UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Compreendendo e prevendo o processo legislativo via ciência de dados

Danilo Amaral de Oliveira

Dissertação de Mestrado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

ICMC USP
SÃO CARLOS

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Danilo Amaral de Oliveira

Compreendendo e prevendo o processo legislativo via ciência de dados

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências – Ciências de Computação e Matemática Computacional. *EXEMPLAR DE DEFESA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. João Porto de Albuquerque
Coorientador: Prof. Dr. Alexandre C. B. Delbem

USP – São Carlos
Novembro de 2018

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

A485c

Amaral de Oliveira, Danilo
Compreendendo e prevendo o processo legislativo via ciência de dados / Danilo Amaral de Oliveira; orientador João Porto de Albuquerque; coorientador Alexandre C. B. Delbem. -- São Carlos, 2018.
102 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2018.

1. Ciência de Dados. 2. Processo Legislativo. 3. Dados Abertos. I. Porto de Albuquerque, João, orient. II. C. B. Delbem, Alexandre, coorient. III. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Danilo Amaral de Oliveira

Understanding and predicting the legislative process via data science

Master dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP, in partial fulfillment of the requirements for the degree of the Master Program in Computer Science and Computational Mathematics. *EXAMINATION BOARD PRESENTATION COPY*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. João Porto de Albuquerque
Co-advisor: Prof. Dr. Alexandre C. B. Delbem

USP – São Carlos
November 2018

Aos inconformados e inconformadas desse mundo.

AGRADECIMENTOS

Agradecer é o mínimo que eu posso fazer para as pessoas que contribuíram e me ajudaram a produzir esse trabalho. É verdade que eu também poderia promover uma festa, porém são muitas pessoas e por isso ficaria muito caro, então deixo somente o meu agradecimento.

Para guiar o leitor, quero deixar os meus agradecimentos em ordem cronológica, todavia a fim de evitar qualquer injustiça gostaria de ressaltar que o primeiro não necessariamente foi o que mais me ajudou ou o último, o menos.

Em primeiro lugar gostaria de agradecer as três pessoas que confiaram e apostaram em mim escrevendo as cartas de recomendação que me ajudaram a conquistar uma vaga na USP que são: o meu antigo coordenador de curso de graduação da UFSCar, o Professor Orides Morandin Junior; o meu orientador de Iniciação Científica o Professor Antonio Francisco do Prado; e o meu primeiro chefe no CPqD, Manuel Gavidea.

Em segundo lugar, gostaria de agradecer ao meu orientador João Porto de Albuquerque e ao meu coorientador Alexandre C. B. Delbem que me deram a oportunidade e a liberdade de explorar as fronteiras entre duas grandes paixões a Ciência da Computação e a Ciência Política, além de toda orientação e tempo investidos em me guiar nesse trabalho.

Em terceiro lugar, gostaria de agradecer a minha família, em especial ao meus pais, Antonio Ferreira e Maria Amália e ao meu irmão Frederico Oliveira que sempre me apoiaram nos momentos mais difíceis.

Em quarto lugar, gostaria aos amigos Dr. Ivan Ervolino e Dr. Leonardo Borges que ajudaram na estruturação e no desenvolvimento do trabalho.

Em quinto lugar, gostaria de agradecer aos membros do grupo de pesquisa AGORA que me apoiaram em diversos momentos durante o curso.

Em sexto lugar, gostaria de agradecer ao Instituto de Ciências Matemáticas e de Computação (ICMC) da USP e aos seus professores e funcionários que trabalham arduamente para criar e manter uma instituição de ensino e pesquisa de excelência e referência internacional.

Em sétimo lugar, aos amigos do sigalei pelo apoio e companheirismo durante a jornada.

Em oitavo lugar, a Ana Cláudia Junqueira pelo apoio e revisão sistemática do texto final.

Por último, mas não menos importante, a todas as pessoas que de alguma maneira me ajudaram com ideias, sugestões, conversas e revisões que influenciaram esse trabalho.

Obrigado!

A tecnologia molda a sociedade.
A sociedade molda a tecnologia.
(Desconhecido)

RESUMO

OLIVEIRA, D. A. **Compreendendo e prevendo o processo legislativo via ciência de dados**. 2018. 102 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2018.

Apenas 10% dos cerca de 1.6 mil novos Projetos de Lei Ordinária apresentados anualmente na Câmara dos Deputados são aprovados. Uma grande quantidade de tempo e de recursos são investidos em análises qualitativas para entender e identificar quais projetos possuem ou não a chance de serem aprovados. Porém, as análises qualitativas manuais são tarefas difíceis e caras devido ao grande volume de projetos e a complexidade do Processo Legislativo. Neste trabalho desenvolvemos um artefato de *software* capaz de identificar as características que influenciam na aprovação ou arquivamento dos projetos visando automatizar a tarefa de análise. O artefato foi desenvolvido em três etapas. Na primeira etapa foi criado um banco de dados de projetos de lei apresentados entre 2003 e 2016 e com 72 variáveis. A base de dados foi dividida em 7 conjuntos, estes submetidos a análises de mineração de dados. Na segunda etapa, por meio do LASSO, selecionamos as 20 características de maior poder preditivo. Na terceira etapa, submetemos essas variáveis a três análises de Aprendizado de Máquina: Regressão Logística, LASSO e Floresta Aleatória. O cruzamento dos resultados das análises de Aprendizado de Máquina, identificou 7 características do Processo Legislativo estatisticamente relevantes para a aprovação ou para o arquivamento de um projeto. Finalmente, analisamos a própria base de dados de projetos de lei para testar a capacidade preditiva do artefato. Por meio da Floresta Aleatória, obtivemos uma performance preditiva média de F1-Score de 0.861 com uma precisão positiva de 0.778. Nossos resultados indicam que é possível utilizar inteligência artificial para prever em parte o comportamento do processo legislativo. O uso do artefato desenvolvido ajudará diferentes grupos sociais e econômicos a se anteciparem a mudanças legislativas.

Palavras-chave: Processo Legislativo, Ciência de Dados, Dados Abertos.

ABSTRACT

OLIVEIRA, D. A. **Understanding and predicting the legislative process via data science**. 2018. 102 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2018.

Only 10% of the approximately 1.6 thousand new ordinary bills presented annually in the Chamber of Deputies of Brazil are approved. A vast amount of time and resources are used in qualitative analysis aiming to understand and identify which bills have the chance of being approved. However, the volume of bills and the complexity of the Legislative Process make these manual qualitative analyzes difficult and expensive. In this work we developed a software artifact capable of identifying variables that influence the approval or archiving of the bill in order to automate the analysis task. The artifact was developed in three stages. In the first stage we created a database of bills submitted between 2003 and 2016 and with 72 variables. We divided the database into 7 sets, which were subjected to data mining analyzes. In the second stage, through the LASSO, we selected the 20 characteristics with the highest predictive power. In the third step, we submitted these variables to three Machine Learning analyzes: Logistic Regression, LASSO and Random Forest. Crossing the results of the Machine Learning analyzes, we identified 7 characteristics of the Legislative Process statistically relevant to the approval or archiving of a bill. Finally, we analyze the database of bills to test the predictive power of the artifact itself. Through the Random Forest, we obtained an average predictive performance of F1-Score equal to 0.861 with a positive precision of 0.778. Our results indicate that it is possible to use artificial intelligence to predict part of the Legislative Process behavior. Our artifact can aid different social and economic groups to anticipate legislative changes.

Keywords: Legislative Process, Data Science, Open Data.

LISTA DE ILUSTRAÇÕES

Figura 1 – Diagrama de Venn para Ciência de Dados 30
Figura 2 – Modelo de um programa baseado em Aprendizado de Máquina em comparação com o modelo tradicional 31
Figura 3 – Contraste entre as tarefas de classificação e regressão. 32
Figura 4 – Taxas de Erros em modelos de Aprendizado de Máquina sub-ajustado versus ajustado de forma ótima versus super-ajustado. 35
Figura 5 – Predição de sobrevivência de um passageiro no Titanic aplicando o Modelo CART 37
Figura 6 – Estrutura do modelo da Floresta Aleatória 38
Figura 7 – Pirâmide de Kelsen, configurada para os tipos de leis brasileiras 38
Figura 8 – Fluxograma Simplificado do Processo Legislativo de um PLO no Congresso Nacional. 40
Figura 9 – Metodologia de Pesquisa de Ciência de Design proposta por Peffers para pesquisas na área de Sistemas de Informação. 56
Figura 10 – Configuração dos experimentos para validar o melhor método de balanceamento de dados. 60
Figura 11 – Total de PLOs aprovados/arquivados na Câmara dos Deputados 66
Figura 12 – Assuntos dos PLOs aprovados e arquivados entre 2003 a 2016 68
Figura 13 – Resultados dos Experimentos realizados conjuntos balanceados (*oversample* e *undersample*) e desbalanceados 69
Figura 14 – Resultados da seleção LASSO de características, para cada biênio 70
Figura 15 – Resultado dos experimentos realizados para escolher a frequência de corte. . 71
Figura 16 – Coeficientes estimados pela Regressão Logística para cada sinal nos 7 conjuntos de dados 73
Figura 17 – Impacto das 20 características nos 6 experimentos. 74
Figura 18 – Importância das características nos 6 experimentos para a decisão de classificação da Floresta Aleatória. 76
Figura 19 – Comparação da “p_velocidade” entre os PLOs aprovados (1) e os PLOs arquivados (0) entre 2003 e 2016. 80
Figura 20 – Análise da característica “p_total_tramitacoes” dos PLOs aprovados e arquivados entre 2003 e 2016. 81
Figura 21 – Projetos por total de pareceres recebidos 83

Figura 22 – Proporção dos PLOs analisados por somente pelas comissões em contraste com os PLOs analisados pelo Plenário. PLOs aprovados/arquivados entre 2003 e 2016. 84
Figura 23 – Regime de Tramitação dos PLOs aprovados/arquivados entre 2003 e 2016. . 85
Figura 24 – Total de dias em tramitação dos PLOs aprovados entre 2003 e 2016. 85
Figura 25 – Histograma dos PLOs pelo seu tempo de tramitação em dias. 86

LISTA DE TABELAS

Tabela 1 – Principais modelos utilizados pelos trabalhos relacionados 53
Tabela 2 – Dados coletados provenientes do Portal de Dados Abertos da Câmara 59
Tabela 3 – Características pré-selecionadas 67
Tabela 4 – Total de vezes que os 20 sinais foram selecionados como estatisticamente significantes. 72
Tabela 5 – Resultados dos experimentos utilizando o modelo da Regressão Logística com penalidade LASSO. 75
Tabela 6 – Resultados dos experimentos utilizando o modelo da Floresta Aleatória . . . 76
Tabela 7 – Comparação dos resultados dos Experimentos de Predição 88
Tabela 8 – VPN (Valor Preditivo Negativo) da Floresta Aleatória 88

LISTA DE ABREVIATURAS E SIGLAS

ABCP	Associação Brasileira de Ciência Política
Anpocs	Associação Nacional de Pós-Graduação e Pesquisa em Ciências Sociais
CART	*Classification And Regression Tree*
FGV	Fundação Getúlio Vargas
LAI	Lei de Acesso à Informação
LASSO	*Least Absolute Shrinkage and Selection Operator*
LOs	Leis Ordinárias
MLG	Modelo Linear Generalizado
OGP	Open Government Partnership
PLO	Projeto de Lei Ordinária
SI	Sistemas de Informação
SMOTE	*Synthetic Minority Over-sampling Technique*

SUMÁRIO

1 INTRODUÇÃO 23
1.1 Contextualização 23
1.2 Problema de Pesquisa 25
1.3 Objetivos 26
1.4 Organização do Trabalho 27

2 FUNDAMENTAÇÃO TEÓRICA 29
2.1 Ciência de Dados 29
2.1.1 Aprendizado de Máquina 31
2.1.1.1 Aprendizado Supervisionado e Aprendizado Não-Supervisionado 32
2.1.1.2 Modelo Linear Generalizado (MLG) 33
2.1.1.3 Regressão Logística 33
2.1.1.4 Sub-ajuste vs Super-ajuste 34
2.1.1.5 Regressão Ridge 35
2.1.1.6 Regressão LASSO 36
2.1.1.7 Árvore de Decisão (Decision Tree) 36
2.1.1.8 Floresta Aleatória (Random Forests) 37
2.2 As leis e o Processo Legislativo 38
2.2.1 As Leis 38
2.2.2 Leis Ordinária 39
2.2.3 O funcionamento do Processo Legislativo 40
2.3 Governo Aberto 41
2.3.1 Dados Abertos Governamentais 42
2.3.2 Portal de Dados Abertos da Câmara dos Deputados 44
2.4 Grupos de Pressão 44

3 TRABALHOS RELACIONADOS 47
3.1 Ciência de Dados no contexto da Ciência Política norte-americana 47
3.2 Ciência de Dados no contexto brasileiro 49

4 METODOLOGIA 55
4.1 Identificação do problema e motivação 57
4.2 Definição dos objetivos da solução 58
4.3 Design e Desenvolvimento 58

4.3.1 Coleta e Balanceamento dos dados 58
4.3.2 Seleção das Características 60
4.3.3 Aplicação dos modelos 61
4.4 Demonstração 61
4.5 Avaliação 62
4.6 Comunicação 64
4.7 Tecnologias e Ferramentas utilizadas 64

5 RESULTADOS 65
5.1 Coleta e balanceamento dos Dados 65
5.2 Seleção de Características 70
5.3 Aplicação da Regressão Logística 72
5.4 Aplicação do LASSO 74
5.5 Floresta Aleatória 76

6 DISCUSSÃO 79
6.1 Objetivo 1: Capacidade Explicativa 79
6.2 Objetivo 2: Capacidade Preditiva 87

7 CONCLUSÃO 91
7.1 Trabalhos Futuros 93

REFERÊNCIAS 95

APÊNDICE A CARACTERÍSTICAS PRÉ-SELECIONADAS 99

CAPÍTULO

1

INTRODUÇÃO

1.1 Contextualização

O Congresso Nacional por meio do Processo Legislativo é responsável por propor e debater projetos que criam e alteram as leis do país, as quais regulamentam os diversos aspectos da vida em sociedade. Sendo assim, as atividades do Congresso Nacional e as suas decisões impactam diretamente grande parte da vida dos cidadãos brasileiros (MENEGUIN, 2010).

Em uma democracia representativa como a vivenciada no Brasil os principais responsáveis por propor e aprovar as leis são os parlamentares que, por sua vez, pertencem a uma pequena parcela de pessoas que são eleitas periodicamente e que na teoria deveriam representar e defender todos os grupos da sociedade, todavia devido a diversos motivos não é incomum que os grupos estejam sub representados e as leis produzidas estejam desconectadas com as reais demandas sociais.

Paralelamente existem grupos sociais que realizam um trabalho sistemático de monitoramento e participação no Processo Legislativo e que são chamados de Grupos de Pressão. O objetivo de um Grupo de Pressão é levar aos parlamentares as suas opiniões e as suas visões sobre os projetos de lei em discussão e com isso contribuir para a produção de uma legislação que contemple as suas demandas. Por isso, é importante que toda a sociedade se organize em grupos para monitorar e participar do Processo Legislativo visando minimizar o risco da produção de leis inócuas e dessincronizadas com os problemas enfrentados pela população (MANCUSO; GOZETTO, 2011).

Os benefícios da participação política para os Grupos de Pressão são evidentes, porém antes de pressionar os parlamentares, é preciso investir tempo e recursos para monitorar e analisar o Processo Legislativo. Em um primeiro momento é necessário acompanhar todas as informações geradas pelo Congresso Nacional para posteriormente entender quais são as ações que o grupo deve tomar para tentar garantir que a sua visão sobre um determinado projeto de lei se torne a

posição vencedora.

Ainda são raras as pesquisas que produzam artefatos de Sistemas de Informação capazes de auxiliar o acompanhamento e a análise do Processo Legislativo no Brasil, por isso a maioria dos grupos ainda depende de métodos subjetivos e manuais para analisar cenários ou para desenhar um prognóstico de um projeto de lei. Por exemplo, entre 10 projetos de lei quais são os projetos que possuem maiores chances de serem aprovados? Dado um projeto de lei que é favorável a um interesse de um grupo quais são os aspectos que são necessários analisar para acelerar a sua aprovação? Nos dias de hoje, as respostas dessas perguntas dependem da habilidade de um analista político em ler e cruzar variáveis relacionadas ao Processo Legislativo. Por sua vez, um processo subjetivo aumenta os custos e compromete a análise, visto que depende da qualidade da pessoa responsável por produzir prognósticos. O atual cenário gera uma assimetria entre os Grupos de Pressão, pois àqueles que possuem mais recursos conseguem contratar mais analistas de alta qualidade, enquanto àqueles grupos com menos recursos não conseguem e por muitas vezes ficam impedidos de participar devido aos custos da sistematização e análise das informações do Processo Legislativo.

A recente sistematização e modernização dos sistemas internos de apoio ao Processo Legislativo do Congresso Nacional juntamente com a formulação de diversas políticas públicas de promoção da transparência governamental, permitiu que diversos conjuntos de dados legislativos gerados pelas atividades parlamentares fossem disponibilizados à sociedade.

Simultaneamente, enquanto é necessário que as tarefas de análise do Processo Legislativo sejam mais objetivas e baratas, a área de Inteligência Artificial e Aprendizado de Máquina avança aceleradamente automatizando trabalhos repetitivos em fábricas ou superando desafios mais ousados que demandam um alto poder cognitivo, por exemplo, como ocorreu em 2017 quando um artefato de IA (o *AlphaIA*) venceu jogadores profissionais em partidas do jogo GO.

Assim sendo, com a motivação de democratizar a atuação dos grupos, nós nos perguntamos: É possível, a partir dos dados, criar um artefato capaz de prover explicações sobre os principais fatores que contribuem para a criação de uma nova lei visando facilitar a participação dos Grupos de Pressão no Processo Legislativo no Congresso Nacional?

Nos Estados Unidos, devido a maturidade do seu sistema democrático e a seu protagonismo na tecnologia, já existem artefatos que usam técnicas modernas de Inteligência Artificial que auxiliam os Grupos de Pressão daquele país a analisar as atividades do seu Congresso Nacional. Como veremos na seção de Trabalhos Relacionados, no Brasil ainda há poucas iniciativas que unem tecnologia e política.

Em síntese, considerando a importância da participação no Processo Legislativo e o desafio de analisar subjetivamente uma grande quantidade de informações e variáveis, nossa proposta foi criar um artefato baseado em dados capaz de explicar e prever o Processo Legislativo no Congresso Nacional. Assim, os mais diversos Grupos de Pressão no Brasil poderão reduzir os

seus custos e ao mesmo tempo aumentar a qualidade da sua atuação. Por fim, em um cenário no qual há mais grupos participando da política espera-se que as leis criadas estejam mais alinhadas às demandas da sociedade.

1.2 Problema de Pesquisa

A Constituição Cidadã promulgada em 1988 produziu o atual Processo Legislativo utilizado pelo Congresso Nacional. Desde então, o Brasil vivencia o período mais longevo de democracia de sua história que possibilitou o amadurecimento das instituições e dos processos aplicados pela administração pública, incluindo o Processo Legislativo.

No decorrer do aprimoramento do Processo Legislativo ocorreu a sua sistematização. O Congresso Nacional, tanto a Câmara dos Deputados como o Senado Federal desenvolveram e implementaram sistemas para apoiar as etapas do Processo Legislativo e aumentar a garantia que as suas regras estão sendo cumpridas. Consequentemente, os sistemas começaram a gerar dados estruturados e não estruturados sobre os projetos aprovados ou arquivados pelos parlamentares.

Os dados legislativos abriram uma nova oportunidade de entender os aspectos e características do Processo Legislativo. No campo da Ciência Política há diversos trabalhos que analisaram o Processo Legislativo e tentaram indicar as principais características que determinam a aprovação de um projeto de lei ou os padrões de comportamento dos parlamentares em determinadas situações. Porém, grande parte deles são baseados em análises qualitativas a partir de uma pequena quantidade de dados setorizados. Por exemplo, o estudo de Calazans (2010) analisa a aprovação da popular Lei da Maria da Penha a partir dos projetos de lei que foram apresentados para combater a violência contra a mulher.

Apesar dos estudos não indicarem uma solução genérica que explica a aprovação de qualquer projeto de lei em qualquer período de tempo, eles são sinais da maturidade do Processo Legislativo, indicando que a construção de leis respeitam um comportamento estável.

Por outro lado, a análise do conjunto de dados e a aplicação de técnicas de Ciência de Dados permite a construção de artefatos genéricos, capazes de gerar *insights* relativos a quaisquer projetos de lei. Ao mesmo tempo, aplicar essas tecnologias não é uma tarefa trivial, visto que há diversos desafios em unir duas áreas completamente diferentes, a Ciência Política e a Ciência da Computação.

O primeiro desafio está relacionado ao volume de projetos, a cada ano em torno de 2 mil novos Projetos de Lei Ordinária são apresentados no Congresso Nacional. Se somarmos todos os projetos apresentados, desde a redemocratização do país contabilizamos mais de 62 mil projetos. O segundo desafio está relacionado a taxa de aprovação, pois apesar do grande volume de projetos apresentados, poucos deles são aprovados. Desde 1990, aproximadamente, somente 9% dos 62 mil projetos de leis apresentados no Congresso Nacional tornaram-se leis.

Outro desafio na análise de dados é a dinamicidade dos parlamentares, os representantes eleitos podem ser substituídos a cada eleição e a todo momento parlamentares saem e voltam de seus cargos. Um outro desafio é a temporalidade da política, a cada 4 anos o Congresso Nacional é renovado e entre esses 4 anos há períodos que são mais propensos ou não a criação de leis.

Ou seja, para aplicar modelos de Aprendizado de Máquina e construir um artefato genérico baseado em dados é preciso considerar ao menos o grande volume de dados, a baixa taxa de aprovação, a dinamicidade e a temporalidade do Processo Legislativo.

Dado o contexto, a motivação, os desafios e os detalhes do problema elegemos a nossa pergunta de pesquisa: é possível criar um artefato capaz de prover explicações e prognósticos, baseadas em dados, do Processo Legislativo no Congresso Nacional?

1.3 Objetivos

Após realizar uma análise inicial do problema, o escopo dessa pesquisa foi reduzido em dois aspectos, no que diz à abrangência do Processo Legislativo e no que diz ao tipo de projeto de lei que está sendo debatido pelo Congresso Nacional.

A princípio, o objetivo era criar um artefato capaz de generalizar as explicações e prognósticos de todo tipo de lei gerado pelo Congresso Nacional. Porém, após analisar as regras e contexto bicameral do Processo Legislativo identificamos que há diferenças substanciais na análise de um projeto de lei proveniente da Câmara dos Deputados e de um proveniente do Senado Federal, por isso, para aumentar a robustez da análise decidimos eleger somente uma das casas.

Ao mesmo tempo, descobrimos que o Processo Legislativo também varia de acordo com o tipo de lei que está sendo criada. O Processo possui um conjunto de regras para um Projeto de Lei Ordinária e um outro para um Projeto de Decreto Legislativo. Nesse caso também restringimos o artefato para somente processar exclusivamente os dados dos Projetos de Lei Ordinária que são os mais comuns.

Para superar os desafios técnicos, utilizamos modelos de Aprendizado de Máquina para trabalhar com um volume maior de dados. Para garantir que o artefato estava conseguindo gerar *insights* de boa qualidade medimos seus resultados a partir de métricas que valorizam a precisão de acerto da classe menor dos dados, no caso, dos projetos aprovados. E por fim, para que o artefato consiga representar o modelo do Processo Legislativo, dividimos e testamos os dados respeitando os ciclos de dois anos que representam a alternância de poder da Câmara dos Deputados, ou seja, o modelo gerado por um biênio deve ser capaz de explicar os dados do biênio subsequente.

