UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências exatas e de tecnologia

Departamento de Computação

Trabalho de conclusão de curso - TCC

Jonathan Santos Silva

Classificação de decisões judiciais sobre dados presentes no diário eletrônico da justiça do trabalho (DEJT)

São Carlos - São Paulo

2023

Jonathan Santos Silva

Classificação de decisões judiciais sobre dados presentes no diário eletrônico da justiça do trabalho (DEJT)

Trabalho de Conclusão de Curso apresentado ao Departamento de Computação como parte dos requisitos para a conclusão da graduação em Egenharia de Computação.

Orientação Profa. Dra. Marcela Xavier

São Carlos - São Paulo

2023

DEDICO

Agradecimentos

Agradeço a meu pai, Manoel Jose da Silva Filho, que nunca mediu esforços para me ensinar sobre o mundo e seus males, obrigado pelo afinco a nossa família. A minha mãe, Ivete Campos dos Santos Silva, que de seu ventre pode dar a vida a mim e a meu irmão, mulher essa que abdicou, também, das virtudes e faculdades da vida, para dedicação exclusiva da família e do lar. Sempre mantive por toda essa trajetória na universidade, os ensinamentos por vocês passados. Tenho orgulho de hoje poder de dedicar esse trabalho e momento a vocês.

"Somos o que fazemos, mas somos, principalmente, o que fazemos para mudar o que somos.

(Eduardo Galeano)

Resumo

A capacidade de classificar de maneira eficiente as decisões jurídicas como aprovadas ou rejeitadas é fundamental para garantir um sistema de justiça justo e eficaz. Neste estudo, apresentamos uma solução para essa tarefa, utilizando técnicas de aprendizado de máquina. A solução proposta envolve a estruturação dos dados, a identificação das unidades, a rotulagem por especialistas e o treinamento de um modelo de aprendizado de máquina. O estudo incluiu uma análise exploratória dos dados fonte e técnicas de pré-processamento de texto para limpeza e normalização dos dados. O modelo proposto alcançou uma taxa de acurácia alta de 96%. Por fim, validamos o modelo utilizando um conjunto de dados externo e casos reais. Os resultados sugerem que o modelo tem potencial para ser uma solução efetiva para classificar decisões jurídicas de forma precisa e rápida.

Palavras-chave: Jurimetria, Documentos Legais, Categorização Textual, Processamento de linguagem Natural (NLP).

Abstract

The ability to efficiently classify legal decisions as approved or rejected is crucial to ensure a fair and effective justice system. In this study, we propose a solution for this task using machine learning techniques. The proposed solution involves data structuring, unit identification, expert labeling, and training of a machine learning model. The study includes an exploratory analysis of the source data and text pre-processing techniques for data cleaning and normalization. The proposed model achieved a high accuracy rate of 92%. Finally, we validated the proposed model using an external dataset and real cases. The results suggest that the proposed model has the potential to be an effective solution for accurately and quickly classifying legal decisions.

Keywords: Jurimetrics, Legal Documents, Text Categorization, Natural Language Processing (NLP).

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

O uso de Inteligência artificial vem beneficiando várias indústrias, em especial a do campo legal/jurídico. Trabalhos na área vem ganhando relevância por apresentarem soluções voltadas a análise de dados não estruturados, análise jurídica e aplicações que auxiliam profissionais da área, na elaboração, resumo e redação de documentos (YU; ALÌ, 2019).

No Brasil, como em outros países, essas tarefas voltadas a redução de trabalho manual, são desafios difíceis de se alcançar, já que analisar e parametrizar informações públicas, como dados de sentenças judiciais, devido à falta de padronização nos documentos, uso de termos genéricos e falta de sintaxes definidas (SERRAS; FINGER, 2021), dificultam o trabalho de engenheiros e cientistas da computação.

Essas informações são extremamente importantes para a jurisprudência brasileira, pois advogados e juízes utilizam de decisões correntes para sustentar seus argumentos (KAUFFMAN; SOARES, 2020), já que compreender a extensão das decisões da corte brasileira é uma tarefa necessária para entrar com ações judiciais, apresentar defesas ou impor recursos. Portanto, possuir essas informações, significa ter objetos argumentativos e de estudo que estabelecem a orientação jurídica de um país.

Atualmente, essas tarefas são realizadas manualmente por especialistas da área, mas o alto volume e a falta de estruturação, as tornam inviáveis (YU; ALÌ, 2019; COLOMBO; BUCK; BEZERRA, 2017).

O tema em questão, que pode ser conferido ao título desse trabalho, se deu pela necessidade de advogados, que entraram em contato com a universidade (UFScar), para extrair informações de dados públicos referentes a decisões judiciais publicadas no diário oficial.

De acordo com conversas feitas, a maior fonte de valor vista nestas informações, está em sua parametrização, já que especialistas na área do direito, utilizam de decisões correntes da corte brasileira, para formular argumentos similares em suas peças.

Portanto, compreender quais são as determinações judiciais referente a casos semelhantes, é um instrumento importante para jurisprudência brasileira.

Em outras palavras, podemos resumir essas parametrizações em:

1. Identificar requisitos que determinam a aprovação de recursos judiciais;
2. Identificar quais decisões judiciais foram aprovadas e rejeitadas;
3. Agrupamento de setores/causas que possuem maior nível de aprovação;

4. Estruturar informações que envolvem. Número do processo, relator, partes envolvidas e conteúdo.

Deste modo, para elaboração desse trabalho, partiremos da validação da qualidade dos dados, onde nos deparamos com o desafio de transformar informações desestruturadas provindas de arquivos PDF, em dados que possamos estruturar de forma que banco de dados e modelos de aprendizado de máquina, possam ter melhor eficácia e partir disso submeter essas informações a modelos de classificação textual, já que dentre todas as parametrizações apresentadas, daremos foco, na resolução do segundo item(Identificar quais decisões judiciais foram aprovadas e rejeitadas).

A organização estrutural do trabalho pode ser vista como: A sessão 3, diz respeito as definições em que o trabalho está imerso, apresentando abstrações fundamentais que esclarecem as decisões tomadas e desenvolvidas no decorrer do projeto. É possível ver, também, uma seção focada em apresentar os trabalhos que formaram nossa base argumentativa para seguir com o tema em questão, esses trabalhos e sua importância podem ser encontrados na seção 3. A prática adotada, aplicando os conhecimentos da seção 3 podem ser vistos seção 5.3 e por fim a conclusão na seção 6

2 Objetivos

2.1 Objetivos Gerais

A finalidade geral deste trabalho é desenvolver e avaliar um modelo de classificação de decisões judiciais presentes no Diário Eletrônico da Justiça do Trabalho (DEJT), por meio de técnicas de aprendizado de máquina. Especificamente, comparar a eficácia de diferentes modelos de classificação em relação à precisão, revocação e medida F1. Além disso, o trabalho visa avaliar o desempenho do modelo desenvolvido em relação à rotulação por especialistas da área do direito. Por fim, o trabalho se insere em um projeto maior de jurimetria que busca reduzir a quantidade de trabalho manual que especialistas da área já executam em seu dia a dia, aumentando a eficiência do processo decisório e garantindo maior celeridade na resolução de conflitos trabalhistas.

