Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Mineração de Dados Legislativos Federais para Análise e Predição de Aprovação de Projetos de Lei

Ilo César Duarte Cabral

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Orientador
Prof. Dr. Glauco Vitor Pedrosa

Brasília
2024

Ficha catalográfica elaborada automaticamente,
com os dados fornecidos pelo(a) autor(a)

Cabral, Ilo C. D.
CC117m Mineração de Dados Legislativos Federais para Análise e Predição de Aprovação de Projetos de Lei / Ilo C. D. Cabral; orientador Glauco V. Pedrosa. -- Brasília, 2024.
91 p.

Dissertação(Mestrado Profissional em Computação Aplicada) -- Universidade de Brasília, 2024.

1. mineração de dados. 2. dados legislativos. 3. aprendizado de máquinas. 4. processamento de linguagem natural. 5. ados desbalanceados. I. Pedrosa, Glauco V., orient. II. Título.

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Mineração de Dados Legislativos Federais para Análise e Predição de Aprovação de Projetos de Lei

Ilo César Duarte Cabral

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Prof. Dr. Glauco Vitor Pedrosa (Orientador)
FGA/UnB

Prof. Dr. Luís Paulo Faina Garcia
CIC/UnB

Dr. Eduardo de Paula Costa
Corteva Agriscience

Prof. Dr. Gladston Luiz da Silva
Coordenador do Programa de Pós-graduação em Computação Aplicada

Brasília, 26 de fevereiro de 2024

Dedicatória

Eu dedico esse trabalho ao meu falecido pai. Ele não mediu esforços na minha educação e digo, sem medo de errar, que estaria realizando um sonho junto comigo.

Agradecimentos

Meu especial agradecimento vai para meu orientador, o Prof. Dr. Glauco Vitor Pedrosa, por sua dedicada orientação e comprometimento na elaboração deste trabalho. Agradeço também aos professores do curso que, em sua totalidade, honraram a nobre arte de ensinar. Agradeço aos demais alunos da turma pelo convívio cortês, e em especial, àqueles com quem dividi esforços nos trabalhos acadêmicos que contribuíram para essa dissertação. Não posso deixar de mencionar o esforço hercúleo dos funcionários da UNB, que aqui represento nas pessoas do Prof. Dr. Gladston Luiz da Silva e Prof. Dr. Marcelo Ladeira pela gestão do curso durante uma pandemia de COVID-19.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), por meio do Acesso ao Portal de Periódicos.

Resumo

A divulgação dos dados legislativos pelo governo brasileiro abriu uma oportunidade para se entender os aspectos relacionados ao processo legislativo. Ao analisar padrões históricos e variáveis relevantes, é possível antecipar resultados legislativos, otimizando o processo decisório. Prever os votos de órgãos deliberativos, por exemplo, pode levar a uma melhor compreensão das políticas governamentais e, assim, gerar estratégias acionáveis, permitindo que legisladores identifiquem questões críticas, aloquem recursos eficientemente e antecipem possíveis impasses. Este trabalho se propôs a investigar modelos para análise e previsão que maximizem o uso de atributos heterogêneos publicamente acessíveis dos dados legislativos para compreender a aprovação/arquivamento de proposições legislativas. Para tal fim, foram desenvolvidos modelos de classificação baseados em algoritmos de aprendizado de máquinas e processamento de linguagem natural sobre os dados categóricos, textuais e de tramitação das Proposições Legislativas, a fim de identificar fatores discriminativos que pudessem influenciar na aprovação de Projetos de Lei e de Emenda. Como contribuição, os modelos de classificação foram avaliados em cinco cenários, utilizando diferentes conjuntos de atributos. Os resultados obtidos mostram um F1-Score de 71%, considerando apenas os dados categóricos das proposições e, ao se agregar os dados de tramitação, é possível obter um F1-Score médio de 90,6%. Os testes realizados demonstram a viabilidade de se prever a aprovação de uma proposição durante seu fluxo no processo legislativo, gerando resultados que agregam conhecimento e levam a uma melhor compreensão dos aspectos relacionados ao processo legislativo brasileiro no âmbito federal.

Palavras-chave: mineração de dados, dados legislativos, aprendizado de máquinas, processamento de linguagem natural, dados desbalanceados

Abstract

The release of legislative data by the Brazilian government opened an opportunity to understand aspects related to the legislative process. By analyzing historical patterns and relevant variables, it is possible to anticipate legislative results, optimizing the decision-making process. Predicting the votes of deliberative bodies, for example, can lead to a better understanding of government policies and thus generate actionable strategies, allowing legislators to identify critical issues, allocate resources efficiently and anticipate possible impasses. This work set out to investigate models for analysis and prediction that maximize the use of publicly accessible heterogeneous data from legislative data to understand the approval/disapproval of legislative proposals. To this end, classification models based on machine learning algorithms and natural language processing were developed on categorical, textual and processing data of Legislative Proposals, in order to identify discriminatory factors that could influence the approval of Bills and Amendment Projects. As a contribution, the classification models were evaluated in five scenarios, using different sets of attributes. The results obtained show an F1-Score of up to 70% considering only the categorical data of the propositions and, when aggregating the processing data, it is possible to obtain an F1-Score of up to 90,6%. The tests carried out demonstrate the feasibility of predicting the approval of a proposition during its flow in the legislative process, generating results that add knowledge and lead to a better understanding of aspects related to the Brazilian legislative process at the federal level.

Keywords: data mining, legislative data, machine learning, natural language processing, imbalanced dataset

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

AM Aprendizado de Máquinas.

BoW Bag-of-Words.

IDF Frequência Inversa de Documentos.

OSS One-Sided Selection.

PL Proposição Legislativa.

PLN Processamento de Linguagem Natural.

SMOTE Synthetic Minority Oversampling Technique.

SVD Singular Value Decomposition.

TCU Tribunal de Contas da União.

TF Frequência do Termo.

TF-IDF Term Frequency–Inverse Document Frequency.

Capítulo 1

Introdução

Este capítulo apresenta a contextualização e motivação para o desenvolvimento deste trabalho, o problema de pesquisa e seus desafios e a organização geral dos restantes capítulos do texto desta dissertação.

1.1 Contextualização

A publicação de dados governamentais em formato aberto contribuiu para aumentar a transparência pública e também permitiu uma maior participação e colaboração da sociedade nas ações e decisões do governo. A divulgação dos dados públicos de governo permitiu que cidadãos assumissem um papel de agente transformador através do monitoramento e fiscalização de atos do governo através das análises das informações divulgadas, possibilitando a geração de valores voltados para transparência e responsabilidade social/fiscal do governo [1].

A divulgação dos dados legislativos, por exemplo, abriu uma oportunidade para se entender os aspectos relacionados ao Processo Legislativo, que é uma área caracterizada por um processo complexo, influenciado por diversos fatores, desde pressões sociais até fatos ideológicos [2]. As leis aprovadas pelo congresso impactam toda a sociedade e, mesmo aquelas que tratam especificamente de práticas sociais, geram reflexos econômicos e oportunidades de negócios para o mercado [3]. É a partir da formulação de proposições legislativas que os agentes políticos instauram as leis, que vão reger a sociedade, regularizando práticas sociais, mercantis, trabalhistas, entre outros.

Uma Proposição Legislativa (PL) consiste em um documento contendo determinado assunto ou tema que estará sujeito a deliberação nas casas legislativas: Câmara dos Deputados e Senado Federal. Assim, são consideradas proposições: a proposta de emenda à Lei Orgânica, o Projeto de Lei, o Projeto de Resolução, a indicação, a moção, a autorização, o requerimento, a emenda, o parecer e o veto.

1.2 Motivação

De acordo com [4], é extremamente importante desenvolver ferramentas para monitorar e analisar os riscos de qualquer projeto apresentado no Congresso. Os custos de monitorização desses projetos são elevados já que envolve a análise de inúmeros projetos propostos. Só na última década, por exemplo, 26.866 projetos de lei foram apresentados. Por isso, ao utilizar modelos computacionais, as entidades interessadas no acompanhamento dos projetos podem priorizar quais devem ser analisados e pensar em estratégias de influência mais eficazes. Exemplos dessas entidades incluem indústrias, empresas comerciais, prestadores de serviços, mercados financeiros, organizações da sociedade civil, governos federais, estaduais e locais e o Tribunal de Contas da União (TCU).

Prever a aprovação de uma PL por meio de Aprendizado de Máquinas oferece benefícios significativos [5]. Ao analisar padrões históricos e variáveis relevantes, os modelos podem antecipar resultados legislativos, otimizando o processo decisório. Isso permite que legisladores identifiquem questões críticas, aloquem recursos eficientemente e antecipem possíveis impasses. Além disso, a previsão contribui para a transparência, envolvendo o público no entendimento dos possíveis desdobramentos das propostas. Ao adotar essa abordagem, os sistemas de Aprendizado de Máquinas oferecem uma ferramenta valiosa para aprimorar a eficácia e eficiência do processo legislativo, promovendo uma governança mais informada e responsiva.

1.3 Questões de Pesquisa

A análise de dados relacionados à situação política no Brasil levanta, naturalmente, certas questões que motivam a mineração dos dados legislativos. Algumas dessas questões que nortearam o desenvolvimento desse trabalho foram:

- Questão n° 1: É possível prever a aprovação de uma PL utilizando-se apenas suas informações textuais?
- Questão n° 2: Qual o poder preditivo dos atributos categóricos na aprovação de uma PL?
- Questão n° 3: Qual o impacto dos dados de tramitação na aprovação da PL?

O grande desafio na previsão de aprovação de PL é lidar com raridade dessas ocorrências. Desde o início de 2001 até o final de 2021 apenas 8,2% das proposições concluídas foram aprovadas. São vários os fatores que dificultam as aprovações, desde conflitos de interesses às exigências do processo legislativo. Sobre este último, o trâmite

prevê várias votações em comissões e possivelmente nos plenários da Câmara e Senado. As votações nos plenários são especialmente difíceis e, dependendo do tipo da proposição, são necessárias maiorias: simples, qualificada (3/5) ou absoluta e elas podem ser em turno único ou dois turnos[1]. Isso significa que a aprovação de um PL exige consenso, em mais de um momento, entre membros do parlamento com vertentes políticas muito diferentes. Além disso, o processo é tão moroso que a maioria das proposições demora, em média, 1.700 dias para serem aprovadas[2].

A maioria dos classificadores da área de Aprendizado de Máquinas (AM) enfrenta sérios problemas em um contexto onde há um desbalanceamento na distribuição de classes [6]. Um conjunto de dados é dito desbalanceado quando nele existe uma clara desproporção entre o número de exemplos de uma ou mais classe em relação às demais classes. O desbalanceamento pode acarretar, entre várias outras consequências, em vieses de classificação e queda de desempenho de modelos preditivos, prejudicando a eficácia e a confiabilidade dos modelos. De acordo com [7], aprender com o conjunto de dados desbalanceados é um dos 10 principais problemas desafiadores na pesquisa de mineração de dados.

Além do desbalanceamento dos dados, outro desafio é como compor um resultado único a partir de tipos de dados diferentes, ou seja, como agregar metadados oriundos das proposições em conjunto com dados extraídos das informações textuais dos documentos que descrevem seu conteúdo. Essa composição de um modelo de previsão que maximiza o uso de dados heterogêneos publicamente acessíveis dos dados legislativos brasileiro é um campo fértil para o avanço de pesquisas.

Considerando as questões de pesquisa e os desafios, este trabalho se propôs à uma análise preditiva sobre os dados das PLs apresentadas na Câmara dos Deputados entre o período de 2001 a 2021, buscando investigar o conjunto de atributos com poder preditivo para aprovação de uma PL. Neste contexto, o presente trabalho apresenta os principais resultados obtidos e os procedimentos metodológicos propostos nesta investigação.

1.4 Objetivos

O objetivo geral deste trabalho foi utilizar modelos computacionais baseados em Aprendizado de Máquinas (AM) e Processamento de Linguagem Natural (PLN) para investigar fatores de previsibilidade de aprovação de uma PL a partir de seus dados cadastrais (textuais e categóricos) e de tramitação.

Para atingir o objetivo geral, alguns objetivos específicos foram definidos:

[1] https://www.politize.com.br/votacoes-no-plenario/
[2] https://www.politize.com.br/projeto-de-lei-processo-legislativo/

- analisar dados associados às proposições que sejam discriminativos, ou seja, que melhor predizem a aprovação de uma PL;
- investigar a composição da informação textual, ou seja, os textos da ementa e do inteiro teor das proposições, como fatores discriminativos para a predição da aprovação de uma PL;
- investigar modificações na base de dados e/ou nos algoritmos de classificação para mitigar o erro de previsão devido ao desbalanceamento dos dados.

1.5 Organização da Dissertação

O texto dessa dissertação está dividido da seguinte forma:

- Capítulo 2 apresenta os conceitos teóricos utilizados para o desenvolvimento do projeto e os trabalhos relacionados à mineração de dados legislativos no contexto nacional e internacional;
- Capítulo 3 discute detalhes do processo legislativo no âmbito federal juntamente com os materiais e métodos utilizados para a realização de testes experimentais;
- Capítulo 4 apresenta e discute os resultados obtidos;
- Capítulo 5 apresenta a conclusão do trabalho desenvolvido com suas contribuições, limitações e trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Este capítulo apresenta uma revisão teórica e conceitual das técnicas e metodologias adotadas neste trabalho para representar os dados e realizar a classificação e validação dos algoritmos baseados em Aprendizado de Máquinas (AM) e Processamento de Linguagem Natural (PLN) na predição de aprovações das Proposições Legislativas, bem como os trabalhos correlatos à temática deste trabalho.

2.1 Mineração de Dados Legislativos

A Mineração de Dados Legislativos é uma área que envolve a análise de informações contidas em documentos legislativos, como leis, regulamentações e registros parlamentares [8]. Esta área emergente combina técnicas de mineração de dados com o vasto conjunto de dados legislativos disponíveis para extrair conhecimento valioso, identificar padrões e entender melhor o cenário político e jurídico.

No contexto da Mineração de Dados Legislativos, os algoritmos de mineração são aplicados aos textos legislativos para descobrir relações, tendências e *insights* que podem ser usados para informar tomadas de decisão, criar políticas mais eficazes e melhorar a compreensão das implicações legais. Isso pode envolver a análise de grandes volumes de dados textuais para identificar termos-chave, entidades e padrões de linguagem.

Uma aplicação importante da Mineração de Dados Legislativos é a automação da análise de grandes conjuntos de leis e regulamentações para identificar alterações relevantes, conflitos ou lacunas [9, 10]. Isso é particularmente valioso em ambientes legislativos complexos, onde existe uma massiva quantidade de informações [11]. Em suma, a Mineração de Dados Legislativos oferece uma abordagem inovadora para explorar e entender dados legislativos, proporcionando benefícios significativos para a tomada de decisões políticas informadas e eficazes.

2.1.1 Mineração de Dados Legislativos no Contexto Brasileiro

No Brasil, os dados legislativos já vêm sendo utilizados por alguns pesquisadores da área da Ciência Política para analisar e explicar características importantes do Processo Legislativo. O trabalho de [12], por exemplo, faz uma análise quali-quanti dos projetos de lei aprovados entre 1999 e 2006 para mensurar o impacto do Presidencialismo de Coalização e, a partir de suas conclusões, apresenta propostas para melhorar a eficiência e eficácia das leis apresentadas pelo Congresso.

O estudo de [13] conduziu uma análise quantitativa de dados legislativos e descobriu que os legisladores de distritos eleitorais com grandes populações de eleitores são mais propensos a adotar novas leis que afetem o país como um todo do que leis que afetam apenas o território do seu próprio distrito eleitoral e que o Senado é menos propenso a produção de política locais do que a Câmara dos Deputados.

Recentemente, alguns trabalhos da área da Ciência de Dados envolvendo dados legislativos foram publicados e que também serviram de inspiração para esse trabalho. Entre eles o trabalho de [14], que desenvolveu um modelo capaz de identificar características que contribuem positiva e negativamente na aprovação de um PLO na Câmara dos Deputados. Nos experimentos, o modelo alcançou capacidade preditiva positiva de 0,579 e capacidade preditiva negativa de 0,991. Os resultados mostram que o modelo é bom para prever se um projeto será suspenso, mas não é confiável para prever se um projeto será aprovado. Este resultado indica que o modelo ainda não considera todos os aspectos importantes do processo legislativo, ou que a abordagem adotada para treinar o modelo não é a ideal.

A grande maioria dos trabalhos envolvendo a mineração de dados legislativos no contexto brasileiro utiliza os dados abertos da Câmara dos Deputados. O trabalho de [15], por exemplo, utilizou os dados das votações para inferir a chance de deputados mudarem seu posicionamento entre governo e oposição. Ele modelou o posicionamento dos parlamentares usando o modelo estatístico denominado de W-NOMINATE para obter os pontos ideais no começo e no fim da legislatura que serviram de entrada para modelos preditivos.

O trabalho de [16] usou algoritmos de mineração e análise heurística para analisar os dados das votações no plenário entre 2015 e 2016. O trabalho identificou o partido com maior número de deputados na Câmara que votam de forma semelhante à sua orientação e o que tem os parlamentares mais fiéis às orientações dadas para as votações. Outro resultado relevante foi a identificação que não há uma influência clara das bancadas estaduais sobre os partidos.

Usando dados abertos da Câmara dos Deputados, o trabalho de [17] desenvolveu um classificador capaz de inferir aprovação ou arquivamento dos Projetos de Lei Ordinária

(PLO). Foram selecionadas 27 características para cada PLO, aprovado ou arquivado, na Câmara dos Deputados, das quais 14 são relacionadas aos projetos de lei e seus trâmites e 13 são relacionadas aos seus respectivos autores. Apesar de não usar textos das proposições, o estudo obteve um F1-Score de 0,861 que pode ser considerado muito bom, diante de um alto desbalanceamento da base. Apenas 10% dos PLO foram aprovados do período da pesquisa (2003-2016).

O trabalho de [18] aplicou Regressão Logística, Florestas Aleatórias e Rede Neural Artificial nos dados das proposições votadas em plenário, obtendo, no melhor resultado, um F1-Score de 96,70% na predição de suas aprovações. O autor ressalta que apenas 1% de todas as proposições apresentadas entre 2000 até e 2020 chegam a ser votadas em plenário.

O Aprovômetro[1] é uma ferramenta que utiliza a inteligência artificial para estimar as chances de aprovação de cada projeto de lei ou proposta em tramitação no Congresso. O algoritmo utiliza décadas de dados do Congresso, além de centenas de variáveis, incluindo o texto do projeto, autores, temas, emendas, tendências econômicas e mudanças de forças políticas para estimar a chance de um projeto se tornar lei com grande precisão. Essa ferramenta previu corretamente o resultado de 97,5% de todos os projetos arquivados, devolvidos ou retirados pelos autores e acertou 72% dos projetos que foram convertidos em lei.

2.1.2 Mineração de Dados Legislativos no Contexto Internacional

No contexto internacional, o trabalho de [5], criou modelos a partir dos textos, títulos e dados de contexto dos projetos de lei do congresso americano e fez uma análise comparativa dos resultados obtidos. Segundo os autores, levando em consideração as métricas de desempenho AUC e MeanBrier, usar apenas texto, supera usar apenas contexto. Enquanto na métrica MeanLogLoss ocorre o inverso. Utilizando texto e contexto juntos, o modelo denominado w2vGLM superou todos os concorrentes em todas as métricas.

O trabalho de [19] propõe um método baseado em grafos envolvendo a mineração de texto de projetos de lei e a conexão social entre os legisladores para prever os votos legislativos.

No trabalho de [20], o texto do projeto e o perfil dos legisladores são utilizados para prever resultados de votações no congresso americano. Os perfil foram modelados usando o Policy Location, um modelo espacial euclidiano. No caso, as políticas ideológicas dos

[1] https://www.jota.info/tudo-sobre/aprovometro

congressistas e dos projetos são representadas como pontos e a estimativas os valores das distâncias entre eles.

Objetivando predizer aprovação de propostas do congresso americano, o trabalho de [21] propôs um processo em 2 etapas, onde na primeira são analisados dois conjuntos de dados: fatores ideológicos dos congressistas e fatores sociais. Estes últimos estão relacionados com a filiação partidária dos representantes e como seu registro de votação anterior se entrelaça com os dos demais. Na segunda etapa, são aplicadas técnicas de PLN nos textos da Wikipedia dos congressistas e Projetos de Lei do Congresso. Finalmente, todas as características extraídas provenientes destas diversas fontes são combinadas para serem utilizadas na previsão de votos.

2.2 Processamento de Linguagem Natural (PLN)

O Processamento de Linguagem Natural (PLN) é um campo interdisciplinar que combina conhecimentos de linguística, ciência da computação e inteligência artificial para permitir que os computadores entendam, interpretem e interajam com a linguagem humana. O principal objetivo do PLN é permitir que as máquinas processem, analisem e gerem texto de maneira semelhante aos humanos.

2.2.1 Pré-processamento de Dados Textuais

Uma das etapas do PLN é a limpeza e padronização nos textos. Essa atividade visa a garantir qualidade no processamento de documentos textuais e contribuir, dentre outras coisas, para diminuir o dicionário de palavras gerados no processo de representação vetorial, uma vez que algumas palavras serão suprimidas e/ou codificadas em um mesmo padrão.

A Figura 2.1 apresenta um processo tipico de pré-processamento de quatro etapas, em que uma frase (sentença) é suscetivamente transformada até se obter uma versão sintética desta frase. Na primeira etapa tem-se a *padronização*, em que o texto é convertido para minúsculo e são retirados os caracteres que não são letras. Na segunda etapa tem-se a remoção de *stopwords*, que elimina palavras irrelevantes. Na prática, são artigos, preposições, pronomes do texto, etc. A *Tokenização* e remoção dos *token* de 3(três) ocorrem na terceira etapa, nela cada *token* corresponde a um termo que aparece no texto. Na última etapa, os *tokens* são reduzidos aos seus respectivos troncos(do inglês steam) e concatenados separados por espaços em branco, formando um novo texto.