Resumindo, a proposta do nosso trabalho, é apresentar um artefato capaz de explicar e prever se um Projeto de Lei Ordinária será aprovado ou não na Câmara dos Deputados, e para

isso o artefato precisará atingir os seguintes objetivos:

1. Indicar as principais características que impactam positiva ou negativamente na aprovação de um Projeto de Lei Ordinária na Câmara dos Deputados;
2. Oferecer um prognóstico, baseado em dados, de um Projeto de Lei Ordinária na Câmara dos Deputados.

1.4 Organização do Trabalho

Para além deste capítulo introdutório, o restante da dissertação está organizado em mais seis capítulos. No segundo capítulo apresentamos a seção de Fundamentação Teórica na qual exploramos os conceitos utilizados na construção do artefato. No terceiro capítulo apresentamos os Trabalhos Relacionados ao tema da nossa pesquisa. No quarto capítulo a Metodologia utilizada para desenvolver o artefato. No quinto capítulo apresentamos os resultados gerados pelo artefato. No sexto capítulo debatemos os resultados gerados pelos artefatos. Por fim, no sétimo capítulo apresentamos as conclusões de pesquisa e indicamos os trabalhos futuros.

CAPÍTULO

2

FUNDAMENTAÇÃO TEÓRICA

2.1 Ciência de Dados

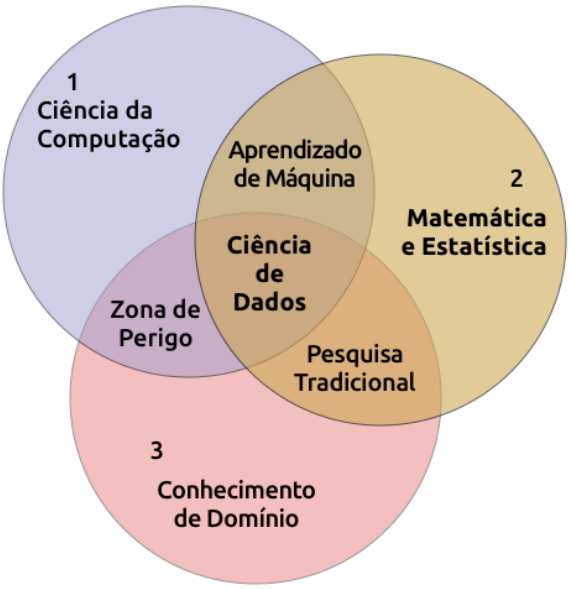
O termo Ciência (do latim *scientia*, traduzido por "conhecimento" ou "saber") é definido formalmente pelo dicionário como o "Conjunto de conhecimentos fundados sobre princípios certos" e informalmente pela Wikipédia como o "meio de adquirir conhecimento baseado no método científico". Assim sendo, podemos entender a área de Ciência de Dados como a aplicação do método científico a partir de um grande volume de dados (*big data*) sobre um determinado tema para produzir conhecimento.

A área da Ciência de Dados é objeto de pesquisa desde 1974, ano que o cientista da computação Peter Naur utilizou o termo "Ciência de Dados" no seu livro *Concise Survey of Computer Methods*, todavia somente nas últimas décadas o assunto ganhou destaque nas comunidades acadêmica e profissional.

O exponencial crescimento de produção de dados (DHAR, 2013) causado pela massificação e a evolução da internet aliado ao aumento da capacidade de processamento dos computadores criou um ambiente favorável para que pesquisadores pudessem utilizar grandes conjuntos de dados para testar e validar as mais diversas hipóteses em diferentes domínios.

Assim como na Ciência em geral, a Ciência de Dados também depende do rigor metodológico e científico. Por isso para produzir resultados válidos, o pesquisador da área de Ciência de Dados, conhecido como cientista de dados, precisa dominar o método científico e possuir habilidades que abrangem Matemática, Estatística, Computação e Bancos de Dados. Estas habilidades juntamente com um profundo entendimento do domínio do problema de pesquisa permite ao cientista de dados formular e testar hipóteses a partir dos dados. O cientista da computação Conway (2010) elaborou um diagrama de Venn exemplificado na Figura 1 sobre essas habilidades.

Figura 1 – Diagrama de Venn para Ciência de Dados



Fonte: Adaptada de Conway (2010).

No primeiro círculo há a área de Ciência de Computação que representa o domínio das técnicas de coleta, manipulação, processamento e armazenamento de dados que são essenciais para criar uma base de dados íntegra e confiável que será utilizada para testar as hipóteses. No segundo círculo há a área da Matemática e Estatística que implica em dominar os métodos e os modelos matemáticos e estatísticos que permite ao cientista de dados identificar e interpretar os padrões escondidos nos conjuntos de dados. No terceiro círculo há a área de Conhecimento de Domínio que representa o domínio e entendimento do contexto dos dados que é crucial para motivar e identificar o problema de pesquisa e ao mesmo tempo formular e testar hipóteses realistas.

Entre o primeiro e o terceiro círculo há a Zona de Perigo que indica que os trabalhos na área de Ciência de Dados que não possuem uma metodologia rigorosa e que aplicam modelos e métodos de modo equivocado geram análises imprecisas ou incorretas. Entre o segundo e o terceiro círculo há a Pesquisa Tradicional, a qual utiliza técnicas tradicionais de matemática e estatística para validar ou não uma hipótese de pesquisa, porém sem utilizar um grande volume de dados ou técnicas modernas de Aprendizado de Máquina. Entre o primeiro e segundo círculo há área de Aprendizado de Máquina que combina modelos estatísticos e matemáticos com a habilidade e capacidade de trabalhar com uma grande volume de dados (*Big Data*).

Entre a intersecção dos três círculos há a área da Ciência de Dados que reúne todas as outras áreas e possui o objetivo de construir artefatos utilizando modelos de Aprendizado de Máquina a partir de um grande volume de dados para explicar um fenômeno ou resolver um determinado problema.

No restante do capítulo apresentamos em maiores detalhes as técnicas e os conceitos das

áreas de Aprendizado de Máquina e do conhecimento do domínio que foram utilizados neste trabalho.

2.1.1 Aprendizado de Máquina

Em 1959 o cientista da computação Arthur Lee Samuel cunhou o termo Aprendizado de Máquina:

> Aprendizado de Máquina é a área da computação que estuda como os computadores podem adquirir a habilidade de aprender sem precisar ser explicitamente programado (Samuel, A. L. 1959).

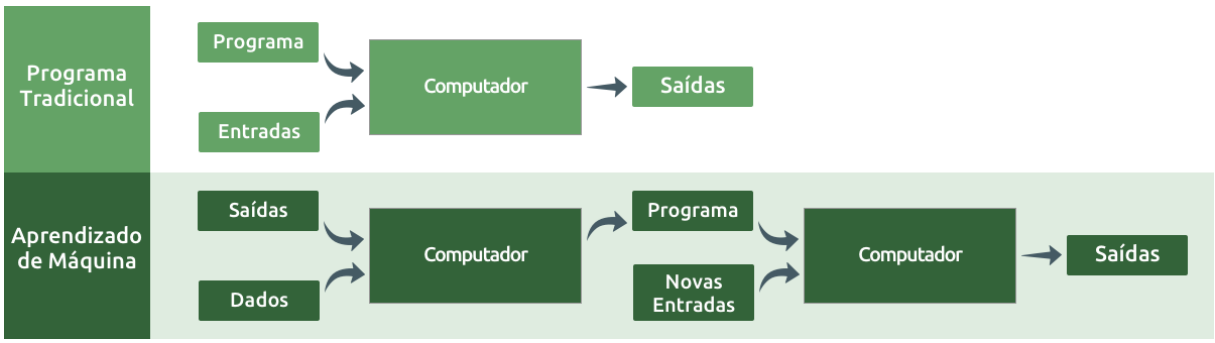
A área de Aprendizado de Máquina é classificada como uma subárea da Inteligência Artificial que estuda o desenvolvimento de modelos e técnicas capazes de "ensinar" a máquina a explicar, reproduzir ou predizer um determinado fenômeno, a partir de um conjunto de dados.

Por exemplo, na Figura 2 comparamos a criação de um programa pela técnica tradicional com um programa criado usando Aprendizado de Máquina.

Um Programa Tradicional de Computador depende das regras previamente definidas e programadas pelo seu criador para gerar uma determinada saída. Ou seja, dado um conjunto de entrada e as regras escritas pelo programador, o programa retorna uma saída. Nesse caso, o computador foi explicitamente programado.

Por outro lado, a Aprendizagem de Máquina permite ao computador aprender as regras e padrões escondidos em um grande volume de dados e assim gerar um programa automaticamente. Esse programa, então, pode ser utilizado para gerar novas saídas de acordo com novos parâmetros de entrada.

Figura 2 – Modelo de um programa baseado em Aprendizado de Máquina em comparação com o modelo tradicional



Fonte: Elaborada pelo autor.

A título de exemplo, para criar um artefato capaz de classificar um e-mail como spam ou não há duas abordagens. Em uma abordagem tradicional, o desenvolvedor poderia criar regras

explícitas que classificaria um e-mail como spam, por exemplo, e-mails que contiverem a palavra "grátis" serão classificados como spam. Assim, com as regras codificadas, o programa então conseguiria classificar o que é ou não spam.

Em contraste, na abordagem de Aprendizado de Máquina, a partir de uma grande quantidade de exemplos de e-mails já rotulados como spam ou não-spam, os modelos de Aprendizado conseguiriam identificar os padrões que definem o que é um spam e a partir deles, classificar novos e-mails como spam ou não. Nesse caso, o desenvolvedor foca em ajustar e validar o modelo para que ele consiga ser o mais eficiente possível, ao invés de se preocupar nas regras que definem o que é ou não spam.

2.1.1.1 Aprendizado Supervisionado e Aprendizado Não-Supervisionado

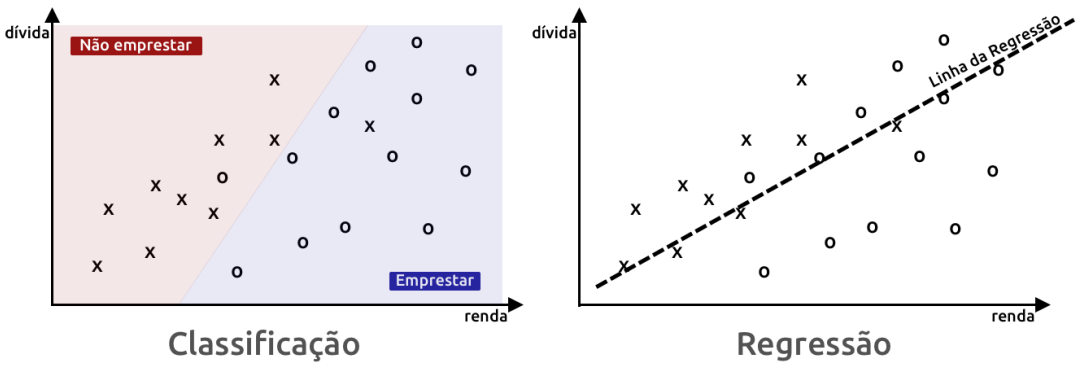
Na área de Aprendizado de Máquina, há dois tipos de aprendizagem. A primeira é a Aprendizagem Supervisionada e a segunda é Aprendizagem Não-Supervisionada.

A Aprendizagem Supervisionada é utilizada quando os conjuntos de dados já estão classificados. Nesse caso, os modelos possuem o desafio de desvendar as relações entre as classificações e o conjunto de dados para poder processar novos exemplos. Por outro lado, na Aprendizagem Não-Supervisionada não se sabe quais classes os dados pertencem, então, os modelos possuem o desafio de identificar os padrões de acordo com a estrutura e relações entre os dados.

Na abordagem da Aprendizagem Supervisionada, há dois tipos de tarefas que podem ser realizadas, a classificação e a regressão, ilustrados na Figura 3.

O objetivo da classificação é descobrir uma função capaz de classificar um item em uma lista de rótulos pré-definido. Já na regressão, o objetivo é identificar uma função capaz de prever um valor para um item e descobrir um relacionamento entre variáveis (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Figura 3 – Contraste entre as tarefas de classificação e regressão.



Fonte: Elaborada pelo autor.

Por exemplo, a partir de uma base de dados de pessoas indicando se elas pagaram ou não seus empréstimos, a classificação poderia ser utilizada para decidir se um novo empréstimo pode ou não ser liberado, enquanto a regressão poderia servir para prever a renda de uma pessoa baseada em um dívida.

Complementando a explicação, no Aprendizado Não-Supervisionado as tarefas mais comuns são o agrupamento e a redução da dimensionalidade. O objetivo do agrupamento é organizar os dados em grupos segundo seu grau de semelhança. Já na redução de dimensionalidade o objetivo é encontrar um subconjunto dos dados capaz de representar os mesmos aspectos do conjunto original para facilitar a manipulação e interpretação dos dados.

2.1.1.2 Modelo Linear Generalizado (MLG)

Modelo Linear Generalizado (MLG) é a generalização teórica proposta para reunir diversos modelos que relacionam uma variável de saída única (y) com outras variáveis de entrada (x) (NELDER; WEDDERBURN, 1972) através de uma função de conexão. Em outras palavras, no MLG o (y) pode ser calculado a partir de uma combinação das variáveis de entrada (x) dada uma função de conexão (f).

Entre os modelos MLG mais conhecidos estão a Regressão Linear, Regressão Logística e Regressão de Poisson. Na nossa pesquisa, utilizaremos somente a Regressão Logística, pois o nosso principal objetivo, é entender o porquê um projeto de lei é aprovado ou não, ou seja, uma classificação binomial e por isso, vamos detalhar os aspectos teóricos somente desse modelo.

2.1.1.3 Regressão Logística

Assim como outros modelos lineares, a Regressão Logística é um modelo que busca, a partir de um conjunto de dados, descrever a relação entre uma variável de resposta com uma ou mais variáveis explicativas (JR; LEMESHOW; STURDIVANT, 2013).

Todavia, a característica diferencial da Regressão Logística é que a sua resposta varia entre o intervalo de 0 à 1. Por isso, apesar da Regressão Logística utilizar uma regressão para encontrar o seu modelo, ela é habitualmente utilizada para classificação binomial, pois sempre a resposta ficará entre duas alternativas.

Por exemplo, podemos modelar a Regressão Logística para o nosso problema. Seja X uma variável associada a um projeto de lei, e **f** um vetor de características do projeto de lei. Seja Y uma variável binomial correspondente à aprovação (Y=1) ou não aprovação (Y=0) do projeto de lei, temos:

$$\Pr(Y = 1 \mid X = x) = \frac{\exp(w^\top f(x))}{1 + \exp(w^\top f(x))} \tag{2.1}$$

na qual, **w** é o valor do coeficiente associado à cada característica do vetor **f**, sendo que a regra preditiva está detalhada na equação 2.2.

$$\hat{y} = \begin{cases} 1, & \text{se } w^\top f(x) > 0 \\ 0, & \text{caso contrário} \end{cases} \tag{2.2}$$

Entre as vantagens da Regressão Logística estão a fácil interpretação e a desnecessidade de hiperparâmetros. No entanto como desvantagem, a Regressão Logística é sensível a *outliers* e atributos correlacionados (MOREIRA; CARVALHO; HORVATH, 2018).

2.1.1.4 Sub-ajuste vs Super-ajuste

Os modelos de Aprendizado de Máquina são poderosos em identificar os padrões e regras escondidas em um grande volume de dados. Porém, é importante ressaltar que o resultado final não depende somente do modelo, mas também da qualidade e quantidade dos dados em questão que podem ter problemas estruturais e que precisam ser corrigidos a priori.

Entre os principais problemas está a falta ou excesso de dados que podem causar os problemas do sub-ajuste ou do super-ajuste.

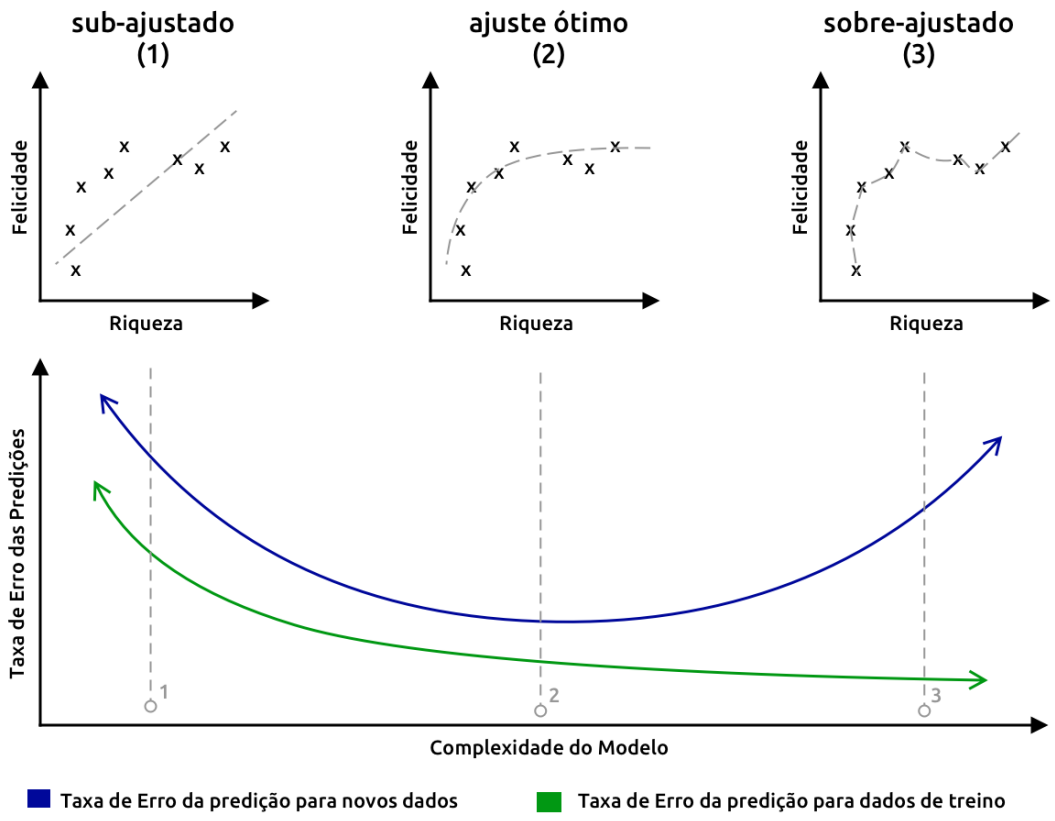
Sub-ajuste (do inglês: *underfitting*) ocorre quando um modelo estatístico não consegue capturar adequadamente a estrutura representada pelos dados. Um modelo sub-ajustado é um modelo no qual estão faltando parâmetros. Sendo assim, tal modelo tenderá a ter um fraco desempenho preditivo (ver Figura 4).

Por outro lado, o super-ajuste (do inglês: *overfitting*) é um problema que ocorre quando o modelo estatístico contém mais parâmetros do que podem ser justificados pelos dados. Sendo assim, tal modelo também tenderá ter um fraco desempenho preditivo em novos dados (ver Figura 4).

No primeiro caso ilustrado na Figura 4 temos um modelo sub-ajustado. Ao tentar ajustar o modelo a partir de poucos parâmetros, o modelo não consegue separar de forma aceitável os dois tipos de dados. Nesse caso, tanto a Taxa de Erro para novos dados quanto para os dados de treino são altas. No outro extremo, à direita da figura, temos um caso de uma modelo super-ajustado. O modelo é ajustado tão bem ao conjunto de dados que ele é ótimo para prever os dados de treino, porém é ineficaz na predição de novos dados. No meio, há o modelo ajustado de forma ótima. Apesar da taxa de erros nos dados do conjunto de treino ser maior que no modelo super-ajustado, ele é consegue prever melhor novos dados pois consegue generalizar melhor a estrutura dos dados.

Para evitar que o modelo fique sub-ajustado, é necessário adicionar novas características ao modelo ou dependendo da base de dados, aumentar o total de exemplos no conjunto de dados. No caso do super-ajuste há duas principais opções para corrigir o problema. A primeira opção é

Figura 4 – Taxas de Erros em modelos de Aprendizado de Máquina sub-ajustado versus ajustado de forma ótima versus super-ajustado.



Fonte: Elaborada pelo autor.

selecionar alguns parâmetros entre todos os parâmetros iniciais para reduzir a complexidade do modelo. A segunda opção é aplicar a regularização dos parâmetros. A regularização é um método que busca penalizar os valores dos parâmetros menos relevantes em detrimento da valorização dos parâmetros mais relevantes para o modelo.

Em seguida, listamos dois métodos de regularização dos parâmetros utilizados em Regressão Logística que foram testados no nosso trabalho afim de evitar o super-ajuste do modelo final.

2.1.1.5 Regressão Ridge

A Regressão Ridge (HOERL; KENNARD, 2015) é um método utilizado para reduzir a multicolinearidade entre variáveis preditoras de um modelo de Regressão.

Quando há muitas características em um modelo, é comum que existam variáveis preditoras que são altamente correlacionadas. Nesse caso, quando houver duas variáveis altamente correlacionadas, a Regressão Ridge penaliza as duas para tornar o resultado mais estável. A penalidade aplicada pela Ridge não anula os coeficientes irrelevantes, mas ela reduz a variância do estimador da regressão.

Entre as desvantagens da regressão Ridge, estão: o número de instâncias deve ser maior que o número de atributos; ela é sensível a *outliers*; os dados devem ser normalizados quando a relação entre os atributos preditivos e os objetivos é não-linear. Por outro lado, as vantagens são: possui forte embasamento matemático; fácil interpretação; gera melhores resultados com atributos correlacionados se compararmos a regressão sem penalidade (MOREIRA; CARVALHO; HORVATH, 2018).

2.1.1.6 Regressão LASSO

A *Least Absolute Shrinkage and Selection Operator* (LASSO) (TIBSHIRANI, 1996) é um método criado para ressaltar as variáveis mais relevantes e penalizar as menos importantes em uma regressão.

Diferentemente da Ridge, a Regressão LASSO zera o coeficiente das variáveis preditivas menos relevantes para o modelo. A penalidade aplicada pelo LASSO, é descrita pela equação 2.3:

$$\max_w \sum_i \log p_w(y_i|x_i) - \lambda ||\mathbf{w}||_1 \tag{2.3}$$

na qual i é o índice de um exemplo do conjunto treino e o segundo termo da equiação é a penalidade LASSO. O valor de λ controla a força da penalidade aplicada. O valor de lambda é determinado automaticamente usando validação cruzada nos dados do modelo.

O método de regularização LASSO possui as mesmas desvantagens que a Ridge porém se destaca nas vantagens, entre elas: a produção de modelos mais fáceis de ser interpretados, pois ela reduz a dimensionalidade, zera atributos irrelevantes e lida melhor com atributos preditivos correlacionados, além de também possuir um forte embasamento teórico (MOREIRA; CARVALHO; HORVATH, 2018).

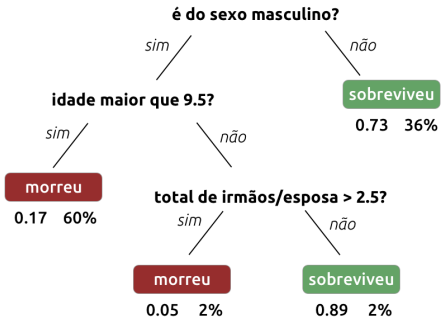
2.1.1.7 Árvore de Decisão (Decision Tree)

A Árvore de Decisão (QUINLAN, 1986) é um dos modelos mais tradicionais de Aprendizado de Máquina que pode ser aplicado tanto a problemas de classificação como problemas de regressão. A partir dos dados, a Árvore de Decisão determina a classe de um objeto ou o valor de uma variável.

As árvores que são capazes de executar tanto a classificação como a regressão são nomeadas de *Classification And Regression Tree* (CART).

Um dos exemplos mais utilizados para mostrar a aplicação prática da CART é o modelo de árvore criado para identificar se uma pessoa sobreviveria ao naufrágio do Titanic que está ilustrado na Figura 5.

