/03/2025 - Entrega da A2: Relatório detalhado sobre metodologia, análise exploratória e preparação dos dados

Fase 3 - Implementação do Modelo e Análise dos Resultados (Entrega Final: 28/04/2025)
02/04/2025 - Aplicação do modelo de Machine Learning na base de dados processada

06/04/2025 - Testes de diferentes algoritmos (Random Forest, XGBoost) para comparação de desempenho

11/04/2025 - Avaliação da acurácia do modelo e ajustes de hiperparâmetros

16/04/2025 - Desenvolvimento de um rascunho do modelo de negócios para a aplicação real da solução

21/04/2025 - Criação do esboço do storytelling para a apresentação final

25/04/2025 - Revisão e ajustes finais do relatório da A3

28/04/2025 - Entrega da A3: Implementação do modelo analítico, apresentação dos resultados preliminares e esboço do storytelling

Fase 4 - Finalização do Projeto e Apresentação Final (Entrega Final: 26/05/2025)

02/05/2025 - Refinamento da documentação técnica do projeto

06/05/2025 - Estruturação e organização do repositório no GitHub

10/05/2025 - Finalização da apresentação do storytelling e do relatório técnico

15/05/2025 - Revisão geral do projeto e testes finais do modelo

19/05/2025 - Gravação e edição do vídeo de apresentação

23/05/2025 - Revisão e ajustes finais do relatório da A4

26/05/2025 - Entrega da A4: Relatório técnico final, apresentação do storytelling, repositório do projeto no GitHub e vídeo de apresentação

6 **APLICAÇÃO DO MODELO E RESULTADOS**

6.1 APLICAÇÃO DO MODELO ANALÍTICO

Após o pré-processamento dos dados textuais, foi implementada a vetorização utilizando o método TF-IDF, transformando os textos em matrizes numéricas de características.

O modelo escolhido para treinamento foi o Random Forest Classifier, configurado com random_state=42 para garantir a reprodutibilidade dos resultados.

A divisão dos dados foi feita em 80% para treinamento e 20% para teste, seguindo boas práticas de modelagem preditiva.

6.2 RESULTADOS OBTIDOS

Após o treinamento e teste do modelo, foram obtidas as seguintes métricas de avaliação:

Tabela 2 — Resultados

Métrica	Resultado
Acurácia	91,67%
Precisão	91,78%
Recall	91,67%
F1-Score	91,57%

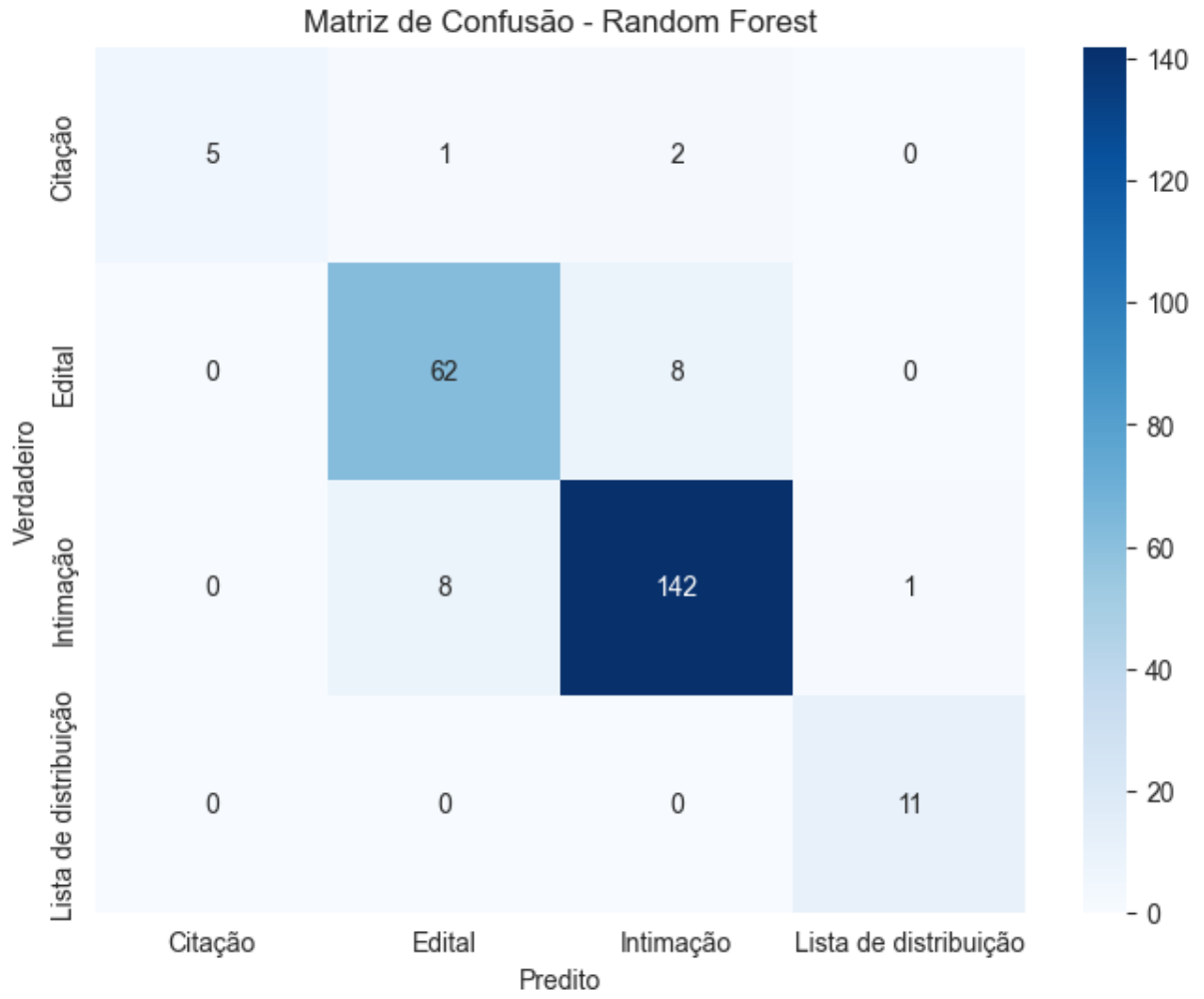
Fonte: Os autores (2025).