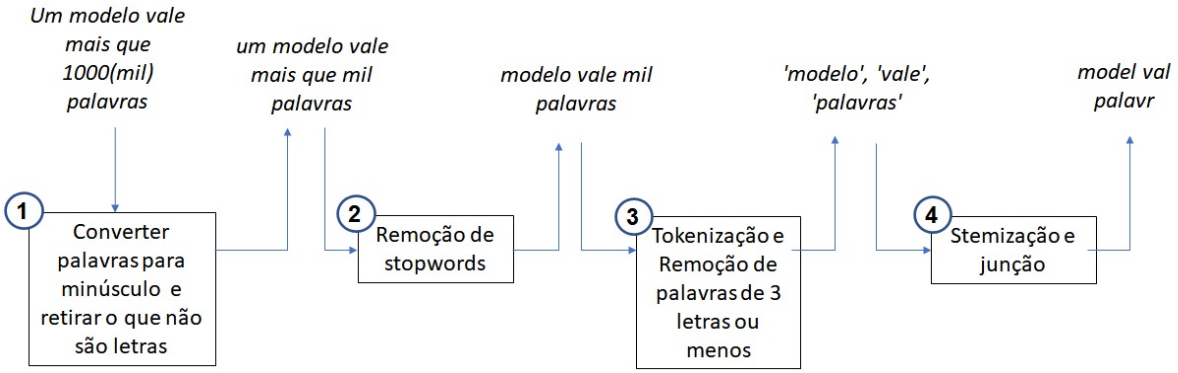


Figura 2.1: Pré-processamento nos dados textuais das PL.

2.2.2 Representação Vetorial de Documentos Textuais

Os algoritmos de Aprendizado de Máquinas (AM) necessitam, em sua grande maioria, de dados quantitativos para realizar suas operações. Por isso, é necessário que documentos textuais sejam, primeiramente, representados em vetores numéricos para que, posteriormente, possam ser utilizados pelos algoritmos de classificação. Na área de Processamento de Linguagem Natural (PLN) existem várias abordagens para tal fim. Essa seção apresenta algumas delas.

Bag-of-Words (BoW)

A técnica Bag-of-Words (BoW) consiste em um processo de transformação do texto em um vetor numérico baseado na presença, contagem ou frequência total das palavras (termos) de um determinado documento. A complexidade da técnica vem tanto em decidir como ocorrerá a formação do vocabulário de palavras conhecidas e como indicar a presença de palavras conhecidas no documento.

Se tratando de problemas onde não importa o posicionamento e o significado dos termos, a técnica Bag-of-Words (BoW) é extremamente simples e eficiente, sendo bastante utilizado em soluções de Processamento de Linguagem Natural (PLN), geralmente envolvendo classificação de texto. Uma das desvantagens desse método é que por ser baseada apenas na contagem e não levar em consideração o posicionamento das palavras, como já foi mencionado anteriormente, há uma perda semântica, ou seja, perda no sentido/significado das palavras. Além disso, se por algum motivo, o texto tenha uma dispersão enorme de palavras, o tamanho do vetor do vocabulário a ser criado também será grande, gerando um acentuado custo computacional.

Term Frequency–Inverse Document Frequency (TF-IDF)

O modelo Term Frequency–Inverse Document Frequency (TF-IDF) mede a importância de uma palavra para um documento com base em uma coleção (ou corpus). Isso implica dizer, que se, por exemplo, a palavra "Computação" aparece repetidamente dentro de um documento, não sendo tão frequente em outros, é um forte indicativo de que esta palavra é relevante para aquele entendimento.

A técnica TF-IDF é composta por dois cálculos: o primeiro computa a Frequência do Termo (TF) normalizada e o segundo computa a Frequência Inversa de Documentos (IDF). Definindo formalmente, considere D o conjunto de documentos (corpus) com um vocabulário de palavras de tamanho n. Seja $d_j \in D$, tal que $d_j = \{x_1, x_2, x_3, ..., x_n\}$, em que x_i denota a quantidade de ocorrências do i-ésimo termo (palavra) em d_j. Formalmente, o cálculo TF da i-ésima palavra é definido como:

$$TF(i) = \frac{x_i}{\sum_{k=0}^{n} x_k} \tag{2.1}$$

O IDF é incorporado para diminuir o peso das palavras que ocorrem mais frequentemente em D e aumentar o peso daquelas que ocorrem raramente. Formalmente, o cálculo IDF da i-ésima palavra é definido como:

$$IDF(i) = \log\left(\frac{|D|}{t_i + 1}\right) \tag{2.2}$$

em que |D| é a quantidade de documentos do corpus e t_i denota a quantidade de documentos em D que contém a i-ésima palavra.
A representação final de cada documento $d \in D$ é dada pelo produto TF e IDF de cada uma das n palavras do vocabulário, ou seja:

$$TF\_IDF(i) = TF(i) \times IDF(i) \tag{2.3}$$

para $i = 1, 2, ..., n$.

2.3 Seleção de Atributos/Características

A seleção de atributos (ou características) consiste em uma seleção de um subconjunto de atributos relevantes (variáveis preditoras) para uso em um modelo de Aprendizado de Máquinas (AM). As técnicas de seleção de atributos são usadas por vários motivos, dentre eles:

- Simplificar modelos para torná-los mais fáceis de interpretação [22];

- Obter tempos de teste/treinamento mais curtos [23];
- Evitar a maldição da dimensionalidade [24]

Existem muitas abordagens para selecionar o(s) atributo(s) no aprendizado de máquina [25]. A seguir são apresentadas algumas abordagens usadas no desenvolvimento deste trabalho.

2.3.1 Eliminação Recursiva de Características (RFE)

A Eliminação Recursiva de Características (*Recursive Feature Elimination* - RFE) é uma técnica de seleção de atributos iterativa e baseia-se na remoção progressiva dos atributos menos importantes em um conjunto de dados. Esta técnica começa com todos atributos disponíveis, então um modelo é treinado e avaliado. Em seguida, os atributos menos importantes são removidos, e o processo é repetido até que o número desejado de abritutos seja atingido.

Durante cada iteração, um estimador é treinado no conjunto de dados atual e os atributos são classificados com base em sua importância. Métodos comuns para essa classificação incluem a análise de coeficientes em modelos lineares ou a importância de atributos em modelos baseados em árvores de decisão. Os atributos menos importantes são removidos, e o processo continua até atingir o número desejado de características ou até que a performance do modelo não melhore significativamente.

A RFE oferece benefícios significativos, incluindo a capacidade de lidar com conjuntos de dados de alta dimensionalidade, melhorar a interpretabilidade dos modelos e reduzir o risco de overfitting. Essa técnica é particularmente útil em situações em que a dimensionalidade dos dados é um desafio, pois ajuda a simplificar o modelo, mantendo as características mais informativas.

No entanto, é essencial ter em mente que a eficácia da RFE pode depender do algoritmo de aprendizado de máquina escolhido e das características específicas do conjunto de dados. Em alguns casos, pode ser necessário ajustar os parâmetros da técnica para otimizar seu desempenho. Em resumo, a RFE é uma ferramenta valiosa, proporcionando uma abordagem sistemática e eficiente para aprimorar modelos de aprendizado de máquina.

2.3.2 Técnicas Baseadas na Redução da Dimensionalidade

As técnicas de redução de dimensionalidade oferecem uma maneira eficiente de reduzir o número de variáveis de entrada (dimensões) antes de aplicar modelos de Aprendizado de Máquina. Além disso, reduzir a quantidade de dimensões diminui os efeitos da Maldição

da Dimensionalidade. A Maldição da Dimensionalidade significa que, se a quantidade de dados para a qual treinar um modelo é fixa, o aumento da dimensionalidade pode levar a um ajuste excessivo. Este problema pode ser evitado trazendo exponencialmente mais dados para cada dimensão adicional.

Análise de Componentes Principais (Principal Component Analysis)

A Análise de Componentes Principais (PCA) é uma técnica estatística que visa a reduzir a dimensionalidade dos dados do conjunto, transformando-os subsequentemente em um conjunto de variáveis denominado de Componentes Principais, preservando ao máximo as informações originais. Em outras palavras, o método transforma ortogonalmente um conjunto de variáveis correlacionadas para um conjunto de valores de variáveis linearmente não correlacionadas(Componentes Principais).

Segundo [26], o objetivo do PCA é explicar a estrutura de variância e covariância entre variáveis através da construção de poucas combinações lineares das variáveis originais, e o que se deseja obter é a "redução do número de variáveis a serem avaliadas e a interpretação das combinações lineares construídas". Dessa forma, a informação contida nas variáveis originais é substituída pela informação contida nos k componentes principais não correlacionados.

Os Componentes Principais são ordenadas pela quantidade decrescente de variabilidade(variância) que explicam. Cada Componentes Principal é gerado para explicar o máximo de variabilidade da parte ainda não explicada, tendo que ser ortogonal às Componentes Principais anteriores. É importante notar que a PCA é sensível à escala dos dados, pelo que se recomenda a sua normalização prévia.

De acordo com [27], as etapas para aplicação do PCA consistem em:

1. Escolher variáveis(i.e., dimensões) a serem avaliadas;
2. Subtrair média de cada dimensão, produzindo um conjunto de dados com média zero;
3. Calcular matriz de covariância;
4. Calcular os autovetores e autovalores da matriz de covariância;
5. Escolher as k componentes principais(i.e., os k autovetores com maior autovalor);
6. Interpretar as informações contidas nas componentes principais.

Decomposição de Valor Singular(Singular Value Decomposition)

A Decomposição de Valor Singular[28] consiste em um processo de fatoração de matrizes capazes de decompor a matriz de documentos $D_{(m \times n)}$ no seguinte formato:

$$D = USV^T \tag{2.4}$$

em que $U_{(m\times m)}$ e $V_{(n\times n)}$ são duas matrizes ortogonais e $S_{(m\times n)}$ é uma matriz diagonal, e m é a quantidade de documentos do corpus e n a quantidade de palavras.

Através da decomposição da matriz D pela Eq. 2.4, é possível reconstruir a matriz D em um espaço p-dimensional (em que $p << n$) considerando a sub-matriz $(m \times p)$ formada pelas primeiras p colunas e as m linhas das respectivas matrizes originais U, S e V. Estas matrizes representam, respectivamente: uma base ortonormal[2] para as colunas de D (autovetores a esquerda); o conjunto de escalares que determinam a relevância de cada autovetor(autovalores); e uma base ortonormal para as linhas de D (autovetores a direita).

A técnica SVD é capaz ainda de "ordenar" a informação contida em D, tornando a "parte dominante" visível, uma vez que os autovetores estão ordenados de forma decrescente pela relevância definida pelos autovalores[29]. O SVD permite o descarte de de dados pouco discriminativos da matriz A, obtendo a melhor aproximação possível para A, considerando um número menor de dimensões linearmente independentes.

Análise Semântica Latente(LSA)

Análise Semântica Latente(LSA) é um mapeamento linear não supervisionado projetado para documentos de texto com base nas técnicas PCA ou SVD. Ele extrai relações entre palavras por meio de seus contextos de uso em documentos, passagens de texto ou sentenças[30]. A análise LSA consiste em quatro etapas principais[30]. Os dois primeiros passos são também usados em modelos de espaço vetorial. A Etapa 3, Redução de Dimensão, é a diferença chave do LSA.

1. Matriz Termo-Documento.
   Uma grande coleção de texto é representada como uma matriz Termo-Documento(Bag-of-Words).

2. Transformação da Matriz Termo-Documento.
   Em vez de trabalhar com frequências de termos brutos, as entradas no documento de termos são muitas vezes transformadas. Um exemplo de transformação é o TF-IDF (Seção 2.2.2).

3. Redução de Dimensão.
   A Decomposição de Valor Singular (Seção 2.4) é aplicada na matriz transformada, ela preserva as k dimensões mais relevantes e define o restante como zero.

[2]Conjunto de vetores, com norma igual a 1, linearmente independentes, capazes de gerarem todos os outros vetores de D.

4. Recuperação de informação no Espaço de Dimensão Reduzida.
   As semelhanças são calculadas entre entidades do novo espaço de dimensão reduzida, ao invés de obtê-las da matriz original de Termo-documento, pois tanto os documentos como termos são representados como vetores dentro do mesmo espaço, dessa forma, as semelhanças entre documento-documento, termo-termo e termo-documento são calculados diretamente. Além disso, termos e/ou documentos podem ser combinados para criar novos vetores dentro do espaço e podem ser comparados da mesma maneira. Por exemplo, em uma consulta para encontrar documentos semelhantes, um novo vetor de pesquisa é criado no centroide(ou seja, média ponderada) dos seus vetores de termo e então comparado aos vetores de documentos. Este processo em que novos vetores são adicionados ao espaço LSA é chamado 'folding-in'. O cosseno ou distância angular entre vetores é comumente usado como medida de semelhança nesta comparação, pois para muitas aplicações de recuperação de informações, ele mostrou-se eficaz, na prática.

2.4 Classificação de Dados

A classificação de dados insere-se dentro da Aprendizagem Supervisionada, cujo objetivo é o desenvolvimento de algoritmos capazes de generalizar dados novos e desconhecidos com base em um conjunto de dados previamente observados chamado de conjunto de treinamento. Os exemplos do conjunto de treinamento estão associados a um rótulo de classe (também conhecida como variável de resposta, ou variável dependente), que se refere à informação que se deseja predizer em novos exemplos. A aprendizagem supervisionada é utilizada para construir algoritmos capazes de predizer o rótulo de classe de novos exemplos com base nas características e nos rótulos de classe observados em outros exemplos similares.

O produto da aprendizagem supervisionada é um classificador que, essencialmente, é uma função capaz de mapear um conjunto de valores de características em um rótulo de classe, e é construído a partir da observação prévia de outros valores para as mesmas características e quais rótulos de classe estão associados a eles.

2.4.1 Modelos de Classificação Supervisionada

Esta seção descreve alguns modelos de classificação utilizados no desenvolvimento deste trabalho.

k-Nearest Neighbors (kNN)

O classificador k-NN funciona com base no princípio da proximidade dos elementos. No caso deste trabalho tem-se uma classificação binária (duas classes), em que um documento, que representa o texto de uma PL, pode ser do tipo "aprovado" ou "reprovado". Formalmente, considerando um conjunto de documentos $D = \{d_1, d_2, d_3, ..., d_m\}$, cada $d_j \in D$ é associado a um rótulo/classe (aprovado ou reprovado) através da seguinte função:

$$\delta(d_j) = \begin{cases} 1, & \text{se } d_j \text{ é um documento aprovado} \\ 0, & \text{caso contrário.} \end{cases} \tag{2.5}$$

Para classificar um novo documento d_q, o algoritmo kNN busca pelos k-documentos em D mais similares à d_q usando alguma medida de similaridade.

Considerando $D^k \subseteq D$ o conjunto dos k-documentos mais similares ao documento d_q, na versão clássica do algoritmo kNN, a classe do novo documento é definida a classe majoritária de seus k-vizinhos, ou seja:

$$kNN_{tradicional}(D^k, d_q) = \arg \max_{c \in \{0,1\}} \sum_{j=1}^{k} I(c, \delta(d_j)) \tag{2.6}$$

em que:

$$I(a, b) = \begin{cases} 1, & \text{se } a = b \\ 0, & \text{caso contrário.} \end{cases} \tag{2.7}$$

Na versão tradicional do algoritmo kNN (Eq. 2.6), a classe de cada vizinho possui o mesmo peso para o classificador. Entretanto, pode-se utilizar pesos diferentes para cada vizinho, dando mais pesos àqueles vizinhos mais próximos e menos peso para os mais distantes. Em outras palavras, a versão ponderada do KNN baseada em distância pode ser definida como:

$$kNN_{ponderado}(D^k, d_q) = \arg \max_{c \in \{0,1\}} \sum_{j=1}^{k} sim(d_q, d_j) \cdot I(c, \delta(d_j)) \tag{2.8}$$

Entretanto, tanto na abordagem kNN-tradicional (Eq. 2.6) quanto na abordagem kNN-ponderado (Eq. 2.8), o problema do desbalanceamento da base de dados continua sendo um problema. Isso significa que, em ambas as abordagens, a classe majoritária irá contribuir com mais vizinhos, tendenciando a definição da classe do novo documento. A Figura 2.2 mostra esse problema: note que dentre os k-vizinhos mais próximos do círculo amarelo existem mais quadrados azuis do que triângulos vermelhos, então o círculo

amarelo será classificado como pertencente à classe dos quadrados azuis, mesmo que dentre seus k-vizinhos estejam todos os triângulos vermelhos da base de dados.

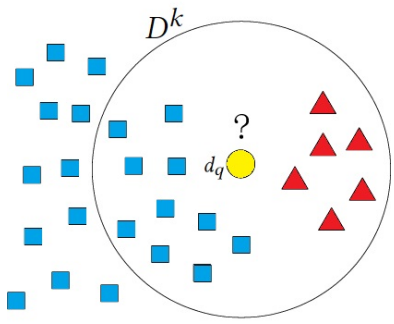


Figura 2.2: Algoritmo kNN.

XGBoost

O XGBoost é um algoritmo de Aprendizado de Máquinas (AM), baseado em árvore de decisão e que utiliza uma estrutura de Gradient Boosting. Gradient Boosting é uma técnica de aprendizado de máquina para problemas de regressão e classificação, que produz um modelo de previsão na forma de um ensemble de modelos de previsão fracos, geralmente árvores de decisão[31]. Ela constrói o modelo em etapas, como outros métodos de *boosting*, e os generaliza, permitindo a otimização de uma função de perda diferenciável arbitrária.

Os algoritmos de aprendizagem baseados em árvores de decisão são considerados um dos melhores e mais utilizados métodos de aprendizagem supervisionada, eles nos fornecem modelos preditivos de alta precisão, estabilidade e facilidade de interpretação. Porém, o *overfitting* é uma das maiores dificuldades para este tipo de modelo. *Overfitting* é um termo usado em estatística para descrever quando um modelo estatístico se ajusta muito bem ao conjunto de dados anteriormente observado, mas se mostra ineficaz para prever novos resultados.

Parar antecipadamente é uma abordagem para treinar modelos complexos de aprendizado de máquina para evitar *overfitting*, esta técnica funciona monitorando o desempenho do modelo que está sendo treinado em um conjunto de dados de teste separado e interrompe o procedimento de treinamento quando o desempenho no conjunto de dados de teste não melhora após um número fixo de iterações de treinamento. Ela evita o *overfitting* ao tentar selecionar automaticamente o ponto de inflexão onde o desempenho no conjunto de dados de teste começa a diminuir enquanto o desempenho no conjunto de dados de treinamento continua a melhorar conforme o modelo começa a se ajustar.

Regressão Logística

A regressão logística é uma técnica de aprendizado de máquinas usada em problemas de classificação binária. Essa técnica de análise de dados usa o modelo logístico para encontrar as relações entre dois fatores de dados. O modelo estatístico, por sua vez, modela a função Logit de um evento como uma combinação linear de uma ou mais variáveis independentes. A função Logit é o inverso da função logística padrão e, ao mesmo tempo, a função quantílica associada à sua distribuição. Portanto, a regressão logística estima os coeficientes na combinação linear(parâmetros) do modelo.
Na Regressão Logística Binária existe uma única variável dependente binária, codificada por uma variável indicadora, onde os dois valores são rotulados como '0' e '1', enquanto as variáveis independentes podem ser binárias ou contínuas. A probabilidade correspondente do valor rotulado como '1' pode variar entre 0 e 1 e a função que converte em probabilidade é a função logística.

Support Vector Machine(SVM)

O SVM é um algoritmo que identifica a Support Vector Machine, que é o melhor hiperplano/linha de separação entre duas classes distintas. Ou seja, encontra a fronteira que melhor segrega duas classes. O código identifica, para cada hiperplano, qual a distância dele para o elemento mais próximo de cada classe. Esse valor é denominado de margem. O objetivo de um SVM é encontrar o hiperplano de maior margem.

Floresta Aleatória

Floresta aleatória é um algoritmo de aprendizado de máquina, que combina a saída de múltiplas árvores de decisão, para chegar a um resultado único. Ele cria, aleatoriamente, árvores de decisão, formando uma floresta, onde cada árvore participará numa espécie de votação, utilizada na escolha do resultado. Os algoritmos de floresta aleatória possuem três hiperparâmetros principais: O tamanho do nó, o número de árvores e o número de recursos amostrados. A partir daí, o classificador de floresta aleatório pode ser usado para resolver problemas de regressão ou classificação.

2.4.2 Avaliação do Desempenho de Algoritmos de Classificação

Avaliar o desempenho de um modelo de classificação significa analisar o quão preciso é o modelo para predizer o rótulo de classe de novos exemplos. Um modelo de classificação ideal é aquele que classifica corretamente o rótulo de classe para todo e qualquer exemplo novo que seja fornecido, porém, na prática, para a maioria dos conjuntos de dados reais,

este cenário é bastante raro. Desta forma, o desempenho de um modelo de classificação também costuma ser medido pela sua taxa de erro de classificação.

A avaliação do desempenho de um classificador requer, além de um conjunto de exemplos de treinamento, um conjunto de exemplos de teste cujos rótulos de classes sejam conhecidos a princípio. Os exemplos de treinamento serão usados para construir o modelo e calibrar a função de decisão, enquanto os exemplos de teste serão utilizados para que o usuário obtenha informações sobre o quão preciso é o modelo desenvolvido para predizer o rótulo de classe de exemplos novos.