2.2 Objetivos Especificos

Este trabalho tenta propor uma solução para o problema, conforme contextualizado na introdução 1, de identificar quais decisões judiciais foram aprovadas e rejeitadas, para isso, precisamos realizar alguns passos, como a estruturação dos dados publicados em arquivos PDF, identificação unitária dos processos para serem submetidos a um processo de rotulagem e por fim, submeter os dados rotulados por especialistas aos modelos de aprendizagem de máquina, esses passos podem ser resumidos nas seguintes etapas:

1. Realizar uma análise exploratória dos dados do DEJT, identificando suas principais características e desafios;
2. Selecionar e aplicar técnicas de pré-processamento de texto para a limpeza e normalização dos dados;
3. Escolher um modelo de aprendizado de máquina adequado para a tarefa de classificação de decisões, considerando suas características e limitações;
4. Treinar e avaliar o modelo desenvolvido utilizando um conjunto de dados de treinamento e teste representativo;
5. Comparar o desempenho do modelo proposto com outros modelos de classificação existentes na literatura;
6. Validar o modelo proposto por meio da aplicação em um conjunto de dados externo e de casos reais.

2.3 Resultados Esperados

É esperado que o modelo de aprendizagem de máquina desenvolvido seja capaz de identificar automaticamente, com base no texto da decisão, a sentença final proferida pelo juiz em relação ao processo. Entre os rótulos adotados, encontram-se Aprovado, Negado, Suspenso e Nenhum.

Para isso, serão empregadas técnicas de processamento de linguagem natural e de aprendizagem de máquina com o propósito de realizar a classificação automática das decisões judiciais. É almejado que o modelo proposto demonstre alta precisão na identificação da decisão final do juiz referente ao processo, a fim de otimizar o trabalho dos advogados e juízes no processo de busca por informações relevantes em grandes volumes de dados.

3 Algoritmos e técnicas relacionadas

3.1 Reconhecimento Óptico de Caracteres (OCR) para a Leitura de Arquivos em Formato PDF

Também conhecida por uma técnica de reconhecimento textual, pode ser utilizada para extração em diferentes propósitos, como imagens, documentos escaneados ou arquivos PDF. Sua utilização elimina trabalho manual de digitação de textos que se encontram em formatos desestruturados, acelerando o desenvolvimento de processos que envolvam análise textual, como em nosso caso, onde a matéria-prima se encontrava no site do diário oficial (DIÁRIO. . . ,), e precisávamos converter as decisões dos juízes para um formato de dados que pudessem ser manipulados e reestruturados.

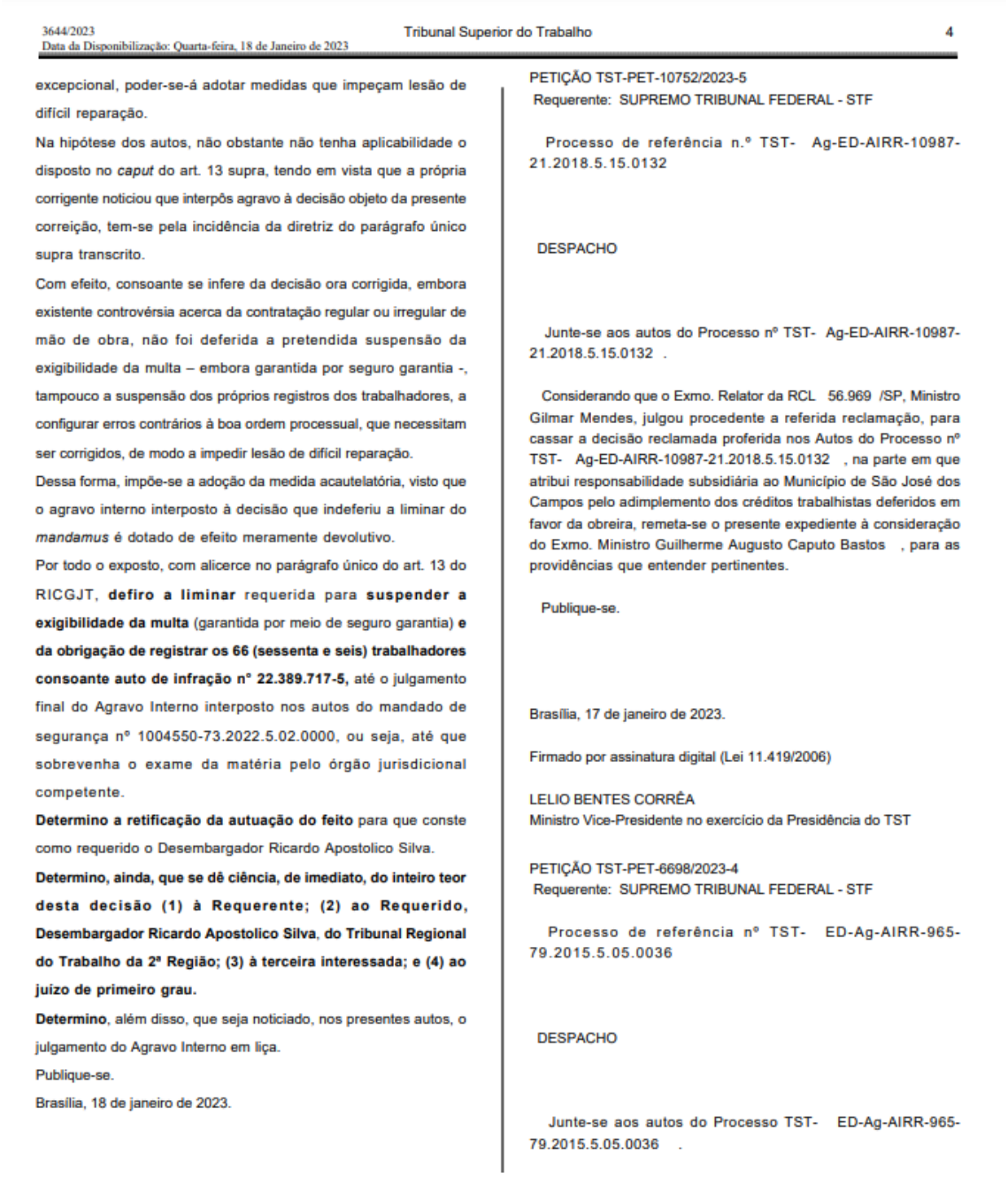
Para isso utilizamos bibliotecas *python* (MuPDF,), que já possuem implementações que facilitam a conversão para texto, em especial, em arquivos como o da figura 1. Onde é necessário um tratamento especial, pois o texto é divido em colunas e é necessário para compreensão e análise mantermos sua ordem lógica.

3.2 Limpeza textual e pré-processamento

Um token pode ser considerado uma parte de um texto, como palavras, caracteres ou frases. Essa etapa, de tokenização, é fundamental em tarefas de processamento de linguagem natural, já que visa dividir o texto em partes menores(tokens), é importante para identificar características e transcrevê-las para uma representação numérica que beneficiarão algoritmos de ML.

3.2.1 Stop Words

Stop Words são palavras ou caracteres que possuem a característica de não carregar significado que possa ser útil para os modelos. A remoção das *Stop Words* ajuda a reduzir a dimensionalidade dos dados, tornando a análise mais fácil e computacionalmente mais eficiente. Em abordagens de agrupamento de características como a que estamos usando, onde os tokens com mais relevância possuem um peso maior, e essas *Stop Words* são muito frequentes, é importante removê-las para evitar ruído nas análises.



3644/2023 Tribunal Superior do Trabalho 4
Data da Disponibilização: Quarta-feira, 18 de Janeiro de 2023

excepcional, poder-se-á adotar medidas que impeçam lesão de difícil reparação.