Figura 5 – Predição de sobrevivência de um passageiro no Titanic aplicando o Modelo CART



Fonte: Elaborada pelo autor.

A partir dos dados dos passageiros ao utilizar a Árvore de Decisão, podemos classificar se um passageiro iria **sobreviver** ou **morrer** no caso de uma classificação. No caso de uma regressão, é possível determinar a chance de sobrevivência de cada pessoa, baseada nas suas características.

Entre as vantagens do modelo da Árvore de Decisão estão: fácil interpretação e fácil aplicação, visto que não há necessidade de pré-processar os dados, pois ele é robusto a *outliers*, dados ausentes, atributos correlacionados e irrelevantes e não precisa de normalização anterior.

Por outro lado, as suas desvantagens estão relacionadas a sua estrutura de decisão. No momento de definição de uma regra para dividir um nó é avaliada localmente sem informações suficientes para saber se a regra garante o ótimo global. Isto pode resultar em uma solução sub-ótima. Outra desvantagem, é a sua bidimensionalidade, a Árvore tentar sempre resolver o problema de forma ilustrada na Figura 5 e alguns problemas podem não ser bem resolvidos com essa estrutura (MOREIRA; CARVALHO; HORVATH, 2018).

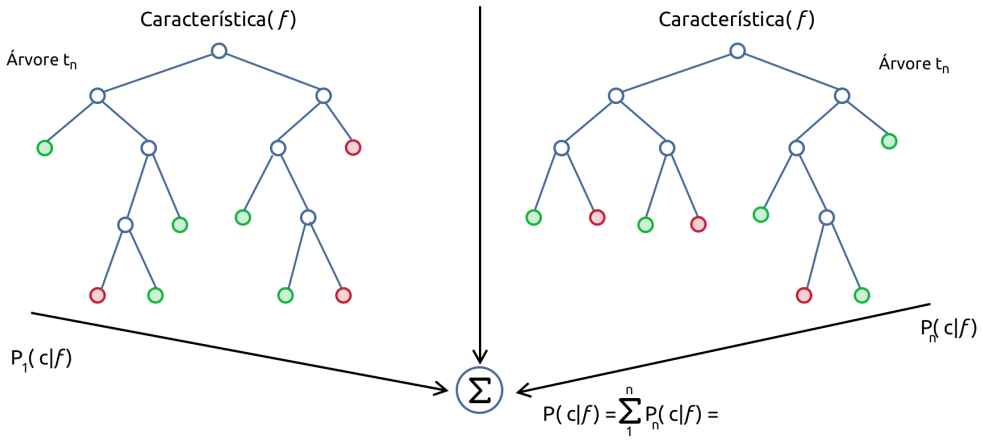
2.1.1.8 Floresta Aleatória (Random Forests)

O modelo da Floresta Aleatória constrói e mistura várias Árvores de Decisão, ver Figura 6, para gerar um modelo final mais acurado e preciso em suas predições (BREIMAN, 2001). Assim como o modelo CART, o modelo da Floresta Aleatória também pode ser utilizado para tarefas de Classificação e de Regressão.

Umas das características do modelo de Florestas Aleatórias é a sua capacidade de ordenar quais características são as mais relevantes em suas predições baseada em testes aleatórios de quais são as variáveis que mais contribuem com a eficácia do resultado final.

Entre as vantagens da Floresta Aleatória estão sua boa performance preditiva em muitos problemas e a fácil manipulação e definição dos hiperparâmetros do modelo. Por outro lado, sua principal desvantagem é o alto custo computacional para gerar e executar várias árvores (MOREIRA; CARVALHO; HORVATH, 2018).

Figura 6 – Estrutura do modelo da Floresta Aleatória



Fonte: Elaborada pelo autor.

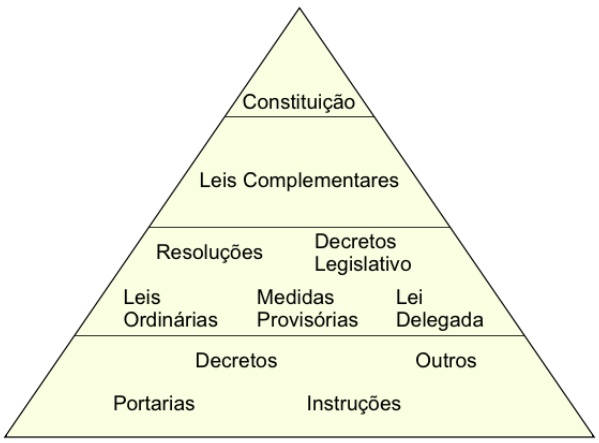
2.2 As leis e o Processo Legislativo

2.2.1 As Leis

Leis são instrumentos jurídicos utilizados para regulamentar as relações sociais e econômicas resultantes da vida em sociedade. No Brasil, não é diferente, temos vários tipos de leis que legislam sobre variados assuntos como por exemplo os impostos até questões relativas a casamento.

Há diversos tipos de leis no sistema jurídico brasileiro, dentre elas, a Constituição é a lei-maior, ou seja, ela descreve os princípios fundamentais que outras leis não podem contradizer. Para compreender os principais tipos de lei do Brasil, utilizamos a abstração hierárquica em forma de pirâmide (KELSEN, 1967), ilustrada na Figura 7, que organiza as leis por nível de importância.

Figura 7 – Pirâmide de Kelsen, configurada para os tipos de leis brasileiras



Fonte: Elaborada pelo autor.

No topo da pirâmide, na Figura 7, temos a Constituição, a lei-maior, que lança as bases e fundamentos para todas as demais legislações.

No segundo nível são apresentadas as Leis Complementares, que servem para tratar e regulamentar alguns assuntos específicos que são explicitados na Constituição (BRASIL, 1988). Exemplos de Leis Complementares são: leis que estabelecem a organização político-administrativa do Estado ou que regulamentam a aposentadoria dos servidores públicos.

No terceiro nível, temos as Leis Ordinárias, as Resoluções, as Medidas Provisórias, os Decretos Legislativos e as Leis Delegadas que regulam assuntos e temas variados que a Constituição expõem direta ou indiretamente em seu texto. A grande parte da legislação pertencem a esse nível.

No quarto nível, na base da pirâmide estão as normas de menor importância em relação as normas que estão em níveis superiores, e essas últimas são aquelas normas destinadas a detalhar as normas muito genéricas ou a complementar vazios e lacunas deixados pelos tipos de lei mencionados anteriormente. São eles os decretos, as portarias, as instruções, além de toda uma série de atos editados pelos órgãos do Estado, em especial pelo Poder Executivo ou Agências Reguladoras, no exercício de sua função típica, que é administrar o país e executar suas leis (PACHECO, 2013).

O Brasil é uma República Federativa, ou seja, é formada por três esferas político-administrativas distintas, sendo elas: a União, os Estados e os Municípios. Para cada esfera, a pirâmide se repete, porém não se sobrepõem. Ou seja, cada esfera hierárquica legisla sobre um escopo diferente, as leis federais abrangem temas gerais de interesse da Nação, as leis municipais tratam de temas locais referentes aos problemas decorrentes da administração do Município e, por último, as leis estaduais dispõem sobre temas que não são cobertos pelas outras duas esferas.

O órgão responsável pela criação das leis na esfera federal é o Congresso Nacional que é uma instituição bicameral, ou seja, é composto por duas Casas Legislativas: a Câmara do Deputados e o Senado Federal. A nível estadual, as Assembleias Estaduais são as responsáveis pelo Processo Legislativo e a nível municipal, são as Câmaras Municipais. A única exceção para essas regras é o Distrito Federal que possui a Câmara Distrital, a qual acumula funções municipal e estadual.

2.2.2 Leis Ordinária

As Leis Ordinárias (LOs) são leis que regulamentam a maioria das relações sociais que não são tratadas pela Constituição ou pelas Leis Complementares. As LOs dispõem sobre os mais variados assuntos e por isso são apresentadas em maior número no Congresso Nacional (ARAÚJO, 2004).

No Congresso Nacional, projetos de LOs podem ser apresentados por qualquer parlamentar da Câmara dos Deputados, do Senado Federal, ou pelo: Presidente da República ou Supremo

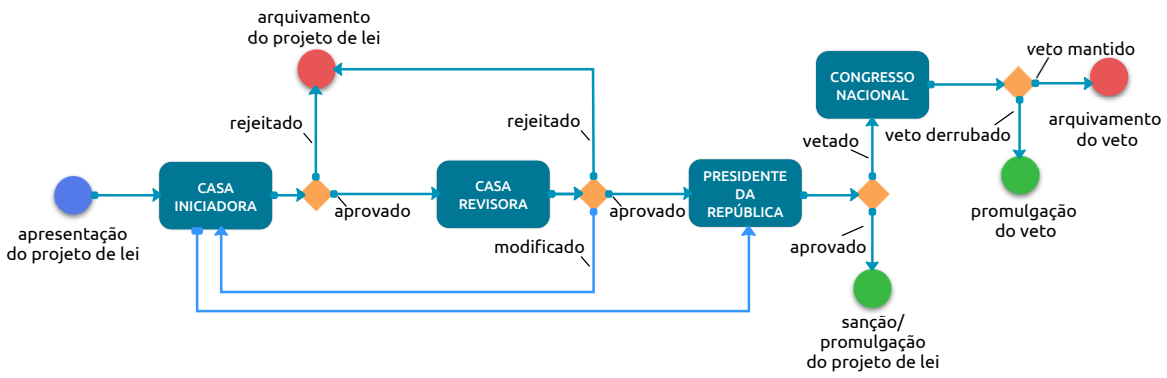
Tribunal Federal ou Tribunais Superiores ou Procurador-Geral da República ou por um conjunto mínimo de cidadãos (BRASIL, 1988). Após a apresentação, os projetos passam pelo Processo Legislativo que é o processo responsável por sistematizar a criação das leis e que está detalhado na próxima seção.

2.2.3 O funcionamento do Processo Legislativo

Entende-se por Processo Legislativo como um conjunto de atos, todos realizados pelos órgãos legislativos objetivando a formação das leis constitucionais, das leis complementares, das leis ordinárias, das resoluções e dos decretos legislativos (SILVA; AFONSO, 2014).

No Congresso Nacional, o Processo Legislativo é utilizado para criar os diversos tipos de leis na esfera federal. Conforme abordamos na seção anterior, o tipo mais comum de lei é um LO sendo que para criar uma LO é necessário aprovar um Projeto de Lei Ordinária (PLO). Na Figura 8, ilustramos o fluxo simplificado do Processo Legislativo para elaboração e análise de um PLO no Congresso Nacional.

Figura 8 – Fluxograma Simplificado do Processo Legislativo de um PLO no Congresso Nacional.



Fonte: Elaborada pelo autor.

A apresentação de um PLO pode ocorrer tanto na Câmara dos Deputados como no Senado Federal. Se o PLO é apresentado à Câmara, essa fica denominada como a Casa Iniciadora e o Senado fica denominado como a Casa Revisora, e vice-versa a depender de onde é apresentado.

Após a apresentação, a Casa Iniciadora deve aprovar ou não o PLO. Se a Casa Iniciadora aprovar, o PLO é encaminhado à Casa Revisora, caso contrário, o PLO é arquivado.

Na segunda etapa, a Casa Revisora, também, precisa aprovar o PLO. Se ela aprovar o PLO, ele é encaminhado ao Presidente da República. Caso ela rejeite o PLO, ele é encaminhado ao arquivo. Caso o PLO seja aprovado com mudanças, o PLO volta à Casa Iniciadora que decidirá se o PLO será enviado ao Presidente com ou sem as mudanças.

O Processo Legislativo possui regras similares na Câmara dos Deputados e no Senado Federal, porém existe uma diferença na composição das casas, enquanto a Câmara possui 513

deputados federais, o Senado possui 81 senadores.

Na terceira etapa, ao chegar no Presidente, ele pode decidir se irá sancionar ou vetar o PLO. Se ele sancionar, o PLO é publicado e se torna uma nova LO. Caso ele decida vetar uma parte ou todo o PLO, os vetos são encaminhados ao Congresso Nacional que deverá decidir se mantém ou não os vetos. Caso o Presidente vete somente uma parte do PLO, o restante do PLO não vetado é publicado, enquanto a parte vetada do PLO é encaminhada ao Congresso Nacional para análise.

É importante ressaltar que as regras do Processo Legislativo variam de acordo com o tipo de lei que está sendo elaborada. Por exemplo, se o projeto de lei é uma Proposta de Emenda à Constituição, ela precisa ser votada em cada casa legislativa duas vezes e o Presidente não pode vetar. Já se o projeto for um PLO, ele é votado somente uma vez em cada casa legislativa e pode ser vetado pelo Presidente.

2.3 Governo Aberto

Em uma sociedade democrática, é esperado que os cidadãos monitorem e participem das decisões dos seus Governos, todavia a dificuldade de acessar informações públicas e a burocracia dificultam esse comportamento. Todavia a recente popularização da internet permitiu aos Governos implementar políticas de disponibilização de dados eficientes, aproximando as decisões políticas dos cidadãos.

No entanto não foi somente a internet que permitiu ao Governo disponibilizar seus dados, a sociedade civil organizada também exerceu e exerce papel importante desde a primeira metade do século XX. Os primeiros movimentos sociais demandando mais transparência começaram ocorrer após a Segunda Guerra Mundial nos Estados Unidos. Neste período o Governo americano era carente de legislações que garantissem o acesso às informações e aos documentos públicos para seus cidadãos. Ao mesmo tempo o país vivenciava o período da Guerra Fria, no qual o segredo e a segurança nacional eram regras. Neste contexto a ASNE (*American Society of Newspaper Editors*) liderou, a partir de 1945, o primeiro movimento social que defendeu explicitamente uma administração pública baseada nos princípios de um "Governo Aberto", ainda que o termo ou seus princípios estavam em construção (YU; ROBINSON, 2012).

O termo "Governo Aberto"foi formalmente cunhado por um ex-congressista norte americano chamado Parks (1957) que escreveu:

> *From the standpoint of the principles of good government under accepted American political ideas, there can be little question but that* ***open government*** *and information availability should be the general rule from which exceptions should be made only where there are substantial rights, interests, and considerations requiring secrecy or confidentiality and these are held by competent authority to overbalance the general public interest in openness and availability. (Parks, 1957)*

Além dos princípios democráticos que um Governo Aberto deve respeitar, pesquisas apontam que há também benefícios sociais e econômicos gerados, a partir de um governo mais transparente, entre eles: aumento da fiscalização das atividades públicas, aumento da participação política do cidadão, aumento da eficiência e eficácia dos serviços públicos e diminuição da corrupção (ANDERSEN, 2009; WONG; WELCH, 2004).

No Brasil, a atual constituição promulgada em 1988 no seu artigo 5º, incisos XIV e XXXIII garantem o direito aos cidadãos de pedir e receber dos órgãos públicos informações de seu interesse (BRASIL, 1988).

Reforçando o dever da transparência nas atividades governamentais, foi sancionada a Lei de Acesso à Informação (LAI) em 2013 (CONTROLADORIA-GERAL, 2011). Essa lei tornou obrigatória a divulgação das informações e dados governamentais em sítios oficiais da rede mundial de computadores preferencialmente em **formato aberto**.

Portanto, como consequência dos movimentos e das políticas públicas sobre a transparência e participação social, os diversos órgãos da administração pública brasileira começaram a disponibilizar dados de interesse público que por sua vez são denominados Dados Abertos Governamentais.

2.3.1 Dados Abertos Governamentais

Dados Abertos Governamentais é um termo utilizado para denominar a "disponibilização, através da Internet, de informações e dados governamentais de domínio público para a livre utilização pela sociedade" (AGUNE; FILHO; BOLLIGER, 2009).

Em 2007 diversos grupos sociais[1] que defendem o direito dos cidadãos em ter acesso irrestrito às informações governamentais, se reuniram nos Estados Unidos e propuseram oito princípios que os órgãos públicos precisam respeitar para que os dados publicados possam ser classificados como Dados Abertos Governamentais. São eles:

1. **Completos**. Todos os dados públicos estão disponíveis. Dado público é o dado que não está sujeito a limitações válidas de privacidade, segurança ou controle de acesso.

2. **Primários**. Os dados são apresentados tais como os coletados na fonte, com o maior nível possível de granularidade e sem agregação ou modificação.

3. **Atuais**. Os dados são disponibilizados tão rapidamente quanto necessário à preservação do seu valor.

4. **Acessíveis**. Os dados são disponibilizados para o maior alcance possível de usuários e para o maior conjunto possível de finalidades.

[1] <https://opengovdata.org/>

5. **Compreensíveis por máquinas**. Os dados são razoavelmente estruturados de modo a possibilitar processamento automatizado.

6. **Não discriminatórios**. Os dados são disponíveis para todos, sem exigência de requerimento ou cadastro.

7. **Não proprietários**. Os dados são disponíveis em formato sobre o qual nenhuma entidade detenha controle exclusivo.

8. **Livres de licenças**. Os dados não estão sujeitos a nenhuma restrição de direito autoral, patente, propriedade intelectual ou segredo industrial. Restrições sensatas relacionadas à privacidade, segurança e privilégios de acesso são permitidas.

Os objetivos desses princípios são estabelecer uma base comum de definições para auxiliar grupos sociais a entenderem e lutarem por pela causa e facilitar o entendimento de quais são as regras que os Governos precisam respeitar ao disponibilizar os dados.

Já no Brasil as últimas três décadas configuraram um período de avanço da modernização, sistematização e informatização dos seus órgãos públicos. Com isso, conjuntos de dados estruturados sobre os mais variados assuntos de interesse da sociedade, como por exemplo, educação, saúde, meio ambiente e segurança, começaram a ser gerados e armazenados em bases gerenciadas pelo Governo.

No primeiro momento, os conjuntos de Dados Governamentais não estavam disponibilizados aos cidadãos que não possuíam acesso. Todavia, na última década diversas políticas públicas para promover a transparência e a publicação de dados foram aprovadas.

Entre os marcos mais relevantes, estão a participação do Brasil na fundação, em 2011, da Open Government Partnership (OGP)[2], a aprovação da LAI[3] e a publicação da política federal de Dados Abertos[4].

Em consequência das políticas e modernização da administração pública, portais de distribuição de dados governamentais foram criados, como por exemplo, o Portal de Dados Abertos do Governo Federal[5] ou o Portal de Dados Abertos da Câmara dos Deputados[6].

[2] http://www.cgu.gov.br/assuntos/transparencia-publica/governo-aberto
[3] http //www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm, acessado em 19 de Julho de 2018
[4] http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2016/decreto/d8777.htm, acessado em 19 de Julho de 2018
[5] Disponível em <http://dados.gov.br/>, acessado em 19 de Julho de 2018.
[6] Disponível em <https://dadosabertos.camara.leg.br/>, acessado em 19 de Julho de 2018

2.3.2 Portal de Dados Abertos da Câmara dos Deputados

Baseado nos princípios de um Governo Aberto, a Câmara dos Deputados[7] e o Senado Federal[8] criaram os seus portais de dados abertos que disponibilizam dados legislativos gerados pelo Processo Legislativo das respectivas casas. Nesse sentido, nossa proposta é utilizar estes dados para subsidiar essa pesquisa e, a partir deles, inferir as principais características mais impactantes na aprovação ou não de um projeto de lei.

Analisando a maturidade e integridade dos dados abertos da Câmara dos Deputados, Cruvinel (2016) mostra que o sistema e a política de Dados Abertos vêm sendo construída e melhorada desde de 2007, e hoje há várias pesquisas e até mesmo aplicações que dependem dos dados disponibilizados pela Câmara. Dessa forma, podemos entender que este portal pode oferecer dados de qualidade necessária para subsidiar o nosso trabalho.

Em casos que os dados não estão disponíveis pelo portal, também há a alternativa de requisitar informação diretamente no canal de pedido de informação ao cidadão tanto da Câmara dos Deputados [9] como do Senado Federal [10].

2.4 Grupos de Pressão

Em uma sociedade plural e diversa é natural que existam inúmeros grupos de pessoas que se juntam com o mesmo propósito, interesse ou visão de mundo. Quando esses grupos de interesse, seja um interesse econômico ou social, se institucionalizam e se formalizam para defender uma causa, eles assumem a classificação de um Grupo de Pressão.

Nessa dissertação iremos utilizar as definições de grupo de interesse e grupo de pressão utilizadas por Farhat (2007):

- **Grupo de interesse** é todo grupo de pessoas físicas e/ou jurídicas, formal ou informalmente ligadas por determinados propósitos, interesses, aspirações ou direitos, divisíveis dos de outros membros ou segmentos de sua união. Esses grupos podem permanecer em estado latente, sem adotar um programa formal de ações destinadas a tornar seus objetivos aceitáveis pelo Estado, pela sociedade ou por determinados segmentos desta.

- **Grupo de pressão** é o grupo de interesses dotados de meios humanos e materiais necessários e suficientes – e da vontade de utilizá-los ativamente – para a promoção dos seus objetivos, até vê-los atingidos. Atua perante toda a sociedade, ou parte dela, ou ainda, diante de órgãos do Estado – Legislativo ou Executivo -, com competência para mudar ou manter o *status quo* referente ao seu interesse.

[7] <http://www2.camara.leg.br/transparencia/dados-abertos>
[8] <http://dadosabertos.senado.leg.br/>
[9] <http://www2.camara.leg.br/transparencia/lei-de-acesso-a-informacao/atendimento>
[10] <http://www.senado.gov.br/transparencia/formtransp.asp>

Um dos aspectos da motivação dessa pesquisa é desenvolver um artefato capaz de auxiliar os mais diversos Grupos de Pressão no monitoramento e influência de um projeto de lei no Congresso Nacional, para que todos possam ter mais oportunidades de pleitear suas demandas. Ao levar os seus pleitos e demandas para os parlamentares, os Grupos de Pressão inserem diversas opiniões e pontos de vistas no debate de criação de uma nova lei. Dessa forma, ao final, espera-se que o embate de ideias aumentem as chances do Congresso Nacional produzir leis mais eficazes (CARVALHO, 2009).

CAPÍTULO

3

TRABALHOS RELACIONADOS

3.1 Ciência de Dados no contexto da Ciência Política norte-americana

Não obstante do recente avanço das políticas públicas relacionadas a transparência governamental e da disponibilização de Dados Abertos Governamentais ainda são raras, no Brasil, pesquisas de Ciência de Dados no domínio da Ciência Política. Mesmo na área da Ciência Política, ainda há poucos estudos que utilizam Dados Abertos. Em 2016 a pesquisadora Possamai analisou o conteúdo dos trabalhos publicados nos encontros entre 2008 e 2016 da Associação Brasileira de Ciência Política (ABCP) e da Associação Nacional de Pós-Graduação e Pesquisa em Ciências Sociais (Anpocs) e descobriu que poucos estudos utilizaram Dados Abertos para formulação e testes de hipóteses ou mesmo para aprofundar o entendimento de fenômenos sociais.

No entanto, nos Estados Unidos esse cenário é distinto. Desde o fim do século XIX pesquisadores norte-americanos já utilizavam dados para explicar fenômenos políticos e sociais em suas pesquisas. Entre as primeiras publicações urgindo a importância dos trabalhos utilizarem dados para entender o comportamento dos parlamentares está o artigo escrito pelo historiador Orin Grant Libby. Libby em 1897 defendeu o uso de dados dos votos dos congressistas para compreender o comportamento político do Congresso Americano ao invés de analisar somente os poucos discursos registrados e que era a forma mais comum, na época, de estudar as atividades legislativas.