As medidas clássicas de avaliação de classificadores se derivam de uma tabela chamada de Matriz de Confusão, que contém a quantidade de classificações corretas contra as classificações preditas para cada classe sobre um conjunto de exemplos, ou seja, ela indica os erros e acertos do modelo comparando com os resultados esperados. Para cada classe é realizada a extração de quatro variáveis:

- TP: verdadeiro positivo, representa o número de exemplos positivos classificados corretamente;
- TN: verdadeiro negativo, representa o número de exemplos negativos classificados corretamente;
- FP: falso positivo, representa o número de exemplo negativos classificados incorretamente;
- FN: falso negativo, representa o número de exemplos negativos classificados incorretamente

A partir destas quatro variáveis definem-se várias métricas de avaliação. Algumas delas utilizadas neste trabalho são:

- Precisão = $\frac{TP}{TP+FP}$
- Revocação = $\frac{TP}{TP+FN}$
- F1-Score = $\frac{2 \times Precisao \times Revocacao}{Precisao+Revocacao}$
- Acurácia = $\frac{TP+TN}{TP+TN+FP+FN}$

A Precisão é a razão entre a quantidade de verdadeiros positivos e o total de estimativas de aprovação, ou seja, é uma medida de eficiência na inferência de aprovação. Ele é importante porque, um classificador ruim poderia inferir uma quantidade muito grande de aprovações equivocadas, o que levaria o usuário a um sobre-esforço de acompanhamento das proposições. No cálculo da Revocação, por sua vez, a quantidade de verdadeiros positivos é dividido pela quantidade de casos rotulados como aprovados, o que denota a

eficácia da inferência de aprovação. Neste caso, um classificador ineficaz acertaria poucas aprovações, o que deixaria o usuário inseguro sobre a utilidade do mesmo. O F1-Score, que é uma média harmônica entre Precisão e Revocação, portanto, leva em consideração tanto a eficácia quanto a eficiência e está muito mais próxima dos menores valores do que uma média aritmética simples. Ou seja, ter um F1-Score baixo, é um indicativo de que ou a precisão ou a revocação estão baixos.

No contexto de dados desbalanceados é importante considerar medidas de desempenho que permitam uma visão específica destes casos. Segundo [32], a medida utilizada para avaliar classificadores em um ambiente desbalanceado deve ser selecionada individualmente, pois cada um desses problemas vem com seu próprio conjunto de desafios, ou seja, as métricas clássicas não são um meio confiável de avaliar um modelo treinado em dados desbalanceado.

2.4.3 Validação Cruzada

A validação cruzada é uma técnica para avaliar a capacidade de generalização de um modelo, a partir de um conjunto de dados [33]. A Figura 2.3 mostra um exemplo de Validação Cruzada k-fold, para k = 5. No método k-fold, o conjunto de dados D é dividido de forma aleatória em k subconjuntos mutuamente exclusivos de igual tamanho, em seguida, um subconjunto é utilizado para teste e os k-1 restantes são utilizados para estimação dos parâmetros. Este processo altera de forma circular o subconjunto de teste k vezes. A cada volta são calculadas os valores de métricas e no final suas médias se tornam mais confiáveis.

O método *leave-one-out* é um caso específico do k-fold, com k igual ao número total de dados N. A cada volta do *loop*, apenas um dado é separado como teste e todos os demais como treinamento, até que todos os dados participem como teste.

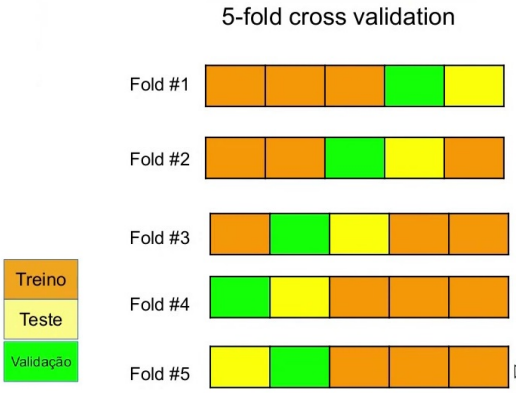


Figura 2.3: Validação Cruzada de 5 partições.

2.5 Desbalanceamento de Dados

Alguns modelos de classificação são projetados para trabalhar com conjuntos de dados balanceados. Um conjunto de dados é dito balanceado quando a quantidade de amostras para todas as classes possíveis é igual ou diferente em apenas uma pequena porcentagem, de maneira que todas as classes estejam igualmente representadas por suas distribuições. Entretanto, em problemas do mundo real, o provável é que os dados disponíveis para análise estejam desbalanceados. Existem vários motivos que levam a este desbalanceamento, por exemplo, em diversas ocasiões comuns, a obtenção de dados de uma determinada classe pode:

- estar condicionada a um evento de ocorrência rara, de maneira que amostras de uma determinada classe são muito mais frequentes do que para outras classes;
- ser bastante cara do ponto de vista econômico, computacional e/ou de tempo, requerendo recursos que podem não estar facilmente disponíveis.

Alguns problemas surgem ao se trabalhar com dados desbalanceados. O primeiro problema trata-se do viés que modelos de classificação costumam apresentar neste cenário em direção aos dados da classe majoritária (aquela que possui a maior quantidade de amostras), uma vez que, quando treinados por conjuntos de dados desbalanceados, os modelos de classificação tenderão a expressar um bom desempenho de classificação para as amostras da classe majoritária e um desempenho menor para amostras da classe minoritária. Diversos estudos na literatura mostram que esta queda de desempenho ocorre não necessariamente devido à diferença na representação das classes em si, mas devido a outros fatores inerentes ao desbalanceamento, como a presença de pequenos disjuntos (small disjuncts), baixa densidade de dados, sobreposição de dados e outros[34].

Outro problema enfrentado ao lidar com conjuntos de dados desbalanceados trata-se da seleção de métricas de desempenho adequadas para verificar o desempenho de classificadores. Métricas de desempenho baseadas em verdadeiros e falsos positivos e negativos, como acurácia, não são apropriadas para avaliar modelos de classificação desenvolvidos sobre dados desbalanceados, uma vez que grandes números de verdadeiros positivos tenderão a acobertar números elevados de falsos positivos e vice-versa, de maneira que um classificador que rotula muito bem amostras da classe majoritária e rotula de maneira mediana amostras da classe minoritária ainda apresentará um bom valor de acurácia.

A maioria dos classificadores enfrenta sérios problemas em um contexto onde há um desbalanceamento na distribuição de classes. Atualmente, existem na literatura muitas abordagens para mitigar problemas de classificação envolvendo conjuntos de

dados desbalanceados [35], que vão desde técnicas mais básicas de subamostragem e sobreamostragem do conjunto de dados até métodos computacionalmente mais sofisticados que combinam redes neurais com modelos *ensemble* e alcançam bons resultados [36]. Métodos de classificação baseados ao nível algorítmico e funções de custo sensíveis também são amplamente utilizados para lidar com problemas relacionados a conjuntos de dados desbalanceados [37].

2.5.1 Abordagens para o Tratamento do Desbalanceamento de Dados

Para mitigar o problema do desbalanceamento de classes, várias estratégias foram propostas na literatura [6]. Essas estratégias podem ser agrupadas em duas grandes categorias: abordagens orientadas a dados e orientadas a algoritmos. A seguir são apresentados alguns conceitos e técnicas dessas duas categorias para o balanceamento de dados.

Métodos Baseados em Amostragem

Métodos baseados em amostragem são os métodos mais simples e mais utilizados para o balanceamento de conjuntos de dados. Esses métodos consistem em re-amostrar os conjuntos de dados para modificar as distribuições de dados para criar um "novo" conjunto de dados balanceado. Para isso, existem duas metodologias particulares: sobreamostragem e subamostragem. O primeiro método visa expandir a classe minoritária e pode causar overfitting devido à criação de várias instâncias semelhantes. A segunda visa reduzir a classe majoritária, o que por sua vez pode resultar na perda de informações relevantes [38].

Existem várias abordagens e implementações de métodos de amostragem, tais como:

- Subamostragem informada: os algoritmos como NearMiss-(1 & 2 & 3) executam subamostragem usando um classificador kNN, selecionando amostras com base na distância das amostras da classe majoritária para as amostras da classe minoritária. A técnica One-Sided Selection (OSS) [39] realiza uma heurística que limpa previamente o conjunto de dados removendo amostras ruidosas e, em seguida, um classificador de 1 vizinho mais próximo é aplicado a todas as amostras;
- Amostragem Sintética com Geração de Dados (sobre-amostragem artificial): esta técnica visa criar novas amostras artificiais da classe minoritária para equilibrar o conjunto de dados. O algoritmo comumente usado é a técnica Synthetic Minority

Oversampling Technique (SMOTE), que usa um algoritmo não supervisionado que cria novas amostras com base na distância entre as amostras;

- Combinação de sobre-amostragem e subamostragem: uma abordagem é realizar sobre-amostragem para gerar amostras sintéticas da classe menor e, em seguida, executar subamostragem para limpar o espaço resultante de amostras ruidosas. A etapa de oversampling é feita usando SMOTE e a limpeza pode ser feita usando o link de Tomek (SMOTETomek) ou vizinhos mais próximos editados (SMOTEENN).
- Métodos Ensemble: EasyEnsemble é uma metodologia não supervisionada que usa subconjuntos aleatórios da classe majoritária. Ele subamostra aleatoriamente o conjunto original para criar um conjunto de dados de conjunto. Outra abordagem bem utilizada é BalanceCascade, um algoritmo supervisionado que iterativamente cria balanceamento e extrai amostras redundantes na classe majoritária para formar um classificador final. Isso difere do método anterior, pois usa um classificador para garantir que as amostras mal classificadas possam ser selecionadas novamente para o próximo subconjunto.

Métodos Baseados ao Nível Algorítmico

Enquanto os métodos baseados em amostragem modificam estruturalmente o conjunto de dados de maneira a equilibrar a quantidade de amostras de cada classe, a ideia da abordagem de balanceamento baseada em nível algoritmo é assumir custos maiores para erros de classificação de amostras da classe minoritária e incorporar tais custos no processo de aprendizado dos algoritmos de classificação já conhecidos. Geralmente, os custos para classificações erradas são definidos em matrizes de custo, cujos valores podem variar para cada domínio e de acordo com opiniões de especialistas. Do ponto de vista dos algoritmos, as estratégias adotadas incluem principalmente métodos sensíveis ao custo e de aprendizagem em conjunto.

Métodos Sensíveis ao Custo

A ideia do aprendizado sensível ao custo é atribuir um alto custo de erro de classificação às classes minoritárias e um baixo custo de erro de classificação às classes majoritárias, e então minimizar o custo total de erro de classificação como objetivo de treinamento e otimização do algoritmo, de modo a melhorar o desempenho da classificação das classes minoritárias. O principal método de aprendizado direcionado sensível ao custo é mudar a estrutura interna do classificador tradicional introduzindo um fator sensível ao custo e transformar o objetivo de treinamento tradicional, de reduzir a taxa de erro de classificação global, pelo objetivo de reduzir o custo do erro de classificação global.

O aprendizado direcionado sensível ao custo torna o algoritmo de classificação tradicional adequado para dados desbalanceados, modificando a estrutura interna do classificador. No entanto, como alguns algoritmos são difíceis de modificar diretamente, o conceito de meta aprendizado sensível ao custo foi proposto. Com a premissa de não alterar o algoritmo de aprendizado existente, o Meta Aprendizado Sensível ao Custo converte o custo de classificação incorreta da classe no peso da amostra. Ele pondera cada amostra conforme o custo de classificação incorreta da classe à qual pertence e, em seguida, reconstrói o conjunto de dados original conforme o peso. A desvantagem é que o processo de reconstrução altera a distribuição das amostras e às vezes perde algumas informações úteis.

Método de Aprendizagem em Conjunto

A ideia principal do método de aprendizagem em conjunto é treinar classificadores bases e depois coordenar seus resultados para obter o resultado. O Aprendizado em Conjunto (Ensemble Learning) é utilizado para classificação de dados desbalanceados, principalmente devido ao seu conceito de integração. Ele combina os resultados dos classificadores base para lidar com a tarefa de classificação de dados desbalanceados. Os métodos de aprendizado em conjunto podem ser divididos em tipos heterogêneos e homogêneos, conforme a relação de categoria entre os classificadores básicos [40].

O aprendizado de conjunto heterogêneo refere-se ao uso de uma variedade de classificadores diferentes para integração, enquanto o aprendizado de conjunto homogêneo refere-se à integração do classificador base com o mesmo tipo de classificador com parâmetros diferentes. Diferentes métodos de integração têm suas próprias vantagens na classificação de dados desbalanceados.

Capítulo 3

Materiais e Métodos

Este capítulo apresenta o procedimento adotado para o desenvolvimento deste trabalho que, sistematicamente, se constitui em quatro etapas consecutivas:

1. Entendimento do Negócio;
2. Obtenção e Tratamento dos Dados;
3. Modelagem dos Dados;
4. Cenários para Análise dos Dados.

A primeira etapa compreende o entendimento dos dados e informações trabalhadas no desenvolvimento desse trabalho. As últimas etapas apresentam as técnicas e cenários utilizados para realizar e avaliar os testes experimentais, que serão descritos no próximo capítulo.

3.1 Entendimento do Negócio

O processo legislativo compreende a elaboração, análise e votação de vários tipos de propostas: leis ordinárias, medidas provisórias, emendas à Constituição, decretos legislativos e resoluções, entre outras, ou seja, a tramitação de um projeto de lei (ou emenda) é o processo que vai desde a sua apresentação até sua discussão e aprovação, ou arquivamento. O site da Câmara[1] do governo Brasileiro apresenta um modelo do processo legislativo interno da Câmara, que está esquematizado na Figura 3.1.

A execução do processo legislativo é comumente chamada de tramitação. Diz-se que uma PL está tramitando quando seu processo está em curso. São definidas várias situações para representar os pontos nos quais as PLs se encontram em um determinado ponto do

[1] https://www.camara.leg.br/entenda-o-processo-legislativo/

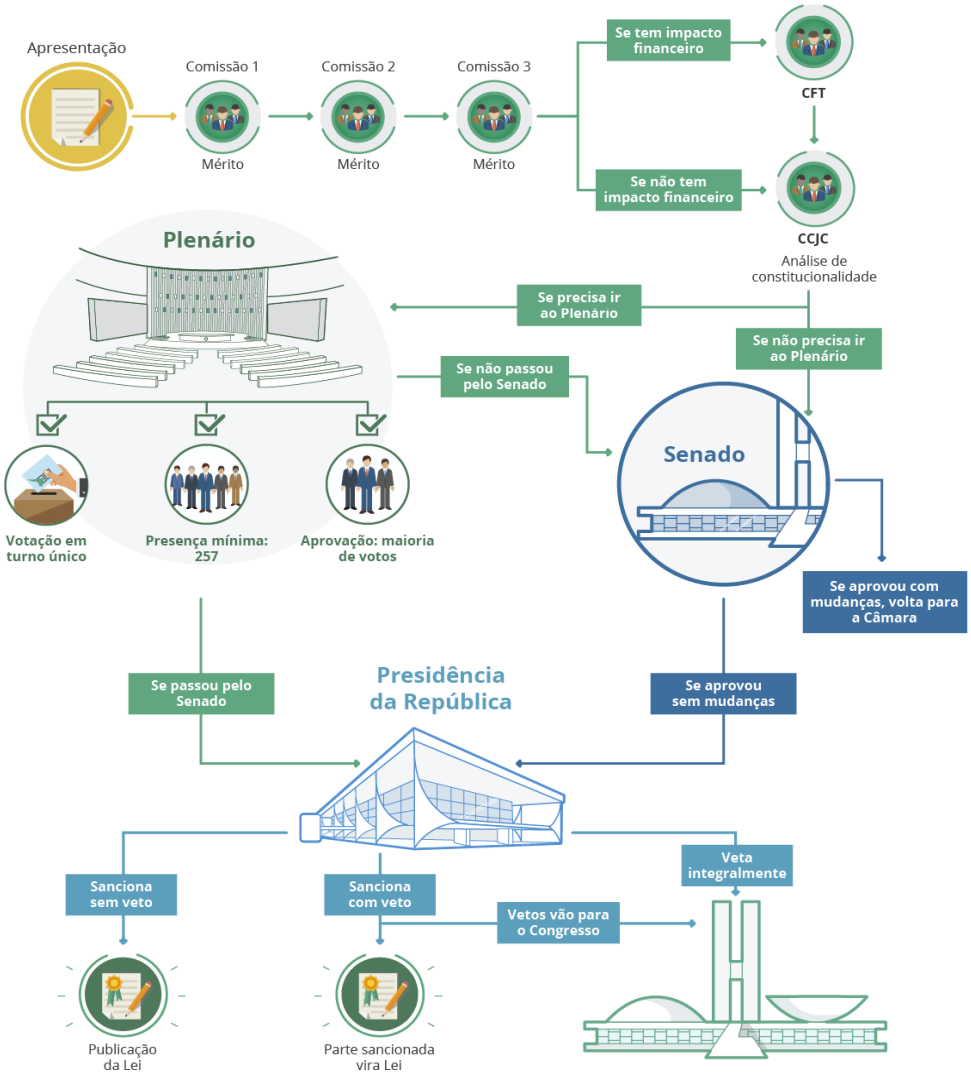


Figura 3.1: Modelo de Tramitação - Câmara do Deputados.

tempo e os trâmites representam as ações que as levam de uma situação para outra, tal como ilustrado na Figura 3.2.

O processo legislativo modelo é o referente à elaboração de lei ordinária, que resumidamente consiste na apreciação de um projeto por uma das câmaras (Casa Iniciadora) e sua revisão pela outra câmara (Casa Revisora) [41]. Entretanto, segundo o que afirma [42], é comumente dito que, a Câmara dos Deputados funciona como a casa iniciadora e o Senado Federal, como revisora. Tal ideia decorre dos mandamentos constitucionais, pois em casos específicos, a Constituição Federal determina que Câmara seja a casa iniciadora e o Senado, a revisora. No que se refere ao papel da Câmara dos Deputados, enquanto casa iniciadora, sua participação se dá nas discussões e votações dos projetos, uma vez que ao Senado, enquanto à casa revisora, compete o papel de 'revisar'

Figura 3.2: Modelo Simplificado de Tramitação.

a decisão de aprovação tomada pela primeira casa.

Um Projeto de Lei Ordinária começa a tramitar na Câmara dos Deputados, exceto quando são apresentados por senador ou comissão do Senado, ele é então distribuído pelo presidente da Câmara dos Deputados para as comissões temáticas que analisaram seu mérito. As análises das Comissões de Finanças e Tributação (CFT) e de Constituição e Justiça e de Cidadania (CCJC) são chamadas de admissibilidade, por arquivarem as proposições inadequadas ao Orçamento ou inconstitucionais. Se forem aprovadas por algumas comissões e rejeitados por outras, vão para o Plenário. Para serem aprovadas no Plenário, é necessário a presença de quórum mínimo de 257 deputados e maioria simples dos votos, em turno único. Nesse passo, pode haver aprovação com destaques a serem votados posteriormente .

Uma proposta de Emenda a Constituição (PEC), por sua vez, pode ser apresentada por no mínimo 171 deputados ou 27 senadores (1/3 do total), pelo presidente da República e por mais da metade das assembleias legislativas. Ela começa a tramitar na Comissão de Constituição e Justiça e de Cidadania (CCJC), que analisa a admissibilidade da proposta. Caso aprovada, o mérito da PEC é analisado por uma comissão especial, que pode alterar a proposta original. A proposta é então, analisada pelo Plenário, onde é votada em dois turnos. A aprovação depende dos votos favoráveis de 3/5 dos deputados (308), em dois turnos de votação. São votados eventuais destaques e a PEC é enviada para a outra casa. Se o texto for aprovado nas duas Casas sem alterações, é promulgado em forma de emenda constitucional em sessão do Congresso Nacional.

Sobre a tramitação é importante entender o significado de Legislatura, o período de quatro anos durante o qual se desenvolvem as atividades legislativas (Constituição Federal, art. 44), que coincide com a duração do mandato dos deputados. Começa em 1º de fevereiro do ano seguinte à eleição e termina em 31 de janeiro após a eleição seguinte. Antes da aprovação do substitutivo ao Projeto de Resolução de Alteração do Regimento Interno (PRC) 190/01, ao final de cada legislatura o parlamentar reeleito fazia um requerimento de desarquivamento das proposições de sua autoria que considerava importantes. As novas regras levam em conta o tempo de tramitação da proposição, em vez da autoria, a principal determina o arquivamento de uma proposição após tramitação por, pelo menos,

três legislaturas completas.

3.2 Obtenção e Preparação dos Dados

A Lei n° 12.527/2012, denominada Lei de Acesso à Informação, regulamentou o direito constitucional de acesso às informações públicas. Ela determinou a divulgação de informações, com o intuito de criar mecanismos que, possibilitem a sociedade acompanhar informações públicas dos órgãos e entidades. Desde então houve notório crescimento da disponibilização de Dados Abertos Governamentais no Brasil, porem com poucas pesquisas de Ciência de Dados derivadas.

Para cada ano, o site de dados abertos da Câmara dos Deputados[2] disponibiliza um arquivo de dados contendo informações das PLs apresentadas naquele ano. Os dados destes arquivos são atualizados diariamente. Alguns dados associados às PLs são: Tipo, Ementa, Regime de Tramitação, Palavras-chave, Data de Apresentação, Órgão, Inteiro Teor, Último Status, Última Situação e Temas. Além disso, a Câmara também disponibiliza arquivos com dados relacionados às Proposições, tais como: Deputados, Frentes Parlamentares, Partidos, etc.

Para a realização desse trabalho foram considerados dados referentes aos anos de 2001 a 2021 e foram selecionadas apenas as proposições dos tipos Proposta de Lei e Proposta de Emenda e com trâmite concluído, ou seja, com situação aprovada ou arquivada/rejeitada. Estes dados foram obtidos a partir de duas fontes de dados do site da Câmara dos Deputados: arquivos CSV e API.

Arquivos CSV

- proposicoes-{ano}.csv: Cada tupla representa uma PL daquele ano. Aqui incluem-se atributos links para baixar arquivo PDF donde se obtêm o inteiro teor.
- proposicoesAutores-{ano}.csv: cada tupla representa um autor de uma PL daquele ano.
- proposicoesTemas-{ano}.csv: cada tupla corresponde a uma área temática na qual uma proposição foi classificada pelo Centro de Documentação e Informação da Câmara.