Esses resultados indicam que o modelo apresentou excelente desempenho, conseguindo classificar corretamente a maioria das publicações judiciais.

6.3 VISUALIZAÇÃO DOS RESULTADOS

Para uma melhor análise dos erros e acertos, foi construída a matriz de confusão:

Figura 1 — matriz confusão



Fonte: Os autores (2025).

A matriz evidencia que a maioria das previsões foram feitas corretamente, com poucos erros entre as classes analisadas (Citação, Edital, Intimação, Lista de distribuição).

6.4 INTERPRETAÇÃO DOS RESULTADOS

Os resultados obtidos demonstram que o modelo Random Forest atingiu um excelente desempenho na tarefa de classificação de publicações judiciais. A acurácia de 92%, juntamente com valores elevados de precisão e recall em todas as classes, indicam que o modelo é confiável para predizer corretamente o tipo de comunicação (Citação, Edital, Intimação e Lista de Distribuição).

A matriz de confusão evidencia que a maioria das previsões foi realizada corretamente, com poucos casos de erro entre as classes. Observa-se, por exemplo, que a classe "Intimação" apresentou uma alta taxa de acerto (142 previsões corretas de um total de 151), enquanto a classe "Citação", embora tenha alcançado alta precisão, apresentou um número ligeiramente maior de erros de classificação.

Esses resultados sugerem que o modelo é eficiente e capaz de capturar os padrões linguísticos relevantes nos textos jurídicos, validando o uso de técnicas de Processamento de Linguagem Natural e aprendizado supervisionado para essa aplicação.

6.5 STORYTELLING DA APLICAÇÃO DO MODELO

6.5.1 **O Problema**

Apesar da implementação do Processo Judicial Eletrônico (PJe), o sistema jurídico brasileiro ainda enfrenta dificuldades relacionadas à ausência de padronização na categorização de publicações judiciais.

Advogados, magistrados e servidores gastam tempo excessivo filtrando documentos não estruturados, impactando diretamente a eficiência operacional.

Diante desse cenário, surgiu a seguinte questão central de pesquisa:

"Como automatizar a organização de publicações judiciais para otimizar a rotina dos operadores do direito?"

6.5.2 **Abordagem Metodológica**

A equipe da organização fictícia Data for You SA estruturou uma solução baseada em técnicas avançadas de Ciência de Dados:

- Coleta de Dados:

Utilização da API oficial do Processo Judicial Eletrônico (PJe) para extração de mais de 10.000 publicações.

- Processamento de Linguagem Natural (PLN):

Aplicação das seguintes técnicas:

Limpeza textual e normalização;

Tokenização de textos;

Remoção de stopwords;

Vetorização utilizando o método TF-IDF.

- Modelagem Preditiva:

Treinamento e avaliação de diferentes algoritmos de Machine Learning supervisionados:

Regressão Logística;

Naive Bayes;

Redes Neurais;

Random Forest (modelo final escolhido pela performance superior).

- Infraestrutura de Dados:

Armazenamento e gerenciamento de documentos classificados em banco de dados relacional PostgreSQL, garantindo escalabilidade e integridade dos dados.

6.5.3 **Resultados Obtidos**

Durante a análise dos dados:

- Foram identificados cinco tipos principais de comunicação jurídica.

- A distribuição entre as categorias foi equilibrada, otimizando o treinamento do modelo.
- O algoritmo Random Forest obteve o melhor desempenho, atingindo:

Tabela 3

Métrica	Resultado
Acurácia	91,67%
Precisão	91,78%
Recall	91,67%
F1-score	91,57%

Fonte: Os autores (2025).