Entre os pioneiros em pesquisas que utilizavam dados para compreender os comportamentos sociais e políticos está o cientista Stuart Rice, presidente Honorário do Instituto Internacional de Estatística. Rice foi um notável estatístico e sociólogo que em 1925 (A.RICE, 1924) publicou um artigo explorando os padrões estatísticos relacionados as opiniões políticas da população. Segundo seu estudo os dados da opinião pública sobre um determinado assunto respeitam

uma distribuição normal. Por exemplo, considerando o tema da proibição da venda de álcool, a opinião das pessoas não ficam restritas em uma escala discreta entre o não ou o sim, pelo contrário, as opiniões das pessoas podem ser classificados entre o sim ou não e também em outras possibilidades intermediárias, como por exemplo, proibir o álcool apenas para menores de idade, permitir a venda de bebidas somente finais de semana, etc. Todavia, o sistema político representativo, até os dias de hoje, confiam em um processo de decisão que limita em duas saídas, ou o sim ou o não. Essa distorção entre o que as pessoas pensam e as possibilidades restritas de decisão política, segundo Rice, causam uma frequência maior de mudanças radicais na opinião das pessoas, pois elas precisam escolher entre as duas opções independentemente se suas opiniões não estão totalmente contempladas pelas alternativas propostas.

Já no século XXI, utilizando técnicas e modelos de Aprendizado de Máquina, está a pesquisa de Yano et al. (YANO; SMITH; WILKERSON, 2012). Yano propôs e validou uma metodologia para prever se um determinado projeto de lei passaria pela primeira fase do Processo Legislativo norte americano, o qual consiste na aprovação do projeto pelas Comissões. Usando dados dos parlamentares autores do projeto e características dos próprios projetos de lei, Yano, a partir de uma metodologia estruturada, criou um modelo aplicando uma Regressão Logística, capaz de explicar quais são os fatores que impactam positiva ou negativamente na aprovação de um projeto de lei em sua primeira fase. Além disso, a Regressão foi capaz de prever se as comissões iriam aprovar ou não o projeto. Ao final, Yano conseguiu criar um modelo que obteve uma taxa média de erro de 9,9%.

Em sintonia com o estudo de Yano, Wang et al. (WANG; VARSHNEY; MOJSILOVI, 2012) também pesquisaram o Processo Legislativo dos Estados Unidos a partir dos dados dos votos dos parlamentares nos projetos de lei. Todavia, diferentemente da proposta do trabalho de Yano, Wang criou um artefato capaz de prever os votos dos parlamentares em um determinado projeto de lei. Baseado nas predições dos votos dos parlamentares, Wang criou um artefato capaz de produzir prognósticos que informam se existe a chance de um determinado projeto de lei ser aprovado ou não no Congresso Nacional e não somente na fase de comissões.

Outro aspecto distinto está relacionado ao fato de Wang et al. (2012) utilizarem uma rede complexa para modelar o Processo Legislativo e realizar predições dos votos dos parlamentares a partir dela. Na rede proposta há dois tipos de vértices e três tipos de arestas conectando os vértices. O primeiro tipo de aresta representa os parlamentares e o segundo, os projetos de lei. Conectando os vértices, o primeiro tipo de aresta realiza a conexão entre os parlamentares, indicando quantos projetos de lei havia apresentados em conjunto. O segundo tipo cria conexões entre os vértices de projetos de lei, indicando a similaridade entre o conteúdo dos projetos. O terceiro tipo, conecta os vértices de parlamentares aos vértices de projetos de lei indicando, como os parlamentares votaram nos projetos. Ao final, o modelo nos testes experimentais alcançou uma performance acima de 90% em seus experimentos.

Nay (2016), também propondo predizer se um projeto de lei seria aprovado ou não

no Congresso Nacional americano, em seu trabalho testou diversos modelos de Aprendizado de Máquina em 70 mil projetos de lei apresentados entre 2001 e 2015. Além de utilizar as características dos projetos e dos parlamentares autores, Nay inovou ao usar conteúdo do projeto para compor o modelo. Além disso, realizou uma série de experimentos com 8 modelos distintos. Ao final, um dos modelos testados por Nay alcançou uma precisão de mais 90%.

Com a mesma proposta do trabalho de Wang, Cheng et al., (2016) propôs pesquisar o problema de prever os votos dos parlamentares Norte Americanos e assim inferir um prognóstico do resultado do projeto de lei. Cheng propôs uma abordagem que maximiza o uso de diferentes tipos de dados, entre eles, os votos dos parlamentares, características do perfil do parlamentar e o texto e aspectos semânticos do projeto de lei para reduzir o erro do modelo final. O modelo foi chamado de DUMHI (*Dual Uncertainty Minimization Heterogenous Information*). Cheng buscou dados de diversas fontes públicas para construir o banco de dados de projetos de lei, perfis dos legisladores e votações. Com os dados, usando um paradigma semi-supervisionado buscou reduzir a incerteza da predição do voto em casos que há dados faltando, por exemplo, um novo legislador que nunca votou. Em comparação com os resultados do estudo de Wang (2012), Cheng (2016) conseguiu gerar melhores indicadores e desenvolveu um modelo mais resiliente. Em seus experimentos, Cheng alcançou uma performance acima de 92%.

Cho et al. (2010) também utilizou dados legislativos em seu estudo, porém ao invés de tentar prever o processo legislativo, tentou buscar entender o comportamento da produção legislativa. Cho utilizou os dados legislativos para construir uma Rede Complexa que representa a composição do Congresso Nacional dos Estados Unidos. Enquanto os nós representam os parlamentares, as arestas representam quantos projetos os parlamentares apresentaram em conjunto. Analisando as propriedades da rede sob a lente de uma Rede de Pequeno Mundo (*Small World*) Cho descobriu que quanto mais coesa a rede for, mais o Congresso americano é capaz de produzir novas leis.

Grande parte dos trabalhos, inclusive os citados anteriormente, estudam o Poder Legislativo federal dos EUA, porém recentemente a pesquisa sobre o comportamento do Processo Legislativo começou a se expandir para a esfera estadual. Os pesquisadores Eidelman et al. (2018) usando o conteúdo de texto e as características de mais de 1 milhão de projetos de lei apresentados nos 50 Estados americanos criaram um modelo capaz de explicar e comparar as principais características que são relevantes para os processos legislativos de cada estado. No final, o modelo também apresentou poder preditivo de aprovação com uma acurácia média de 86%.

3.2 Ciência de Dados no contexto brasileiro

No Brasil, apesar de não encontrarmos trabalhos da área de Ciência de Dados com o objetivo de entender e prever o resultado do Processo Legislativo do Congresso Nacional,

encontramos diversos estudos da área da Ciência Política que usam dados e análises quantitativas para subsidiar suas análises qualitativas e explicar características importantes do Processo Legislativo.

Gomes (2013), utilizando dados do Processo Legislativo, realizou uma análise quantitativa e qualitativa dos projetos de lei apresentados e aprovados entre 1999 e 2006 no Congresso Nacional. A partir de suas análises, o autor discute conceitos políticos, os aspectos das prerrogativas do Poder Legislativo e como eles são representados pelos dados. Por exemplo, Gomes analisa o impacto do sistema de Presidencialismo de Coalizão no Processo Legislativo e constata que há predominância da atuação do Executivo em todas as vias legislativas. Em uma importante observação, Gomes afirma:

> Os padrões observados reafirmam o papel das regras na produção de políticas, sugerindo que o sistema não é caótico e que as instituições políticas do Brasil têm possibilitado a superação de impasses na produção de políticas públicas de modo negociado no período posterior à Constituição Federal de 1988 (Gomes, 2013).

A partir dessa afirmação, podemos assumir que o Processo Legislativo brasileiro está maduro e respeita determinadas regras e padrões, dessa forma podemos assumir que seus dados possuem uma qualidade boa para sustentar hipóteses e testes. No fim, a partir de suas descobertas e conclusões o autor realiza sugestões para melhorar a eficiência e eficácia da produção de leis.

Ricci (2003) em seu trabalho estudou o conteúdo dos projetos de lei aprovados pelo Congresso Nacional de 1991 até 2002. A partir da classificação dos projetos de lei obedecendo três critérios: generalidade, efeitos e impacto territorial, ele organizou todos os projetos que viraram normas, nas seguintes categorias:

Geral projetos que afetam a todos os cidadãos e/ou pelo menos uma unidade organizacional que atua no âmbito nacional.

Secional projetos que afetam uma "categoria social específica, claramente identificável, grande, suficientemente diversificada e ativa numa área relevante para a sociedade", assim como unidades organizacionais que servem essas categorias.

Microssecional - Impacto amplo projetos que afetam grupos/categorias sociais específicos e/ou unidades organizacionais correlatas presentes ou ativas sobre o território nacional.

Microssecional - Impacto concentrado projetos afetam grupos/categorias sociais delimitados e/ou unidades organizacionais que agem em um determinado território.

Individual projetos afetam indivíduos ou grupos muito pequenos, assim como coisas/unidades organizacionais restritas.

Simbólicas projetos que criam normas cuja importância é relativa e definida pelo simbolismo da disposição adotada.

Então a partir desses dados, postulou três hipóteses sobre a produção legislativa. A primeira hipótese testou se os parlamentares brasileiros privilegiam criar novas leis para beneficiar e satisfazer interesses da região onde concentrava a maioria dos seus eleitores. A segunda testou se os parlamentares provenientes de distritos com alta quantidade de eleitores, priorizam criar novas leis que impactavam toda a nação, ao invés de leis que impactam somente o espaço territorial da sua base eleitoral. Por último, a terceira hipótese testou se o Senado era menos propenso a produção de política locais do que a Câmara dos Deputados. Ao final do artigo, usando dados e testes estatísticos (Qui-quadrado), o autor concluiu que a primeira hipótese era falsa e a segunda e a terceira eram verdadeiras.

Tadeu (2014), em sua dissertação realizou uma análise sobre a atuação e importância das Frentes Parlamentares no Congresso Nacional. As Frentes Parlamentares são grupos suprapartidários que reúnem deputados e senadores para defender uma pauta específica. Por exemplo, são frentes parlamentares a FPA (Frente Parlamentar da Agropecuária) e a FPE (Frente Parlamentar Evangélica). Tadeu conclui que as Frentes abrem um novo canal para aproximar-se das demandas da sociedade e que podem influenciar no Processo Legislativo, se bem estruturadas e utilizadas na arena política.

Bimestralmente, o Centro de Tecnologia e Sociedade da universidade Fundação Getúlio Vargas (FGV) gera um relatório com análises descritivas sobre as atividades legislativas do Congresso Nacional. O relatório está em sua segunda edição (SGANZERLA *et al.*, 2018) e seu objetivo é entender, a partir dos dados de uma forma mais granular, qual foi o comportamento do Congresso Nacional e quais foram os assuntos mais debatidos no Processo Legislativo no período estudado.

As pessoas normalmente utilizam um referencial espacial para declarar um posicionamento político, por exemplo, é comum escutar a expressões “sou de esquerda” ou “sou de direita”. A teoria espacial do voto é a formalização da ideia que é possível organizar os posicionamentos políticos em um campo espacial. Leoni (2002) usando os dados de votação entre os anos de 1991 a 1999 e a partir da teoria espacial do voto e aplicando o modelo estatístico W-NOMINATE estimou as posições ideológicas dos deputados federais. Seu estudo revelou uma estrutura ideológica basicamente unidimensional. Segundo Leoni, uma dimensão conseguiu explicar a maior parte das votações na Câmara dos Deputados, pois ao adicionar dimensões adicionais não se constatou de forma significativa o aumento da capacidade explicativa do modelo estatístico.

Recentemente, alguns trabalhos da área de Ciência de Dados aplicados ao domínio da política foram publicados e também serviram de inspiração para esse trabalho. Entre eles, tem-se o de Baptista (2015) que em sua dissertação utilizou dados de votação dos parlamentares para desenvolver um artefato para descobrir quais dos parlamentares da Câmara dos Deputados

têm maior chance de ter mudado de posicionamento em um período, se tornando governo ou oposição. Inicialmente, Baptista utilizou o modelo estatístico W-NOMINATE para modelar o posicionamento dos parlamentares, a partir dos dados de votação. Então, com os pontos ideais no começo e no fim da legislatura, conseguiu criar um modelo preditivo capaz de identificar se naquele período de tempo, o parlamentar havia mudado ou não de posicionamento, ou seja, se ele migrou do governo para a oposição ou vice-versa.

Além das iniciativas acadêmicas, há projetos que trabalham com dados legislativos para outras finalidades, como jornalismo e cidadania. Uma das ferramentas, mais conhecidas e utilizadas no jornalismo político, é o Basômetro [1] que é uma sistema de visualização interativa que permite identificar o apoio dos parlamentares ao governo e acompanhar como eles se posicionaram nas votações legislativas. Dantas et al. (2014) reuniram diversas análises que utilizaram os dados e *insights* do Basômetro para produzir diversos estudos que no final se tornou um livro chamado "Análise política & jornalismo de dados: ensaios a partir do Basômetro".

O Radar Parlamentar[2] é uma iniciativa do grupo de software livre da USP e que é, segundo a definição em seu site: “aplicativo que ilustra as semelhanças entre partidos políticos com base na análise matemática dos dados de votações que ocorrem na casa legislativa. As semelhanças são apresentadas em um gráfico bidimensional, em que círculos representam partidos ou parlamentares, e a distância entre esses círculos representa o quão parecido os mesmos votam”. O objetivo do Radar Parlamentar é oferecer aos cidadãos uma nova forma de entender a política, bem como, permitir uma fácil acompanhamento de como votam seus representantes eleitos.

[1] <http://estadaodados.com/basometro/>
[2] <http://radarparlamentar.polignu.org/>

Em resumo, na Tabela 1 relacionamos os trabalhos discutidos nessa seção com os modelos computacionais e estatísticos utilizados para analisar os dados legislativos.

Tabela 1 – Principais modelos utilizados pelos trabalhos relacionados

País	Trabalho	Estatística Descritiva	Modelo Linear Generalizado (MLG)	Redes Complexas	W-Nominate	Support Vector Machine (SVM)	Stochastic Gradient Boosting (GBM)	Floresta Aleatória	Bayesiano
EUA	Yano et al. 2008		x						
EUA	Wang et al. 2012			x					
EUA	Nay 2016		x				x	x	
EUA	Cheng et al. 2016								x
EUA	Cho et al. 2010			x					
EUA	Eidelman et al. 2018					x			x
Brasil	Gomes 2013	x							
Brasil	Ricci 2003	x							
Brasil	Tadeu 2004	x							
Brasil	Sganzerla 2018	x							
Brasil	Leoni 2002				x				
Brasil	Baptista 2015		x		x	x	x	x	

Analisando a Tabela 1 percebemos que os trabalhos publicados nos Estados Unidos exploraram mais o uso de modelos de Aprendizado de Máquina na Ciência Política. No Brasil, grande parte dos trabalhos ainda usam estatísticas simples e descritas para entender o comportamento dos parlamentares, porém recentemente alguns trabalhos começam a aplicar técnicas mais modernas, e nessa linha entra o nosso trabalho.

Resultados preliminares dessa pesquisa foram publicadas no Simpósio Brasileiro de Sistemas de Informação de 2018 em um artigo com o título “Compreendendo e Prevendo o Processo Legislativo na Câmara dos Deputados do Brasil” (OLIVEIRA; ALBUQUERQUE; DELBEM, 2018).

Por fim, a inovação deste trabalho foi utilizar técnicas, metodologias e ferramentas de SI e a partir delas, melhorar o modelo previamente publicado que seja capaz de entender e prever o resultado do Processo Legislativo na Câmara dos Deputados.

CAPÍTULO

4

METODOLOGIA

As pesquisas da área de Sistemas de Informação (SI) possuem como objetivo principal a expansão do conhecimento sobre a aplicação de tecnologias da informação em organizações humanas visando gerar um impacto positivo e aumentar o bem-estar econômico e social (editorial da revista ISR-Information System Research[1]).

Na comunidade de SI há dois paradigmas distintos, e ao mesmo tempo complementares, que são comumente aceitos e utilizados para moldar a metodologia aplicada as suas pesquisas. Os paradigmas são a **Ciência de Design** e a **Ciência Comportamental** (MARCH; SMITH, 1995).

A Ciência Comportamental possui raízes nas Ciências Humanas. Seu objetivo é guiar as pesquisas para que elas produzam teorias capazes de explicar e prever fenômenos relacionados a interação entre os artefatos de SI e as organizações humanas (HEVNER *et al.*, 2004). Por exemplo, a pesquisa "Impacto do uso de infográficos como materiais de aprendizagem e suas correlações com satisfação, estilos de aprendizagem e complexidade visual" (LYRA; ISOTANI, 2017) é um exemplo de um trabalho que utiliza o paradigma da Ciência Comportamental, pois a autora visa explicar o porquê as visualizações e infográficos podem ser utilizados para promover e potencializar a transferência de conhecimento em contraste com os métodos tradicionais de ensino.

Por outro lado, a Ciência de Design possui raízes nas Ciências Exatas. É um paradigma que molda a pesquisa para que no fim ela produza resultados capazes de resolver um problema. O objetivo desse paradigma é produzir, a partir de um problema real enfrentando por organizações humanas, inovações representadas por artefatos de SI que resolvam o problema proposto (HEVNER *et al.*, 2004). Por exemplo, a pesquisa "Data + Intuition: A Hybrid Approach to Developing Product North Star Metrics" (CHEN; FU, 2017) é um exemplo que aplica o paradigma de Ciência de Design, pois propõe uma metodologia para construir um artefato que usa dados e modelos de

[1] <https://pubsonline.informs.org/page/isre/editorial-statement>

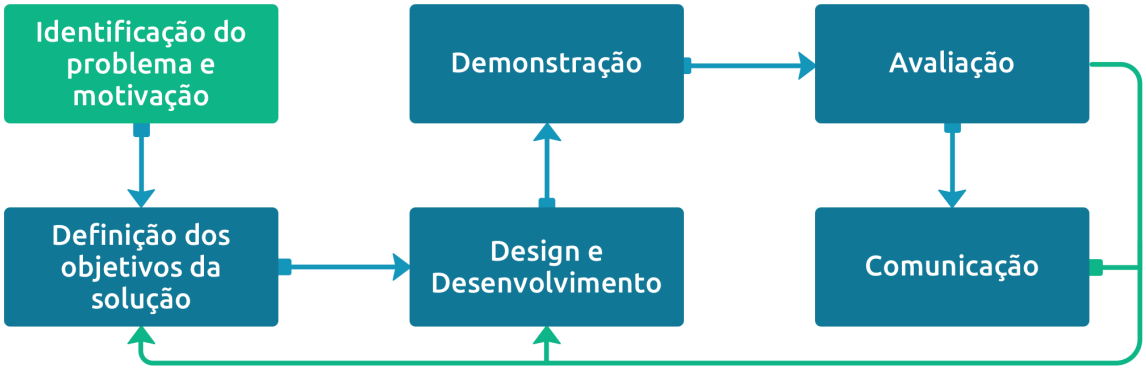
Aprendizado de Máquina para solucionar o problema da escolha de quais métricas uma empresa deve definir para melhorar o seu produto e para guiar seus esforços de desenvolvimento.

Enquanto o paradigma da Ciência Comportamental busca a *verdade* o paradigma da Ciência de Design busca a *utilidade*. Por exemplo, no primeiro estudo a autora tenta a partir de sua pesquisa entender se infográficos são eficazes para transmissão do conhecimento, já no segundo estudo, os autores criam um artefato para ajudar as empresas a definirem quais indicadores devem ser utilizados para monitorar a operação da empresa.

Considerando as duas alternativas, este trabalho utilizará o paradigma de Ciência de Design. Nós desenvolvemos, a partir do problema real enfrentado pelos Grupos de Pressão que precisam entender e prever Processo Legislativo, um artefato capaz de prover explicações e predições, utilizando técnicas/ferramentas da área de Ciência de Dados.

Para garantir que um artefato de SI resolva um problema de uma forma válida, é necessário utilizar uma metodologia para guiar o trabalho. Nesse trabalho, nós seguimos a metodologia de Pesquisa de Ciência de Design adaptada para a área de SI (PEFFERS *et al.*, 2007) exemplificada na Figura 9.

Figura 9 – Metodologia de Pesquisa de Ciência de Design proposta por Peffers para pesquisas na área de Sistemas de Informação.



Fonte: Adaptada de Peffers *et al.* (2007).

A metodologia de Pesquisa de Ciência de Design ilustrada na Figura 9 foi proposta por Peffers em uma iniciativa de padronizar as metodologias utilizadas em pesquisas de SI que utilizam o paradigma da Ciência de Design. A metodologia é composta por 6 etapas e envolve os passos necessários para produzir uma solução válida em SI para um problema enfrentando por alguma organização humana. A seguir, detalhamos as 6 etapas da metodologia e como elas foram aplicados ao nosso estudo.

4.1 Identificação do problema e motivação

A primeira etapa da metodologia abrangeu tarefas nas quais exploramos o contexto da pesquisa com o intuito de entender e definir os aspectos e o escopo do problema. Os aspectos e o escopo foram definidos e se transformaram nos requerimentos que a solução proposta ao final da pesquisa deveria satisfazer. Nesta etapa, também, foram exploradas a relevância e a importância do problema, para assim decidir se justificava investir recursos para resolvê-lo.

Em um primeiro momento, pensamos em propor uma metodologia capaz de gerar um artefato usando modelos de Aprendizado de Máquina que poderia explicar e prever o resultado final do Processo Legislativo de todos os tipos de Projetos de Lei que são debatidos no Congresso Nacional a partir dos dados.

Contudo, após o estudo mais detalhado do Processo Legislativo, o escopo do problema foi reduzido em dois aspectos. O primeiro aspecto limitado foi o tipo de Projeto de Lei que o artefato seria capaz de processar. Ao invés de tentar criar um artefato que trabalhasse com todos os tipos de Projetos de Leis que passam pelo Processo Legislativo no Congresso Nacional, decidimos focar somente no PLO. O segundo aspecto de limitação está relacionado à abrangência do Processo Legislativo. Conforme explicado em seção anterior, o Processo Legislativo é bicameral envolvendo duas casas, a Câmara dos Deputados e o Senado Federal e no nosso trabalho decidimos especializar o artefato em explicar e prever o resultado de um PLO somente na Câmara dos Deputados.

No primeiro aspecto, decidimos reduzir o escopo do artefato para abranger exclusivamente um tipo de Projeto de Lei, pois cada tipo de projeto possui regras de tramitação distintas no Processo Legislativo. Assim, entendemos que seria necessário criar um artefato para cada tipo aumentando a complexidade da pesquisa o que poderia resultar na redução da qualidade dos *insights* gerados. Escolhemos analisar especificamente o PLO porque é o tipo de projeto com maior número de exemplares e que suas regras afetam diretamente a vida dos cidadãos, pois regulam os mais variados assuntos.

No segundo aspecto, decidimos especializar o modelo para explicar e prever o Processo Legislativo somente no âmbito da Câmara dos Deputados, pois da mesma forma que há variações do Processo Legislativo para tipos de projeto, há também variações das regras entre a Câmara dos Deputados e o Senado Federal, e por isso seria necessário criar dois modelos, sendo um para cada casa. Decidimos pela Câmara dos Deputados, ao invés do Senado Federal, pois ela é a casa de origem da maioria dos projetos de lei apresentados e aprovados. Considerando os projetos de lei que se tornaram leis desde 2003, 73% deles foram originários na Câmara dos Deputados.

Em suma, criamos um artefato para explicar e prever e o Processo Legislativo, para Projetos de Lei Ordinária na Câmara dos Deputados. Com isso, esperamos que esse artefato facilite a participação de mais Grupos de Pressão nas decisões políticas. Porquanto entendemos que quanto mais grupos pautando as decisões dos parlamentares, maior a chance da produção de

leis estar consoante com as demandas e os problemas da sociedade.

4.2 Definição dos objetivos da solução

Na segunda etapa, o propósito foi determinar objetivos realísticos para a pesquisa com base nos aspectos e requerimentos do problema e no conhecimento acumulado na etapa anterior. A metodologia proposta por Peffers delimita que há dois tipos de objetivos, quantitativos ou qualitativos.

Diante do contexto de otimizar o tempo e melhorar a tomada de decisão dos Grupos de Pressão na Câmara dos Deputados e o problema de entender e prever a aprovação ou não de um PLO na Câmara dos Deputados, propomos dois objetivos para a nossa pesquisa.