API

- /proposicoes/{id}/tramitacoes: cada invocação ao *end point* retorna uma lista com os dados das tramitações da PL identificado pelo id. Cada tramitação só foi importada uma vez. O fato das PLs poderem ser tramitadas da mesma forma

[2]https://dadosabertos.camara.leg.br/ swagger/api.html#staticfile

mais de uma vez foi, portanto, desprezado no processo. Importante ressaltar que, a tramitação conclusiva do trâmite de cada PL foi excluída do atributo 'tramitação', ou seja, foram excluídas aquelas tramitações que representam arquivamento, a saber: 'Devolução ao autor', 'Arquivamento', 'Vetado totalmente' e 'Retirada pelo Autor' assim com a que representa aprovação: 'Transformação em Norma Jurídica' e 'Transformado em Norma Jurídica com Veto Parcial'. Caso contrário, ter-se-ia campos dummy correspondentes a essas tramitações com valores iguais a 1. Nesse cenário, os classificadores tenderiam a acertar 100% de suas estimativas, pois ter-se-ia uma relação direta entre o valor desses campos e a aprovação ou arquivamento da PL. A 'apresentação da proposta' também foi retirada do atributo por inutilidade, afinal todas as proposições teriam o valor 1 nesse atributo.

A Figura 3.3 mostra esquematicamente os dados das PLs que foram obtidos diretamente do site da Câmara e utilizados neste trabalho. As informações categóricas de cada PL foram binarizadas, ou seja, os valores qualitativos foram representados de forma numérica com valores 0 (zero) ou 1 (um), criando-se assim o conjunto de atributos *dummy*, representados em azul na Figura 3.3. Exemplificando, 'tipo de projeto' era um atributo categórico com domínio: 'projeto_lei', 'projeto_emenda', 'projeto_lei_complementar' e 'projeto_lei_conversao'. No processo de binarização, ele foi substituído pelos atributos *dummy*: 'projeto_lei', 'projeto_emenda', 'projeto_lei_complementar' e 'projeto_lei_conversao'. Sendo assim, uma PL com valor 'projeto_lei' no atributo 'tipo de projeto', passou a ter o valor 1 no atributo 'projeto_lei' e zero nos atributos: 'projeto_emenda', 'projeto_lei_complementar' e 'projeto_lei_conversao'. Os conjuntos de campos *dummy* gerados a partir dos campos categóricos: tramitacao, autor e tema, estão listados, respectivamente, nos anexos I.1, II.1 e III.1.

A aprovação (ou reprovação) de uma PL foi definida a partir da informação do seu 'último status': foram consideradas aprovadas aquelas PLs com valor 1 no atributo 'Transformado em Norma Jurídica', já aquelas PLs com os valores iguais a 1 nos atributos 'Devolvida ao Autor', 'Retirado pelo Autor', 'Vetado totalmente' e 'Arquivada' foram consideradas reprovadas.

Os dados da PL foram gravados em uma base MySQL e tratados via comandos SQL. Nos casos onde não existia uma padronização nos dados, foi definido manualmente um valor padrão e os valores equivalentes foram atualizados. Por exemplo, no atributo que corresponde ao autor da PL o 'Partido Comunista do Brasil' está representado como: 'PCdoB' ou 'PC do B' e para padronizar esses valores foi adotado a sigla 'PCdoB'. Os dados ausentes foram preenchidos como 'ZZ-Indeterminado'. Outro exemplo é o atributo 'Regime', nele, entre outros casos, a 'Urgencia' estava representado como: 'Urgência (Art.

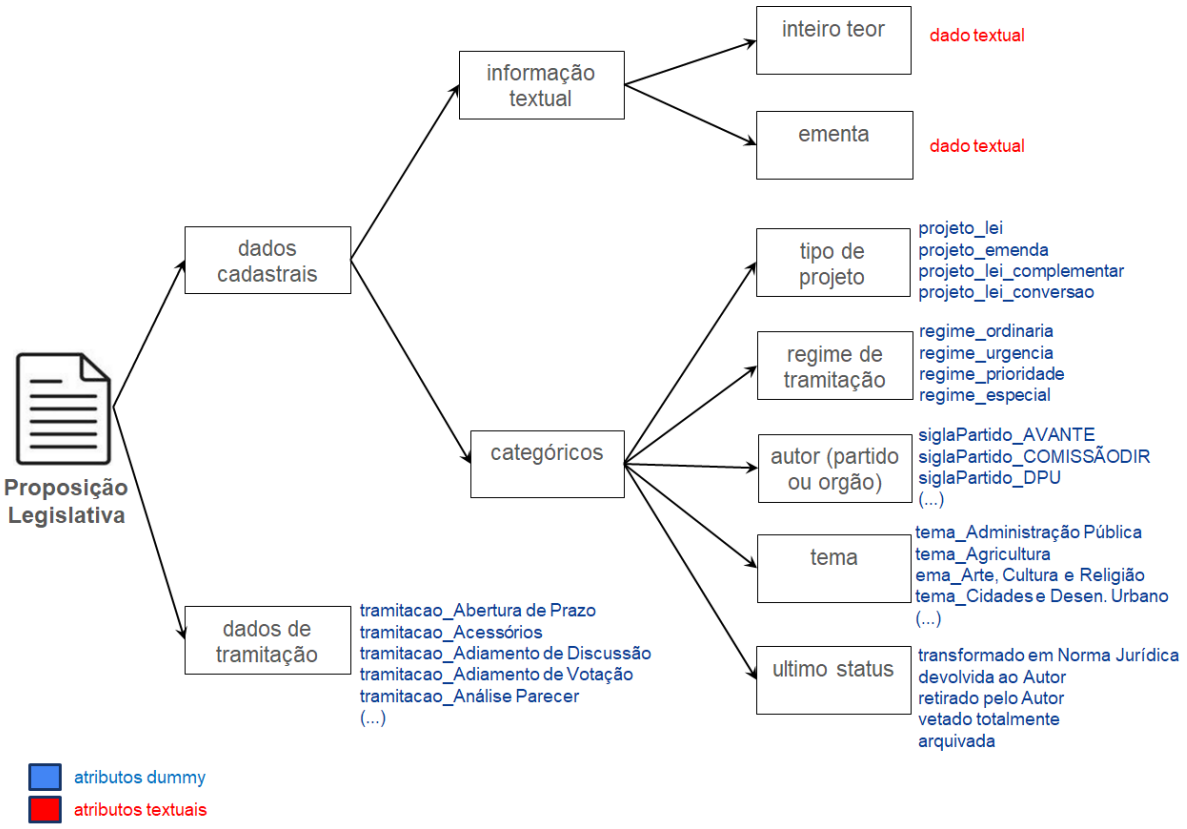


Figura 3.3: Atributos das PL.

64, CF)', 'Urgência (Art. 155, RICD)', 'Urgência (Art. 154, RICD)' ou 'Urgência', em todos esses casos foi adotado o padrão 'urgencia'.

Os atributos textuais, por sua vez, foram pré-processados em quatro etapas, conforme o processo da Figura 2.1. Essa 'limpeza' e padronização nos textos é uma abordagem realizada no PLN para garantir a qualidade no processamento de documentos textuais e ela contribui, entre outras coisas, para diminuir o dicionário gerado, uma vez que algumas palavras serão suprimidas e/ou codificadas em um mesmo padrão.

3.3 Modelagem dos Dados

A partir da base completa foram geradas bases intermediárias para cada período de Legislatura, conforme apresentado na Tabela 3.1. Essa divisão das proposições em períodos legislativos permite gerar previsões baseadas na composição da câmara que reflitam as tendências (políticas, econômicas, financeiras e culturais) da época. Por exemplo, é de se esperar que proposições de temas com assuntos atuais, como direito da mulher e LGBTQI+, sejam os assuntos mais citados nas proposições aprovadas em período mais recentes, do que em períodos mais antigos, por exemplo, em 2001.

Base	Período	Proposições			Aprov(%)
		Reprovadas	Aprovadas	Total	
BD_Completa	2001-2021	25.753	2.296	28.049	8,2
BD_Leg56	2019-2021	1.343	251	1.594	15,7
BD_Leg55	2015-2018	4.067	378	4.445	8,5
BD_Leg54	2011-2014	4.623	348	4.971	7,0
BD_Leg53	2007-2010	5.865	543	6.408	8,5
BD_Leg52	2003-2006	6.716	557	7.273	7,7
BD_Leg51	2001-2002	3.139	219	3.358	6,5

Tabela 3.1: Bases de Dados

Para avaliar o desempenho dos classificadores, as bases intermediárias foram divididas em:

- Base de Treinamento: dados correspondentes ao período legislativo de 2001 a 2018;
- Base de Teste: dados correspondentes ao período legislativo de 2019 a 2021.

A taxa de desbalanceamento (reprovadas por aprovadas) na base de teste é de 5,35 e na a base de treino de 11,93.

Balanceamento dos Dados

Foram criados e avaliados 3 conjuntos de dados para investigar como o desbalanceamento poderia influenciar na predição dos classificadores, a saber:

1. Sem reamostragem: base desbalanceada.
2. Subamostragem: as PLs da classe majoritária (reprovadas) são removidas até que se tenha uma base balanceada.
3. Sobreamostragem: as PLs da classe minoritária (aprovadas) são duplicadas até que se tenha uma base balanceada.

A subamostragem e sobreamostragem foram realizadas apenas na base de treinamento e a base de teste foi preservada sem balanceamento, para todas as técnicas avaliadas.

Seleção de hiperparâmetros

A seleção dos hiperparâmetros ocorreu em duas etapas: revisão da literatura e busca exaustiva. A primeira etapa teve o objetivo de identificar os principais hiperparâmetros que são utilizados em cada técnica e qual a faixa de variação para os hiperparâmetros contínuos. A Tabela 3.2 apresenta a relação dos classificadores utilizados e seus hiperparâmetros, com o domínio utilizado nos testes. Na busca exaustiva, foram geradas

10 combinações de valores dos hiperparâmetros e atribuído um número de versão para cada uma delas. Foram implementados dois testes: uma classificação usando as bases de treinamento e testes e uma validação cruzada k-fold com k=5 na base de treinamento, para cada estratégia de balanceamento. Para os hiperparâmetros não mencionados, foi mantido o padrão da biblioteca no Python.

Classificador	**Hiperparâmetros**
XGBoost	use_label_encoder: False early_stopping_rounds: 5 ou 10 eval_metric: 'error', 'logloss1, 'auc' ou 'rmse' max_depth: 2, 3, 4 ou 5 subsample: 0,4, 0,5 ou 1 scale_pos_weight: 11,5(apenas base desbalanceada)
Regressão Logística	penalty: 'l1' e l2 dual: False C = 0,4, 0,5, 0,6, 0,7, 0,8 ou 0,9 fit_intercept: True ou False solver: 'lbfgs', 'newton-cg' ou 'liblinear'
kNN	weights: 'uniform' ou 'distance' algorithm: 'ball_tree', 'auto', 'kd_tree' ou 'brute' leaf_size: 25, 30, 35 ou 40 p: 1 ou 2 k: 12, 14, 15, 17 ou 20
SVM	C: 0,5, 1,0 e 1,5 kernel: 'rbf', 'linear', 'poly' ou 'sigmoid' degree: 2, 3 ou 4 gamma: 'scale' ou 'auto' coef0 = 0,0 ou 0,5
Floresta Aleatória	bootstrap: True ou False n_estimators: 100, 150, 200 ou 250 random_state: None, 1 ou 2 criterion: 'gini' ou 'entropy' max_depth = None, 6 ou 8

Tabela 3.2: Domínio dos Hiperparâmetros dos Classificadores

Modelagem dos Dados Textuais

Depois de pré-processados, os dados textuais das PLs foram vetorizados utilizando-se a abordagem TF-IDF (Seção 2.2.2) que mede a importância de uma palavra para um documento com base em uma coleção (ou corpus).

Entretanto, a vetorização de cada documento através do modelo TF-IDF gerou vetores esparsos n-dimensionais, em que n é a quantidade de palavras do corpus (vocabulário). Isso significa que, para cada documento, o vetor gerado foi de alta dimensionalidade

e, por causa disso, para não comprometer a eficiência dos testes experimentais, a dimensionalidade do vetor foi reduzida utilizando a técnica Singular Value Decomposition (SVD) (Seção 2.4)

3.4 Cenários para Análise dos Dados

Para compreender e avaliar o poder preditivo dos atributos das PLs na sua aprovação/reprovação, foram implementados e avaliados 5 (cinco) cenários, que serão detalhados a seguir.

3.4.1 Cenário n° 1: Predição baseada nos dados textuais

A Figura 3.4 mostra esquematicamente o fluxo da arquitetura desenvolvida para avaliar a aprovação das PLs utilizando apenas seus dados textuais. As setas em amarelo referem-se a geração do modelo, enquanto em verde a aplicação na produção. O primeiro passo é a representação vetorial dos textos, utilizando a abordagem TF-IDF. Em seguida, para reduzir a dimensionalidade dos vetores gerados, foi aplicada a técnica Singular Value Decomposition (SVD).

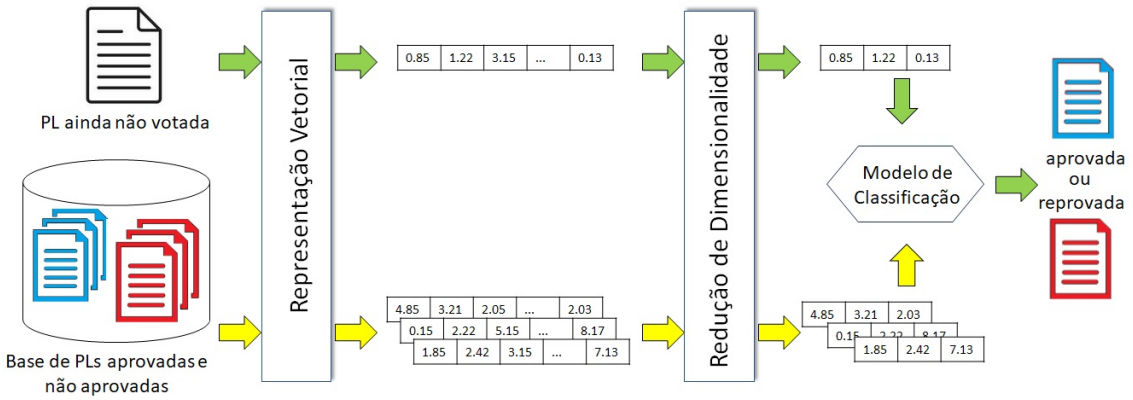


Figura 3.4: Arquitetura implementada - Cenário n° 1.

Técnicas Avaliadas

Uma vez que os documentos das PLs estejam representados em vetores compactos, o próximo passo foi utilizar classificadores para prever a aprovação (ou reprovação) desses documentos. Para isso, comparou-se o desempenho dos seguintes versões:

- $kNN_{tradicional}$ (Eq. 2.6)
- $kNN_{ponderado}$ (Eq. 2.8)

- $XGBoost$
- $XGBoost_{balanceado}$

A diferença do $XGBoost_{balanceado}$ para o $XGBoost$ é apenas a configuração do hiperparâmetro 'scale_pos_weight' que trata desbalanceamento nos dados.

Além dessas quatro versões de classificadores foi desenvolvida uma nova proposta, denominada de $kNN_{proposto}$ que modifica o $kNN_{ponderado}$ para:

1. lidar com o desbalanceamento da base de dados
2. ponderar documentos de PLs que são mais próximos no tempo

A proposta do $kNN_{proposto}$ para lidar com o desbalanceamento consiste em utilizar um fator de desbalanceamento, aumentando a distância dos k-vizinhos pertencentes à classe majoritária, que no caso deste trabalho se refere à classe das proposições reprovadas. Esse fator de desbalanceamento é uma constante "α" multiplicada na distância de similaridade dos documentos $d_j \in D$ que pertencem à classe majoritária:

$$fator_{desbalanceamento}(d_j) = \begin{cases} 1, & \text{se } \delta(d_j) = 1 \\ \alpha & \delta(d_j) = 0 \end{cases} \tag{3.1}$$

Para ponderar os documentos mais próximos no tempo, foi definido um fator temporal para mensurar o quão distante dois documentos estão no tempo, ou seja, o quão distante duas propostas foram apresentadas na Câmara. Em outras palavras, o fator temporal entre dois documentos d_i e d_q é dado pela diferença do ano de suas proposições, formalmente definida como:

$$fator_{temporal}(d_j, d_q) = \log\left(\frac{Periodo}{|Ano(d_j) - Ano(d_q)| + 1}\right) \tag{3.2}$$

em que, $Ano(d_i)$ se refere ao ano que o documento d_i foi proposto e $Periodo$ se refere ao lapso temporal calculado pela diferença do ano do documento mais antigo contido na base de dados até o documento mais novo. Quanto menor o valor do fator-temporal, mais distante no tempo dois documentos estarão.

Formalmente, integrando o fator temporal (Eq. 3.2) e o fator de desbalanceamento (Eq. 3.1) tem-se a seguinte proposta:

$$\begin{aligned} kNN_{proposto}(D^k, d_q) = \arg\max_{c \in \{0,1\}} \sum_{j=1}^{k} sim(d_q, d_j) \cdot I(c, \delta(d_j)) \\ \cdot fator_{desbalanceamento}(d_j) \cdot fator_{temporal}(d_j, d_q) \end{aligned} \tag{3.3}$$

Definição dos Hiperparâmetros e Dimensões

Versões do KNN

Para a execução das técnicas $kNN_{tradicional}$ e $kNN_{ponderado}$ é necessário definir o valor de K. Esse valor foi definido através da força-bruta na base de treinamento: variou-se o valor de K dentro do intervalo 1 e 100 e, para cada valor, foi calculado o F1-Score. Na base de teste, utilizou-se o valor de K que obteve o melhor valor de F1-Score identificado no treinamento. A Figura 3.5 mostra um exemplo dos valores de F1-Score obtidos para cada K.

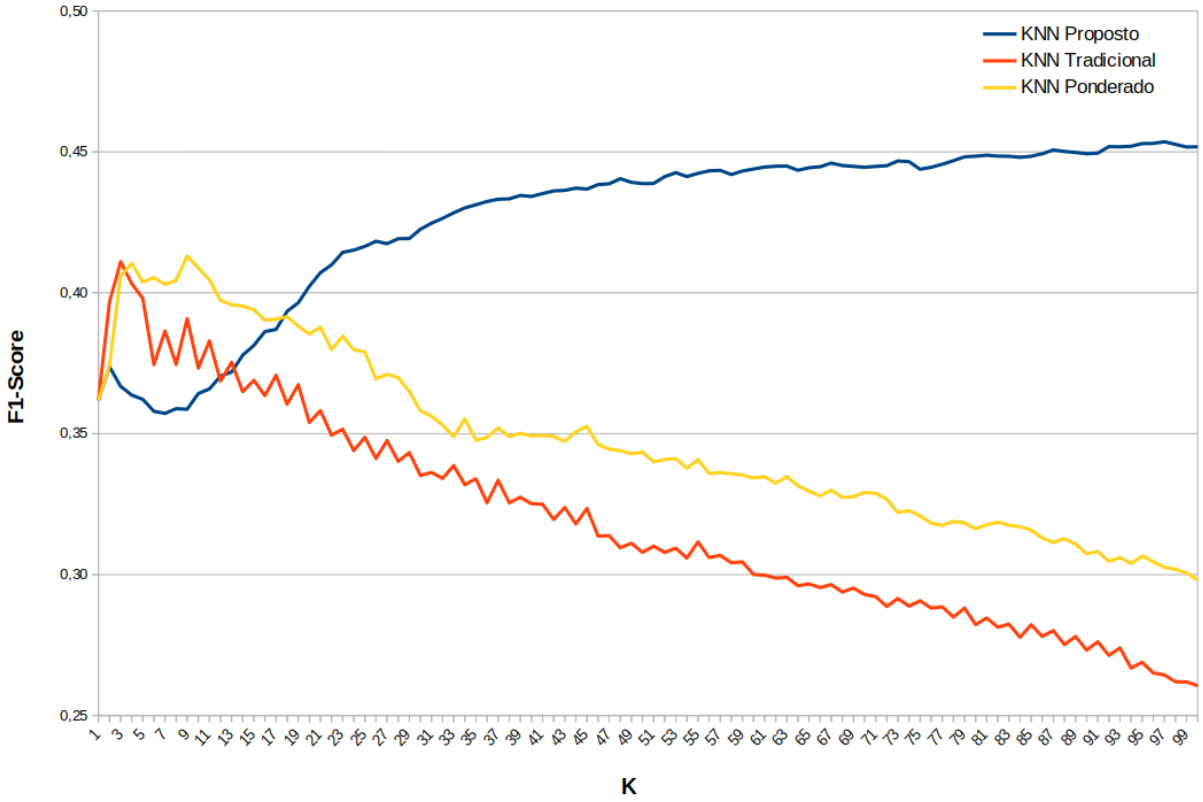
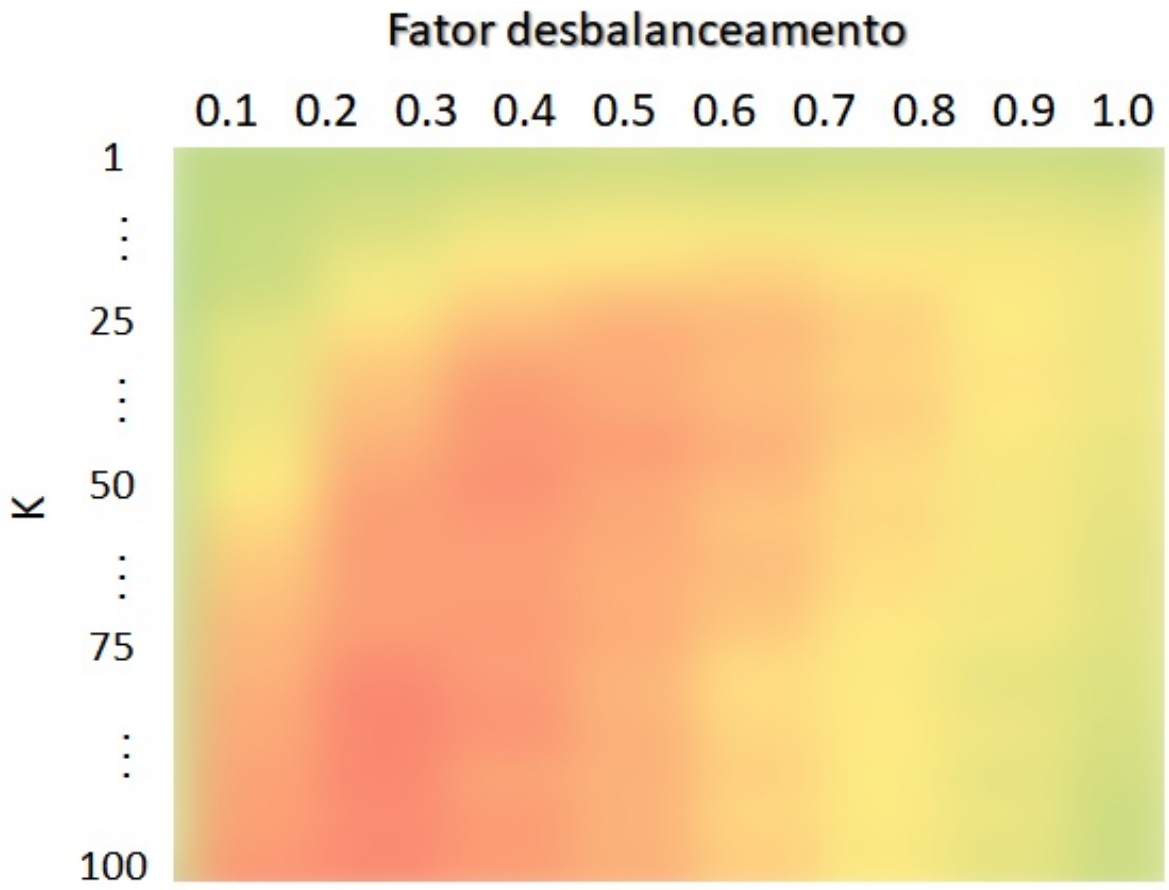


Figura 3.5: F1-score contra K - Cenário n° 1.