Na hipótese dos autos, não obstante não tenha aplicabilidade o disposto no *caput* do art. 13 supra, tendo em vista que a própria corrigente noticiou que interpôs agravo à decisão objeto da presente correição, tem-se pela incidência da diretriz do parágrafo único supra transcrito.

Com efeito, consoante se infere da decisão ora corrigida, embora existente controvérsia acerca da contratação regular ou irregular de mão de obra, não foi deferida a pretendida suspensão da exigibilidade da multa – embora garantida por seguro garantia -, tampouco a suspensão dos próprios registros dos trabalhadores, a configurar erros contrários à boa ordem processual, que necessitam ser corrigidos, de modo a impedir lesão de difícil reparação.

Dessa forma, impõe-se a adoção da medida acautelatória, visto que o agravo interno interposto à decisão que indeferiu a liminar do *mandamus* é dotado de efeito meramente devolutivo.

Por todo o exposto, com alicerce no parágrafo único do art. 13 do RICGJT, **defiro a liminar** requerida para **suspender a exigibilidade da multa** (garantida por meio de seguro garantia) **e da obrigação de registrar os 66 (sessenta e seis) trabalhadores consoante auto de infração nº 22.389.717-5,** até o julgamento final do Agravo Interno interposto nos autos do mandado de segurança nº 1004550-73.2022.5.02.0000, ou seja, até que sobrevenha o exame da matéria pelo órgão jurisdicional competente.

Determino a retificação da autuação do feito para que conste como requerido o Desembargador Ricardo Apostolico Silva.

Determino, ainda, que se dê ciência, de imediato, do inteiro teor desta decisão (1) à Requerente; (2) ao Requerido, Desembargador Ricardo Apostolico Silva, do Tribunal Regional do Trabalho da 2ª Região; (3) à terceira interessada; e (4) ao juízo de primeiro grau.

Determino, além disso, que seja noticiado, nos presentes autos, o julgamento do Agravo Interno em liça.

Publique-se.

Brasília, 18 de janeiro de 2023.

PETIÇÃO TST-PET-10752/2023-5
Requerente: SUPREMO TRIBUNAL FEDERAL - STF

Processo de referência n.º TST- Ag-ED-AIRR-10987-21.2018.5.15.0132

DESPACHO

Junte-se aos autos do Processo nº TST- Ag-ED-AIRR-10987-21.2018.5.15.0132 .

Considerando que o Exmo. Relator da RCL 56.969 /SP, Ministro Gilmar Mendes, julgou procedente a referida reclamação, para cassar a decisão reclamada proferida nos Autos do Processo nº TST- Ag-ED-AIRR-10987-21.2018.5.15.0132 , na parte em que atribui responsabilidade subsidiária ao Município de São José dos Campos pelo adimplemento dos créditos trabalhistas deferidos em favor da obreira, remeta-se o presente expediente à consideração do Exmo. Ministro Guilherme Augusto Caputo Bastos , para as providências que entender pertinentes.

Publique-se.

Brasília, 17 de janeiro de 2023.

Firmado por assinatura digital (Lei 11.419/2006)

LELIO BENTES CORRÊA
Ministro Vice-Presidente no exercício da Presidência do TST

PETIÇÃO TST-PET-6698/2023-4
Requerente: SUPREMO TRIBUNAL FEDERAL - STF

Processo de referência nº TST- ED-Ag-AIRR-965-79.2015.5.05.0036

DESPACHO

Junte-se aos autos do Processo TST- ED-Ag-AIRR-965-79.2015.5.05.0036 .

Figura 1 – Exemplo de como as decisões judiciais são publicados. Fonte: Diário eletrônico da justiça do trabalho

3.2.2 Noise removal

Muitos textos possuem pontuações e caracteres especiais que são importantes para a interpretação humana, mas podem ser desconsiderados pelos algoritmos de aprendizado de máquina. No do tipo de dados que estamos trabalhando, alguns caracteres especiais utilizados para referência de leis, números de processos e informações presentes em cabeçalhos e rodapés de arquivos PDF não são relevantes para a classificação das decisões judiciais, portanto, podem ser desconsiderados.

3.2.3 Stemming

O processo de *Stemming* consiste na redução de palavras à sua forma raiz. Em NLP (Processamento de Língua Natural), o *Stemming* é utilizado para pré-processar dados textuais visando reduzir a dimensionalidade dos dados e eliminar as variações morfológicas das palavras que apresentam mesmo significado. Esta técnica contribui para a simplificação dos dados textuais e eliminação de diferenças irrelevantes para os modelos nas formas das palavras, evitando assim, interferências na precisão dos modelos.

3.2.4 Lemmatization

Embora *lemmatization* e *stemmimg* sejam muito próximos em seus objetivos, as duas técnicas se diferenciam, pelo fato de que quando aplicamos técnicas de *lemmatization*, sempre obtemos como resultado, palavras que existem no dicionário, enquanto ao reduzir uma palavra a sua versão de *stem*, isso não é verdade. É importante se atentar a essa característica, pois a perca da gramática pode resultar em ambiguidade, o que pode levar a interpretações errôneas e, além disso, a perda de informação, já que o *stemming* pode ocultar a forma e o significado da palavra, o que também pode prejudicar a precisão da análise do texto.

3.3 Rotulagem de dados

A rotulagem de dados é o processo de atribuir etiquetas ou rótulos, que representam a classe que aquela informação pertence, essa etapa é importante, pois contribui para que métodos de classificação, como mencionado em 3.4.1, consigam identificar particularidades nos dados que sejam suficientes para atribuir as devidas classes.

Utilizar um conjunto de dados rotulados que será exposto a técnicas de aprendizado de máquina, é uma técnica também chamada de aprendizado supervisionado, é o tipo mais comum de modelo, e pelo fato de possuir referências sobre quais são as classes corretas, é possível no processo de avaliação da qualidade dos métodos desenvolvidos e treinados, avaliar seu desempenho com maior segurança. Para esse trabalho, por estarmos utilizando,

informações de um domínio específico, relacionado a processos do direito trabalhista, advogadas da área nos ajudaram no processo de rotulação.

Dentre as classes escolhidas para rotulagem, elegemos:

- **Aprovado**. Uma decisão judicial aprovada é aquela em que o tribunal decide a favor da parte que fez a solicitação ou da parte que apresentou a ação judicial. Isso significa que a decisão favorece essa parte, sendo vista como uma vitória para ela.
- **Negado**. Uma decisão judicial negada é aquela em que o tribunal decide contra a parte que fez a solicitação ou a parte que apresentou a ação judicial.
- **Suspenso**: Uma decisão judicial suspensa é aquela em que o tribunal adia temporariamente a decisão final sobre o caso. Isso pode acontecer por vários motivos, como a necessidade de mais evidências ou informações antes de decidir final.
- **Nenhum**: Uma decisão judicial nenhuma é aquela em que o tribunal não decide o final sobre o caso. Utilizamos essa classe para remover ruídos, como, processos de despacho, que possuem forma, mas não conteúdo de decisão (nada decide, somente faz o processo andar). Isso pode acontecer se o tribunal decidir que não tem jurisdição sobre o caso, se o caso for retirado pelas partes envolvidas, ou se houver outras razões pelas quais o caso não pode ser julgado.

3.4 Introdução à classificação textual

A classificação textual é uma tarefa de aprendizado de máquina que consiste em atribuir uma categoria ou rótulo predefinido a um determinado texto. Ela é amplamente utilizada em várias aplicações, como detecção de spam, análise de sentimentos e detecção de língua.