O primeiro objetivo definido foi criar um artefato capaz de prover explicações, baseadas em dados, sobre quais características e comportamentos impactam de forma positiva ou negativa na aprovação de um PLO na Câmara dos Deputados. Para o atingimento do objetivo será necessário gerar e analisar evidências estatísticas que indicam as características.

O segundo objetivo definido foi garantir que o mesmo artefato produza prognósticos que indicam se um PLO será aprovado ou não na Câmara dos Deputados. Ou seja, o modelo deve ser capaz de indicar se um determinado projeto de lei possui ou não chances de aprovação, a partir de suas características. Para garantir o cumprimento do segundo objetivo, será necessário gerar testes e avaliar a qualidade dos resultados gerados pelo modelo final.

4.3 Design e Desenvolvimento

A terceira etapa envolveu a parte prática da pesquisa. Nesta etapa os requisitos funcionais e não funcionais do artefato foram definidos e posteriormente o próprio artefato foi desenvolvido, utilizando peças de software. Nesse estudo desenvolvemos o artefato em três subetapas: coleta e balanceamento dos dados, seleção de características, e aplicação dos modelos.

4.3.1 Coleta e Balanceamento dos dados

Através do portal de Dados Abertos da Câmara dos Deputados foram coletados[2] dados sobre os PLOs aprovados ou arquivados entre 2003 e 2016 na Câmara dos Deputados. Os dados foram limpos e organizados em biênios, conforme mostra a Tabela 2.

Os dados foram organizados em biênios, pois a cada dois anos são realizadas eleições para renovar o comando da Câmara dos Deputados. O comando, representado pela Mesa Diretora, possui um grande poder no Processo Legislativo na Câmara dos Deputados, segundo o seu Regimento Interno. Por consequência organizamos os conjuntos de dados em biênios, para

[2] Dados coletados no dia 31 Maio de 2018

Tabela 2 – Dados coletados provenientes do Portal de Dados Abertos da Câmara

#	Legislatura	Ano 20xx	Arquivados	Aprovados	Taxa de Aprovação
1	52	03-04	3206	206	6,04%
2	52	05-06	933	251	21,20%
3	53	07-08	5062	385	7,07%
4	53	09-10	1588	488	23,51%
5	54	11-12	3369	290	7,93%
6	54	13-14	1343	277	17,10%
7	55	15-16	4144	373	8,26

garantir que o artefato possa ser treinado em um conjunto e pudesse ser testado no conjunto temporalmente subsequente. Assim, pudemos averiguar se o modelo é capaz de explicar e prever o resultado independentemente de quem esteja dirigindo a Câmara dos Deputados.

Os projetos foram estruturados, inicialmente em 27 características. As características pré-selecionadas foram levantadas a partir da análise dos autores sobre o que poderia ser relevante para o modelo baseado no estudo do problema de pesquisa, das regras do Processo Legislativo e na leitura dos Trabalhos Relacionados. Entre as 27, a característica pertencia ou ao projeto de lei ou ao autor principal do projeto de lei. A lista de características e seus tipos foram listadas no Apêndice A.

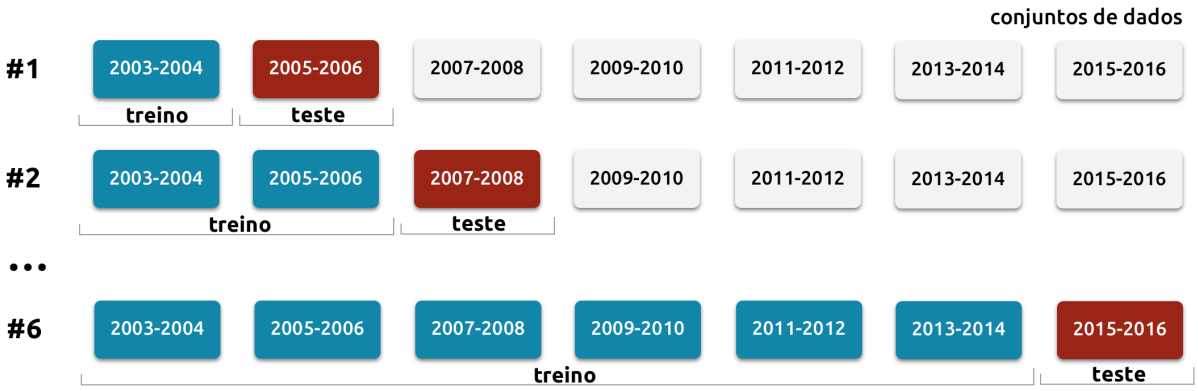
Ao analisar os dados, notou-se um evidente desbalanceamento entre os projetos aprovados e arquivados. Há mais itens arquivados do que itens aprovados. Para evitar que o modelo final tenda sempre classificar um projeto como arquivado e outros problemas gerados pelo desbalanceamento (HE; GARCIA, 2009), foi necessário balancear os dados.

Porém, antes de balancear os dados, foi preciso transformar as características nominais em ordinais para que elas pudessem ser utilizadas e analisadas. No fim, as 27 características geraram 72 sinais, ou seja, cada projeto de lei foi sistematizado em 72 variáveis.

Após transformar os dados, testamos duas abordagens para balancear os dados, o *'oversample'* e o *'undersample'*. Na primeira abordagem *oversample*, utilizamos o método *Synthetic Minority Over-sampling Technique* (SMOTE) para sintetizar novas instâncias artificiais de projetos aprovados baseadas nas instâncias dos reais projetos aprovados (CHAWLA *et al.*, 2002). Por outro lado, na abordagem *undersample* escolhemos aleatoriamente a mesma quantidade de projetos arquivados em relação ao total de projetos aprovados para compor o conjunto que treinou o modelo. Para verificar qual abordagem era melhor, realizamos 6 experimentos de classificação conforme ilustra a Figura 10.

No primeiro experimento, treinamos uma Regressão Logística com penalidade LASSO com dados entre 2003 e 2004 utilizando todas as características inicialmente mapeadas e testamos no conjunto subsequente com dados dos anos de 2005 e 2006. No próximo experimento, treinamos o modelo com dados entre os anos 2003 e 2006 e testamos com dados entre os anos

Figura 10 – Configuração dos experimentos para validar o melhor método de balanceamento de dados.



Fonte: Elaborada pelo autor.

de 2007 e 2008. No total foram realizados 6 experimentos, sendo que os 6 experimentos foram repetidos nas duas abordagens. Por fim, comparamos os resultados de cada abordagem para eleger a melhor método para o balanceamento de dados.

4.3.2 Seleção das Características

Após a finalização do balanceamento dos dados, decidimos garantir a qualidade das características que seriam utilizadas no modelo, pois, além da intuição dos autores, não havia evidências sobre a relação das características com a aprovação ou não dos PLOs.

Considerando os fatos que não havia evidências sobre a relação das características com a aprovação ou não dos PLOs e que colocar características irrelevantes no modelo podem invalidar os seus resultados e aumentar a complexidade desnecessariamente, as características foram submetidas a uma etapa de seleção.

No primeiro momento, aplicamos um processo de seleção de características em cada biênio com dados balanceados listado na Tabela 2. O processo envolveu executar uma validação cruzada e usando a LASSO para identificar quais eram os sinais mais relevantes (FONTI, 2017) para cada biênio.

Ao final, o LASSO selecionou uma quantidade de características para cada biênio. A fim de evitar que características que foram relevantes somente em um biênio fossem selecionadas, desenvolvemos então a fórmula de seleção 4.1:

$$\hat{m} = \begin{cases} 1, & \text{se frequência} >= N \\ 0, & \text{caso contrário} \end{cases} \tag{4.1}$$

seja m, uma variável indicando se a característica irá ser selecionada (m=1) ou não (m=0),

seja a **frequência** o total de vezes que a variável foi selecionada pelo LASSO nos conjuntos bienais e seja **N** a nota de corte que indica quantas vezes a variável precisa ser selecionada para ser utilizada. Caso a característica seja selecionada em N ou mais conjuntos de dados, ela será utilizada no modelo final, caso contrário ela será descartada.

Para determinar a nota de corte **N**, os experimentos foram executados variando a nota de corte de 1 a 7. Ao final, a nota de corte foi selecionada baseada na comparação dos resultados.

4.3.3 Aplicação dos modelos

Nesta subetapa, após finalizar o balanceamento e a seleção de características, os dados começaram a ser minerados usando modelos de Aprendizado de Máquina.

No primeiro momento, aplicamos uma Regressão Logística em cada biênio para entender o padrão e a correlação entre as variáveis independentes com a variável dependente. Nesse sentido, para cada biênio com os sinais selecionados foi aplicada uma Regressão Logística com o intuito de identificar quais delas se comportam da mesma forma independentemente do período, dando sustentação a uma posterior análise explicativa.

De forma complementar, utilizando a mesma configuração ilustrada pela Figura 10 aplicamos os modelos da Regressão Logística com penalidade LASSO e da Floresta Aleatória nos dados.

A Regressão Logística com penalidade LASSO foi capaz de gerar dois tipos de contribuições. Por meio dos resultados dos testes, calculamos o impacto das características definido na Fórmula 4.2 que complementou o entendimento das características relevantes e em um segundo momento por meio dos testes de classificação medimos e entendemos se o modelo foi capaz de prever a aprovação ou não de um PLO.

Da mesma forma, a Floresta Aleatória contribuiu para esse trabalho. As variáveis importantes derivadas da Floresta Aleatória complementaram o aspecto explicativo do Processo Legislativo e os resultados dos testes de classificação nos ajudaram a comparar qual é o melhor modelo preditivo, a Floresta ou a Regressão Logística com LASSO.

4.4 Demonstração

Nesta etapa executamos a metodologia proposta e geramos os resultados que foram utilizados na etapa posterior de avaliação.

A primeira tarefa envolveu coletar e estruturar os dados da Câmara dos Deputados. Após realizar uma análise exploratória dos dados, transformamos as 27 características em 72 sinais candidatos ao modelo. Acompanhando a metodologia, os primeiros passos envolveram balancear os dados para posteriormente selecionar os sinais mais relevantes para o modelo final.

Ao determinar o modelo final, treinamos a Regressão Logística nos sete conjuntos de dados bienais para entender o comportamento das variáveis no decorrer do tempo. Analisamos dois aspectos, primeiro a significância dos sinais nos conjuntos de dados e posteriormente os valores dos seus coeficientes.

Em seguida, rodamos os experimentos configurados na Figura 10 com os modelos LASSO e a Floresta Aleatória para complementar a capacidade explicativa do artefato e para medir a capacidade preditiva do modelo final.

4.5 Avaliação

Nessa subetapa, nós observamos e analisamos se o artefato resolveu o problema proposto. As tarefas nessa etapa envolveram comparar o objetivos propostos com os resultados gerados pela demonstração do artefato.

Para validar a capacidade explicativa desse artefato, realizamos um análise qualitativa dos resultados gerados pela etapa anterior. O primeiro passo foi analisar os resultados computados pelos experimentos, nos quais foram aplicados a Regressão Logística. Para cada experimento foram analisados os coeficientes da Regressão e a significância das variáveis. Esses aspectos nos ajudaram a entender se os sinais escolhidos para o modelo final eram significantes e possuíam o mesmo comportamento ao longo do tempo.

Já nos experimentos nos quais utilizamos o LASSO conseguimos interpretar o comportamento do modelo, por meio da Fórmula de Impacto complementando a análise da capacidade explicativa. A Fórmula de Impacto é uma fórmula capaz de indicar quais características estão relacionadas a aprovação, ou não, dos projetos de lei (YANO; SMITH; WILKERSON, 2012). A fórmula calcula o impacto de cada característica na decisão final para a classificação de y. Por exemplo, para a característica j, temos:

$$\frac{\mathrm{w}_j}{N}\sum_{i=1}^{N} f_j(x_i) \tag{4.2}$$

na qual i é o índice dos exemplos do conjunto de teste, e N representa o total de exemplos no mesmo conjunto de teste. A fórmula de impacto indica quanto na média a característica em questão impactou a classe de Y. No nosso estudo, foram realizados 6 experimentos, conforme ilustrados na Figura 10, portanto, o impacto de uma característica será a média dos seus impactos calculados para cada experimento. Com os dados de impacto de cada sinal, também conseguimos estudar o comportamento do modelo e entender se ele é estável ao longo do tempo.

Por último, completando a análise da capacidade explicativa, utilizamos o *ranking* das variáveis mais importantes gerado pela Floresta Aleatória. Com esses dados conseguimos analisar quais foram os sinais mais importantes que a Floresta considerou na sua tomada de decisão pela aprovação ou arquivamento de um Projeto de Lei.

Para avaliar a capacidade preditiva dos modelos utilizamos duas medidas, a precisão e a F1-Score em detrimento da tradicional acurácia simples. Conforme mostrado na Tabela 2 os conjuntos de dados do nosso estudo são desbalanceados, havendo muito mais projetos arquivados do que aprovados. Dessa forma, a acurácia simples não é a ideal para avaliar os resultados nos nossos experimentos de classificação, pois há predominância da classe negativa (arquivamento) nos conjuntos de testes, 90% dos projetos são arquivados.

Hipoteticamente, se utilizarmos um modelo simplista que indica sempre que um projeto será arquivado, ele poderia alcançar alta acurácia de 90%, imprimindo uma falsa impressão de eficácia. Por isso, para avaliar os modelos utilizamos as duas métricas citadas anteriormente, a F1-score e a precisão.

A F1-score é uma métrica que indica o quão bem o modelo consegue prever os exemplos positivos e a confiabilidade de sua predição positiva. Ela é composta pela média harmônica da precisão (P) e do *recall* (R), dada pela fórmula 4.3:

$$2 * (\frac{P \times R}{P + R}) \tag{4.3}$$

A precisão é a métrica que indica a confiabilidade da predição positiva de um modelo. Por exemplo, se a precisão de um modelo é alta, podemos entender que quando ele classifica positivamente o item, então a qualidade desse prognostico é muito boa. A métrica precisão é calculada pela fórmula 4.4:

$$\frac{VP}{VP + FP} \tag{4.4}$$

na qual o *VP* significa Verdadeiros Positivos e *FP* significa os Falsos Positivos.

A *recall* é a métrica que indica o quão eficiente o modelo é em prever positivamente os exemplos. Um *recall* alto significa que o modelo é capaz de identificar grande parte dos exemplos positivos do conjunto teste. A *recall* é calculada pela fórmula 4.5

$$\frac{VP}{VP + FN} \tag{4.5}$$

na qual o *VP* significam Verdadeiros Positivos e *FN* significam os Falsos Negativos.

De forma complementar, também vamos analisar os resultados pelo Valor Preditivo Negativo (VPN). Esse VPN é a métrica que indica a confiabilidade da predição negativa do modelo. Se o VPN é alto, por exemplo 0.99, indica que o modelo, quando classificar negativamente o exemplo,a chance é grande de corresponder a verdade. O VPN é dado pela fórmula 4.6:

$$\frac{VN}{VN + FN} \tag{4.6}$$

na qual o *VN* significa Verdadeiros Negativos e *FN* significa os Falsos Negativos.

Resumindo os aspectos da avaliação do modelo, na nossa base de dados desbalanceados, na qual há muitos mais exemplos negativos que positivos, torna muito mais difícil para os modelos de Aprendizado de Máquina identificarem com eficácia e eficiência os exemplos positivos. Por isso, para averiguar a qualidade do modelo, iremos, no primeiro momento avaliar os resultados pela F1-score. Não basta acertar todos os verdadeiros positivos, por isso vamos complementar a análise da qualidade dos modelos por meio da precisão que indica o quão confiável é uma predição positiva. Por fim, de forma complementar, para avaliar a exatidão negativa dos modelos iremos usar o VPN.

4.6 Comunicação

Nesta etapa, criamos a estrutura do trabalho e reportamos os resultados e suas conclusões. Para comunicar efetivamente sobre todas as etapas da pesquisa, desde a identificação do problema, seu tratamento e desfecho, o nosso trabalho espeitou as diretrizes propostas por Gregor (2013) para criação de publicações na área de SI.

4.7 Tecnologias e Ferramentas utilizadas

Por fim, apresentamos as principais ferramentas, linguagens e bibliotecas utilizadas no desenho, desenvolvimento e avaliação dos artefatos para garantir a transparência e a reprodutibilidade da pesquisa.

Para gerar os resultados entre as etapas de coleta e a estruturação de Dados, criamos *scripts* utilizando a linguagem Python[3] e bibliotecas que facilitam a manipulação de dados, entre elas, a numpy[4] e a pandas [5]. Para salvar os dados, escolhemos um banco não relacional chamado Couchdb[6]. A partir das tarefas envolvendo o balanceamento e aplicação de modelos de Aprendizado de Máquina, utilizamos *scripts* e bibliotecas disponíveis em R[7].

[3] <https://www.python.org/>
[4] <http://www.numpy.org>
[5] <https://pandas.pydata.org>
[6] <http://couchdb.apache.org/>
[7] <https://www.r-project.org>

CAPÍTULO

5

RESULTADOS

Nesta seção detalhamos as evidências e os resultados gerados pelas etapas de "Design e Desenvolvimento", de "Demonstração" e de "Avaliação" previstas na metodologia de Ciência de Design que sustentarão a discussão sobre como o artefato final alcançou os dois objetivos definidos para essa pesquisa.

5.1 Coleta e balanceamento dos Dados

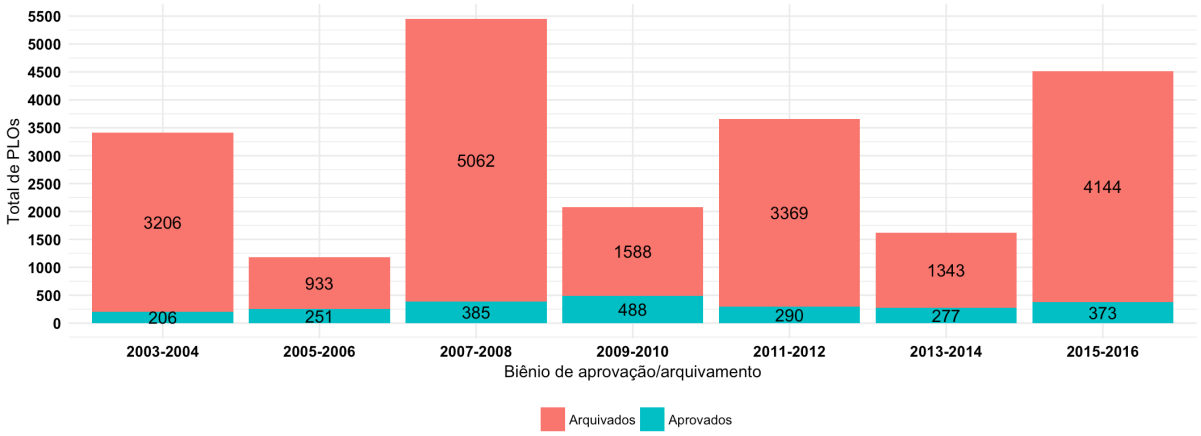
Após finalizar a limpeza e estruturação dos dados coletados, o conjunto de dados contabilizou 21.915 PLOs que foram aprovados ou arquivados entre os anos de 2003 e 2016.

Conforme nós esclarecemos na seção anterior, os dados dos PLOs foram organizados em biênios, pois o modelo deve ser capaz de generalizar os padrões dos dados independentemente de quem esteja ocupando a presidência da Câmara, para que assim o modelo possa ser robusto e válido. Ao final dessa primeira etapa de organização, contabilizamos 7 conjuntos de dados.

Assim como identificamos na Tabela 2, a Figura 11 corrobora visualmente com a desproporção entre os PLOs aprovados e arquivados em cada biênio. Há muito mais projetos arquivados do que aprovados. Calculando a proporção exata, a partir dos 21.915 projetos, descobrimos que aproximadamente 10% dos PLOs foram aprovados. Em outras palavras, o conjunto de dados é desbalanceado, contendo muito mais exemplares de PLOs arquivados que aprovados.

A variação entre a aprovação e arquivamento de projetos a cada biênio é outro aspecto relevante para observação. É possível notar que no primeiro biênio de um mandato mais projetos são arquivados que no segundo biênio. Por exemplo, nos anos de 2003 e 2004 foram arquivados 3206 projetos, enquanto nos anos de 2005 e 2006 foram arquivados 933, figurando então uma redução de 71%, sendo que esse padrão se repete se compararmos com os biênios 2007-2008 e 2009-2010. Contudo, para os projetos aprovados, a taxa de comparação se mantém relativamente

Figura 11 – Total de PLOs aprovados/arquivados na Câmara dos Deputados



Fonte: Dados da pesquisa.

constante, pois a variação é menor. Por exemplo, no biênio 2003-2004 foram aprovados 206 projetos enquanto em 2005-2006 foram 251, indicando um variação menor, de 20%.

Após a entender a distribuição entre os projetos aprovados e os arquivados a cada biênio, partimos para analisar e entender as características do conjunto de dados. No total foram pré-selecionadas 27 características, conforme mostra a Tabela 3. Uma explicação mais detalhada das características pode ser encontrada no Anexo A. Entre as 27 características pré-selecionadas, há 11 do tipo nominal e 16 do tipo ordinal.

Para utilizar as variáveis nominais e as variáveis ordinais no mesmo modelo, transformamos as características em sinais que por sua vez são valores que as representam numericamente. No caso das características ordinais, não foi necessário aplicar nenhum método de transformação, mas no caso das nominais, aplicamos o processo de transformação dos valores nominais em valores ordinais.

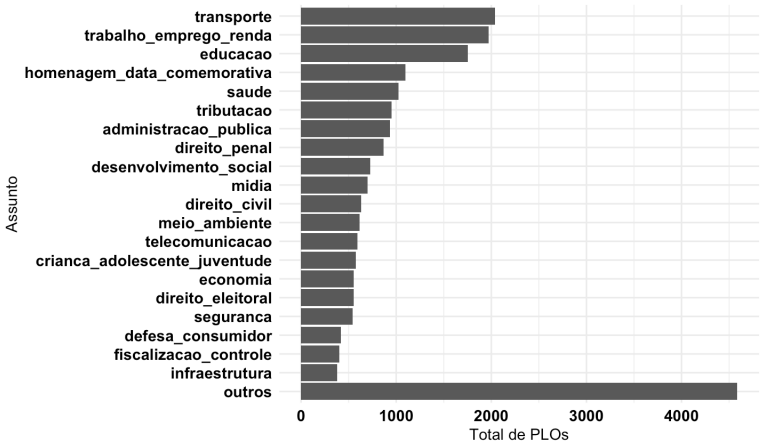
Após o processo de transformação, o modelo inicial contabilizou 72 sinais. Considerando todos os sinais, as características ordinais contribuíram com 16 sinais e as características nominais com 56 sinais. Entre as características nominais há 8 nominais boolianas que contribuíram com 8 sinais, e 3 características categóricas que contribuíram com 48 sinais. Essas três são a "l_tipo_-autor", a "p_assunto" e a "p_sessao_legislativa". A característica "l_tipo_autor" gerou 3 sinais, a "p_assunto" gerou 42 sinais e a "p_sessao_legislativa" gerou 3 sinais.