Já na execução do $kNN_{proposto}$, são necessários dois parâmetros: valor de K e o valor do fator de desbalanceamento (Eq. 3.1). Para escolher os melhores valores para esses dois parâmetros, foram realizados sucessivos testes, usando-se a base de treinamento e variando-se o valor do fator de desbalanceamento e o valor de K. A Figura 3.6 mostra um exemplo de mapa de calor obtido variando-se os dois parâmetros. As cores quentes indicam valores altos para o F1-Score. Nota-se, pela Figura, que os melhores valores são obtidos para valores baixos do fator de desbalanceamento e altos de K.

A redução da dimensionalidade do vetor TF-IDF também ocorreu por força-bruta, iniciando-se com uma dimensionalidade igual a 100 (com incrementos de 100) e para



Obs: Cores mais quentes representam F1-Score maiores.

Figura 3.6: Mapa de calor do $kNN_{proposto}$ - Cenário n° 1

cada dimensionalidade foi calculado o F1-Score de cada uma das técnicas. Os resultados obtidos podem ser observados na Figura 3.7. Por questões de desempenho nos casos onde houve empate no maior F1-Score, a dimensionalidade escolhida foi a menor.

Versões do XGBoost

As duas versões foram parametrizadas conforme tabela 3.3. Manteve-se os demais hiperparâmetros conforme padrão do componente.

Síntese

A Tabela 3.4 sumariza os parâmetros e dimensionalidades que geraram os melhores valores de F1-Score.

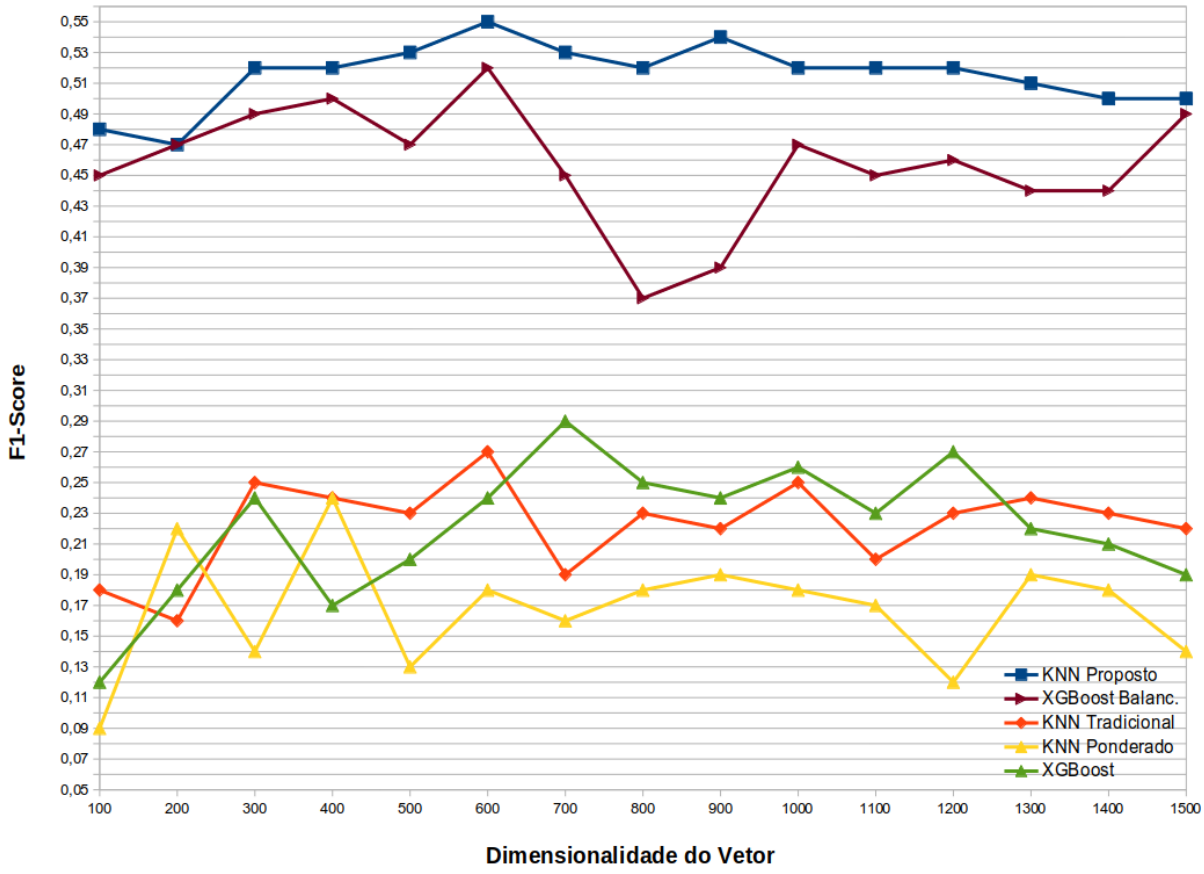


Figura 3.7: Valores de F1-Score com SVD - Cenário n° 1.

3.4.2 Cenário n° 2: Predição baseada nos dados categóricos

Os dados categóricos das PLs se referem às informações cadastrais (não-textuais) da PL que são fornecidas no início do processo legislativo, com a exceção do 'regime de tramitação' que pode ser alterado durante o processo legislativo.

A Figura 3.8 mostra esquematicamente o fluxo da arquitetura desenvolvida para avaliar a aprovação de PL a partir dos seus dados categóricos. Nele, observa-se dois fluxos: o de treinamento, onde as PL com trâmite concluído são submetidas ao classificador para gerar o modelo e o segundo, de produção, onde as PL em tramitação são submetidas ao modelo para gerar as previsões de aprovação. Os dois processos rodam periodicamente, cada um de acordo com sua necessidade, seja de atualizar o modelo ou de obter novas previsões baseadas nos novos dados de tramitação.

Cinco classificadores foram utilizados para comparação no Cenário n° 2, são eles: XGBoost, Regressão Logística, kNN, SVM e Floresta Aleatória. Para cada um deles foram executadas 10 classificações e Validações Cruzadas, sempre variando os parâmetros dos classificadores e mantendo 5 divisões nas validações. Cada classificação recebeu um número de versão e a de maior resultado de F1-Score representou o classificador

Versão	Parâmetros
XGB	use_label_encoder: False early_stopping_rounds: 10 eval_metric: 'logloss'
XGB_{bal}	use_label_encoder: False early_stopping_rounds: 10 eval_metric: 'logloss' scale_pos_weight: 11,5

Tabela 3.3: Parâmetros das Versões do XGBoost - Cenário n° 1.

Métrica	kNN_{prop}	kNN_{trad}	kNN_{pond}	XGB_{bal}	XGB
Dimensionalidade	600	600	400	600	700
k	97	4	5	-	-
Fator Desbalanceamento	0,12	-	-	-	-
scale_pos_weight	-	-	-	11,5	-

Tabela 3.4: Parâmetros e Dimensionalidades Melhores Resultados - Cenário n° 1.

na comparação com os demais.

Na comparação entre as melhores versões de cada classificador consideraram-se os seguintes critérios:

1. F1-Score Obtido na Classificação: valor resultante do processo de classificação usando a base de teste;
2. F1-Score da Validação Cruzada: média aritmética dos F1-Score obtidos nas 5 divisões da Validação Cruzada usando a base de treinamento.

A Tabela 3.5 mostra os valores dos hiperparâmetros das melhores versões.

3.4.3 Cenário n° 3: Predição combinando os dados textuais e categóricos

Neste 3° cenário, a proposta é avaliar como os dados textuais (Cenário n° 1) e os dados categóricos (Cenário n° 2) poderiam, juntos, contribuir para previsibilidade da aprovação/reprovação de uma PL. Para isso, foram implementadas e analisadas duas abordagens para combinar os resultados:

1. Classificação em Dados Compostos;
2. Classificação por Comitê.

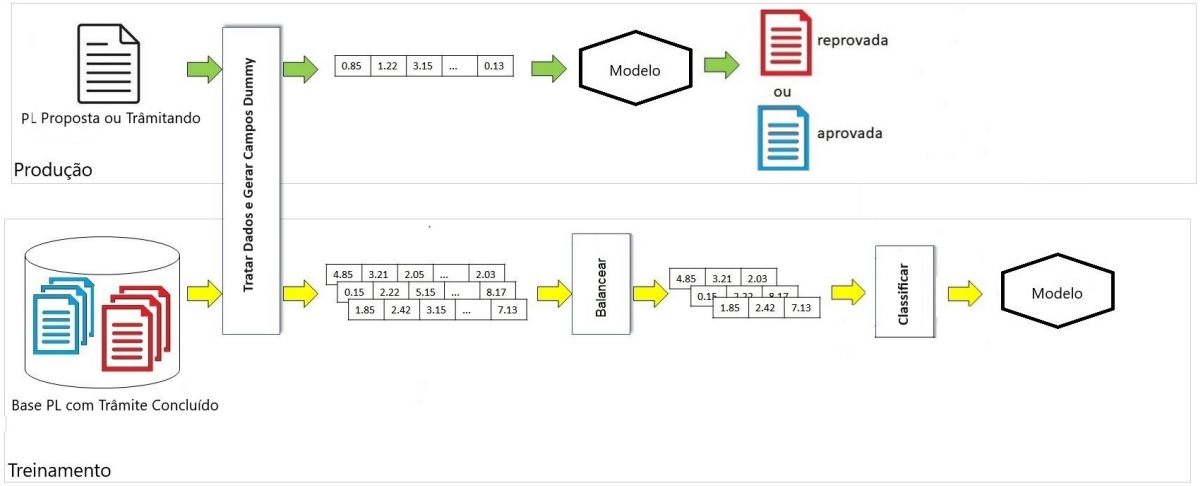


Figura 3.8: Arquitetura implementada - Cenário n° 2.

Classificação em Dados Compostos

Nessa abordagem, foi adicionado um novo atributo chamado de 'aprovado_texto' ao Cenário n° 2, contendo o valor do *label* de aprovação inferido no Cenário n° 1, compondo assim os dados compostos. Mateve-se a arquitetura e os hiperparâmetros da classificação de dados categóricos, adicionando apenas mais esse atributo nos dados. O modelo da Figura 3.9 apresenta uma visão deste processo.

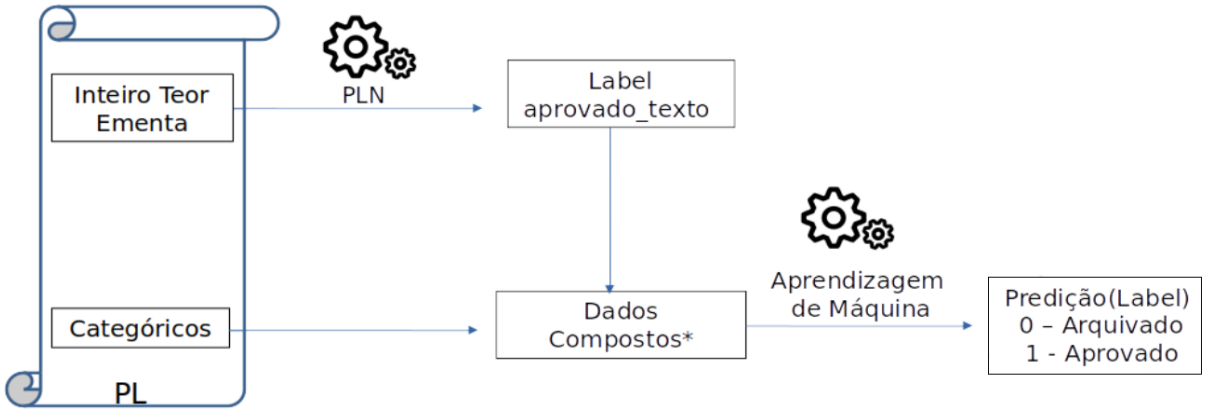


Figura 3.9: Arquitetura para classificar dados compostos - Cenário n° 3.1.

Classificação por Comitê de Classificadores

Nesta abordagem foi utilizado um comitê dos classificadores $kNN_{proposto}$, para os dados textuais, e o XGBoost, para os dados categóricos. A partir dos resultados individuais de cada classificado foi realizada uma votação, considerando os respectivos rótulos de inferência de aprovação para cada PL. O critério adotado, ilustrado na Figura 3.10, foi

Classificador	Hiperparâmetros
XGBoost	use_label_encoder = False early_stopping_rounds = 10 eval_metric = 'error' max_depth = 4 subsample = 1 scale_pos_weight = 11,5
Regressão Logística	penalty = 'l1' dual = False C = 0,4 fit_intercept = False solver = 'liblinear'
kNN	weights = 'uniform' algorithm = 'ball_tree' leaf_size = 30 p = 1 k = 17
SVM	C = 0,5 kernel = 'rbf' degree = 3 gamma = 'scale' coef0 = 0,0
Floresta Aleatória	bootstrap = True n_estimators = 1 random_state = 'gini' criterion = 'gini' max_depth = none

Tabela 3.5: Hiperparâmetros das melhores versões - Cenário n° 2.

que, basta que um dos rótulos tenha valor '1' (aprovado), então o rótulo do comitê também será '1', ou seja, apenas no caso onde os dois valores forem '0' (reprovado), o rótulo do comitê será '0' (reprovado).

Os parâmetros que geraram os resultados do comitê constam na tabela 3.6.

3.4.4 Cenário n° 4: Predição baseada nos dados categóricos e de tramitação

Neste cenário foram adicionados os atributos *dummy* das tramitações aos dados categóricos, mantendo-se inalterado o restante da arquitetura da Figura 3.8. A lista completa dos atributos de tramitação estão disponíveis no Anexo I.1. A Tabela 3.7 mostra, para cada classificador, os valores dos hiperparâmetros das suas melhores versões.

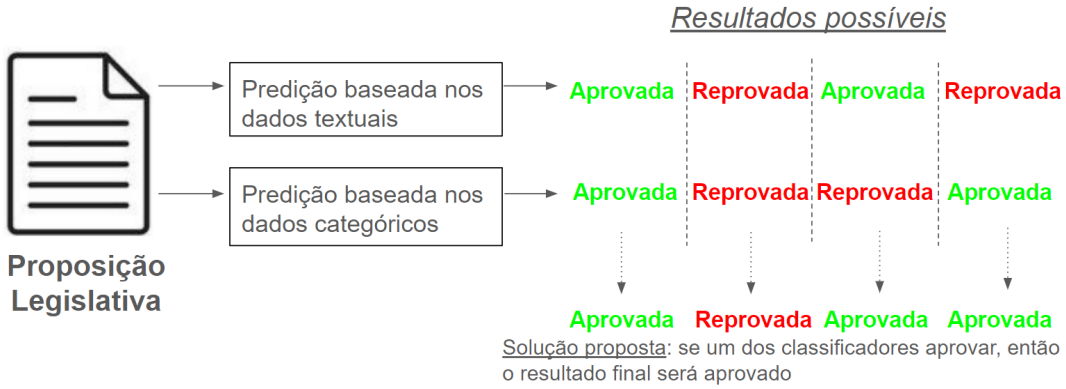


Figura 3.10: Regra utilizada para o Comitê de Classificadores - Cenário n° 3.2.

Classificador	Hiperparâmetros
XGBoost	use_label_encoder = False early_stopping_rounds = 10 eval_metric = 'error' max_depth = 4 subsample = 1
$kNN_{proposto}$	Coef. Balanceamento= 0,31 Dimensinalidade = 600

Tabela 3.6: Hiperparâmetros dos Classificadores do Comitê - Cenário n° 3.2.

3.4.5 Cenário n° 5: Simulação de aplicação em produção

O maior desafio dste cenário foi projetar um resultado para o F1-Score em produção. Na base de treinamento todas as PL foram concluídas e os atributos dummy mais relevantes tem valor 1, enquanto, em produção, cada PL terá estes campos conforme sua situação no processo. Não se pode, portanto, esperar que o desempenho do modelo em produção se assemelhe ao obtido em treinamento. Felizmente, as tramitações possuem o atributo 'data_hora' que permite selecionar todas as tramitações da PL até uma data específica, permitindo a simulação da aplicação do modelo nas bases de produção. Para esse fim, definiu-se 3 cortes temporais na base de teste e os respectivos intervalos das bases de treinamento (veja Tabela 3.8). A cada treinamento selecionou-se, entre as 10 versões do XGBoost, a que gerou o maior valor de F1-Score. Estas versões estão listadas na tabela 3.9 e seus respectivos hiperparâmetros na tabela 3.10 O valor do F1-Score foi calculado como uma média aritmética dos valores obtidos em cada um dos Cortes Temporais.

Classificador	Hiperparâmetros
XGBoost	use_label_encoder = False early_stopping_rounds = 10 eval_metric = 'error' max_depth = 5 subsample = 1
Regressão Logística	penalty='l2' dual = False C = 0,9 fit_intercept = True solver = 'lbfgs'
kNN	weights = 'uniform' algorithm = 'ball_tree' leaf_size = 30 p = 1 k = 17
SVM	C=1.0 kernel = 'poly' degree = 2 gamma = 'scale' coef0 = 0,0
Floresta Aleatória	bootstrap = True n_estimators = 100 random_state = 'entropy' max_depth = None

Tabela 3.7: Hiperparâmetros das melhores - Cenário n° 4.

Corte Temporal	Treinamento		Teste	
	Ano Inicial	Ano Final	Data Inicial	Data Final
2019-12-31	2001	2018	01/01/19	31/12/19
2020-12-31	2001	2019	01/01/20	31/12/20
2021-12-31	2001	2020	01/01/21	31/12/21

Tabela 3.8: Cortes Temporais - Cenário n° 5

Corte Temporal	Desbalanceada		Subamostrada		Sobreamostrada	
	Cat*	Comp**	Cat*	Comp**	Cat*	Comp**
2019-12-31	2	2	4	2	1	2
2020-12-31	7	7	9	7	9	7
2021-12-31	2	4	2	2	4	2

Cat* - dados categóricos.
Comp** - dados compostos.

Tabela 3.9: Versões do XGBoost Selecionadas nos Treinamentos - Cenário n° 5.

Versão	Parâmetros
1	use_label_encoder: False early_stopping_rounds: 10 eval_metric: 'error' max_depth: 3 subsample: 0,4 scale_pos_weight: 11,5 *
2	use_label_encoder: False early_stopping_rounds: 10 eval_metric: 'logloss max_depth: 3 subsample: 1 scale_pos_weight: 11,5 *
4	use_label_encoder: False early_stopping_rounds: 10 eval_metric: 'rmse' max_depth: 3 subsample: 1 scale_pos_weight: 11,5 *
7	use_label_encoder: False early_stopping_rounds: 10 eval_metric: 'error' max_depth: 5 subsample: 1 scale_pos_weight: 11,5 *
9	use_label_encoder: False early_stopping_rounds: 10 eval_metric: 'error' max_depth: 4 subsample: 1 scale_pos_weight: 11,5 *

* aplicado apenas em bases desbalanceadas.

Tabela 3.10: Parâmetros das Versões selecionadas - Cenário n° 5.

Capítulo 4

Resultados e Discussões

Este capítulo apresenta e discute os resultados obtidos a partir dos cenários definidos no capítulo anterior.

4.1 Cenário n° 1: Predição baseada nos dados textuais

Para o Cenário n° 1 apenas os dados textuais das PLs foram considerados e cinco classificadores foram avaliados, são eles: $kNN_{proposto}$ (Eq. 3.3), $kNN_{tradicional}$ (Eq. 2.6), $kNN_{ponderado}$ (Eq. 2.8), XGBoost e $XGBoost_{balanceado}$. Neste Cenário, avaliou-se apenas a base desbalanceada.

A Tabela 4.1 mostra os valores obtidos para cada um dos classificadores. Todos tiveram valores altos de acurácia, porém essa avaliação não é pertinente, como discutido em 2.4.2. Mais relevante foram os baixos valores de Revocação que denotam muitos erros na classificação dos documentos da classe minoritária (documentos aprovados). O melhor resultado de Revocação foi obtido pelo $kNN_{proposto}$, que conseguiu classificar corretamente 55% destes documentos. Por outro lado, acertou apenas 55% das vezes que inferiu aprovação, obtendo uma precisão inferior a dos demais classificadores. A métrica escolhida para avaliação neste trabalho foi o F1-Score e, nesse caso, o $kNN_{proposto}$ obteve valor de F1-Score bem superior aos demais, com exceção do $XGBoost_{balanceado}$ onde a diferença foi de apenas 5,8%. Isso mostra que o $kNN_{proposto}$ permitiu aumentar o poder discriminativo do algoritmo kNN ao prever a aprovação de PL.