A classificação textual é possível devido ao avanço do processamento de linguagem natural e das técnicas de aprendizado de máquina. Através do uso de algoritmos de aprendizado supervisionado, os modelos são treinados com conjunto de dados rotulados e aprendem a identificar padrões e características que são típicos de cada categoria. Desta forma, quando apresentado com um novo texto, o modelo consegue atribuir a ele a categoria provável.

3.4.1 Métodos de classificação textual

Uma das tarefas mais importantes para classificar textos é escolher um modelo adequado. Dentre eles, os mais comuns são: *Decision Trees*, *Rule-based Classifiers*, *SVM Classifiers*, *Neural Network Classifiers* e Modelos probalisticos como bayesian (KOWSARI

et al., 2019a). Com base nas revisões bibliográficas realizadas, elegemos modelos que serão utilizados no trabalho, dentre eles:

3.4.2 SVM Classifiers

O objetivo do SVM (Support Vector Machine) é encontrar o hiperplano ótimo que separe as classes de texto com base nas suas características, como a presença de certas palavras ou frases. O SVM utiliza uma técnica chamada de kernel para transformar os dados de entrada em um espaço dimensional maior, permitindo a criação de uma separação nítida entre as classes. O SVM é conhecido por ter uma boa capacidade de generalização, ou seja, de fazer previsões precisas em dados desconhecidos, e é amplamente utilizado em aplicações de classificação textual, como a detecção de spam e a análise de sentimentos (NOBLE, 2006).

3.4.3 Arvore de decisão

As Árvores de Decisão funcionam criando uma série de perguntas sobre as características dos dados de entrada, como a presença de certas palavras ou frases, para tomar decisões sobre a classificação final. A árvore é construída a partir da raiz, e as decisões são tomadas em cada nó até chegar a uma folha, que contém a classificação final. As Árvores de Decisão são conhecidas por serem fáceis de entender e interpretar, além de serem capazes de lidar com dados com múltiplas características e classes. No entanto, elas tendem a criar árvores complexas e não lineares, o que pode levar a *overfitting*, ou seja, um desempenho ruim em dados desconhecidos, (LOH, 2011).

3.4.4 Classificador KNN

O KNN (K-Nearest Neighbors) é um algoritmo de aprendizado de máquina supervisionado que funciona comparando uma amostra desconhecida com as k amostras mais próximas dela em termos de suas características, como a presença de certas palavras ou frases. A classe da amostra é então determinada pela classe mais comum entre as k amostras mais próximas. O KNN é conhecido por ser fácil de implementar e entender, e por ser capaz de lidar com dados de alta dimensionalidade. No entanto, ele pode ser computacionalmente custoso, com tendência a ser afetado por ruído e *outliers* nos dados. Além disso, a escolha do valor de k pode afetar significativamente o desempenho do algoritmo (GUO et al., 2003).

3.4.5 Logistic Regression

A Regressão Logística é um algoritmo de aprendizado de máquina supervisionado usado para classificação textual. Ele funciona aplicando uma função logística a um conjunto

de características dos dados de entrada, como a presença de certas palavras ou frases, para prever a probabilidade de uma amostra pertencer a cada classe. A previsão final é feita através de uma função de threshold, onde as amostras são classificadas como pertencentes à classe cuja probabilidade é maior. A Regressão Logística é conhecida por ser rápida e fácil de implementar, além de ser capaz de lidar com múltiplas classes. No entanto, ela não é tão eficiente quanto outros algoritmos, como o SVM, em lidar com dados de alta dimensionalidade e complexidade (KLEINBAUM, 1994).

3.5 Métricas de avaliação

Pelo fato de modelos de aprendizagem muitas vezes serem apresentados como caixas pretas, sua interpretabilidade pode se tornar difícil até para especialistas da área. Avaliar adequadamente o funcionamento de um modelo, é uma tarefa fundamental, principalmente, quanto resultados incorretos podem causar consequências negativas (CARVALHO; PEREIRA; CARDOSO, 2019). Sabendo disso, técnicas de medição que produzem saídas padronizadas, podem ser utilizadas para preservar a forma que análises sobre os resultados são feitas. Dentre as mais conhecidas e que buscamos aplicar nesse trabalho, temos. **Precisão**, **revocação** e **medida F1**.

3.6 Conclusão

Em resumo, a classificação textual é uma tarefa importante de aprendizado de máquina que consiste em atribuir uma categoria ou rótulo predefinido a um determinado texto. Para isso, é necessário seguir uma série de etapas que envolvem desde o pré-processamento do texto até a avaliação dos resultados obtidos com o modelo.

No pré-processamento, é comum realizar etapas como tokenização, remoção de stop words, normalização, stemming, entre outras técnicas que visam aprimorar a qualidade dos dados e facilitar a análise. Isso pode ser feito tanto de forma manual como automática, dependendo do caso.

Para avaliar a eficácia do modelo de classificação, são utilizadas diversas métricas de avaliação, tais como a acurácia, precisão, recall e F1-score. Essas métricas ajudam a compreender como o modelo está se saindo em relação aos dados utilizados para treinamento e teste, permitindo ajustar o modelo de forma a melhorar sua performance.

Outro aspecto importante é o OCR (Optical Character Recognition), que consiste em transformar imagens com texto em arquivos digitais que podem ser editados e pesquisados. Esse processo pode ser útil em tarefas de classificação textual, já que permite analisar documentos que estejam em formato de imagem ou pdf.

E por fim, a escolha dos diferentes métodos que podem ser utilizados, como SVM Classifiers, Árvore de Decisão, Classificador KNN e Regressão Logística. Cada um desses métodos apresenta vantagens e desvantagens em relação a sua eficiência em lidar com dados de alta dimensionalidade, complexidade e generalização. Além disso, a rotulagem de dados é fundamental para o sucesso da tarefa de classificação textual, uma vez que contribui para que os modelos possam identificar particularidades nos dados que sejam suficientes para a realização da classificação.

4 Trabalhos Correlatos

Nesta seção, serão apresentados alguns dos principais trabalhos que favoreceram a formulação deste projeto de conclusão de curso. A busca por pelos artigos foi realizada majoritariamente utilizando ferramentas de busca como Google Scholar, utilizando os seguintes termos: *Jurimetria*, *Legal Documents*, *Legal Domain*, *Text Classification* e *LegalAI*.

Com o intuito de realizar as pesquisas, foi mantido um conjunto de dois pontos de foco, entre os quais se destacam:

1. Trabalhos Relacionados a Processamento de Linguagem Natural e Jurisprudência;
2. Trabalhos que Abordam Problemas de Processamento de Dados Textuais no Âmbito Jurídico.

4.1 Trabalhos Relacionados a Processamento de Linguagem Natural e Jurisprudência

No que diz respeito aos artigos coletados sobre o item 1, três deles se mostraram mais pertinentes. O primeiro, intitulado "*Predicting Brazilian Court Decisions*"(LAGE-FREITAS et al., 2022), é o trabalho que mais se assemelha à problemática que estamos tentando resolver. Em suas análises, os autores constataram que profissionais do direito utilizam decisões judiciais passadas como guias para suas decisões e previsões, o que pode ser útil na prática jurídica. No entanto, prever resultados jurídicos é difícil porque requer a análise de vastas quantidades de documentos jurídicos, e os sistemas são complexos. Este estudo apresenta uma nova metodologia para prever decisões judiciais no Brasil utilizando algoritmos de aprendizado de máquina, alcançando um F1-score de 80,2

Outro estudo relevante que se encaixou no item 1 foi "*Challenges When Using Jurimetrics in Brazil—A Survey of Courts*"(COLOMBO; BUCK; BEZERRA, 2017). Os autores constataram que a jurimetria é a aplicação de métodos estatísticos ao direito. Isso envolve a extração e organização de dados dos tribunais, o que pode ser um desafio devido à estrutura não padronizada e à linguagem natural dos dados brutos. Devido a isso, a jurimetria é uma área multidisciplinar que requer experiência em direito, estatística e ciência da computação. No Brasil, existem desafios adicionais devido à heterogeneidade dos diferentes sistemas judiciais, falta de padronização, interpretação e implementação de leis de dados abertos. Foi realizada uma pesquisa para avaliar a prontidão dos tribunais brasileiros para implementar um sistema jurimétrico, sendo identificadas questões de

privacidade e técnicas. Para isso, os autores propuseram um roteiro para enfrentar esses desafios por meio de tecnologia e políticas públicas.