Tabela 3 – Características pré-selecionadas

Característica	**Categoria**	**Detalhes**	**Sinais gerados**
l_autores_grupo_mesclado	nominal	Booleano, é "sim"ou "não".	1
l_estado	ordinal	Intervalo: [0, 77]	1
l_governo	nominal	Booleano, é "sim"ou "não".	1
l_lideranca	nominal	Booleano, é "sim"ou "não".	1
l_liderancas_total	ordinal	Intervalo: [0, -]	1
l_mandatos	ordinal	Intervalo: [0, -]	1
l_membro_comissao	nominal	Booleano, é "sim"ou "não".	1
l_partido_proporcao	ordinal	Intervalo: [0, 1]	1
l_presidente_flag	nominal	Booleano, é "sim"ou "não".	1
l_tema_projeto_score	ordinal	Intervalo: [0, -]	1
l_tema_relatorias_score	ordinal	Intervalo: [0, -]	1
l_tipo_autor	nominal	Há 4 tipos de autor	3
l_vice_presidente_partido_flag	nominal	Booleano, é "sim"ou "não".	1
p_apensado_flag	nominal	Booleano, é "sim"ou "não".	1
p_assunto	nominal	Há 43 tipos de assuntos	42
p_casa_revisora_flag	nominal	Booleano, é "sim"ou "não".	1
p_emendas	ordinal	Intervalo: [0, -]	1
p_total_pareceres	ordinal	Intervalo: [0, -]	1
p_processo	ordinal	Intervalo: [0, 2]	1
p_regime	ordinal	Intervalo: [1, 3]	1
p_sessao_legislativa	nominal	Há 4 tipos de sessão	3
p_total_apensados	ordinal	Intervalo: [0, -]	1
p_total_autores	ordinal	Intervalo: [0, -]	1
p_total_dias	ordinal	Intervalo: [0, -]	1
p_total_orgaos	ordinal	Intervalo: [0, -]	1
p_total_tramitacoes	ordinal	Intervalo: [0, -]	1
p_velocidade	ordinal	Intervalo: [0, -]	1
		Total de sinais:	**72**

A característica que gerou mais sinais após a sua transformação foi a "p_assuntos". Ao analisar a Figura 12 percebemos que os três temas mais apresentados são: transporte; trabalho, emprego e renda; educação. Os 20 principais temas podem ser analisados na Figura 12.

Figura 12 – Assuntos dos PLOs aprovados e arquivados entre 2003 a 2016

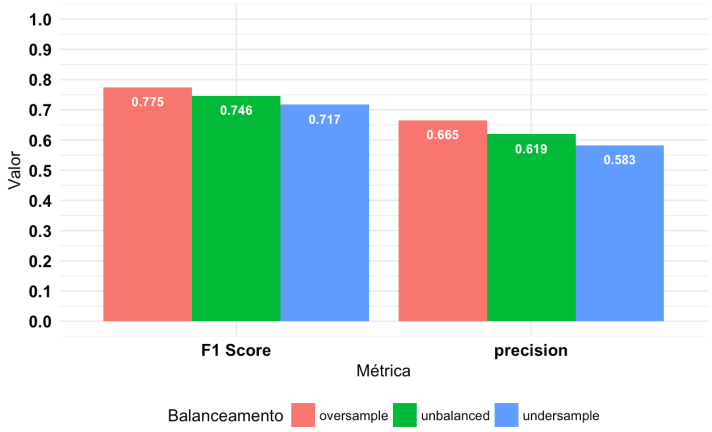


Fonte: Elaborada pelo autor.

Após finalizar a exploração da distribuição e variação das características pré-selecionadas e a transformação das variáveis, partimos para o balanceamento dos dados. Foram testados dois métodos, o método *oversample* e o método *undersample*. A título de comparação também realizamos testes com o conjunto desbalanceado, para dessa forma conseguirmos verificar o ganho preditivo ao balancear os dados.

Para cada método de balanceamento foram executados os 6 experimentos detalhados na Figura 10 utilizando um modelo de Regressão Logística com penalidade LASSO, usando todos os 72 sinais inicialmente selecionados para o modelo. Ao final, a média dos resultados foram compilados e ilustrados na Figura 13.

Figura 13 – Resultados dos Experimentos realizados conjuntos balanceados (*oversample* e *undersample*) e desbalanceados



Fonte: Dados da pesquisa.

Analisando os resultados dos experimentos ilustrados na Figura 13, tanto pela métrica F1 Score como na precisão o método *oversample* de balanceamento gerou melhores resultados. Dessa forma, podemos entender que ao balancear os dados pelo método *oversample* o modelo é capaz de identificar com mais eficiência e confiabilidade quais projetos serão aprovados. Assim sendo, apoiado nesses resultados, a partir desse momento, em qualquer tarefa de treinamento de todos os modelo utilizaremos os dados balanceados através do método *oversample*.

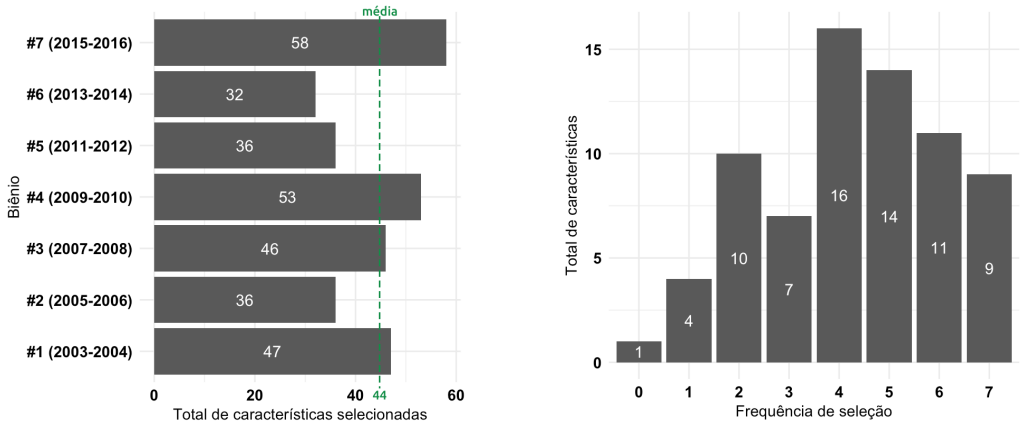
5.2 Seleção de Características

As 27 características listadas na Tabela 3 e detalhadas no Anexo A, selecionadas inicialmente para compor o modelo, foram escolhidas a partir da análise qualitativa dos trabalhos relacionados, e também do estudo das regras do Processo Legislativo no contexto do problema de pesquisa. Por isso, não havia qualquer evidência sobre alguma correlação ou causalidade entre as características escolhidas e o resultado da aprovação ou arquivamento. Portanto, para construir um modelo válido capaz de prover explicações sobre o Processo Legislativo e ao mesmo tempo potencializar a sua performance preditiva, adicionamos uma etapa de seleção de características.

Entre as 27 características, 13 são de natureza nominal enquanto 14 são de natureza ordinal. Conforme citamos anteriormente, para utilizar as características nominais em modelos de Aprendizado de Máquina, aplicamos o processo de transformação *dummy*. Ao final dessa transformação, o conjunto totalizou 72 sinais para cada PLO.

Após o processo de transformação de variáveis nós executamos, para cada conjunto de dados, uma seleção de sinais usando o LASSO. Para cada biênio, o LASSO selecionou os sinais que ele considerou mais relevantes, por exemplo, ao analisar a Figura 14a, inferimos que no conjunto #7, que contém os projetos arquivados e aprovados entre os anos de 2015 e 2016, o LASSO selecionou 58 sinais.

Figura 14 – Resultados da seleção LASSO de características, para cada biênio



(a) Total de características selecionadas em cada biênio

(b) Frequência de seleção baseada na nota de corte

Fonte: Elaborada pelo autor.

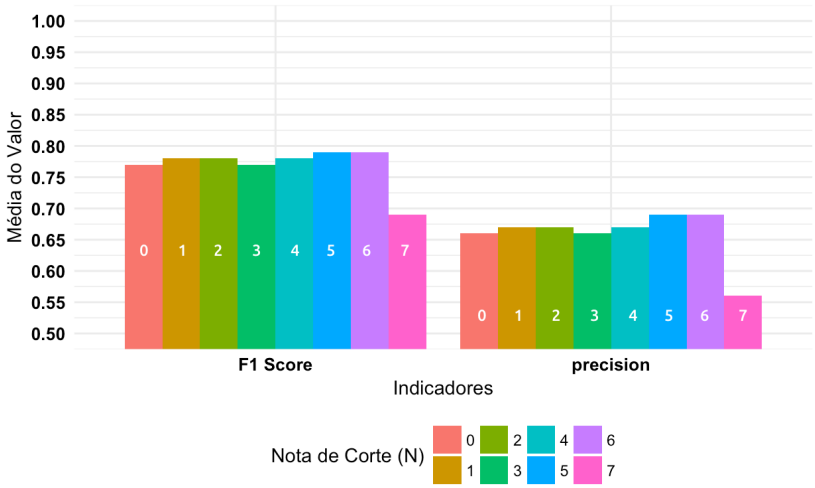
Analisando com mais detalhes a Figura 14a, podemos notar que em média 44 sinais dos 72 possíveis foram selecionadas. O conjunto que mais selecionou foi o conjunto do biênio #7 (2015-2016) que selecionou 58 sinais. O conjunto que menos selecionou foi o conjunto do biênio do #6 que selecionou 32 sinais.

Paralelamente, a Figura 14b mostrou que a maior parte dos sinais aparecem 4 vezes, ou mais vezes. Há um sinal que não foi selecionado em algum conjunto enquanto há 9 sinais que

foram selecionados nos 7 biênios.

Para determinar a frequência de corte, ou seja, quantas vezes minimamente um sinal precisa ser selecionado para poder ser utilizado no modelo final, ou seja, para determinar o N da fórmula 4.1 realizamos e comparamos os resultados gerados pelos experimentos configurados na Figura 10 variando a nota de corte de 0 a 7. Por exemplo, para a nota de corte 5, usamos os 34 sinais que apareceram 5 ou mais vezes, ou para a nota de corte 6, usamos os 20 sinais que apareceram 6 ou 7 vezes.

Figura 15 – Resultado dos experimentos realizados para escolher a frequência de corte.



Fonte: Dados da pesquisa.

Analisando a Figura 15, percebemos que os ganhos entre as notas 0 a 6 são pequenos tanto pelo F1 Score como pela precisão. As notas 5 e 6 empatam e geram melhores resultados que as notas 0, 1, 2, 3 e 4. Dessa forma, em um primeiro momento decidimos descartar as notas de 0 a 4. Por outro lado, a performance da nota de corte 7 é muito inferior as suas anteriores, por isso, também descartamos a nota 7.

As notas 5 e 6 empatam na performance, todavia diferem no total de sinais selecionados. A nota 6, mesmo com menos sinais consegue gerar os mesmo resultados que a nota 5. Por isso, para facilitar a explicação do artefato final e reduzir a sua complexidade, a nota de corte escolhida foi a 6, ou seja, o modelo final utilizará somente os 20 sinais que apareceram em pelo menos 6 dos 7 conjuntos de dados.

5.3 Aplicação da Regressão Logística

Nesta seção, utilizamos uma Regressão Logística sem penalidade em cada um dos 7 conjuntos de dados usando o modelo dos 20 sinais selecionados para entender se esse consegue representar padrões de comportamento robustos que se mantenham durante o tempo.

Em um primeiro momento, para cada biênio, treinamos uma Regressão Logística sem penalidades e removemos os sinais que não foram classificados como estatisticamente significantes. Dessa forma, conseguimos contabilizar quanto cada um dos vintes sinais foram considerados estatisticamente significantes, ou seja, cada sinal foi selecionado nenhuma ou até sete vezes.

A Tabela 4 mostra o total de vezes que a Regressão identificou um sinal estatisticamente significante nos 7 conjuntos de dados. Ao final, 19 dos 20 sinais foram selecionadas pelo menos uma vez. O única que não foi selecionado foi o "p_sessao_legislativa_3". Por outro lado, 5 sinais foram significativos nos 7 conjuntos.

Tabela 4 – Total de vezes que os 20 sinais foram selecionados como estatisticamente significantes.

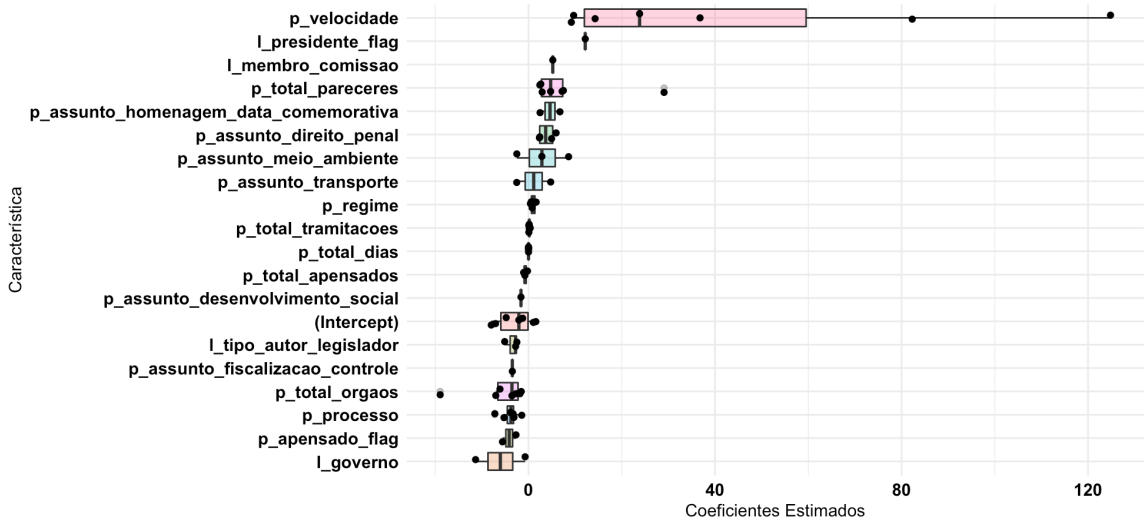
Sinal	**Total de seleções**
p_velocidade	7
p_total_tramitacoes	7
p_total_pareceres	7
p_total_orgaos	7
p_processo	7
p_regime	6
p_total_dias	5
p_assunto_direito_penal	4
p_total_apensados	3
p_assunto_meio_ambiente	3
l_tipo_autor_legislador	3
p_assunto_transporte	2
p_assunto_homenagem_data_comemorativa	2
p_apensado_flag	2
l_governo	2
p_assunto_fiscalizacao_controle	1
p_assunto_desenvolvimento_social	1
l_presidente_flag	1
l_membro_comissao	1
l_sessao_legislativa_3	0

Não basta o sinal ser significante em todos os conjuntos, é preciso que ele também apresente um comportamento estável ao logo do tempo. Ou seja, que um sinal em todos os conjuntos possua o mesmo sinal do coeficiente, um sinal deve apresentar um padrão positivo para aprovação em todos os conjuntos ou vice-versa.

Para analisar os 20 sinais tabulamos os valores dos coeficientes calculados nos 7 conjuntos

de dados e os ilustramos na Figura 16.

Figura 16 – Coeficientes estimados pela Regressão Logística para cada sinal nos 7 conjuntos de dados



Fonte: Dados da pesquisa.

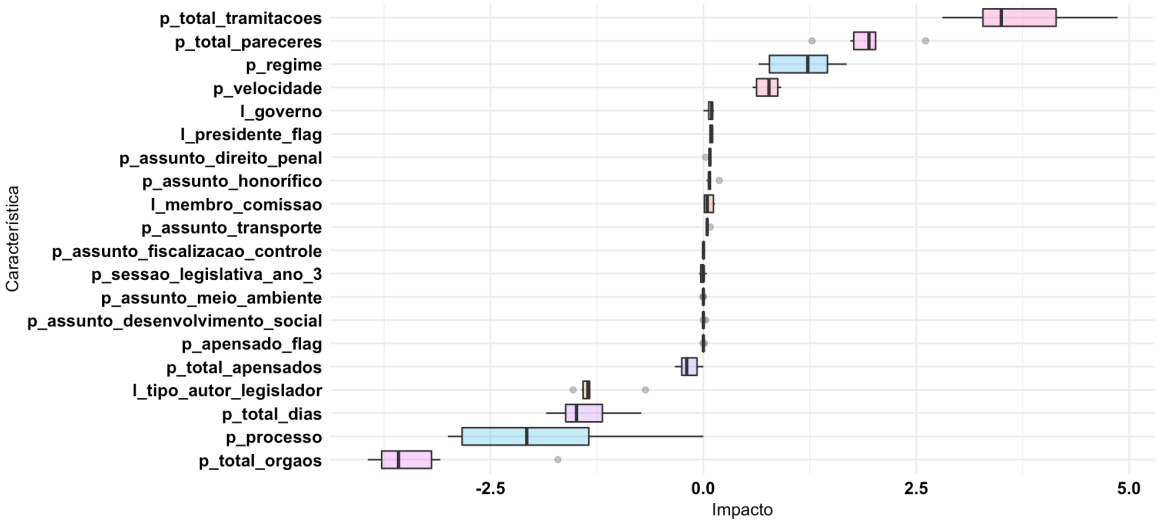
Analisando a Figura 16, percebemos que todos os 19 sinais, em geral, apresentam coeficientes com o mesmo padrão. Por exemplo, o sinal “p_velocidade” contribui positivamente na aprovação de um PLO em todos os 7 conjuntos. Por outro lado, o sinal “p_processo” nos 7 casos apresenta coeficientes negativos contribuindo para um prognóstico de arquivamento de um PLO.

5.4 Aplicação do LASSO

Posteriormente rodamos os experimentos ilustrados na Figura 10 aplicando o LASSO. Nesses experimentos adicionamos novos resultados para compreender e prever o Processo Legislativo. Os resultados nos permitiram calcular o impacto dos sinais que ajudaram na compreensão do processo legislativo, ao mesmo tempo que geraram a F1 score e a precisão, e essas por sua vez nos permitiram analisar a capacidade preditiva do modelo.

Após calcular o impacto de cada sinal para cada experimento tabulamos os resultados e os ilustramos na Figura 17. A Figura mostra que há quatro sinais mais impactantes para aprovação dos PLOs, enquanto há quatro mais impactantes para arquivamento.

Figura 17 – Impacto das 20 características nos 6 experimentos.



Fonte: Dados da pesquisa.

Os sinais mais impactantes para a aprovação são: “p_total_tramitacoes”, “p_total_-pareceres”, “p_regime” e “p_velocidade”. Em contrapartida, os sinais que mais contribuem para o arquivamento de um PLO são “p_total_orgaos”, “p_processo”, “p_total_dias” e “l_tipo_autor_-legislador”.

Após calcular os impactos de cada sinal que contribuem para análise explicativa do modelo, calculamos o F1 Score e a precisão que ajudaram na avaliação da capacidade preditiva do mesmo. Os resultados estão consolidados na Tabela 5.

Tabela 5 – Resultados dos experimentos utilizando o modelo da Regressão Logística com penalidade LASSO.

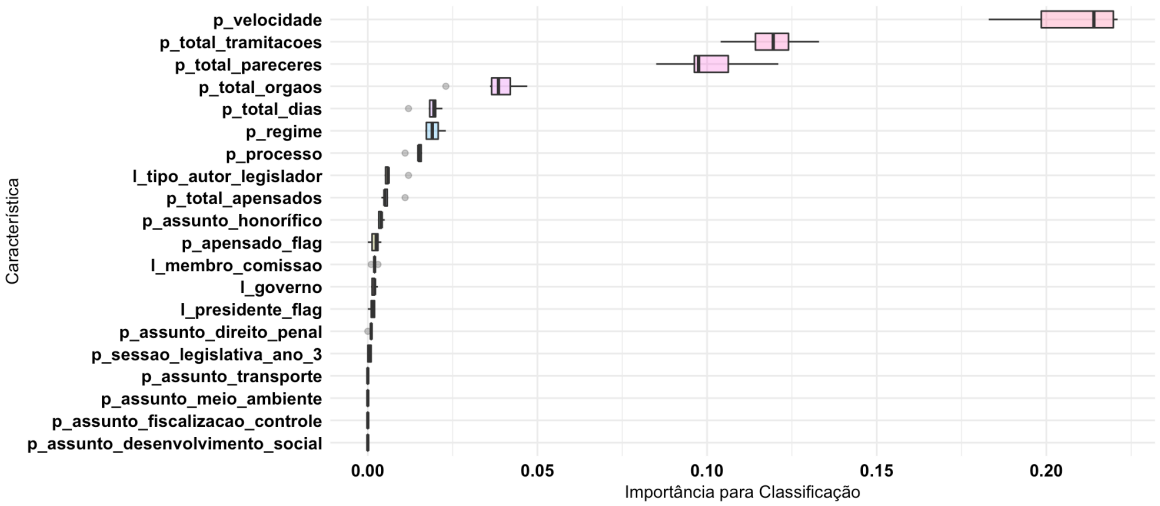
Experimento	precisão	F1 Score
1	0,842	0,845
2	0,586	0,736
3	0,767	0,851
4	0,608	0,742
5	0,666	0,790
6	0,654	0,767
Média:	**0,687**	**0,789**

Ao analisar a Tabela 5 a média do F1 Score ficou em 0,789 enquanto a média da precisão ficou em 0,687.

5.5 Floresta Aleatória

Assim como na etapa anterior, repetimos os experimentos da Figura 10 utilizando o modelo da Floresta Aleatória. Para cada experimento usando a Floresta Aleatória, nós calculamos a importância dos sinais que estão ordenados e ilustrados na Figura 18.

Figura 18 – Importância das características nos 6 experimentos para a decisão de classificação da Floresta Aleatória.



Fonte: Dados da pesquisa.

A importância dos sinais são utilizados para complementar a análise explicativa do modelo. Ao analisar a lista de importância dos sinais, as sete mais relevantes são “p_velocidade”, “p_total_tramitacoes”, “p_total_pareceres”, “p_total_orgaos”, “p_total_dias”, “p_regime” e “p_-processo”.

Para averiguar a capacidade preditiva da Floresta Aleatória, também computamos as métricas precisão e F1 Score dos experimentos na Tabela 6.

Tabela 6 – Resultados dos experimentos utilizando o modelo da Floresta Aleatória

Experimento	Precisão	F1 Score
1	0,784	0,860
2	0,621	0,764
3	0,822	0,896
4	0,819	0,893
5	0,844	0,906
6	0,777	0,848
Média:	**0,778**	**0,861**

Ao analisar os indicadores da tabela acima, percebemos inicialmente que havia um ganho de performance da Floresta Aleatória se comparado com o modelo anterior. O LASSO alcançou

uma precisão média de 0.687, enquanto a Floresta alcançou 0.778. Já no quesito F1 Score, a Floresta alcançou uma média 0.861 e o LASSO 0.789.

CAPÍTULO

6

DISCUSSÃO

Nessa seção debatemos como os resultados produzidos por meio da aplicação dos modelos de Aprendizado de Máquina substanciam o atingimento dos dois objetivos propostos pela nossa pesquisa. Em um primeiro momento, analisamos qualitativamente os dados e identificamos quais são as principais características que indicam a aprovação ou não de um PLO na Câmara dos Deputados. A partir dessas análises, além de estudarmos o comportamento dessas características ao longo do tempo, nós geramos *insights* de atuação para os Grupos de Pressão. Por fim, analisamos os resultados dos experimentos preditivos e verificamos a qualidade do prognóstico produzido pelos modelos.

6.1 Objetivo 1: Capacidade Explicativa

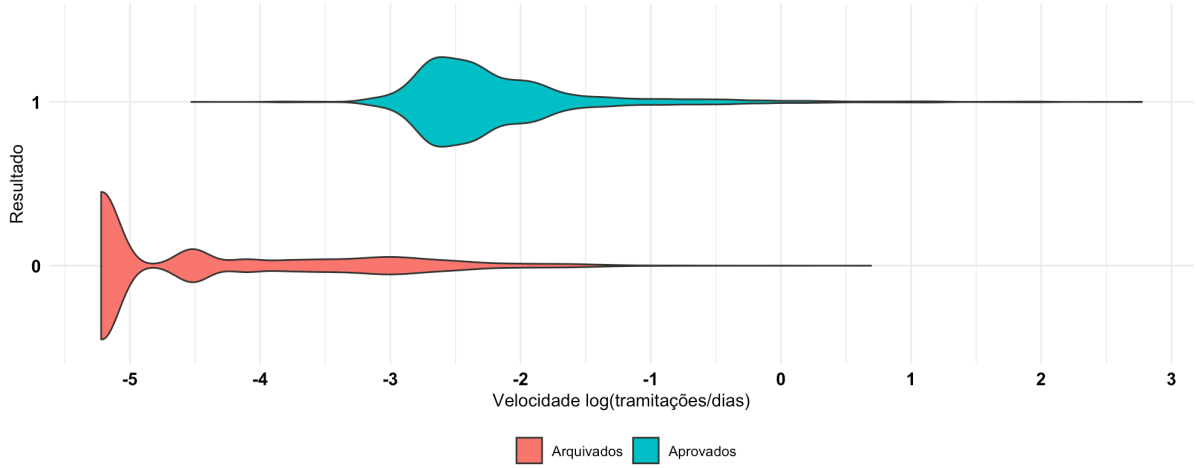
Ao analisarmos os três resultados: a Tabela 4 que lista os sinais estatisticamente significantes identificados pela Regressão Logística; a ordem de importância das características identificadas pela Floresta Aleatória ilustrada na Figura 18; e o impacto dos 20 sinais calculados pelo LASSO ilustrado na Figura 17, pudemos perceber uma correlação entre esses resultados. Os 7 sinais mais importantes identificados pela Floresta Aleatória, também são os 7 sinais mais estatisticamente significantes identificados pela Regressão Logística que, por sua vez, também são os 7 sinais que possuem maior impacto na decisão de aprovação ou arquivamento pelo LASSO.