Todavia os valores obtidos não serem ótimos, os resultados mostram a exequibilidade do conteúdo textual ser um atributo de suporte para a previsibilidade de aprovação/reprovação de uma PL. Uma explicação intuitiva para este fato é possibilidade do texto da proposição se tratar de um tema similar à outra proposição apresentada no

Métrica	kNN_{prop}	kNN_{trad}	kNN_{pond}	XGB_{bal}	XGB
Acurácia	0,86	0,86	0,86	0,86	0,87
Precisão	0,55	0,66	0,81	0,55	0,89
Revocação	0,55	0,17	0,14	0,49	0,17
F1-score	**0,55**	0,27	0,24	0,52	0,29

Tabela 4.1: Comparação das métricas - Cenário n° 1.

passado. Assim, é possível que a decisão da Câmara se repita para proposições com conteúdo similares.

4.2 Cenário n° 2: Predição baseada nos dados categóricos

Para o Cenário n° 2 apenas os dados categóricos das PLs foram consideradores e cinco classificadores foram avaliados: XGBoost, SVM, kNN, Regressão Logística e Floresta Aleatória. Executou-se testes experimentais, com todos os cinco classificadores, usando: a base desbalanceada, sobreamostrada e subamostrada. O melhor valor de F1-Score foi obtido na base subamostrada e, por esse motivo, os resultados a seguir serão apresentados considerando apenas esta base.

A Figura 4.1 mostra, para cada um dos classificadores avaliados, os valores de F1-Score obtidos na Classificação e Validação Cruzada.

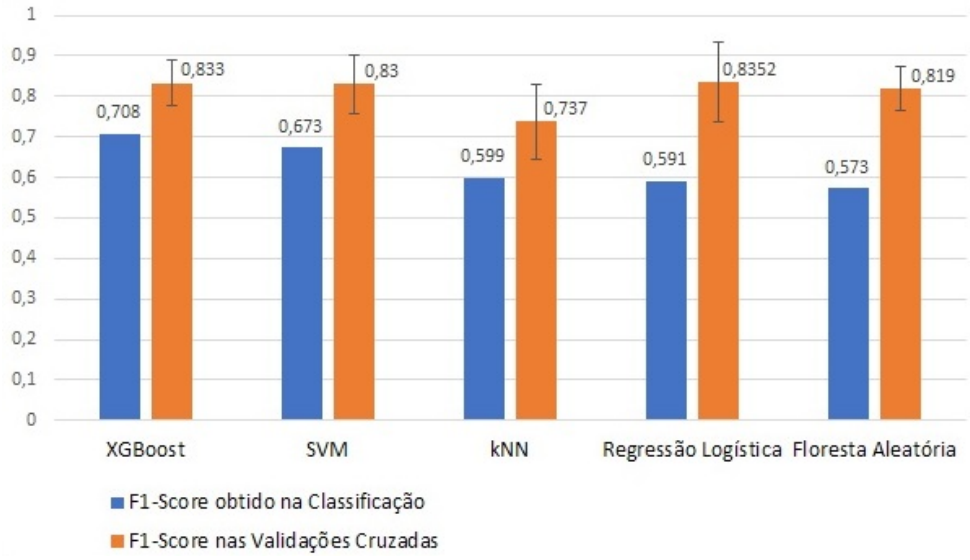


Figura 4.1: Comparação da classificação (F1-Score) - Cenário n° 2.

O XGBoost obteve o melhor desempenho na classificação da base de teste dentre os classificadores avaliados. Por esse motivo, foi realizado um novo experimento com o

classificador XGBoost para investigar se a seleção de atributos utilizando a técnica RFE, descrita na Seção 2.3.1, poderia aumentar o desempenho deste classificador.

A Tabela 4.2 mostra os atributos selecionados pela técnica RFE. Essa técnica gera um ranking de atributos e apenas os atributos ranqueados com valor 1 foram considerados na classificação.

Tramitacao	**Considerado**	**Ranking**
siglaPartidoAutorCorrigido_Órgão do Poder Legislativo	True	1
siglaPartidoAutorCorrigido_Órgão do Poder Judiciário	True	1
projeto_lei_conversao	True	1
regime_ordinaria	True	1
regime_urgencia	True	1
siglaPartidoAutorCorrigido_Órgão do Poder Executivo	True	1
tema_Homenagens e Datas Comemorativas	True	1
tema_Direito Civil e Processual Civil	False	2
...	...	...

Obs: A relação completa está no Anexo IV.1.

Tabela 4.2: Ranking RFE - Cenário n° 2.

A Tabela 4.3 mostra os resultados do classificador XGBoost antes e após a seleção de atributos. Houve o aumento de apenas 1%, tanto no F1-Score quanto na Acurácia. Essa diferença mostrou-se pequena, assim como nos resultados nas demais métricas. Porém, este acréscimo de 1% no F1-Score não se verificou no conjunto de dados desbalanceado e no sobreamostrado. Nesses dois conjuntos, os valores das métricas não se alteraram. Uma desvantagem observada na seleção de atributos foi no tempo total de processamento. Sem aplicar a técnica o tempo gasto foi de 269 milissegundos, contra 180.407 aplicando.

	XGBoost	
	sem RFE	com RFE
Acurácia	0,91	0,92
Precisão	0,77	0,81
Revocação	0,63	0,61
F1-Score	0,69	0,70

Tabela 4.3: Impacto do RFE - Cenário n° 2.

4.3 Cenário n° 3: Predição combinando os dados textuais e categóricos

O objetivo do Cenário n° 3 é avaliar se os dados textuais e categóricos das PLs podem ser combinados para, juntos, aumentarem o poder preditivo dos classificadores. Um dos desafios para a implementação desse cenário é identificar uma maneira de combinar esses dados. Por isso, a investigação foi dividida em duas abordagens:

- Cenário n° 3.1: neste cenário foi adicionado aos atributos categóricos da PL um novo atributo chamado 'aprovado_texto', gerado a partir da classificação da PL utilizando seus dados textuais, tal como obtido no Cenário n° 1;
- Cenário n° 3.2: neste cenário foi realizado um comitê dos classificadores $kNN_{proposto}$ e XGBoost. O primeiro é utilizado para classificar a PL usando seus dados textuais e o segundo seus dados categóricos. Para combinar os resultados utilizou-se a seguinte regra: se pelo menos um dos classificadores atribuírem o rótulo 'aprovado' à PL, então o resultado da PL será inferido como 'aprovado'.

A seguir os resultados obtidos em cada um desses sub-cenários serão apresentados.

4.3.1 Cenário n° 3.1: Classificação por dados compostos

Adicionou-se aos dados categóricos um novo atributo chamado de 'aprovado_texto' gerado a partir do Cenário n° 1. Apenas o classificador XGBoost foi utilizado, visto seu desempenho superior dentre os classificadores avaliados no Cenário n° 2. O melhor F1-Score foi obtido usando-se a base sobreamostrada, por esse motivo, os resultados a seguir serão apresentados considerando apenas essa base.

A Tabela 4.4 mostra o desempenho do classificador XGBoost sem o atributo 'aprovado_texto' e com o mesmo. Nota-se que não houve melhoria nas métricas com a adição desse novo atributo, por isso, o resultado conclusivo foi de que o atributo textual incorporado aos dados categóricos não aumentou o poder preditivo do classificador.

	XGBoost	
	Categórico	Composto
Acurácia	0,91	0,90
Precisão	0,78	0,69
Revocação	0,62	0,67
F1-Score	0,69	0,68

Tabela 4.4: Métricas Dados Compostos - Cenário n° 3.1.

Para investigar se, de fato, o atributo 'aprovado_texto' seria relevante na classificação, foi realizada uma seleção dos atributos utilizando a técnica RFE, tal como realizado no Cenário n° 2. A Tabela 4.5 mostra que não houve nenhuma alteração nos valores das métricas antes e após a seleção de atributos. Entretanto, nota-se na Tabela 4.6 que o atributo 'aprovado_texto' foi considerado como relevante. Ao se analisar o ranking de relevâncias individuais dos atributos pela técnica RFE, o atributo 'aprovado_texto' aparece na 8° posição, tal como ilustrado na Figura 4.2.

	XGBoost	
	sem RFE	com RFE
Acurácia	0,90	0,90
Precisão	0,68	0,68
Revocação	0,69	0,69
F1-Score	0,68	0,68

Tabela 4.5: Métricas Dados Compostos com RFE - Cenário n° 3.1.

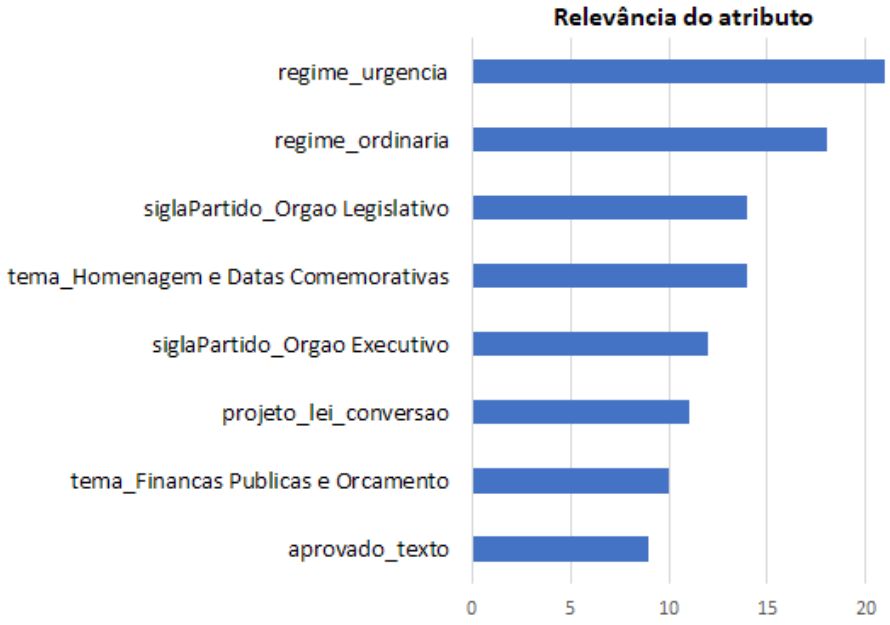


Figura 4.2: Relevância(F1-Score) dos atributos - Cenário n° 3.1

4.3.2 Cenário n° 3.2: Classificação por Comitê de Classificadores

O comitê de classificadores foi formado pelo $kNN_{proposto}$ e XGBoost, o primeiro baseado nos dados textuais e o segundo nos dados categóricos. Os dois classificadores foram escolhidos porque obtiveram os melhores resultados nos seus respectivos cenários. Os

Tramitacao	Considerado	Ranking
projeto_lei	True	1
tema_Arte, Cultura e Religião	True	1
tema_Agricultura, Pecuária, Pesca e Extrativismo	True	1
tema_Administração Pública	True	1
siglaPartidoAutorCorrigido_Órgão do Senado Federal	True	1
siglaPartidoAutorCorrigido_Órgão do Poder Legislativo	True	1
siglaPartidoAutorCorrigido_Órgão do Poder Judiciário	True	1
siglaPartidoAutorCorrigido_Órgão do Poder Executivo	True	1
siglaPartidoAutorCorrigido_ZZ-Indeterminado	True	1
siglaPartidoAutorCorrigido_Sociedade Civil	True	1
siglaPartidoAutorCorrigido_SD	True	1
tema_Cidades e Desenvolvimento Urbano	True	1
siglaPartidoAutorCorrigido_S.PART.	True	1
siglaPartidoAutorCorrigido_REDE	True	1
siglaPartidoAutorCorrigido_PV	True	1
siglaPartidoAutorCorrigido_PTdoB	True	1
siglaPartidoAutorCorrigido_PTN	True	1
...	...	...
siglaPartidoAutorCorrigido_AVANTE	True	1
aprovado_texto	True	1
...	...	...

Obs: A relação completa está no Anexo V.1.

Tabela 4.6: Ranking RTE - Cenário n° 3.1.

testes foram executados nas bases desbalanceadas. A Tabela 4.7 mostra os resultados cada classificador e do comitê. Nota-se que, comparado com o XGBoost, o valor de Revocação melhorou no comitê por conta da maior quantidade de verdadeiros positivos, porém, a Precisão caiu por conta do aumento da quantidade de falsos positivos. No final, a combinação dos dois classificadores não obteve melhoria do F1-Score.

Classificador	**Dados**	**Acurácia**	**Precisão**	**Revocação**	**F1-Score**
$kNN_{proposto}$	Textuais	0,88	0,72	0,34	0,46
XGBoost	Categóricos	0,92	0,81	0,63	0,71
Comitê	—	0,91	0,75	0,67	0,71

Tabela 4.7: Métricas Resultantes - Cenário n° 3.2

4.4 Cenário n° 4: Predição baseada nos dados categóricos e de tramitação

Para o Cenário n° 4 foram considerados os dados de tramitação e os dados categóricos das PLs. Novamente, cinco classificadores foram avaliados: XGBoost, SVM, kNN, Regressão Logística e Floresta Aleatória. Executou-se testes experimentais, com todos os cinco classificadores, usando: a base desbalanceada, sobreamostrada e subamostrada. O melhor valor de F1-Score foi obtido na base sobreamostrada e, por esse motivo, os resultados a seguir serão apresentados considerando apenas esta base.

A Figura 4.3 mostra, para cada um dos classificadores avaliados, os valores de F1-Score obtidos durante a Validação Cruzada e Classificação.

Neste cenário, houve um empate técnico entre os classificadores e selecionou-se o XGBoost para os testes.

Ao se analisar a relevância dos atributos usando a técnica RFE, a Figura 4.4 mostra que dos 19 atributos mais importantes do conjunto considerado na classificação, 12 deles são atributos relacionados à tramitação. Isso mostra que, de fato, os dados de tramitação são fortes indicativos de aprovação/reprovação de uma PL. Note que o atributo 'tramitacao_Remessa à Sanção' foi o mais relevante, já que a aprovação só depende da sansão presidencial, este fato confirma a relevância dos dados das tramitações para a assertividade da inferência de aprovação de uma PL e leva a conclusão que, quanto mais avançadas estiverem os processos das PL no fluxo legislativo, maior será a assertividade do classificador.

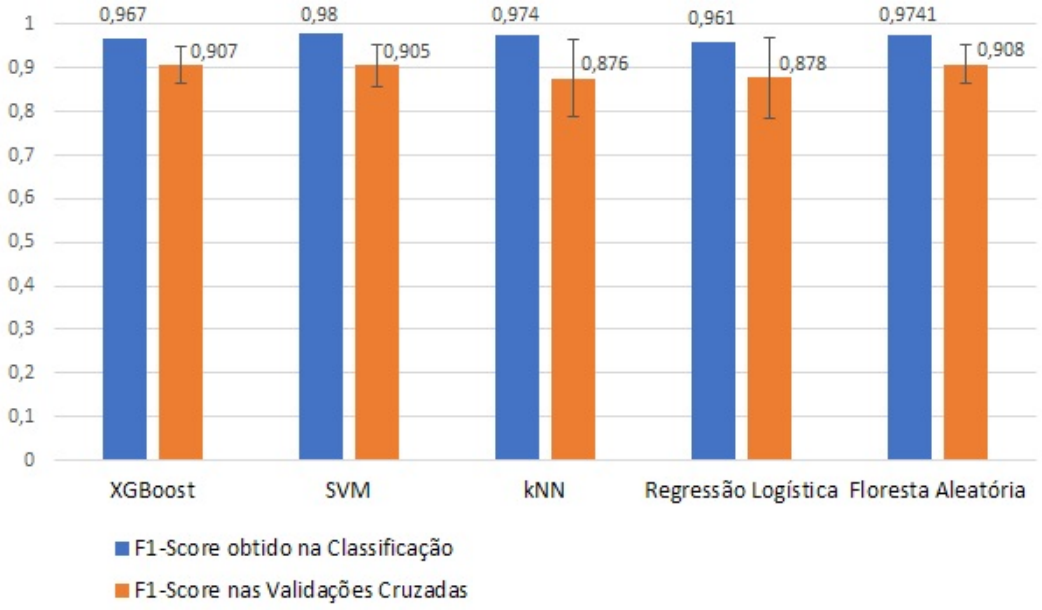


Figura 4.3: Resultados dos classificadores - Cenário n° 4.

4.5 Cenário n° 5: Simulação em produção

Em um cenário real as PL estarão em diferentes estágios no Processo Legislativo. Neste contexto, pode-se ter dois extremos: proposições recém-publicadas e sem dados de tramitação e, do outro lado, as que aguardam apenas a tramitação final. Por isso, nesse Cenário n° 5 investiga-se o poder preditivo dos classificadores em diferentes cortes temporais do trâmite das PL, para identificar se, de fato, quanto mais avançadas estiverem, mais precisas serão as predições de aprovação/arquivamento.

Para simular o comportamento dos atributos de tramitação, criou-se 3 bases de testes entre 2019 e 2021. As bases de treinamento variaram conforme o corte temporal. Elas se iniciaram em 2001 e vão até o ano anterior ao corte, simulando assim o treinamento do modelo a cada início de ano.

A Tabela 4.8 mostra os valores de F1-Score obtidos pelo classificador XGBoost para cada corte temporal analisado. Foram avaliadas as bases desbalanceada, subamostrada e sobreamostrada, e também com e sem o atributo 'aprovado_texto'.

Nota-se que os Corte Temporais permitem identificar perfeitamente a situação das proposições no tempo, porém só estão representadas na base de teste as proposições concluídas até a data da carga. Como a data de carga foi pouco superior ao final de 2021, as PL de 2021 selecionadas tiveram sua tramitação muito mais rápida que os 14,04 meses da média da base de treinamento (2019-2021). Em contrapartida, as PL selecionadas de 2019 tiveram muito mais tempo para serem concluídas. Consequentemente, no corte de 2019-12-31 o trâmite de muitas das PL de 2019 estavam longe de serem concluídas,

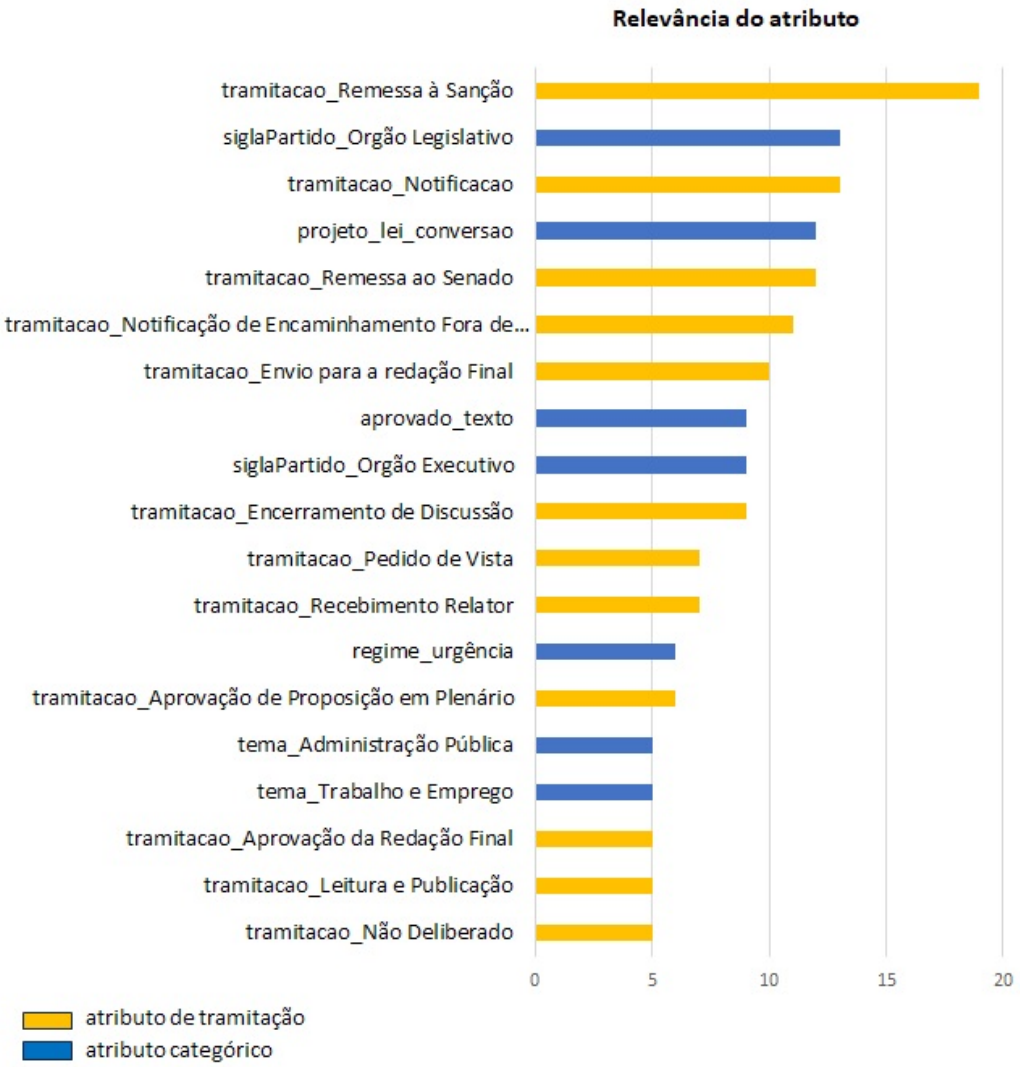


Figura 4.4: Relevância individual dos atributos - Cenário n° 4

resultando em um F1-Score menor. Já no corte de 2021-06-30 as PL de 2021 já tinham seus processos concluídos ou bem adiantados, o que refletiu num F1-score mais alto.

Em termos de processo, as PL que estão a um passo da conclusão de seus processos tem os destinos de seus processos legislativos mais previsíveis (Arquivamento ou Transformação em Norma Jurídica). Estes resultados confirmam a relevância dos dados das tramitações para a assertividade da inferência de aprovação e leva a conclusão que, quanto mais avançadas estiverem os processos das PL, maiores a eficácia e eficiência na previsão do modelo. Esta conclusão é intuitiva, uma vez que, uma PL passa por diversas avaliações e/ou votações até chegar a aprovação, ou reprovação. Por exemplo, ser aprovada nas comissões são indicativos de aprovação.