O tericeiro artigo (KASTELLEC, 2010) trata da importância das regras legais para a consistência do direito e como elas ajudam juízes e outros profissionais a tomarem decisões. O autor argumenta que um método estatístico chamado árvores de classificação pode ajudar a estudar o campo legal, através da análise de padrões de fatos e melhor captura e relação entre fatos do caso e decisões tomadas pelo corpo jurídico. O artigo aplica o método de árvore de classificação a casos decididos pela Suprema Corte dos Estados Unidos e tribunais de apelação para ilustrar suas vantagens e prossegue discutindo a importância das regras legais para a construção de bons modelos de aprendizado com alta precisão.

4.2 Trabalhos que Abordam Problemas de Processamento de Dados Textuais no Âmbito Jurídico.

Sobre os trabalhos do item 2. Foi possível catalogar outros três artigos que fossem de nosso interesse, dentre eles. "*What's Inside the Black Box? AI Challenges for Lawyers and Researchers*", (YU; ALÌ, 2019), com provocações muito importantes sobre o impacto da inteligência artificial (IA) na profissão jurídica, incluindo o uso de ferramentas de pesquisa em IA que permitem a análise de grandes conjuntos de dados e a identificação de padrões. No entanto, a complexidade dos sistemas de IA pode levar a pesquisas defeituosas, especialmente quando realizadas por pesquisadores com pouca experiência em tecnologia da informação ou por pessoas que não possuem conhecimento técnico da área do direito. O artigo também destaca os perigos potenciais da confiança cega na imparcialidade, confiabilidade e infalibilidade da IA jurídica, que está sujeita a vieses inerentes aos algoritmos e aos conjuntos de dados usados. O Regulamento Geral de Proteção de Dados da União Europeia reconhece os efeitos da tomada de decisão algorítmica nos "direitos fundamentais e liberdades das pessoas naturais"e aborda a questão dos possíveis abusos da IA.

Outro estudo relevante que discute o tema do presente trabalho, o qual foi conduzido pelo artigo, "*AI in legal services: new trends in AI-enabled legal services*", (KAUFFMAN; SOARES, 2020), onde os autores discutem como a inteligência artificial (IA) está transformando a indústria jurídica, trazendo novas ferramentas e recursos que melhoram os serviços jurídicos e o acesso à justiça. Juízes e advogados movidos a IA, análise de documentos, pesquisa jurídica e automação de práticas são alguns exemplos de como a IA está sendo usada no setor jurídico. No entanto, o desenvolvimento e uso da IA na indústria jurídica são limitados por desafios relacionados a dados, algoritmos e implementação. Portanto, é necessário realizar pesquisas interdisciplinares para enfrentar esses desafios e garantir o

desenvolvimento e implementação éticos de soluções de IA como serviços para melhorar a entrega de serviços jurídicos.

O terceiro trabalho catalogado, intitulado, "*Text Classification Algorithms: A Survey*", (KOWSARI et al., 2019b), explica o processo de classificação textual, que envolve a conversão de dados não estruturados, como os processos que estamos utilizando nesse projeto, em um formato estruturado usando técnicas de aprendizado de máquina. Os passos envolvidos no processamento são. O primeiro é a extração de características, onde métodos como TF-IDF e Word2Vec são usados para converter os dados em um espaço de características estruturado. O segundo passo é a redução da dimensionalidade, onde técnicas como PCA e NMF são usadas para reduzir a complexidade computacional dos dados. O próximo passo é selecionar uma técnica de classificação, que pode incluir métodos como regressão logística, SVM e k-vizinhos mais próximos. Finalmente, o desempenho do classificador é avaliado usando métodos como precisão, pontuação F e ROC.

Outro passo importante para o processamento de dados textuais é sua limpeza, etapa também conhecida como pré-processamento, que se refere às etapas realizadas para preparar os dados para análise. Isso inclui técnicas para remover palavras desnecessárias (stop words), corrigir a capitalização, lidar com gírias e abreviações, remover ruídos, corrigir erros ortográficos e consolidar diferentes formas de palavras (stemming e lematização). A tokenização também é usada para quebrar o texto em elementos significativos chamados tokens. Todas essas etapas são importantes para uma classificação e mineração de texto efetivas.

4.3 Conclusão

Com base nos trabalhos recolhidos, podemos inferir que as aplicações de Inteligência Artificial no campo jurídico, por meio do LegalAI, podem trazer benefícios significativos, como a redução de tarefas tediosas e demoradas para profissionais jurídicos e a possibilidade de servir como referência para aqueles não familiarizados com o domínio jurídico. Além disso, a pesquisa nesta área tem sido significativa nas últimas décadas, com desenvolvimentos em aprendizado profundo levando a um melhor desempenho de tarefas. A seção apresentada, também mostra haver uma ampla gama de trabalhos relacionados à jurimetria, processamento de dados textuais no campo legal, e classificação textual, disponíveis, tanto em nível nacional quanto internacional, que podem ser utilizados como base para projetos de conclusão de curso ou outras pesquisas na área.

5 Metodologia Desenvolvida e Experimentos

Conforme mencionado na introdução 1, o presente trabalho faz parte do objetivo geral do sistema e marca o início do projeto. Acreditamos que a realização do reconhecimento textual de arquivos não estruturados e a classificação de decisões judiciais são um excelente ponto de partida e fornecerão uma base sólida para o progresso das demais etapas mencionadas na seção 1, que compreendem:

1. Identificar requisitos que determinam a aprovação de recursos judiciais;
2. Identificar quais decisões judiciais foram aprovadas e rejeitadas;
3. Agrupamento de setores/causas que possuem maior nível de aprovação;
4. Estruturar informações que envolvem. Número do processo, relator, partes envolvidas e conteúdo.

A imagem apresentada na Figura 2 ilustra as etapas do processo de desenvolvimento adotadas para a construção do sistema proposto neste trabalho. Essas etapas são cruciais para garantir um desenvolvimento eficiente e de qualidade, abrangendo desde a definição dos requisitos até a implantação do sistema.