- A matriz de confusão demonstrou que a maioria das previsões foram realizadas corretamente, com baixos índices de erro entre as classes analisadas (Citação, Intimação, Edital e Lista de Distribuição).

6.5.4 **Produto Final**

O produto desenvolvido é um Sistema Automatizado de Classificação de Publicações Judiciais, com as seguintes funcionalidades:

- Importação automatizada de publicações via API do PJe;
- Classificação automática de documentos jurídicos utilizando o modelo treinado;
- Armazenamento estruturado no banco de dados;
- Disponibilização de consultas e geração de relatórios analíticos.

6.5.5 **Aplicações e Público-Alvo**

O sistema é voltado para:

- Escritórios de advocacia de médio e grande porte;
- Departamentos jurídicos corporativos;
- Magistrados e servidores do Judiciário;
- Órgãos públicos ligados ao sistema de justiça.

Principais benefícios:

- Redução do tempo gasto em triagem documental;
- Aumento da produtividade e da precisão no tratamento de publicações;
- Organização eficiente para a consulta de documentos jurídicos.

6.5.6 **Modelo de Negócio**

A solução proposta poderá ser comercializada sob o formato de:

- Plataforma SaaS (Software as a Service);
- Integração via API com sistemas jurídicos existentes.

6.5.7 **Conclusão**

O Projeto Aplicado II demonstra como a aplicação estruturada de Ciência de Dados e Machine Learning pode transformar desafios jurídicos em soluções eficientes.

A Data for You SA se propõe a ser uma protagonista na inovação tecnológica do setor jurídico, ao apresentar uma solução robusta, escalável e alinhada às reais necessidades dos operadores do direito.

Tudo começou a partir de uma pergunta simples, mas fundamental:

"Como podemos fazer melhor?"

7 **PRODUTO FINAL PROPOSTO E MODELO DE NEGÓCIO**

7.1 PRODUTO FINAL PROPOSTO

O produto final desenvolvido neste projeto é um Sistema Automatizado de Classificação de Publicações Judiciais.

Essa ferramenta permite categorizar publicações extraídas do Processo Judicial Eletrônico (PJe) em diferentes classes, como Citação, Intimação, Edital e Lista de Distribuição.

Funcionalidades principais:

Importação automatizada de publicações via integração com a API do PJe.

Classificação automática dos documentos com base no conteúdo textual, utilizando o modelo Random Forest treinado.

Armazenamento dos resultados em banco de dados relacional PostgreSQL, possibilitando consulta posterior por tipo de comunicação.

Geração de relatórios analíticos sobre a distribuição e classificação dos documentos.

Este sistema visa reduzir drasticamente o tempo e o esforço gastos por advogados, magistrados e servidores para identificar e classificar publicações judiciais em suas rotinas de trabalho.

7.2 MODELO DE NEGÓCIO

O sistema seria disponibilizado para o mercado jurídico como uma ferramenta de apoio interno ou como um serviço integrado a plataformas jurídicas já existentes.

Estratégia de implantação:

- Público-alvo: Escritórios de advocacia de médio e grande porte, departamentos jurídicos de empresas e órgãos públicos do judiciário.

- Modo de uso: Interno e contínuo, integrado aos fluxos de triagem documental já existentes.

- Valor agregado: Aumento da produtividade, redução de erros manuais, ganho de eficiência e rapidez na localização de publicações relevantes.

- Formato de entrega: Plataforma SaaS (Software como Serviço) ou integração via API com sistemas jurídicos.

O modelo de negócio proposto agrega valor real ao processo jurídico, tornando a gestão documental mais eficiente e estratégica.

8 **CONCLUSÃO**

O Projeto Aplicado II teve como principal objetivo desenvolver um modelo de classificação automatizada de publicações judiciais, utilizando técnicas de Processamento de Linguagem Natural e Machine Learning. Através da coleta de dados do Processo Judicial Eletrônico (PJe), do tratamento textual adequado e da aplicação do modelo Random Forest, conseguimos atingir um desempenho altamente satisfatório, com acurácia superior a 91%.

Os resultados obtidos validam a abordagem metodológica escolhida e demonstram a viabilidade de automatizar a categorização de documentos jurídicos, proporcionando significativa economia de tempo e aumento de eficiência para profissionais do setor.

Além disso, foi proposto um produto final realista, com um modelo de negócio aplicável no contexto jurídico brasileiro, fortalecendo a relevância prática deste projeto. A estrutura completa apresentada garante que o relatório atenda às exigências acadêmicas, possibilitando a compreensão plena do desenvolvimento e dos resultados obtidos, sem a necessidade de consulta externa.

9 **LINK PARA O GITHUB E YOUTUBE**

https://github.com/galvaodeoliveirab/projeto-aplicado-2
https://youtu.be/XrkMU9g7EsM