Ao analisar os coeficientes dos 7 sinais calculados pela Regressão Logística e os seus impactos calculados na aplicação do LASSO, podemos inferir que 4 sinais estão diretamente relacionados à aprovação, enquanto 3 sinais estão diretamente relacionados ao arquivamento. Os 4 sinais relacionados à aprovação são: “p_velocidade”, “p_total_tramitacoes”, “p_total_pareceres” e “p_regime”, enquanto os três relacionados ao arquivamento são: “p_total_orgaos”, “p_processo”, e “p_total_dias”. A seguir exploramos em detalhes o comportamento desses 7 sinais ao longo do tempo.

A “p_velocidade” indica quantas tramitações ocorreram nos últimos 6 meses de um PLO antes dele ser aprovado ou arquivado. Como há muitos valores entre 0 e 1 e alguns valores *outliers* maiores que 1, na Figura 19 apresentamos a “p_velocidade” em escala logarítmica para facilitar a análise da distribuição dos valores desse sinal.

Ao analisar a Figura 19, percebemos que a velocidade dos PLOs aprovados, em geral, é maior que a velocidade dos PLOs arquivados. Essa proporção explica os valores positivos dos coeficientes do sinal “p_velocidade” calculados pela Regressão Logística e que estão indicados na Figura 16.

Figura 19 – Comparação da “p_velocidade” entre os PLOs aprovados (1) e os PLOs arquivados (0) entre 2003 e 2016.



Fonte: Dados da pesquisa.

Na lista de impacto dos sinais, ilustrado na Figura 17 a “p_velocidade” é somente a quarta mais impactante para aprovação. Isso ocorre devido ao fato da velocidade possuir valores baixos, pois a maior parte dos projetos possuem velocidades que variam entre 0 a 1, ver Figura 19. Como o impacto de uma variável é calculado pelo seu efeito médio no modelo LASSO em sua decisão de classificação da variável dependente (ver fórmula 4.2) o baixo valor dessa variável causou, portanto, o baixo valor do impacto.

Completando a análise desse sinal, a Floresta Aleatória na Figura 18 indicou como sinal mais importante a “p_velocidade”, ou seja, segundo o modelo da Floresta Aleatória esta é a primeira variável que precisa ser analisada para entender se o projeto está em vias de ser aprovado ou não.

Do ponto de vista do Processo Legislativo, o projeto só tramita e dessa forma aumenta a sua velocidade, se esse for pautado para ser debatido nas reuniões que ocorrem nas Comissões da Câmara dos Deputados ou para ser votado nas sessões do Plenário. Assim, uma forma de antecipar a identificação dos projetos com maiores chances de ser aprovados, é analisar a pauta dos eventos que ocorrem tanto no Plenário quanto àqueles que ocorrem nas Comissões da Câmara

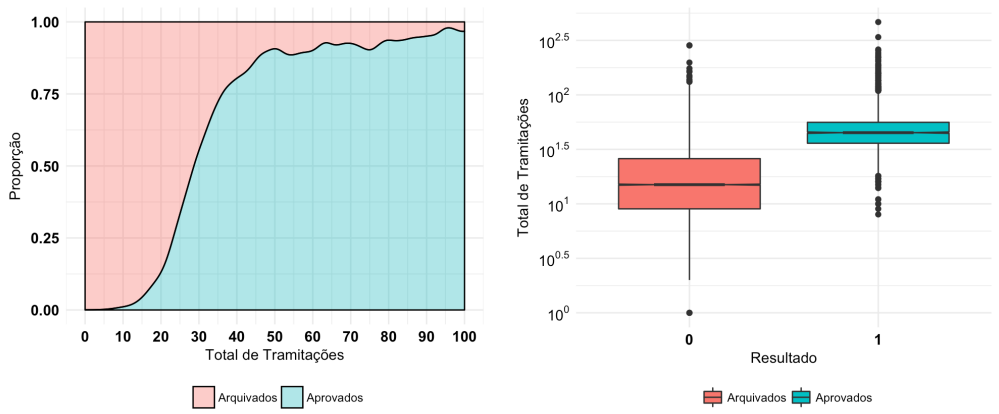
dos Deputados.

Do ponto de vista dos Grupos de Pressão, podemos entender que caso um grupo possua um interesse em aprovar um projeto, é importante criar uma estratégia para aumentar a velocidade de tramitação do mesmo, isso significa que ele deve criar ações de influência para manter um determinado PLO na pauta da Câmara dos Deputados. Por outro lado, os grupos contrários a um determinado PLO podem pensar em estratégias que tirem a atenção do assunto para que dessa forma ele não seja pautado e assim a sua velocidade de tramitação seja reduzida.

Em seguida, a "p_total_tramitacoes" indica o total de tramitações que um projeto sofreu até o momento da sua aprovação ou arquivamento. Ao analisar os coeficientes calculados da "p_total_tramitacoes" na Regressão Logística, verificou-se que todos são positivos. A média na Regressão Logística do coeficiente nos 7 experimentos foi de 0.18. Como os coeficientes são todos positivos, podemos interpretar que quanto maior o número de tramitações, maior a chance de um PLO ser aprovado.

Ao analisar as Figuras 20a e 20b é possível notar a evidente diferença entre o número total de tramitações ocorridas entre projetos arquivados e aprovados.

Figura 20 – Análise da característica "p_total_tramitacoes" dos PLOs aprovados e arquivados entre 2003 e 2016.



(a) Proporção de PLOs por tramitação (b) Total de tramitações nos PLOs

Fonte: Dados da pesquisa.

Visto que não há limites quantitativos para apresentação de PLOs pelos parlamentares, e ao mesmo tempo, a quantidade de legisladores e tempo são recursos limitados, não é esperado que todos os PLOs recebam o mesmo número de tramitações, pois cada tramitação pode ser entendida como resultado de uma decisão que os parlamentares estão tomando sobre um determinado projeto, e essas decisões consomem pouco dos recursos finitos disponíveis.

No entanto, ao analisar a Figura 20a, percebe-se que há muito mais PLOs arquivados que receberam somente entre 0 a 10 tramitações, do que aprovados. Será que todas essas propostas

que não receberam tramitações são irrelevantes e por isso não mereciam um debate mais profundo dos parlamentares?

Considerando, nesse caso, que os projetos são irrelevantes e que por isso não compensaria investir tempo analisando as propostas, será que os parlamentares estão sabendo utilizar bem seu poder constitucional de propor novas leis? Caso contrário, se há projetos relevantes que não estão sendo analisados, será que as regras e o modelo do Processo Legislativo não precisam ser revistos para garantir que o máximo de propostas sejam discutidas?

Independente da eficiência do Processo Legislativo e dos questionamentos acima, é possível afirmar que para um PLO ser aprovado é necessário que ele sofra tramitações, ou seja, que caminhe pelo Processo Legislativo. Todavia poucos PLOs são aprovados e considerando que grande parte dos projetos arquivados não recebem ao menos 10 tramitações, podemos entender então que o padrão de vida de um PLO ao ser apresentado é ser "engavetado", ou seja, não avançar no Processo Legislativo até o momento do seu arquivamento.

Do ponto de vista do Grupo de Pressão, o sinal "p_tramitacoes" reforça o *insight* de atuação que o sinal "p_velocidade" produziu. Caso algum Grupo de Pressão esteja interessado em aprovar um projeto, eles não podem esperar passivamente que os projetos do seu interesse caminhem pelo Processo Legislativo. É preciso realizar um trabalho constante de influência para aumentarem as chances do projeto avançar no Processo Legislativo, e consequentemente produzir a sua aprovação. Por outro lado, se o interesse de um Grupo de Pressão é o arquivamento do PLO, então uma postura mais reativa pode ser mais interessante, sendo mais efetivo monitorar o total de tramitações e a sua velocidade para garantir que o PLO não está sendo pautado.

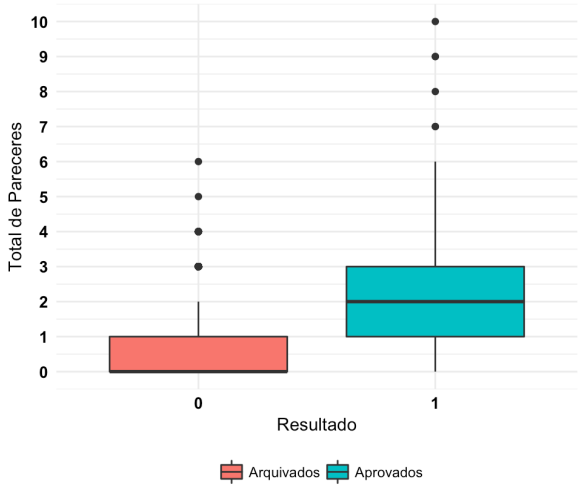
Ao analisar a variável a "p_total_pareceres", que indica o total de pareceres que um PLO recebeu, corroboramos o debate realizado até o momento. Na Figura 21 podemos perceber que poucos projetos arquivados receberam um parecer, ou seja, os parlamentares sequer analisam o mérito de grande parte dos PLOs apresentados.

Novamente, colocamos a questão, como poucos PLOs são analisados, será que propostas importantes para o país estão sendo ignoradas pelos parlamentares?

O quarto sinal analisado é o "p_total_orgaos", que indica quantos órgãos internos da Câmara precisarão analisar os projetos durante o Processo Legislativo. Considera-se aqui, como órgãos internos da Câmara dos Deputados, as Comissões e o Plenário. Ressaltando que, em regra, os projetos são finalizados nas comissões, não tendo necessidade de passarem pelo Plenário da Casa. Assim, segundo o quarto sinal analisado, em média 2 ou 3 comissões temáticas analisam e votam um determinado PLO. Diferentemente dos sinais analisados até o momento, todos os coeficientes do sinal "p_total_orgaos" apresentaram valores negativos na Regressão Logística (ver Figura 16). Em complementação, também foi o sinal que mais impactou para decisão pelo arquivamento no LASSO (ver Figura 17).

Dessa forma, podemos entender que quanto maior número de órgãos para analisar o

Figura 21 – Projetos por total de pareceres recebidos



Fonte: Dados da pesquisa.

projeto, maior número de passos o PLO precisará passar, assim sendo, quanto mais órgãos, maior a tendência do projeto ser arquivado, pois o caminho a ser percorrido pelo projeto é maior.

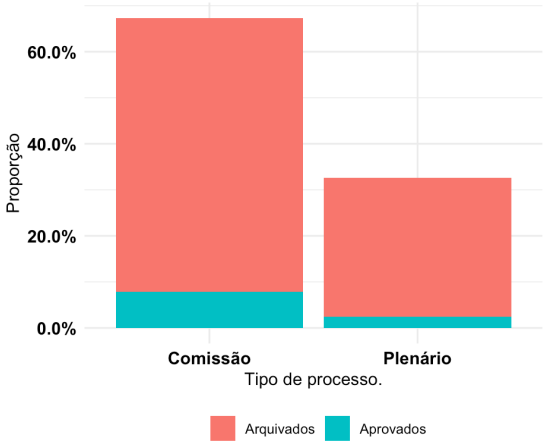
Após a apresentação do PLO, a Mesa Diretora determina quais Comissões deverão analisar o projeto e se esse será votado em Plenário ou não. Assim, nesse momento, os grupos favoráveis ao projeto podem utilizar dessa janela de decisão para pedir um recurso, visando reduzir a quantidade de órgãos e tentar impedir a análise do PLO pelo Plenário, já grupos contrários ao projeto podem entrar com recursos para aumentar o caminho a ser percorrido, e tentar garantir que o PLO passe pelo Plenário.

Assim como citado anteriormente, nem todos os projetos precisam passar pelo crivo do Plenário, alguns deles tramitam de forma conclusiva pelas Comissões, ou seja, eles podem ser aprovados pela Câmara dos Deputados se simplesmente forem aprovados pelas Comissões. Para cada projeto há o sinal "p_processo" que indica se o projeto irá passar somente pelas Comissões (0) ou se também irá precisar ser votado pelo plenário (2). Assim como o sinal anterior, todos os coeficientes do "p_processo" são negativos (ver Figura 16).

Ao analisar a Figura 22 percebe-se que grande parte dos projetos tramitam de forma conclusiva. A decisão relativa se um PLO deve ser analisado ou não de forma conclusiva pelas Comissões cabe ao Presidente da Câmara, sendo que a sua decisão é balizada pelo Regimento Interno da Câmara dos Deputados. Ao mesmo tempo, essa decisão não é imutável, pois qualquer parlamentar pode apresentar um recurso requerendo que um PLO seja analisado pelo Plenário.

Da mesma forma que o *insight* anterior, podemos entender que ao adicionar o Plenário no Processo, o caminho a ser percorrido aumenta e, por isso, a chance de aprovação é reduzida. Além disso, a taxa de aprovação dos projetos que passam pelo Plenário é menor que a taxa de aprovação dos projetos que somente passam pelas Comissões.

Figura 22 – Proporção dos PLOs analisados por somente pelas comissões em contraste com os PLOs analisados pelo Plenário. PLOs aprovados/arquivados entre 2003 e 2016.



Fonte: Dados da pesquisa.

Diante desse cenário, Grupos de Pressão favoráveis a um determinado PLO, devem evitar o Plenário, por outro lado, grupos contrários ao projeto devem convencer algum parlamentar para que ele apresente um recurso pedindo que o PLO seja analisado pelo Plenário.

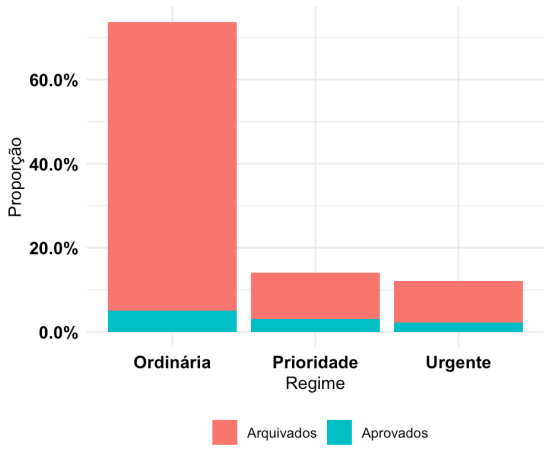
Selecionado como estatisticamente significante em 6 conjuntos, ver Tabela 4, está o sexto sinal "p_regime", que indica se um projeto está sob regime ordinário (valor 1), prioritário (valor 2) ou de urgência (valor 3). Todos os coeficientes do "p_regime" foram positivos (ver Figura 16), assim como o seu impacto (ver Figura 17). Por isso, podemos entender que um PLO em regime urgência possui mais chances de ser aprovado que um projeto em regime prioritário, que, por sua vez, possui mais chances que um projeto em regime ordinário.

Ao analisar a distribuição entre os regimes de tramitação dispostos na Figura 23, podemos perceber que há menos projetos em urgência, quando comparado ao total de projetos em prioridade, e há menos projetos em prioridade comparado a projetos em regime ordinário. Todavia a proporção entre projetos aprovados e arquivados nos regimes de urgência e prioridade é muito maior que nos projetos no regime ordinário.

Do ponto de vista dos Grupos de Pressão, todos aqueles projetos em regime de prioridade ou de urgência devem demandar ações de influência, pois além dos PLOs em urgência e prioridade possuírem mais chances de serem aprovados, o tempo de tramitação dos projetos em urgência é menor que o tempo médio de aprovação de PLOs em prioridade os quais, por sua vez, possuem um tempo menor de tramitação que os PLOs em regime ordinário (ver Figura 24). Novamente, os Grupos de Pressão contrários aos PLOs em urgência ou em prioridade precisam criar estratégias para que os parlamentares votem pela rejeição, por outro lado, os Grupos de Pressão favoráveis precisam garantir que os parlamentares votem a favor do PLO.

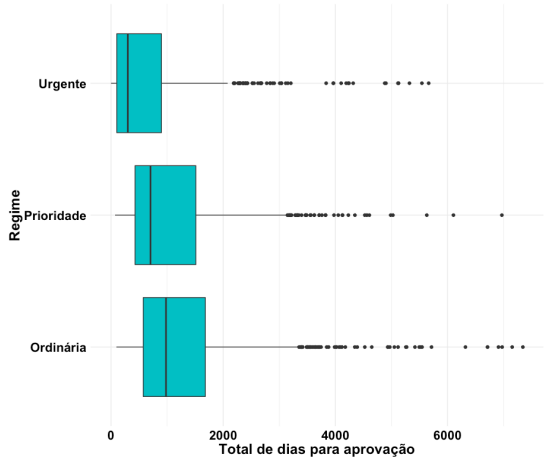
A diferença entre o tempo médio de aprovação de um PLO em urgência em comparação

Figura 23 – Regime de Tramitação dos PLOs aprovados/arquivados entre 2003 e 2016.



Fonte: Dados da pesquisa.

Figura 24 – Total de dias em tramitação dos PLOs aprovados entre 2003 e 2016.



Fonte: Dados da pesquisa.

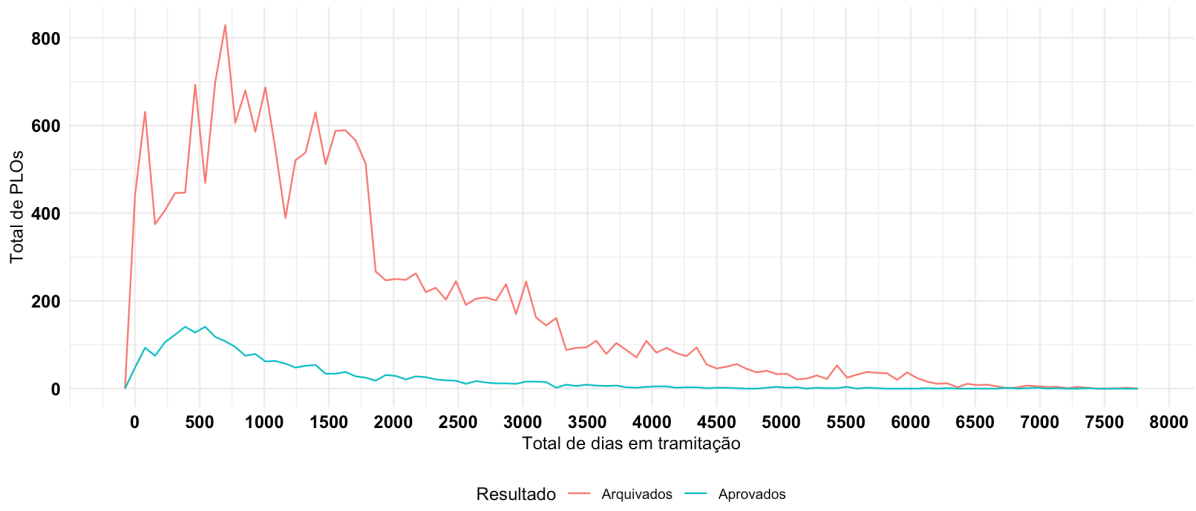
a um PLO em regime ordinário é esperado, visto que, quando um projeto está sob regime de urgência ou sob regime de prioridade, os prazos para decisões são mais curtos e o projeto possui prioridade na ordem de votação. É importante ressaltar que o regime de urgência possui prazos mais curtos e maior prioridade na ordem de votação do que o regime de prioridade.

Por último, o sétimo sinal, o "p_total_dias", indica quantos dias passaram entre o início e fim da tramitação do PLO. Esse sinal foi estatisticamente significativo em 5 conjuntos na Regressão Logística. Ao plotar o histograma do total de dias dos PLOs, temos a Figura 25.

Ao analisar a linha dos PLOs aprovados, o pico é em torno de 500 dias, a partir desse tempo a quantidade dos PLOs aprovados caem. Dessa forma, podemos entender que quanto mais tempo passar, após o período de 500 dias, menor as chances de um PLO ser aprovado.

Dessa forma, para os grupos interessados em aprovar um determinado PLO, é importante

Figura 25 – Histograma dos PLOs pelo seu tempo de tramitação em dias.



Fonte: Dados da pesquisa.

tentar fazê-lo o quanto antes, pois grande parte dos projetos são aprovados nos primeiros 500 dias. Por outro lado, para os grupos contrários é recomendado pensar em estratégias para tirar a atenção do projeto, tornando-o obsoleto e assim tirando sua atratividade aos olhos dos parlamentares.

Os outros 13 sinais, apesar de terem sido selecionadas pelo LASSO por demonstrarem um certo poder preditivo, foram selecionados de forma significativa em 4 ou menos conjuntos (ver Tabela 4). Das 13, 6 sinais, são relacionados ao tema do projeto, são eles: "p_assunto_-direito_penal", "p_assunto_meio_ambiente", "p_assunto_transporte", "p_assunto_homenagem_-data_comemorativa", "p_assunto_fiscalizacao_controle" e "p_assunto_desenvolvimento_social".

Não é possível afirmar que os sinais relacionados aos assuntos dos PLOs demonstram um comportamento padrão no Processo Legislativo, pois, além de não serem classificados como estatisticamente significantes em grande parte dos conjuntos, seus coeficientes não apresentam estabilidade.

Por exemplo, o "p_assunto_meio_ambiente", que foi selecionado três vezes, nos biênios 2003-2004 e 2009-2010, obteve coeficientes positivos, já em 2011-2012, o coeficiente foi negativo. Ou seja, nos primeiros biênios, o Congresso demonstrou a tendência de aprovar projetos sobre o tema meio ambiente, já em 2011-2012, mostrou-se o contrário.

Já o "p_assunto_direito_penal", que foi selecionado 4 vezes como significante na Regressão Logística, apresentou coeficientes positivos em todas as ocasiões. Isso demonstra uma predisposição da Câmara dos Deputados em aprovar projetos sobre esse tema, todavia, nada impede que um futuro Congresso, com outra composição, inverta essa lógica.

Os outros 7 sinais são: "p_total_apensados", "l_tipo_autor_legislador", "p_apensado_-flag", "l_governo", "l_presidente_flag", "l_membro_comissao" e "l_sessao_legislativa".

Entre os 7 sinais acima, o sinal "l_tipo_legislador", o qual indica que o projeto é de autoria de um parlamentar, é o mais significante estatisticamente, sendo que foi selecionado em 3 conjuntos pela Regressão Logística. Apesar de apresentar valores baixos nos coeficientes da Regressão, foi o quarto sinal que mais impactou pela decisão de arquivamento do LASSO. É importante notar que 98% dos projetos arquivados nos 7 conjuntos de dados foram de autoria de parlamentares. Assim, podemos interpretar que esse fator de impacto está inflacionado por esta proporcionalidade.

Todavia, podemos relacionar o impacto negativo do "l_tipo_legislador" a estudos da Ciência Política. Diversos trabalhos publicados já ressaltaram o baixo protagonismo dos parlamentares no Processo Legislativo. O nosso modelo político, chamado de Presidencialismo de Coalizão, partilha com o Poder Executivo significativos poderes e influência no pauta legislativa. Assim grande parte dos projetos aprovados ou pautados nas reuniões são ou de autoria ou de interesse do Executivo (RODRIGUES, 2004).