Apesar da importância estatística da validação cruzada, o processo de geração de subconjuntos não considerou critérios temporais, o que pode ter levado a inconsistências

Corte Temporal	Desbalanceada		Subamostrada		Sobreamostrada	
	Cat*	Comp**	Cat*	Comp**	Cat*	Comp**
2019-12-31	77,1	77,6	80,8	80,8	76,9	80,0
2020-12-31	91,3	91,3	91,1	91,1	91,6	91,1
2021-12-31	100,0	100,0	100,0	100,0	100,0	100,0
Média	**89,5**	**89,6**	**90,6**	**90,6**	**89,5**	**90,4**

Cat* - dados categóricos.
Comp** - dados compostos.

Tabela 4.8: Resultados nos Cortes Temporais - Cenário n° 5.

como dados de teste anteriores aos de treinamento. No processo de classificação isso não ocorre, pois os dados de teste são de uma legislatura posterior aos usados no treinamento. Por conta disso, o resultado da classificação foi considerado mais significativo. O cenário selecionado para ser aplicado em produção foi o da base subamostrada, como pode ser observado na tabela 4.8.

Capítulo 5

Conclusão

Este trabalho apresentou um estudo e investigação visando agregar conhecimento e propiciar melhor compreensão dos aspectos relacionados ao Processo Legislativo no Brasil. A proposta deste trabalho constituiu em desenvolver modelos baseados em Aprendizado de Máquinas (AM) para prever a aprovação de uma Proposição Legislativa (PL). Para tal fim, foram utilizados dados públicos, obtidos no site da Câmara dos Deputados, de proposições apresentadas pelos parlamentares em duas décadas, especificamente entre os anos de 2001 a 2021. Selecionaram-se apenas as proposições dos tipos Proposta de Lei e Proposta de Emenda e com trâmite concluído, ou seja, com situação aprovada ou arquivada/rejeitada. Testes experimentais foram realizados comparando diferentes cenários e algoritmos de classificação.

Uma das questões de pesquisa norteadora deste trabalho foi a de que os dados textuais das PL seriam relevantes na previsibilidade de sua aprovação/reprovação. Esse fato se mostrou factível através da proposta de um novo classificador, denominado $kNN_{proposto}$, que é uma versão evoluída do algoritmo kNN. O classificador $kNN_{proposto}$, considerou três importantes aspectos, ao inferir a aprovação de uma proposição legislativa: (i) o fator de desbalanceamento da base de dados, (ii) o fator temporal em que as proposições foram apresentadas e (iii) a similaridade entre o elemento de pesquisa e seus K vizinhos. Resultados experimentais mostraram que essas três modificações contribuíram para aumentar o desempenho do classificador kNN. Embora o resultado do valor de 55% obtido pelo classificador $kNN_{proposto}$ seja excelente, mostrou-se útil na previsibilidade de aprovação/reprovação das Proposição Legislativa (PL).

Outra questão de pesquisa investigada foi a identificação do poder preditivo dos atributos categóricos na aprovação de uma PL, que são as informações registradas no momento de cadastro do projeto. Nos testes realizados foi obtido um valor de F1-Score de 71% considerando apenas os atributos categóricos. Esse resultado corrobora com os achados de outros trabalhos correlacionados, mostrando a relevância prognóstica

dos atributos categóricos no modelo de classificação. Dentre o conjunto de atributos analisado, o regime da proposição mostrou alta relevância. É importante ressaltar que a previsibilidade do modelo não é muito alta, pois não existem atributos que consigam capturar a intenção dos parlamentares em aprovar ou não, um determinado projeto. Todavia, o modelo de classificação considerando somente os atributos categóricos podem facilitar a análise do cenário e o pensamento de estratégias de influência praticamente no início de seus trâmites, em um momento muito difícil pra a predição humana.

Por fim, a última questão de pesquisa levantada neste trabalho foi uma análise do impacto dos atributos de tramitação na previsibilidade de aprovação/reprovação de uma PL. Os resultados mostram que, no que diz respeito ao Processo Legislativo, aquelas Proposição Legislativa (PL) que estão a poucos passos de encerra-lo tem seus destinos mais previsíveis. O F1-Score de 90,6% previstos para a produção confirmam a relevância dos dados de tramitação para a confiança nas inferências de aprovação de uma PL. O crescimento da métrica a medida que o corte temporal se aproxima da data da carga, por sua vez, indicam que, quanto mais avançado o processo legislativo, mais assertivas serão as previsões do modelo.

Em decorrência da pesquisa desenvolvida neste trabalho, um artigo foi publicado em conferências Qualis da área da computação:

- Cabral, Ilo e Pedrosa, Glauco. *A Classifier-Based Approach to Predict the Approval of Legislative Propositions.* In: 25th International Conference on Enterprise Information Systems (ICEIS), 2023, Prague. p. 303.

5.1 Limitações

É importante ressaltar que a política é uma área em constante transformação e a maneira como os agentes políticos são influenciados e tomam decisões se modificam a todo momento, seja por mudança cultural, por mudanças nas regras do Processo Legislativo ou até mesmo pela pressão das Redes Sociais. Por isso, os modelos de classificação avaliados precisarão ser revisitados com periodicidade, a fim de verificar se os achados aqui apresentados continuam sendo válidos.

Apesar dos modelos desenvolvidos não serem ótimos, eles fornecem suporte relevante para grupos que precisam monitorar o Processo Legislativo, fornecendo uma previsibilidade com base em propostas do passado. Assim, é possível focar os esforços na análise das propostas, das oportunidades e riscos a seus negócios e na definição de estratégias para lidar com elas, promovendo uma governança mais informada e responsiva.

Ademais, os experimentos e os resultados obtidos neste trabalho estão circunscritos ao escopo de apenas uma parte do Processo Legislativo de uma casa legislativa (Câmara dos

Deputados) e de Projetos de Lei. Todavia, a metodologia empregada pode ser replicada, por exemplo, para criar modelos que possam ser estendidos para outros tipos de Projetos.

5.2 Trabalhos Futuros

A partir dos resultados obtidos, foram definidos alguns caminhos possíveis para o avanço de novas pesquisas:

- Desenvolver modelos de previsão com base em outros tipos de Proposições Legislativas, além das Propostas de Lei e Propostas de Emenda;
- Investigar novos modelos de classificação que possibilitem capturar a intenção dos parlamentares sobre um determinado tema. Assim, baseado nessa informação será possível antecipar os votos dos legisladores e, no momento da apresentação da PL, seria possível avaliar um prognóstico de maior previsibilidade, mesmo sem nenhum dado de tramitação obtido;
- Comitê de Classificadores: a proposta é formar um comitê com dois classificadores, um treinado com os dados categóricos das proposições e outro com os dados de tramitação. Nesse caso, a proposta é avaliar a ponderação do peso de cada um desses classificadores durante as etapas do Processo Legislativo;
- Análise de Sentimento: a ideia seria avaliar como e se a Análise de Sentimentos, sobre os temas das proposições, podem contribuir positivamente nas suas inferências de aprovação.

Referências

[1] Ribeiro, Claudio Jose Silva e RF de Almeida: *Dados abertos governamentais (open government data): instrumento para exercício de cidadania pela sociedade*. Encontro Nacional de Pesquisa em Ciência da Informação, 12:2568–2580, 2011. 1

[2] Cruvinel, Gustavo Warzocha Fernandes: *Mais transparência: desvendando os dados abertos da Câmara dos Deputados*. Editora Dialética, 2022. 1

[3] Tabak, Benjamin Miranda: *A análise econômica do direito: proposições legislativas e políticas públicas*. Revista de informação legislativa, 52(205):321–345, 2015. 1

[4] Oliveira, Danilo Amaral de, João Porto de Albuquerque e Alexandre C. B. Delbem: *Compreendendo e prevendo o processo legislativo na câmara dos deputados do brasil*. Em *Anais do XIV Simpósio Brasileiro de Sistemas de Informação*, páginas 175–182, Porto Alegre, RS, Brasil, 2018. SBC. `https://sol.sbc.org.br/index.php/sbsi/article/view/5085`. 2

[5] Nay, John J.: *Predicting and understanding law-making with word vectors and an ensemble model*. PLOS ONE, 12(5):e0176999, maio 2017, ISSN 1932-6203. `http://dx.doi.org/10.1371/journal.pone.0176999`. 2, 7

[6] Barella, Victor Hugo: *Técnicas para o problema de dados desbalanceados em classificação hierárquica*. Tese de Doutoramento, Universidade de São Paulo, 2015. 3, 21

[7] Yang, Qiang e Xindong Wu: *10 challenging problems in data mining research*. Int. J. Inf. Technol. Decis. Mak., 5:597–604, 2006. 3

[8] Cate, Fred H: *Government data mining: The need for a legal framework*. Harv. CR-CLL Rev., 43:435, 2008. 5

[9] Lee, Yin Harn: *United kingdom copyright decisions and legislative developments 2014*. IIC-International Review of Intellectual Property and Competition Law, 46(2):226–237, 2015. 5

[10] Sengupta, Souvik e Vishwang Dave: *Predicting applicable law sections from judicial case reports using legislative text analysis with machine learning*. Journal of Computational Social Science, 5(1):503–516, 2022. 5

[11] Savelyev, Alexander I: *The issues of implementing legislation on personal data in the era of big data*. Law: J. Higher Sch. Econ., página 43, 2015. 5

[12] Barros Correia Gomes, F. de e E. Câmara: *Produção Legislativa no Brasil: Visão Sistêmica e Estratégica no Presidencialismo de Coalizão*. Temas de interesse do Legislativo. Edições Câmara, ISBN 9788540200999. https://books.google.com.br/books?id=-RfhDwAAQBAJ. 6

[13] Ricci, Paolo: *O conteúdo da produção legislativa brasileira: leis nacionais ou políticas paroquiais?* Dados, 46(4):699–734, 2003. https://doi.org/10.1590/s0011-52582003000400003. 6

[14] Oliveira, Danilo Amaral de, Joao Porto de Albuquerque e Alexandre C. B. Delbem: *Understanding and predicting the legislative process in the chamber of deputies of brazil*. SBSI'18, New York, NY, USA, 2018. Association for Computing Machinery, ISBN 9781450365598. https://doi.org/10.1145/3229345.3229371. 6

[15] Baptista, Vitor Márcio Paiva de Sousa: *Um modelo para a detecção das mudanças de posicionamento dos deputados federais*. Mestrado, Universidade Federal da Paraíba, João Pessoa/PB, Brasil, 2015. 6

[16] Carvalho, Laura Solano de: *Análise heurística do padrão de votação dos deputados em proposições na câmara dos deputados entre 2015 e 2016*, 2016. 6

[17] Oliveira, Danilo Amaral de: *Compreendendo e prevendo o processo legislativo via ciência de dados*. Mestrado, Universidade de São Paulo, São Carlos/SP, Brasil, 2018. 6

[18] Brito, Ranniery Dias de: *Análise e predição nas votações de leis federais na câmara dos deputados*, 2022. 7

[19] Wang, Jun, Kush R. Varshney e Aleksandra Mojsilović: *Legislative prediction via random walks over a heterogeneous graph*. Em *Proceedings of the 2012 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, abril 2012. https://doi.org/10.1137/1.9781611972825.94. 7

[20] Cheng, Yu, Ankit Agrawal, Huan Liu e Alok Choudhary: *Legislative Prediction with Dual Uncertainty Minimization from Heterogeneous Information*, volume 10, páginas 361–369. junho 2015, ISBN 978-1-61197-401-0. 7

[21] Karimi, Hamid, Tyler Derr, Aaron Brookhouse e Jiliang Tang: *Multi-factor congressional vote prediction*. Em *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, páginas 266–273, 2019. 8

[22] Ziegler, Andreas: *An introduction to statistical learning with applications. r. g.james, d.witten, t.hastie, and r.tibshirani (2013). berlin: Springer. 440 pages, isbn: 978-1-4614-7138-7*. Biometrical journal, 58(3):715–716, 2016, ISSN 0323-3847. 10

[23] "Liu, Huan": *"Feature Selection"*, páginas "402–406". "Springer US", "Boston, MA", "2010", ISBN "978-0-387-30164-8". "https://doi.org/10.1007/978-0-387-30164-8_306". 11

[24] Kramer, Mark A.: *Nonlinear principal component analysis using autoassociative neural networks*. AIChE journal, 37(2):233–243, 1991, ISSN 0001-1541. 11

[25] Kononenko, Igor e Se June Hong: *Attribute selection for modelling*. Future Generation Computer Systems, 13(2-3):181–195, 1997. 11

[26] Johnson, R. e Wichern: *Applied multivariate statistical analysis*. 2002. 12

[27] Smith, L. I.: *A tutorial on principal components analysis*. Relatório Técnico, Department of Computer Science, University of Otago, New Zealand, 2002. 12

[28] Karl, Andrew, James Wisnowski e W Heath Rushing: *A practical guide to text mining with topic extraction*. Wiley Interdisciplinary Reviews: Computational Statistics, 7(5):326–340, 2015. 12

[29] *6. Singular Value Decomposition*, páginas 57–74. `https://epubs.siam.org/doi/abs/10.1137/1.9780898718867.ch6`. 13

[30] Dumais, S. T.: *Latent semantic analysis*. Em *Annual Review of Information Science and Technology*, páginas 188–230, 2005. 13

[31] Juez-Gil, Mario, Álvar Arnaiz-González, Juan J. Rodríguez e César García-Osorio: *Experimental evaluation of ensemble classifiers for imbalance in big data*. Applied Soft Computing, 108:107447, 2021, ISSN 1568-4946. `https://www.sciencedirect.com/science/article/pii/S1568494621003707`. 16

[32] Brzezinski, Dariusz, Jerzy Stefanowski, Robert Susmaga e Izabela Szczech: *On the dynamics of classification measures for imbalanced and streaming data*. IEEE Transactions on Neural Networks and Learning Systems, 31(8):2868–2878, 2020. 19

[33] Kohavi, Ron: *A study of cross-validation and bootstrap for accuracy estimation and model selection*. 14, março 2001. 19

[34] He, Haibo e Edwardo A. Garcia: *Learning from imbalanced data*. IEEE Transactions on Knowledge and Data Engineering, 21(9):1263–1284, 2009. 20

[35] Wang, Le, Meng Han, Xiaojuan Li, Ni Zhang e Haodong Cheng: *Review of classification methods on unbalanced data sets*. IEEE Access, 9:64606–64628, 2021. 21

[36] Li, Xing e Lei Zhang: *Unbalanced data processing using deep sparse learning technique*. Future Generation Computer Systems, 125:480–484, 2021. 21

[37] Barot, Pratikkumar e Harikrishna Jethva: *Imbtree: Minority class sensitive weighted decision tree for classification of unbalanced data*. International Journal of Intelligent Systems and Applications in Engineering, 9(4):152–158, 2021. 21

[38] Liu, Fen e Quan Qian: *Cost-sensitive variational autoencoding classifier for imbalanced data classification*. Algorithms, 15(5), 2022, ISSN 1999-4893. `https://www.mdpi.com/1999-4893/15/5/139`. 21

[39] Batista, Gustavo E. A. P. A., Andre C. P. L. F. Carvalho e Maria Carolina Monard: *Applying one-sided selection to unbalanced datasets.* Em *Lecture Notes in Computer Science*, páginas 315–325. Springer Berlin Heidelberg, 2000. https://doi.org/10.1007/10720076_29. 21

[40] Zhang, Xuesong, Yan Zhuang, Wei Wang e Witold Pedrycz: *Transfer boosting with synthetic instances for class imbalanced object recognition.* IEEE Transactions on Cybernetics, 48(1):357–370, 2018. 23

[41] Carneiro, André Corrêa de Sá; Santos, Luiz Cláudio Alves dos; Nóbrega Netto Miguel Gerônimo da. Brasília : Edições Câmara, 2021. 25

[42] Moreira, Cláudia Cristina Pacheco: *Proposições de iniciativa parlamentar: projetos de lei apresentados no senado federal entre 1987 e 2005.* Mestrado, Instituto de Ciências Políticas, Universidade de Brasília, Brasília, Brasil, 2006. 25

Anexo I

Campos Dummy da Tramitação

tramitacao_Não Informado
tramitacao_Apresentação de Proposição
tramitacao_Desapensação
tramitacao_Leitura e publicação
tramitacao_Apensação
tramitacao_Não Apensação
tramitacao_Distribuição
tramitacao_Redistribuição
tramitacao_Envio para a redação Final
tramitacao_Declarada insubsistância
tramitacao_Criação de Comissão Temporária
tramitacao_Constituição de Comissão Temporária
tramitacao_Instalação de Comissão
tramitacao_Despacho à CCJR - Redação Final
tramitacao_Despacho à Promulgação
tramitacao_Despacho à Sanção
tramitacao_Remessa ao Senado Federal
tramitacao_Despacho de Apensação
tramitacao_Despacho de Desapensação
tramitacao_Despacho de sujeita a Arquivamento
tramitacao_Despacho de sujeita a Devolução
tramitacao_Devolução ao autor
tramitacao_Despacho de Arquivamento
tramitacao_Despacho de Desarquivamento

tramitacao_Remessa à Presidência da República
tramitacao_Remessa a Ministério
tramitacao_Despacho de Não Acolhimento
tramitacao_Desarquivamento - Errata
tramitacao_Inclusão em Pauta
tramitacao_Inclusão de Urgência em Pauta
tramitacao_Recebimento na Mesa solicitando resposta
tramitacao_Recebimento de Resposta
tramitacao_Devolução ao Relator
tramitacao_Devolução à Mesa para Novo Despacho
tramitacao_Encaminhamento
tramitacao_Encaminhamento - art. 52, § 6º do RICD
tramitacao_Expedição de Documento
tramitacao_Aprovação de Recurso
tramitacao_Apresentação de Recurso
tramitacao_Rejeição de Recurso
tramitacao_Apresentação de Requerimento
tramitacao_Aprovação de Requerimento
tramitacao_Aprovação de Urgência (154, 155 ou 64 CF)
tramitacao_Rejeição de Proposicao
tramitacao_Prejudicado
tramitacao_Providência Interna
tramitacao_Retirada pelo Autor
tramitacao_Proposição Devolvida ao Autor
tramitacao_Notificação de Despacho
tramitacao_Leitura de Parecer em substituição à Comissão
tramitacao_Leitura e publicação da Redação Final
tramitacao_Leitura e publicação de Parecer favorável de admissibilidade de PEC
tramitacao_Leitura e publicação do Parecer - Urgência (154,155 ou 64 CF)
tramitacao_Discussão
tramitacao_Encaminhamento da Votação
tramitacao_Discussão (Inicio e Continuacao)
tramitacao_Adiamento de Discussão
tramitacao_Discussão (Plenário)
tramitacao_Obstrução Discussão (Plenário)

tramitacao_Recebimento de Emenda de Plenário
tramitacao_Encerramento de Discussão
tramitacao_Votação
tramitacao_Votação em 1º turno
tramitacao_Votação em 2º turno
tramitacao_Adiamento de Votação
tramitacao_Aprovação em 1º turno
tramitacao_Aprovação em 1º turno - com redação
tramitacao_Aprovação da Redação Final
tramitacao_Aprovação
tramitacao_Aprovação de Proposição Interna
tramitacao_Discussão da Materia pelos Deputados
tramitacao_Aprovação de Proposição em Plenário
tramitacao_Verificação de Votação
tramitacao_Retirada de Pauta
tramitacao_Transformação em Norma Jurídica
tramitacao_Rejeição de Requerimento
tramitacao_Retirada de Requerimento
tramitacao_Apresentação de Proposição interna - Comissão
tramitacao_Análise Parecer
tramitacao_Designação de Relator
tramitacao_Redistribuição a Relator
tramitacao_Parecer do Relator
tramitacao_Parecer do Relator - Emendas
tramitacao_Parecer do Relator - Manifestação
tramitacao_Parecer do Relator-Parcial
tramitacao_Parecer do Relator - Revisão
tramitacao_Designação de Relator do Vencedor
tramitacao_Designação de Relator Parcial
tramitacao_Designação de Relator Revisor
tramitacao_Leitura e publicação do Parecer
tramitacao_Cancelamento de Parecer
tramitacao_Rejeição do Parecer do Relator
tramitacao_Aprovação do Parecer
tramitacao_Apresentação do Relatório Prévio

tramitacao_Aprovação do Relatório Prévio ou Parcial
tramitacao_Rejeição do Relatório Prévio
tramitacao_Apresentação do Relatório Final
tramitacao_Abertura de Prazo
tramitacao_Prorrogação de Prazo
tramitacao_Reabertura de Prazo
tramitacao_Prejudicado Requerimento
tramitacao_Recebimento de Emenda
tramitacao_Recebimento de Emenda a Substitutivo
tramitacao_Pedido de Vista
tramitacao_Manifestação de Voto
tramitacao_Declaração de Voto em Separado
tramitacao_Declaração de Prejudicialidade
tramitacao_Prorrogação de prazo para conclusão de CPI
tramitacao_Conclusão de CPI
tramitacao_Manifestação pela Prejudicialidade
tramitacao_Manifestação pela Incompetência
tramitacao_Transformada em Nova Proposição
tramitacao_Recebimento
tramitacao_Recebimento - Relator
tramitacao_Arquivamento
tramitacao_Desarquivamento
tramitacao_Notificação de Apensação
tramitacao_Notificação de Desapensação
tramitacao_Recebimento - Redação Final
tramitacao_Recebimento - Relator (Sem Manifestação)
tramitacao_Recebimento - Parecer Preliminar
tramitacao_Dispensada a Redação Final
tramitacao_Encerramento de Prazo
tramitacao_Saída de membro da comissão
tramitacao_Devolução de Vista
tramitacao_Novo despacho
tramitacao_Publicação de Proposição
tramitacao_Publicação de Despacho
tramitacao_Publicação de Parecer