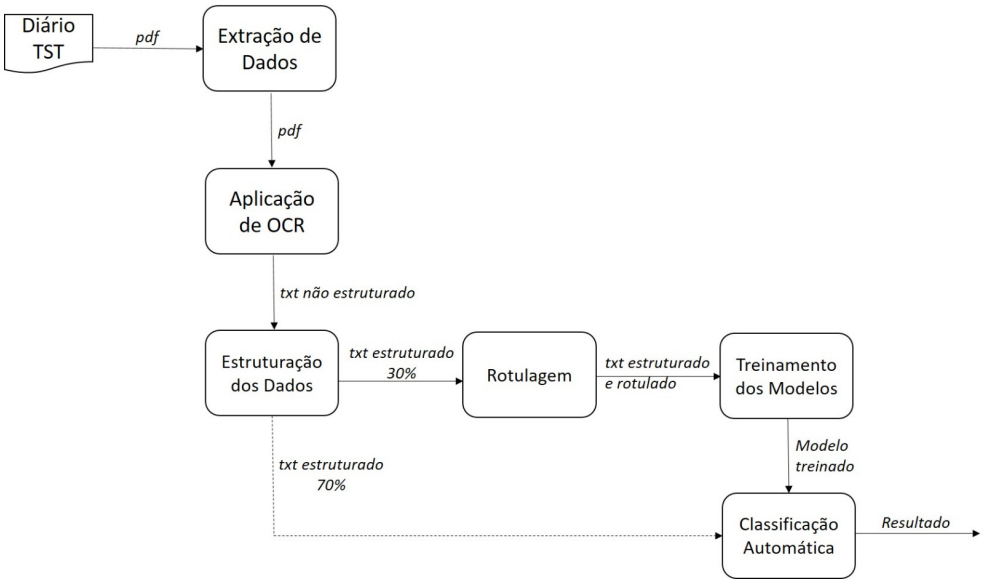
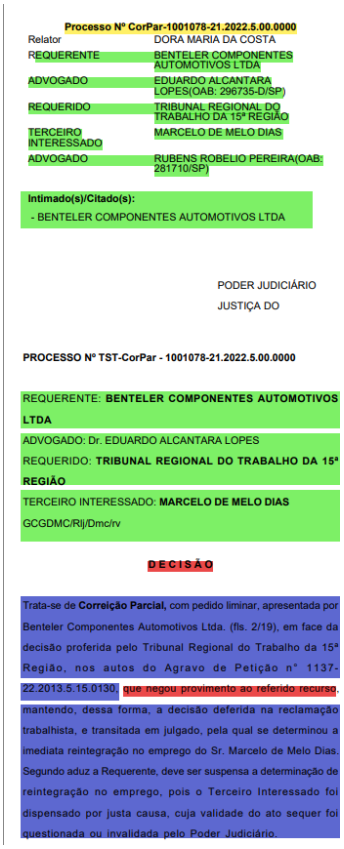


Figura 2 – Etapas do projeto. Fonte: Autor

5.1 Método para estruturação dos dados

Desta maneira, após realizar o *download* manual dos dados através do site do diário eletrônico da justiça do trabalho, com informações que contemplam o período de 09/01/2023 até 27/01/2023, utilizamos a biblioteca de *OCR* mencionada na seção, 3.1, para tornar os textos manipuláveis, após, focamos na quebra de sua granularidade, identificando por Expressões Regulares, como apresentado na figura 3, sub-tópicos de interesse(Em especial o corpo do texto e seu identificador único, o número do processo). Em amarelo, o identificador único para cada decisão, em verde as partes envolvidas, em azul, o que denominamos como o corpo do texto, que será a matéria-prima para aplicação das técnicas mencionadas em 3.4.1. Já em vermelho, a saída esperada pelo modelo, denotando a sentença judicial. Que não se resume à apenas esta. Mas todas listadas em 3.3.



Processo Nº CorPar-1001078-21.2022.5.00.0000

Relator	DORA MARIA DA COSTA
REQUERENTE	BENTELER COMPONENTES AUTOMOTIVOS LTDA
ADVOGADO	EDUARDO ALCANTARA LOPES(OAB: 296735-D/SP)
REQUERIDO	TRIBUNAL REGIONAL DO TRABALHO DA 15ª REGIÃO
TERCEIRO INTERESSADO	MARCELO DE MELO DIAS
ADVOGADO	RUBENS ROBELIO PEREIRA(OAB: 281710/SP)

Intimado(s)/Citado(s):
- BENTELER COMPONENTES AUTOMOTIVOS LTDA

PODER JUDICIÁRIO
JUSTIÇA DO

PROCESSO Nº TST-CorPar - 1001078-21.2022.5.00.0000

REQUERENTE: **BENTELER COMPONENTES AUTOMOTIVOS LTDA**
ADVOGADO: Dr. EDUARDO ALCANTARA LOPES
REQUERIDO: **TRIBUNAL REGIONAL DO TRABALHO DA 15ª REGIÃO**
TERCEIRO INTERESSADO: **MARCELO DE MELO DIAS**
GCGDMC/Rlj/Dmc/rv

DECISÃO

Trata-se de **Correição Parcial,** com pedido liminar, apresentada por Benteler Componentes Automotivos Ltda. (fls. 2/19), em face da decisão proferida pelo Tribunal Regional do Trabalho da 15ª Região, nos autos do Agravo de Petição nº 1137-22.2013.5.15.0130, que negou provimento ao referido recurso, mantendo, dessa forma, a decisão deferida na reclamação trabalhista, e transitada em julgado, pela qual se determinou a imediata reintegração no emprego do Sr. Marcelo de Melo Dias. Segundo aduz a Requerente, deve ser suspensa a determinação de reintegração no emprego, pois o Terceiro Interessado foi dispensado por justa causa, cuja validade do ato sequer foi questionada ou invalidada pelo Poder Judiciário.

Figura 3 – Identificadores. Amarelo: Identificador único do processo; Verde: Partes envolvidas; Azul: Corpo do texto que será submetido a métodos de ML; Vermelho: Saída esperada, apontando a decisão judicial. Fonte: Autor

5.2 Método para classificação dos dados

Após realizarmos a fragmentação da granularidade dos dados, movemos as informações para a ferramenta de rotulagem *Label Studio*, conforme a figura, 4, optamos por sua utilização, pois precisávamos de uma ferramenta rapidamente instalável, que possibilitasse

a criação de interfaces de usuário personalizadas e utilização de modelos de rotulagem flexíveis.

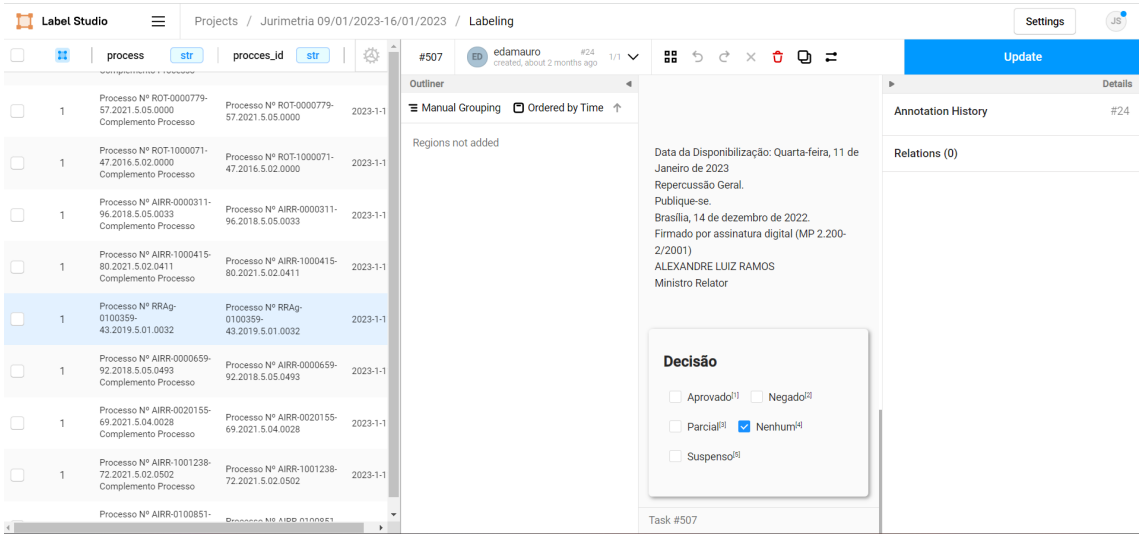


Figura 4 – Label Studio. Fonte: Autor

A escolha atendeu nossas expectativas, onde teríamos diversos estagiários do curso de direito, rotulando os dados paralelamente e ao final, exportaríamos as categorizações feitas, como é possível observar abaixo 5.1. Onde os campos úteis, foram, *process*, nosso eixo X e *decision* o eixo Y.