tramitacao_Recebimento de autógrafos
tramitacao_Remessa á Promulgação
tramitacao_Remessa à Sanção
tramitacao_Submeta-se a Plenário
tramitacao_Publicação de Documento
tramitacao_Perda de eficácia
tramitacao_Reentrada na comissão por arquivamento
tramitacao_Desarquivamento a Pedido
tramitacao_Desarquivamento de Ofício
tramitacao_Conferência Sinopse 2001
tramitacao_Revisão de despacho
tramitacao_Adoção de Proposição pela Comissão
tramitacao_Notificação de Encaminhamento Fora de Fluxo
tramitacao_Deliberação* (inativa)
tramitacao_Pronta para ordem do dia* (inativa)
tramitacao_Despacho* (inativa)
tramitacao_Notificação de Apoiamento
tramitacao_Decisão da Presidência
tramitacao_Notificação - Pasta genérica
tramitacao_Notificação (Sinopse) - Revisão da Ementa / Indexação
tramitacao_Notificação (CeDi) - Legislação Citada
tramitacao_Notificações
tramitacao_Pela Recusa
tramitacao_Recebimento de Retorno
tramitacao_Não Deliberado
tramitacao_Tramitação de Proposição Acessória
tramitacao_Devolução à CCP
tramitacao_Criação de TVRs
tramitacao_Transformado em Norma Jurídica com Veto Parcial
tramitacao_Vetado Totalmente
tramitacao_Ofício de Devolução ao Autor
tramitacao_Questão de Ordem
tramitacao_Oficio Conferência Apoiamento
tramitacao_Despacho
tramitacao_Arquivamento - Art.133 do RI

tramitacao_Deferido o requerimento de retirada pelo autor
tramitacao_Autorização de abertura de prazo recursal
tramitacao_Pareceres Favoráveis nas Comissões
tramitacao_Pareceres Contrários quanto ao Mérito
tramitacao_Parecer pela Inadequação Financeira e/ou Orçamentária
tramitacao_Parecer pela Inconstitucionalidade ou Injuridicidade
tramitacao_Contra Declaração de Prejudicialidade
tramitacao_Contra Despacho de Devolução ao Autor
tramitacao_Contra Despacho pelo Indeferimento
tramitacao_Arquivamento da representação, por inépcia/ausência de justa causa.
tramitacao_Arquivamento da representação, por improcedência.
tramitacao_Despacho Revisto
tramitacao_Pela perda do mandato de deputado federal.
tramitacao_Volta a aguardar criação de nova Comissão Especial.
tramitacao_Ratificação de Parecer
tramitacao_Remessa ao Congresso Nacional
tramitacao_Relatório de Conferência de Assinaturas
tramitacao_Ação de Relatoria
tramitacao_Notificacao para Publicação Intermediária
tramitacao_Encerramento de Comissão Temporária
tramitacao_Substituição de Versão
tramitacao_Apresentação de Quadro Analítico
tramitacao_Não Acolhimento
tramitacao_Providência Interna - Comunicação de Dilatação de Prazo
tramitacao_Providência Interna - Comunicação de resposta ostensiva
tramitacao_Providência Interna - Comunicação de resposta reservada
tramitacao_Leitura de proposição em Plenário
tramitacao_Instauração de Processo - COÉTICA
tramitacao_Sorteio para Escolha do Relator
tramitacao_Início da Instrução Probatória
tramitacao_Encerramento da Instrução Probatória
tramitacao_Sessão Solene
tramitacao_Comissão Geral
tramitacao_Votação (Plenário)
tramitacao_Obstruçao Votação (Plenário)

tramitacao_Destaques (Plenário)
tramitacao_Votação (Outros Requerimentos)
tramitacao_Aprovação de Proposição (Plenário)
tramitacao_Rejeição de Proposição (Plenário)
tramitacao_Manutenção do texto (Plenário)
tramitacao_Supressão do texto (Plenário)
tramitacao_Verificação de Votação (Plenário)
tramitacao_Prejudicialidade (Plenário)
tramitacao_Retirada pelo Autor (Plenário)
tramitacao_Não Acolhimento (Plenário)
tramitacao_Encaminhamento (Plenário)
tramitacao_Encerramentos
tramitacao_Acessórios
tramitacao_Saída de Relator da Comissão - Sem Parecer Apresentado
tramitacao_Saída de Relator da Comissão - Com Parecer Apresentado
tramitacao_Retorno de Relator à Comissão
tramitacao_Aprovação de Requerimento Procedimental
tramitacao_Rejeição de Requerimento Procedimental
tramitacao_Prejudicado Requerimento Procedimental
tramitacao_Retirada de Requerimento Procedimental
tramitacao_Matéria apreciada em Plenário
tramitacao_Notificação do Representado
tramitacao_Abertura de Prazo para Defesa Escrita
tramitacao_Encerramento de Prazo Para Apresentação da Defesa Escrita
tramitacao_Apresentação do Plano de Trabalho
tramitacao_Ofício do Senado Federal
tramitacao_Ofício do Congresso Nacional
tramitacao_Leitura do Oficio CN de encaminhamento de MPV
tramitacao_Suspensão de Tramitação
tramitacao_Aprovação no Senado Federal
tramitacao_Aprovação no Senado Federal - Com Emendas

Tabela I.1: Campos Dummy de Tramitação

Anexo II

Campos Dummy do Autor

siglaPartidoAutorCorrigido_AVANTE
siglaPartidoAutorCorrigido_CIDADANIA
siglaPartidoAutorCorrigido_COMISSÃO DIRETORA
siglaPartidoAutorCorrigido_COMISSÃO ESPECIAL
siglaPartidoAutorCorrigido_COMISSÃO EXTERNA
siglaPartidoAutorCorrigido_COMISSÃO MEDIDA PROVISÓRIA
siglaPartidoAutorCorrigido_COMISSÃO MISTA PERMANENTE
siglaPartidoAutorCorrigido_COMISSÃO PARLAMENTAR DE INQUÉRITO
siglaPartidoAutorCorrigido_COMISSÃO PERMANENTE
siglaPartidoAutorCorrigido_CONSELHO
siglaPartidoAutorCorrigido_DEM
siglaPartidoAutorCorrigido_DPU - Defensoria Pública da União
siglaPartidoAutorCorrigido_MISTA CPI
siglaPartidoAutorCorrigido_MISTA ESPECIAL
siglaPartidoAutorCorrigido_MPU - Ministério Público da União
siglaPartidoAutorCorrigido_NOVO
siglaPartidoAutorCorrigido_PAN
siglaPartidoAutorCorrigido_PATRIOTA
siglaPartidoAutorCorrigido_PCdoB
siglaPartidoAutorCorrigido_PDT
siglaPartidoAutorCorrigido_PEN
siglaPartidoAutorCorrigido_PERMANENTE DO SENADO FEDERAL
siglaPartidoAutorCorrigido_PFL
siglaPartidoAutorCorrigido_PHS

siglaPartidoAutorCorrigido_PL
siglaPartidoAutorCorrigido_PMB
siglaPartidoAutorCorrigido_PMDB
siglaPartidoAutorCorrigido_PMN
siglaPartidoAutorCorrigido_PMR
siglaPartidoAutorCorrigido_PODE
siglaPartidoAutorCorrigido_PP
siglaPartidoAutorCorrigido_PPB
siglaPartidoAutorCorrigido_PPS
siglaPartidoAutorCorrigido_PR
siglaPartidoAutorCorrigido_PRB
siglaPartidoAutorCorrigido_PRONA
siglaPartidoAutorCorrigido_PROS
siglaPartidoAutorCorrigido_PRP
siglaPartidoAutorCorrigido_PRTB
siglaPartidoAutorCorrigido_PSB
siglaPartidoAutorCorrigido_PSC
siglaPartidoAutorCorrigido_PSD
siglaPartidoAutorCorrigido_PSDB
siglaPartidoAutorCorrigido_PSDC
siglaPartidoAutorCorrigido_PSL
siglaPartidoAutorCorrigido_PSOL
siglaPartidoAutorCorrigido_PST
siglaPartidoAutorCorrigido_PT
siglaPartidoAutorCorrigido_PTB
siglaPartidoAutorCorrigido_PTC
siglaPartidoAutorCorrigido_PTN
siglaPartidoAutorCorrigido_PTdoB
siglaPartidoAutorCorrigido_PV
siglaPartidoAutorCorrigido_REDE
siglaPartidoAutorCorrigido_REPUBLIC
siglaPartidoAutorCorrigido_S.PART.
siglaPartidoAutorCorrigido_SD
siglaPartidoAutorCorrigido_Sociedade Civil
siglaPartidoAutorCorrigido_ZZ-Indeterminado

siglaPartidoAutorCorrigido_Órgão do Poder Executivo
siglaPartidoAutorCorrigido_Órgão do Poder Judiciário
siglaPartidoAutorCorrigido_Órgão do Poder Legislativo
siglaPartidoAutorCorrigido_Órgão do Senado Federal

Tabela II.1: Campos Dummy de Autor

Anexo III

Campos Dummy do Tema

tema_Administração Pública
tema_Agricultura, Pecuária, Pesca e Extrativismo
tema_Arte, Cultura e Religião
tema_Cidades e Desenvolvimento Urbano
tema_Ciência, Tecnologia e Inovação
tema_Comunicações
tema_Defesa e Segurança
tema_Direito Civil e Processual Civil
tema_Direito Constitucional
tema_Direito Penal e Processual Penal
tema_Direito e Defesa do Consumidor
tema_Direito e Justiça
tema_Direitos Humanos e Minorias
tema_Economia
tema_Educação
tema_Energia, Recursos Hídricos e Minerais
tema_Esporte e Lazer
tema_Estrutura Fundiária
tema_Finanças Públicas e Orçamento
tema_Homenagens e Datas Comemorativas
tema_Indústria, Comércio e Serviços
tema_Meio Ambiente e Desenvolvimento Sustentável
tema_Política, Partidos e Eleições
tema_Previdência e Assistência Social

tema_Processo Legislativo e Atuação Parlamentar
tema_Relações Internacionais e Comércio Exterior
tema_Saúde
tema_Sem Tema
tema_Trabalho e Emprego
tema_Turismo
tema_Viação, Transporte e Mobilidade

Tabela III.1: Campos Dummy de Tema

Anexo IV

Ranking RTE sobre os dados categóricos

Tramitacao	Considerado	Ranking
siglaPartidoAutorCorrigido_Órgão do Poder Legislativo	True	1
siglaPartidoAutorCorrigido_Órgão do Poder Judiciário	True	1
projeto_lei_conversao	True	1
regime_ordinaria	True	1
regime_urgencia	True	1
siglaPartidoAutorCorrigido_Órgão do Poder Executivo	True	1
tema_Homenagens e Datas Comemorativas	True	1
tema_Direito Civil e Processual Civil	False	2
tema_Arte, Cultura e Religião	False	2
tema_Finanças Públicas e Orçamento	False	2
regime_prioridade	False	2
siglaPartidoAutorCorrigido_MPU - Ministério Público da União	False	2
siglaPartidoAutorCorrigido_COMISSÃO DIRETORA	False	3
tema_Educação	False	3
siglaPartidoAutorCorrigido_COMISSÃO MEDIDA PROVISÓRIA	False	3
siglaPartidoAutorCorrigido_PL	False	3
projeto_lei_complementar	False	3
siglaPartidoAutorCorrigido_PTB	False	4
tema_Viação, Transporte e Mobilidade	False	4

tema_Direitos Humanos e Minorias	False	4
tema_Economia	False	4
regime_especial	False	4
tema_Política, Partidos e Eleições	False	5
siglaPartidoAutorCorrigido_PSD	False	5
siglaPartidoAutorCorrigido_PPS	False	5
siglaPartidoAutorCorrigido_PCdoB	False	5
siglaPartidoAutorCorrigido_PFL	False	5
siglaPartidoAutorCorrigido_PRONA	False	6
tema_Direito e Justiça	False	6
projeto_lei	False	6
siglaPartidoAutorCorrigido_COMISSÃO PERMANENTE	False	6
projeto_emenda	False	6
siglaPartidoAutorCorrigido_PPB	False	7
tema_Administração Pública	False	7
siglaPartidoAutorCorrigido_PTC	False	7
siglaPartidoAutorCorrigido_PT	False	7
tema_Trabalho e Emprego	False	7
tema_Meio Ambiente e Desenvolvimento Sustentável	False	8
siglaPartidoAutorCorrigido_PR	False	8
siglaPartidoAutorCorrigido_PRB	False	8
tema_Direito Penal e Processual Penal	False	8
siglaPartidoAutorCorrigido_S.PART.	False	8
tema_Estrutura Fundiária	False	9
siglaPartidoAutorCorrigido_PODE	False	9
siglaPartidoAutorCorrigido_PP	False	9
tema_Previdência e Assistência Social	False	9
tema_Indústria, Comércio e Serviços	False	9
tema_Esporte e Lazer	False	10
tema_Energia, Recursos Hídricos e Minerais	False	10
tema_Saúde	False	10
siglaPartidoAutorCorrigido_PSDB	False	10
siglaPartidoAutorCorrigido_PMDB	False	10
siglaPartidoAutorCorrigido_PDT	False	11

tema_Defesa e Segurança	False	11
tema_Cidades e Desenvolvimento Urbano	False	11
siglaPartidoAutorCorrigido_SD	False	11
tema_Processo Legislativo e Atuação Parlamentar	False	11
tema_Sem Tema	False	12
siglaPartidoAutorCorrigido_AVANTE	False	12
tema_Comunicações	False	12
siglaPartidoAutorCorrigido_REPUBLIC	False	12
tema_Relações Internacionais e Comércio Exterior	False	12
tema_Direito Constitucional	False	13
siglaPartidoAutorCorrigido_CIDADANIA	False	13
siglaPartidoAutorCorrigido_COMISSÃO EXTERNA	False	13
siglaPartidoAutorCorrigido_DEM	False	13
siglaPartidoAutorCorrigido_CONSELHO	False	13
siglaPartidoAutorCorrigido_Órgão do Senado Federal	False	14
siglaPartidoAutorCorrigido_COMISSÃO ESPECIAL	False	14
siglaPartidoAutorCorrigido_PTN	False	14
siglaPartidoAutorCorrigido_COMISSÃO MISTA PERMANENTE	False	14
siglaPartidoAutorCorrigido_COMISSÃO PARLAMENTAR DE INQUÉRITO	False	14
tema_Direito e Defesa do Consumidor	False	15
siglaPartidoAutorCorrigido_DPU - Defensoria Pública da União	False	15
siglaPartidoAutorCorrigido_PSB	False	15
siglaPartidoAutorCorrigido_PSC	False	15
siglaPartidoAutorCorrigido_ZZ-Indeterminado	False	15
siglaPartidoAutorCorrigido_PRTB	False	16
tema_Agricultura, Pecuária, Pesca e Extrativismo	False	16
tema_Turismo	False	16
tema_Ciência, Tecnologia e Inovação	False	16
siglaPartidoAutorCorrigido_PROS	False	16
siglaPartidoAutorCorrigido_PST	False	17
siglaPartidoAutorCorrigido_PMR	False	17
siglaPartidoAutorCorrigido_PRP	False	17

siglaPartidoAutorCorrigido_PMB	False	17
siglaPartidoAutorCorrigido_PMN	False	17
siglaPartidoAutorCorrigido_PTdoB	False	18
siglaPartidoAutorCorrigido_PSOL	False	18
siglaPartidoAutorCorrigido_MISTA CPI	False	18
siglaPartidoAutorCorrigido_MISTA ESPECIAL	False	18
siglaPartidoAutorCorrigido_PV	False	18
siglaPartidoAutorCorrigido_PAN	False	19
siglaPartidoAutorCorrigido_REDE	False	19
siglaPartidoAutorCorrigido_NOVO	False	19
siglaPartidoAutorCorrigido_PSL	False	19
siglaPartidoAutorCorrigido_PATRIOTA	False	19
siglaPartidoAutorCorrigido_PEN	False	20
siglaPartidoAutorCorrigido_PSDC	False	20
siglaPartidoAutorCorrigido_PHS	False	20
siglaPartidoAutorCorrigido_PERMANENTE DO SENADO FEDERAL	False	20
siglaPartidoAutorCorrigido_Sociedade Civil	False	20
tema_Comunicações	False	21

Tabela IV.1: Ranking RFE dados categóricos

Anexo V

Ranking RTE sobre os dados compostos

Tramitacao	Considerado	Ranking
projeto_lei	True	1
tema_Arte, Cultura e Religião	True	1
tema_Agricultura, Pecuária, Pesca e Extrativismo	True	1
tema_Administração Pública	True	1
siglaPartidoAutorCorrigido_Órgão do Senado Federal	True	1
siglaPartidoAutorCorrigido_Órgão do Poder Legislativo	True	1
siglaPartidoAutorCorrigido_Órgão do Poder Judiciário	True	1
siglaPartidoAutorCorrigido_Órgão do Poder Executivo	True	1
siglaPartidoAutorCorrigido_ZZ-Indeterminado	True	1
siglaPartidoAutorCorrigido_Sociedade Civil	True	1
siglaPartidoAutorCorrigido_SD	True	1
tema_Cidades e Desenvolvimento Urbano	True	1
siglaPartidoAutorCorrigido_S.PART.	True	1
siglaPartidoAutorCorrigido_REDE	True	1
siglaPartidoAutorCorrigido_PV	True	1
siglaPartidoAutorCorrigido_PTdoB	True	1
siglaPartidoAutorCorrigido_PTN	True	1
siglaPartidoAutorCorrigido_PTC	True	1
siglaPartidoAutorCorrigido_PTB	True	1
siglaPartidoAutorCorrigido_PT	True	1
siglaPartidoAutorCorrigido_PST	True	1

siglaPartidoAutorCorrigido_PSOL	True	1
siglaPartidoAutorCorrigido_PSL	True	1
siglaPartidoAutorCorrigido_REPUBLIC	True	1
siglaPartidoAutorCorrigido_PSDC	True	1
tema_Ciência, Tecnologia e Inovação	True	1
tema_Defesa e Segurança	True	1
tema_Trabalho e Emprego	True	1
tema_Sem Tema	True	1
tema_Saúde	True	1
tema_Relações Internacionais e Comércio Exterior	True	1
tema_Processo Legislativo e Atuação Parlamentar	True	1
tema_Previdência e Assistência Social	True	1
tema_Política, Partidos e Eleições	True	1
tema_Meio Ambiente e Desenvolvimento Sustentável	True	1
tema_Indústria, Comércio e Serviços	True	1
tema_Homenagens e Datas Comemorativas	True	1
tema_Comunicações	True	1
tema_Finanças Públicas e Orçamento	True	1
tema_Esporte e Lazer	True	1
tema_Energia, Recursos Hídricos e Minerais	True	1
tema_Educação	True	1
tema_Economia	True	1
tema_Direitos Humanos e Minorias	True	1
tema_Direito e Justiça	True	1
tema_Direito e Defesa do Consumidor	True	1
tema_Direito Penal e Processual Penal	True	1
tema_Direito Constitucional	True	1
tema_Direito Civil e Processual Civil	True	1
tema_Estrutura Fundiária	True	1
tema_Turismo	True	1
siglaPartidoAutorCorrigido_PSDB	True	1
siglaPartidoAutorCorrigido_PSC	True	1
siglaPartidoAutorCorrigido_MISTA ESPECIAL	True	1
siglaPartidoAutorCorrigido_MISTA CPI	True	1

siglaPartidoAutorCorrigido_DPU - Defensoria Pública da União	True	1
siglaPartidoAutorCorrigido_DEM	True	1
siglaPartidoAutorCorrigido_CONSELHO	True	1
siglaPartidoAutorCorrigido_COMISSÃO PERMANENTE	True	1
siglaPartidoAutorCorrigido_COMISSÃO PARLAMENTAR DE INQUÉRITO	True	1
siglaPartidoAutorCorrigido_COMISSÃO MISTA PERMANENTE	True	1
siglaPartidoAutorCorrigido_COMISSÃO MEDIDA PROVISÓRIA	True	1
siglaPartidoAutorCorrigido_COMISSÃO EXTERNA	True	1
siglaPartidoAutorCorrigido_MPU - Ministério Público da União	True	1
siglaPartidoAutorCorrigido_COMISSÃO ESPECIAL	True	1
siglaPartidoAutorCorrigido_CIDADANIA	True	1
siglaPartidoAutorCorrigido_AVANTE	True	1
aprovado_texto	True	1
regime_especial	True	1
regime_prioridade	True	1
regime_urgencia	True	1
regime_ordinaria	True	1
projeto_lei_conversao	True	1
projeto_lei_complementar	True	1
projeto_emenda	True	1
siglaPartidoAutorCorrigido_COMISSÃO DIRETORA	True	1
siglaPartidoAutorCorrigido_PSD	True	1
siglaPartidoAutorCorrigido_NOVO	True	1
siglaPartidoAutorCorrigido_PATRIOTA	True	1
siglaPartidoAutorCorrigido_PSB	True	1
siglaPartidoAutorCorrigido_PRTB	True	1
siglaPartidoAutorCorrigido_PRP	True	1
siglaPartidoAutorCorrigido_PROS	True	1
siglaPartidoAutorCorrigido_PRONA	True	1
siglaPartidoAutorCorrigido_PRB	True	1

siglaPartidoAutorCorrigido_PR	True	1
siglaPartidoAutorCorrigido_PPS	True	1
siglaPartidoAutorCorrigido_PPB	True	1
siglaPartidoAutorCorrigido_PP	True	1
siglaPartidoAutorCorrigido_PAN	True	1
siglaPartidoAutorCorrigido_PODE	True	1
siglaPartidoAutorCorrigido_PMN	True	1
siglaPartidoAutorCorrigido_PMDB	True	1
siglaPartidoAutorCorrigido_PMB	True	1
siglaPartidoAutorCorrigido_PL	True	1
siglaPartidoAutorCorrigido_PHS	True	1
siglaPartidoAutorCorrigido_PFL	True	1
siglaPartidoAutorCorrigido_PERMANENTE DO SENADO FEDERAL	True	1
siglaPartidoAutorCorrigido_PEN	True	1
siglaPartidoAutorCorrigido_PDT	True	1
siglaPartidoAutorCorrigido_PCdoB	True	1
siglaPartidoAutorCorrigido_PMR	True	1
tema_Viação, Transporte e Mobilidade	True	1

Tabela V.1: Ranking RFE dados compostos