Listing 5.1 – Saída de dados, Label Studio


```
[
  {
    "process_id": "Processo N  Ag-ED-ROT-0100729-84.2020.5.01.0000",
    "process": "Processo N  Ag-ED-ROT-0100729-84.2020.5.01...",
    "extract_at": "2023-1-11",
    "title": "Diario_3639__11_1_2023.pdf",
    "book": "Caderno do Tribunal Superior do Trabalho - Judici rio",
    "id": 491,
    "decision": "Nenhum",
    "annotator": 5,
    "annotation_id": 31,
    "created_at": "2023-01-24T14:23:56.847207Z",
    "updated_at": "2023-01-24T14:23:56.847244Z",
    "lead_time": 100.867
  }
  ...
]
```

É importante esclarecer, que essa ferramenta foi implantada em máquinas virtuais, utilizando a provedora de nuvem *Linode*, que disponibilizam $ 100 dólares de créditos por 3 meses para elaboração desse trabalho. Decidimos executar o *Label-Studio* em máquinas em nuvem, para que os advogados e estagiários, pudessem rotular os dados acessando a ferramenta através de seus navegadores web.

Após isso, utilizando as funções disponibilizadas pelo *scikit-learn*, para aplicar limpezas nos dados, como citado em, 3.2, em seguida aplicamos, com a mesma biblioteca, os métodos listados em 3.4.1. Os detalhes de cada modelo e seus resultados podem ser encontrados na sessão 5.3.

5.3 Resultados experimentais

Os modelos apresentados estão sendo utilizados para classificação de textos em categorias pré-definidas. Todos os modelos seguem a mesma estrutura de pipeline, que inclui as etapas de vetorização (CountVectorizer), transformação TF-IDF (TfidfTransformer) e classificação utilizando diferentes algoritmos de aprendizado de máquina.

O modelo 1 utiliza o algoritmo de classificação KNN. O modelo 2 utiliza o algoritmo de classificação SGDClassifier, que é um classificador linear que utiliza gradiente descendente estocástico. O modelo 3 utiliza o algoritmo de classificação LogisticRegression, que é um modelo de regressão logística. O modelo 4 utiliza o algoritmo de classificação DecisionTreeClassifier, que é um modelo de árvore de decisão.

Os modelos foram treinados utilizando o conjunto de dados de treinamento (X_train e y_train) com 2/3 dos dados e avaliados utilizando o conjunto de dados de teste (X_test e y_test), com o restante, 1/3 dos dados catalogados. A distribuição dos dados pode ser observado na imagem 5.

Para realizar a distribuição das informações, foram realizadas os seguintes tratamentos. Com a função train_test_split da biblioteca sklearn.model_selection dividimos o conjunto de dados em conjuntos de treinamento e teste. O objetivo dessa função é permitir o treinamento e avaliação de modelos de aprendizado de máquina de forma mais precisa e eficiente.

Utilizamos a opção test_size=0.33 para definir a proporção dos dados a serem usados como conjunto de teste. Também utilizamos o parâmetro random_state=42 para garantir a reprodutibilidade dos resultados e shuffle=True para embaralhar os dados antes da divisão.

Além disso, utilizamos o parâmetro stratify=desicion_content_list para garantir que as proporções das classes nos conjuntos de treinamento e teste fossem balanceadas. Isso foi necessário, uma vez que a distribuição de classes dos dados não era uniforme, como

é possivel observar em 5.

Para a vetorização dos textos, foi utilizada a técnica de CountVectorizer, que cria um vocabulário com as palavras mais frequentes e conta a frequência de cada uma delas nos textos. Em seguida, foi aplicada a transformação TF-IDF (Term Frequency-Inverse Document Frequency), que calcula a importância de cada palavra no texto com base em sua frequência e em quantos textos ela aparece.

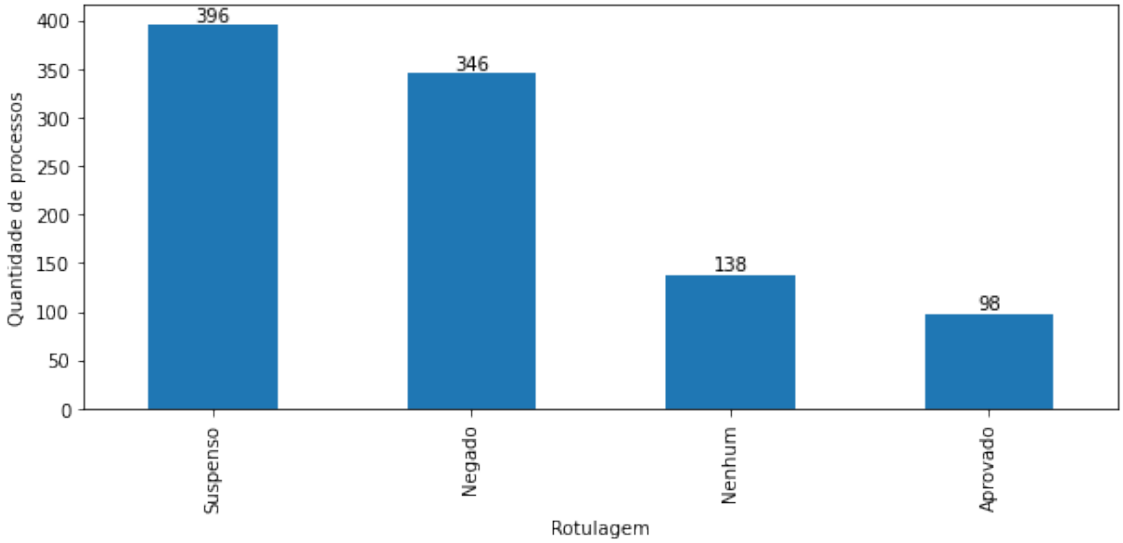


Figura 5 – Distribuição dos processos. Fonte: Autor

A acurácia e o relatório de classificação (classification_report) foram utilizados para avaliar o desempenho dos modelos e as médias da predição de cada rotulagem é possível observar nos dados apresentados na tabela 1.

Em resumo, os modelos apresentados foram criados visando classificar textos em categorias pré-definidas utilizando diferentes algoritmos de aprendizado de máquina, sendo avaliados utilizando as métricas de acurácia e relatório de classificação, os resultados podem ser observado na tabela 1

Modelo	**Precisão**	**Revocação**	**F1**	**Acurácia**
SVM	0.93	0.88	0.90	0.93
Regressão Logística	0.96	0.93	0.95	0.96
Árvores de Decisão	0.91	0.92	0.91	0.94
KNN	0.74	0.72	0.73	0.81

Tabela 1 – Tabela de métricas dos modelos.

Os links para as implementações podem ser encontrados no repositório GitHub em <https://github.com/perebaj/playground/tree/main/etl_dejt> e também no Colab do Google Drive em <https://colab.research.google.com/drive/1iivSVTwML7vmdnc_CtK05G28_7CD9fuq?usp=sharing>

5.4 Conclusão

Com base na análise dos resultados obtidos, conclui-se que a utilização do LegalAI, por meio de técnicas de processamento de linguagem natural, pode ser aplicada com êxito na estruturação de decisões judiciais. A partir do reconhecimento de texto utilizando OCR e da aplicação de métodos de limpeza, foi possível alcançar resultados satisfatórios em relação à estruturação dos dados.

Ademais, o uso da ferramenta de rotulagem Label Studio possibilitou a categorização dos dados de maneira rápida e eficiente, viabilizando, assim, o treinamento e a validação dos modelos de classificação.

A tabela de métricas 1 apresenta o desempenho de quatro modelos de classificação em relação a quatro métricas: precisão, revocação, F1-score e acurácia.

Pode-se inferir que os modelos de SVM e regressão logística apresentaram desempenho superior em todas as métricas quando comparados aos modelos de árvores de decisão e KNN. Além disso, o modelo de regressão logística obteve a maior pontuação em todas as métricas, o que indica que esteve em melhor desempenho geral entre os modelos avaliados.

Por outro lado, o modelo KNN apresentou a menor pontuação em todas as métricas, o que sugere que esteve em pior desempenho entre os modelos avaliados. Contudo, é importante salientar que as métricas apresentadas dependem do conjunto de dados e do problema em questão, logo, os resultados podem variar em contextos distintos.

Em síntese, os resultados obtidos indicam que o LegalAI, através de técnicas de processamento de linguagem natural, é uma ferramenta promissora para a estruturação de decisões judiciais. O uso de técnicas de reconhecimento de texto e classificação de dados pode fornecer aos profissionais do Direito uma abordagem mais eficiente e automatizada para o processamento de documentos jurídicos.

6 Conclusões finais

Com base nos resultados obtidos neste trabalho de conclusão de curso, podemos concluir que a classificação textual é uma tarefa importante e desafiadora em aprendizado de máquina, especialmente quando aplicada ao campo jurídico. O uso de técnicas de processamento de linguagem natural e LegalAI pode ser uma solução eficiente para a estruturação e categorização de decisões judiciais.

Durante os experimentos, foram avaliados quatro modelos de classificação: SVM, árvore de decisão, KNN e regressão logística. Os resultados mostraram que os modelos SVM e regressão logística apresentaram um desempenho superior em todas as métricas de avaliação, com destaque para a regressão logística, que obteve a melhor pontuação em todas as métricas.

A acurácia do modelo de regressão logística foi de 96,6%, enquanto a precisão, recall e F1-score foram de 96,5%, 96,6% e 96,5%, respectivamente. Isso indica que o modelo foi capaz de classificar corretamente a grande maioria dos documentos avaliados, apresentando uma alta precisão e uma baixa taxa de falsos positivos e falsos negativos.

Esses resultados demonstram que o uso de LegalAI e técnicas de processamento de linguagem natural pode ser uma solução eficiente e promissora para a classificação de documentos jurídicos. No entanto, é importante destacar que os resultados obtidos dependem do conjunto de dados utilizado e do contexto em que a classificação está sendo aplicada.

6.1 Trabalhos Futuros

Com base nos objetivos gerais e nas etapas propostas para a solução do problema de identificação de decisões judiciais aprovadas e rejeitadas, alguns objetivos futuros podem ser traçados. Por exemplo, um objetivo seria melhorar a eficiência do processo de rotulagem de dados, explorando técnicas de rotulagem semi-supervisionada ou aprendizado ativo.

Outro objetivo seria expandir a análise exploratória dos dados do DEJT, explorando outras características e desafios do conjunto de dados, como a variação da distribuição de decisões judiciais ao longo do tempo ou a influência de diferentes regiões geográficas.

Em relação à conclusão gerada, é importante ressaltar que a solução proposta apresentou resultados promissores na classificação de decisões judiciais aprovadas e rejeitadas, com um desempenho superior em relação a outros modelos de classificação existentes na literatura. Além disso, o modelo proposto pode ser aplicado em outros conjuntos de dados de decisões judiciais com estrutura semelhante ao DEJT.

Referências

CARVALHO, D. V.; PEREIRA, E. M.; CARDOSO, J. S. Machine learning interpretability: A survey on methods and metrics. *Electronics*, Multidisciplinary Digital Publishing Institute, v. 8, n. 8, p. 832, jul. 2019. Citado na página 20.

COLOMBO, B. A.; BUCK, P.; BEZERRA, V. M. Challenges when using jurimetrics in Brazil—A survey of courts. *Future Internet*, Multidisciplinary Digital Publishing Institute, v. 9, n. 4, p. 68, out. 2017. Citado 2 vezes nas páginas 11 e 22.

DIÁRIO Eletrônico da Justiça do Trabalho. <https://dejt.jt.jus.br/dejt/f/n/diariocon>. Acessado: 2023-1-28. Citado na página 15.

GUO, G. et al. KNN Model-Based approach in classification. In: *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*. [S.l.]: Springer Berlin Heidelberg, 2003. p. 986–996. Citado na página 19.

KASTELLEC, J. P. The statistical analysis of judicial decisions and legal rules with classification trees. *J. Empir. Leg. Stud.*, Wiley, v. 7, n. 2, p. 202–230, jun. 2010. Citado na página 23.

KAUFFMAN, M. E.; SOARES, M. N. AI in legal services: new trends in AI-enabled legal services. *Service Oriented Computing and Applications*, v. 14, n. 4, p. 223–226, dez. 2020. Citado 2 vezes nas páginas 11 e 23.

KLEINBAUM, D. G. Introduction to logistic regression. In: _____. *Logistic Regression: A Self-Learning Text*. New York, NY: Springer New York, 1994. p. 1–38. ISBN 978-1-4757-4108-7. Disponível em: <https://doi.org/10.1007/978-1-4757-4108-7_1>. Citado na página 20.

KOWSARI, K. et al. Text classification algorithms: A survey. *Information*, MDPI, v. 10, n. 4, p. 150, 2019. Citado na página 19.

KOWSARI, K. et al. Text classification algorithms: A survey. *Information*, Multidisciplinary Digital Publishing Institute, v. 10, n. 4, p. 150, abr. 2019. Citado na página 24.

LAGE-FREITAS, A. et al. Predicting brazilian court decisions. *PeerJ Comput Sci*, peerj.com, v. 8, p. e904, mar. 2022. Citado na página 22.

LOH, W.-Y. Classification and regression trees. *WIREs Data Mining and Knowledge Discovery*, John Wiley & Sons, Ltd, v. 1, n. 1, p. 14–23, jan. 2011. Citado na página 19.

MuPDF. <https://mupdf.com/>. : 2023-1-29. Citado na página 15.

NOBLE, W. S. What is a support vector machine? *Nat. Biotechnol.*, Springer Science and Business Media LLC, v. 24, n. 12, p. 1565–1567, dez. 2006. Citado na página 19.

SERRAS, F. R.; FINGER, M. verBERT: Automating brazilian case law document multi-label categorization using BERT. In: *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*. [S.l.]: SBC, 2021. p. 237–246. Citado na página 11.

YU, R.; ALÌ, G. S. What's inside the black box? AI challenges for lawyers and researchers. *Legal Information Management*, Cambridge University Press, v. 19, n. 1, p. 2–13, mar. 2019. Citado 2 vezes nas páginas 11 e 23.