Já os outros 6 sinais, "p_total_apensados", "p_apensado_flag", "l_governo", "l_presidente_flag", "l_membro_comissao" e "l_sessao_legislativa", foram considerados poucas vezes como estatisticamente significante pela Regressão Logística, por isso, é difícil tirar conclusões robustas a partir deles.

6.2 Objetivo 2: Capacidade Preditiva

Após analisar as variáveis que influenciam a aprovação ou não de um projeto de lei, passamos a analisar se o modelo final, gerado pelos 20 sinais selecionados, é capaz de prever se um PLO será aprovado ou não.

Na Figura 11, identificamos que há muito mais projetos arquivados do que aprovados, por isso, naturalmente, é mais fácil identificar quais projetos possuem mais chances de serem arquivados. Assim, para verificar se um modelo seria capaz de prever de forma eficiente o prognóstico positivo de um PLO, escolhemos dois indicadores: o F1 Score e a precisão, conforme explanado na seção de Metodologia. Para realizar os experimentos, também seguimos a configuração detalhada na Figura 10.

Ao analisar os resultados expostos na Tabela 7, percebemos que a capacidade preditiva do modelo da Floresta Aleatória obteve um resultado superior ao LASSO, tanto pela métrica F1 Score como pela precisão. Enquanto a média da F1 Score na LASSO foi de 0.789, na Floresta Aleatória foi de 0.861, representando uma variação positiva de 10%.

Ao analisar o indicador da precisão, os dois modelos conseguiram gerar melhores resultados quando testados na segunda metade da legislatura, que, no caso, são os biênios de 2005-2006, 2009-2010 e 2013-2014. Nesses biênios, a proporção de projetos arquivados é menor se comparado aos primeiros biênios de cada legislatura (2007-2008, 2011-2012 e 2015-2016),

Tabela 7 – Comparação dos resultados dos Experimentos de Predição

Experimento	Anos de Treino	Anos de Teste	F1 Score		precisão	
-	-	-	LASSO	Floresta Aleatória	LASSO	Floresta Aleatória
#1	2003-2004	2005-2006	0,845	0,860	0,842	0,784
#2	2003-2006	2007-2008	0,736	0,764	0,586	0,621
#3	2003-2008	2009-2010	0,851	0,896	0,767	0,822
#4	2003-2010	2011-2012	0,742	0,893	0,608	0,819
#5	2003-2012	2013-2014	0,790	0,906	0,666	0,844
#6	2003-2014	2015-2016	0,767	0,848	0,654	0,777
-	-	**Média**	**0,789**	**0,861**	**0,687**	**0,778**

tornando a tarefa de identificação dos PLOs aprovados relativamente mais fácil. Inclusive na precisão, a Floresta Aleatória também apresentou uma média maior que o LASSO, sendo que a Floresta apresentou uma média de 0.778, enquanto o LASSO alcançou 0.687.

De forma geral, dada a dificuldade na qual somente 10,4% dos PLOs são aprovados, considerando o modelo da Floresta Aleatória que alcançou um média de precisão de 0.778 e um F1-Score de 0.861 é um significativo avanço para identificar qual daqueles projetos possuem mais chances de serem aprovados na Câmara dos Deputados, visto que, atualmente, não há outro modelo publicado na literatura acadêmica no contexto do Brasil.

Quando o modelo da Floresta Aleatória classifica positivamente um PLO, podemos considerar a exatidão dessa informação de 70%. Ao mesmo tempo, quando o mesmo modelo classifica negativamente, ou seja, quando ele diz que o projeto será arquivado, a exatidão dessa dado é de 99,5% (calculado pelo VPN), ver tabela 8.

Tabela 8 – VPN (Valor Preditivo Negativo) da Floresta Aleatória

Experimento	Anos de Treino	Anos de Teste	VPN
#1	2003-2004	2005-2006	0,986
#2	2003-2006	2007-2008	0,999
#3	2003-2008	2009-2010	0,995
#4	2003-2010	2011-2012	0,998
#5	2003-2012	2013-2014	0,995
#6	2003-2014	2015-2016	0,994
-	-	**Média**	**0,995**

Do ponto de vista do Grupo de Pressão, se a Floresta Aleatória classifica negativamente, então existe grandes chances do projeto ser arquivado. Caso o grupo seja favorável ao projeto, este deve criar uma estratégia que visa estimular as variáveis que relacionam positivamente a aprovação de um PLO, por exemplo, sua velocidade. Se o grupo for contrário, basta monitorar e acompanhar a sua tramitação observando as variáveis descritas na seção anterior.

Por outro lado, se o modelo classifica positivamente um PLO, a exatidão dessa informação é menor, de 70%. Assim, se um projeto for classificado como positivo pela Floresta, os grupos

favoráveis aos projetos podem, a partir dessa informação, priorizar esse projeto e analisar as variáveis que estão causando essa maior chance, bem como criar uma estratégia visando potencializá-las. Ao mesmo tempo, os grupos contrários também podem utilizar essa informação para priorização e pensar estratégias para neutralizar o projeto.

É importante ressaltar que a preditividade temporal do modelo não é muito alta, pois os sinais utilizados no modelo final são valores que são gerados como resultado das ações dos parlamentares. Ou seja, o modelo não consegue capturar a intenção dos parlamentares em aprovar ou não um determinado projeto de lei. Quando o modelo indica que um projeto possui altas chances de ser aprovado e ainda não o foi, há pouco tempo para reação, visto que as ações de aprovação já estão sendo executadas. Dessa forma, os Grupos de Pressão possuem pouco tempo para reagir, todavia, o modelo a partir de suas variáveis podem facilitar a análise do cenários e o pensamento de estratégias de influência.

CAPÍTULO

7

CONCLUSÃO

Nesse trabalho, dentro do contexto da área de Ciência de Dados, utilizamos: dados legislativos, uma metodologia clara e modelos de Aprendizagem de Máquina para compreender e prever a aprovação de um Projeto de Lei Ordinária na Câmara dos Deputados.

A primeira etapa envolveu coletar e sistematizar os dados dos PLOs aprovados e arquivados entre 2003 e 2016 em 72 características (sinais). Para respeitar a alternância do poder da Câmara, os dados foram organizados em biênios, sendo assim, no final totalizaram 7 conjuntos de dados. Posteriormente, como havia muito mais projetos arquivados do que aprovados, balanceamos os dados. Em seguida, por meio do LASSO selecionamos apenas as características mais relevantes, gerando um modelo com 20 variáveis (sinais).

A partir dos 20 sinais selecionados, aplicamos a Regressão Logísticas para entender quais, entre elas, rejeitavam a hipótese nula, ou seja, quais eram estatisticamente significante ao longo do tempo. Ao final, identificamos 7 sinais mais relevantes para todos os conjuntos de dados. Enquanto 4 sinais contribuíram positivamente à aprovação, houve outros 3 sinais que contribuíram para o arquivamento de um PLO. Os outros 13 sinais que apesar de terem sido selecionadas pelo LASSO para o modelo final não demonstraram ser estatisticamente significantes, dessa forma eles não contribuíram para a compreensão do Processo Legislativo.

Os 4 sinais relacionados à aprovação encontrados foram: “p_total_tramitacoes”, “p_total_pareceres”, “p_regime” e “p_velocidade”. Por outro lado, os 3 sinais relacionados ao arquivamento encontrados foram: “p_total_orgaos”, “p_processo” e “p_total_dias”.

Ao mesmo tempo que identificamos e analisamos os sinais mais relevantes para analisar as chances de um PLO ser aprovado ou não, também geramos *insights* de ações estratégicas que podem ser utilizados por Grupos de Pressão interessados em influenciar positivamente ou não a aprovação de um determinado PLO.

Após compreender o Processo Legislativo para um PLO na Câmara dos Deputados,

partimos para avaliação preditiva do modelo. Realizamos experimentos utilizando o LASSO e a Floresta Aleatória nos conjuntos de dados. A Floresta Aleatória apresentou melhores resultados, sendo que o modelo alcançou o F1 Score de 0.861 e na métrica precisão alcançou 0.778. Com isso pudemos afirmar que a Floresta Aleatória consegue identificar com uma boa exatidão os projetos que serão aprovados, e ao mesmo tempo há uma boa confiabilidade quando o modelo prediz positivamente.

De forma complementar, também analisamos a qualidade da classificação negativa da Floresta Aleatória, a qual apresentou um resultado ótimo alcançando a métrica VPN de 0.995, indicando que quando classificado negativamente, a chance da Floresta Aleatória estar correta é de quase 100%.

Com as informações e *insights* gerados pelo nosso artefato, os mais diversos Grupos de Pressão podem entender melhor as regras do Processo Legislativo e potencializar a sua participação. Além isso esses Grupos poderão avaliar objetivamente o contexto de um PLO, e verificar se vale a pena ou não investir tempo e recursos tentando influenciar as decisões dos parlamentares.

É importante lembrar que a política está em constante transformação e a forma como os políticos são influenciados e tomam decisões evoluem a todo momento, seja por mudança cultural da população, seja por novas regras do Processo Legislativo ou seja ainda pelas novas tecnologias de comunicação, por exemplo, as Redes Sociais. Por isso, será necessário revisitar esse modelo e as características de tempos em tempos e refazer os testes para verificar se ainda o modelo continua válido.

Apesar do artefato final ainda não ser ótimo, ele fornece informações importantes para os grupos que precisam monitorar e influenciar o Processo Legislativo. Ademais, os resultados alcançados no trabalho foram mais consistentes e robustos que os resultados alcançados em trabalho preliminar publicado pelos mesmos autores no Simpósio Brasileiro de Sistemas de Informação no ano de 2018 (OLIVEIRA; ALBUQUERQUE; DELBEM, 2018).

Outro ponto de atenção é que as 20 variáveis que compõem o modelo final são dados gerados a partir da reação às atividades dos grupos políticos ou legisladores, por exemplo, uma tramitação aumenta a "p_total_tramitacoes", aumentando as chances de um PLO ser aprovado. Todavia, esse modelo não captura a intenção das atividades dos políticos, mas sim, as variáveis que são resultados delas, por isso, a predição temporal de um PLO é limitada, sendo que o artefato não captura a intenção dos parlamentares em aprovar ou não um projeto de lei.

Ao final desse trabalho, podemos afirmar que a metodologia e resultados utilizados por essa pesquisa gerou um artefato, utilizando os modelos da Regressão Logística e da Floresta Aleatória, capaz de explicar e prever a aprovação de um Projeto de Lei Ordinária na Câmara dos Deputados, ou seja, podemos afirmar que os dois objetivos propostos foram cumpridos.

Os atuais resultados estão circunscritos ao escopo que remete somente para uma parte do

Processo Legislativo de uma casa legislativa, no caso para Projetos de Lei Ordinária na Câmara dos Deputados. Todavia a metodologia pode ser replicada, por exemplo, para criar um artefato que possa prever se esses PLOs seriam aprovados no Senado Federal, ou mesmo, se um Projeto de Lei Ordinária seria aprovado na Assembleia Legislativa de São Paulo.

Com a metodologia já desenhada e validada o tempo para criação de um artefato é reduzido facilitando assim que novas pesquisas com outros objetos de estudo possam surgir, por isso, afirmamos que essa é a terceira contribuição desse trabalho.

7.1 Trabalhos Futuros

Nos próximos trabalhos pretendemos explorar outros meios que otimizem a predição de um projeto de lei ser aprovado ou não. É possível, assim como apresentado na seção de Trabalhos Relacionados, utilizar Redes Complexas para criar modelos mais resilientes e preditivos que capturem a intenção dos parlamentares sobre um determinado tema, assim, baseado nos dados será possível antecipar os votos dos legisladores. Dessa forma, já no momento da apresentação do projeto, com esse tipo de modelo seria possível calcular um prognóstico de maior qualidade.

REFERÊNCIAS

AGUNE, R. M.; FILHO, A. S. G.; BOLLIGER, S. P. Governo aberto SP: Disponibilização de Bases de Dados e Informações em Formato Aberto. **III Congresso Consad de Gestão Pública**, p. 1–36, 2009. Citado na página 42.

ANDERSEN, T. B. E-Government as an anti-corruption strategy. **Information Economics and Policy**, Elsevier B.V., v. 21, n. 3, p. 201–210, 2009. ISSN 01676245. Disponível em: <http://dx.doi.org/10.1016/j.infoecopol.2008.11.003>. Citado na página 42.

ARAÚJO, P. M. O bicameralismo importa ? Uma análise da produção legislativa no Brasil a partir do Senado Federal. p. 1–26, 2004. Disponível em: <http://www.bilhetim.com.br/upload/arq{_}arquivo/55.> Citado na página 39.

A.RICE, S. The political vote as a frequency distribution of opinion. **Journal of the American Statistical Association**, Taylor & Francis, v. 19, n. 145, p. 70–75, 1924. Citado na página 47.

BAPTISTA, V. M. P. d. S. Universidade Federal da Paraíba Centro de Informática Programa de Pós-Graduação em Informática Um modelo para a detecção das mudanças de posicionamento dos deputados federais Vitor Márcio Paiva de Sousa Baptista. 2015. Citado na página 51.

BRASIL. Constituição da República Federativa do Brasil. v. 2, n. 00101023023054, p. 4, 1988. Citado nas páginas 39, 40 e 42.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, Oct 2001. ISSN 1573-0565. Disponível em: <https://doi.org/10.1023/A:1010933404324>. Citado na página 37.

CALAZANS, M.; CORTES, I. O processo de criação aprovação implementação da Maria da Penha. **Brasil. Lei Maria da Penha comentada em uma perspectiva jurídico-feminista.**, p. 39–63, 2010. Citado na página 25.

CARVALHO, F. d. S. O Papel Dos Grupos De Interesse E Pressão Na Formatação E Fortalecimento Da Democracia Brasileira :. **Aurora**, n. 5, p. 32–39, 2009. Citado na página 45.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. Smote: synthetic minority over-sampling technique. **Journal of artificial intelligence research**, v. 16, p. 321–357, 2002. Citado na página 59.

CHEN, A. C.; FU, X. Data + intuition: A hybrid approach to developing product north star metrics. In: **Proceedings of the 26th International Conference on World Wide Web Companion**. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2017. (WWW '17 Companion), p. 617–625. ISBN 978-1-4503-4914-7. Disponível em: <https://doi.org/10.1145/3041021.3054199>. Citado na página 55.

CHENG, Y.; AGRAWAL, A.; LIU, H.; CHOUDHARY, A. Legislative Prediction with Dual Uncertainty Minimization from Heterogeneous Information. **Statistical Analysis and Data Mining: The ASA Data Science Journal**, 2016. ISSN 19321872. Citado na página 49.

CHO, W. K. T.; FOWLER, J. H. Legislative Success in a Small World: Social Network Analysis and the Dynamics of Congressional Legislation. **Journal of Politics**, v. 72, n. 1, p. 124–135, 2010. ISSN 0022-3816. Citado na página 49.

CONTROLADORIA-GERAL, D. U. Acesso à informação pública: uma introdução à lei nº 12.527, de 18 de novembro de 2011. **Brasília (DF)**, 2011. Citado na página 42.

CONWAY, D. **THE DATA SCIENCE VENN DIAGRAM**. 2010. Disponível em: <http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>. Citado nas páginas 29 e 30.

CRUVINEL, G. W. F. Dados governamentais abertos: um diagnóstico da demanda na Câmara dos Deputados. v. 1, p. 159, 2016. ISSN 1098-6596. Citado na página 44.

DANTAS, H.; TOLEDO, J. R. de; TEIXEIRA, M. A. **Análise política & jornalismo de dados: ensaios a partir do Basômetro**. [S.l.: s.n.], 2014. 228 p. Citado na página 52.

DHAR, V. Data science and prediction. **Commun. ACM**, ACM, New York, NY, USA, v. 56, n. 12, p. 64–73, dez. 2013. ISSN 0001-0782. Disponível em: <http://doi.acm.org/10.1145/2500499>. Citado na página 29.

EIDELMAN, V.; KORNILOVA, A.; ARGYLE, D. How Predictable is Your State? Leveraging Lexical and Contextual Information for Predicting Legislative Floor Action at the State Level. 2018. Disponível em: <http://arxiv.org/abs/1806.05284>. Citado na página 49.

FARHAT, S. **Lobby, o que é, como se faz: ética e transparência na representação junto a governos**. [S.l.]: Editora Peirópolis, 2007. Citado na página 44.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. Knowledge Discovery and Data Mining: Towards a Unifying Framework. **Int Conf on Knowledge Discovery and Data Mining**, p. 82–88, 1996. Disponível em: <http://www.aaai.org/Papers/KDD/1996/KDD96-014>. Citado na página 32.

FONTI, V. Feature Selection using LASSO. 2017. Citado na página 60.

GOMES, F. D. B. C. **Produção Legislativa no Brasil: Visão Sistêmica e Estratégica no Presidencialismo de Coalizão**. [s.n.], 2013. ISBN 9788540200999. Disponível em: <http://bd.camara.gov.br/bd/handle/bdcamara/13756>. Citado na página 50.

GREGOR, S.; HEVNER, A. R. Positioning and Presenting Design Science Research for Maximum Impact. **MIS Quarterly**, v. 37, n. 2, p. 337–355, 2013. ISSN 02767783. Citado na página 64.

HE, H.; GARCIA, E. A. Learning from imbalanced data. **IEEE Transactions on Knowledge and Data Engineering**, v. 21, n. 9, p. 1263–1284, 2009. ISSN 10414347. Citado na página 59.

HEVNER, A. R.; MARCH, S. T.; PARK, J.; RAM, S. Design Science in Information Systems Research. **MIS Quarterly**, v. 28, n. 1, p. 75–105, 2004. ISSN 0276-7783. Citado na página 55.

HOERL, A. E.; KENNARD, R. W. Biased Estimation for Ridge Regression : Problems Nonorthogonal. v. 42, n. 1, p. 80–86, 2015. Citado na página 35.

JR, D. W. H.; LEMESHOW, S.; STURDIVANT, R. X. **Applied logistic regression**. [S.l.]: John Wiley & Sons, 2013. v. 398. Citado na página 33.

KELSEN, H. **Pure theory of law**. [S.l.]: Univ of California Press, 1967. Citado na página 38.

LEONI, E. Ideologia, democracia e comportamento parlamentar: a Camara dos Deputados (1991-1998). **Dados**, scielo, v. 45, p. 361 – 386, 00 2002. ISSN 0011-5258. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0011-52582002000300002&nrm=iso>. Citado na página 51.

LIBBY, O. G. A plea for the study of votes in congress. **Annual Report of the American Historical Association for 1896**, U.S. Government Printing Office, p. 323–334, 1897. Citado na página 47.

LYRA, K.; ISOTANI, S. Impacto do uso de infográficos como materiais de aprendizagem e suas correlações com satisfação, estilos de aprendizagem e complexidade visual. p. 46, 2017. ISSN 2316-8889. Disponível em: <http://www.br-ie.org/pub/index.php/wcbie/article/view/7363>. Citado na página 55.

MANCUSO, W. P.; GOZETTO, A. C. O. Lobby : instrumento democrático de representação de interesses ? **Organicom**, v. 8, n. 14, p. 118–128, 2011. ISSN 2238-2593. Citado na página 23.

MARCH, S. T.; SMITH, G. F. Design and natural science research on information technology. **Decision Support Systems**, v. 15, n. 4, p. 251–266, 1995. ISSN 01679236. Citado na página 55.

MENEGUIN, F. B. Avaliação de impacto legislativo no Brasil. **Centro de Estudos da Consultoria do Senado**, 2010. Disponível em: <http://escholarship.org/uc/item/8ts831r2>. Citado na página 23.

MOREIRA, J.; CARVALHO, A.; HORVATH, T. **A General Introduction to Data Analytics**. Wiley, 2018. ISBN 9781119296263. Disponível em: <https://books.google.com.br/books?id=9BtiDwAAQBAJ>. Citado nas páginas 34, 36 e 37.

NAY, J. J. Predicting and Understanding Law-Making with Machine Learning. p. 1–12, 2016. Disponível em: <http://arxiv.org/abs/1607.02109>. Citado na página 48.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. **Journal of the Royal Statistical Society. Series A (General)**, [Royal Statistical Society, Wiley], v. 135, n. 3, p. 370–384, 1972. ISSN 00359238. Disponível em: <http://www.jstor.org/stable/2344614>. Citado na página 33.

OLIVEIRA, D.; ALBUQUERQUE, J. P. de; DELBEM, A. Compreendendo e prevendo o processo legislativo na câmara dos deputados do brasil. jan 2018. Disponível em: <http://XXXXX/180107.pdf>. Citado nas páginas 53 e 92.

PACHECO, L. B. **Como se fazem as leis**. [S.l.: s.n.], 2013. 81 p. ISBN 9788540200449. Citado na página 39.

PARKS, W. THE OPEN GOVERNMENT PRINCIPLE: APPLYING THE RIGHT TO KNOW UNDER. **The George Washington Law**, v. 26, 1957. ISSN 02729490. Citado na página 41.

PEFFERS, K.; TUUNANEN, T.; ROTHENBERGER, M. A.; CHATTERJEE, S. A design science research methodology for information systems research. **Journal of Management Information Systems (JMIS)**, v. 24, n. 3, p. 45–78, 2007. ISSN 0742-1222. Citado na página 56.

POSSAMAI, A. J. Dados Abertos no Governo Federal Brasileiro: Desafios de transparência e interoperabilidade. **Avaliação**, p. 1–309, 2016. Citado na página 47.

QUINLAN, J. R. Induction of decision trees. **Mach. Learn.**, Kluwer Academic Publishers, Hingham, MA, USA, v. 1, n. 1, p. 81–106, mar. 1986. ISSN 0885-6125. Disponível em: <http://dx.doi.org/10.1023/A:1022643204877>. Citado na página 36.

RICCI, P. O conteúdo da produção legislativa brasileira: leis nacionais ou políticas paroquiais? **Dados**, v. 46, n. 4, p. 699–734, 2003. ISSN 0011-5258. Citado na página 50.

RODRIGUES, G. D. M. Produção legislativa: poder executivo & congresso nacional. 2004. Citado na página 87.

SAMUEL, A. L. Some Studies in Machine Learning Using the Game of Checkers. **IBM Journal of Research and Development**, v. 3, n. 3, p. 210–229, 1959. ISSN 0018-8646. Disponível em: <http://ieeexplore.ieee.org/document/5392560/>. Citado na página 31.

SGANZERLA, R.; CERDEIRA, P. d. C.; VASCONCELLOS, F.; CUNHA, B.; CASTRO, M.; CARABETTA, J.; SALES, A.; SCOVINO, F. Congresso em Números 2018 Relatório de Atividades do Congresso Nacional. 2018. Citado na página 51.

SILVA, D.; AFONSO, J. Curso de direito constitucional positivo. 37ª edição, revista e atualizada até a emenda constitucional n. 76, de 28.11. 2013. **São Paulo: Editora Malheiros**, 2014. Citado na página 40.

TADEU, G.; SILVA, R.; DAVID, O.; FLEISCHER, V. Impactos das Frentes Parlamentares na Dinâmica do Congresso Nacional durante a Presidência do Partidos dos Trabalhadores (52 a 54 Legislatura). 2014. Citado na página 51.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society. Series B (Methodological)**, JSTOR, p. 267–288, 1996. Citado na página 36.

WANG, J.; VARSHNEY, K. R.; MOJSILOVI, A. Legislative Prediction via Random Walks over a Heterogeneous Graph. **Sdm**, p. 1095–1106, 2012. Citado na página 48.

WONG, W.; WELCH, E. Does E-Government Promote Accountability? A Comparative Analysis of Website Openness and Government Accountability. **Governance**, v. 17, n. 2, p. 275–297, 2004. ISSN 0952-1895. Disponível em: <http://doi.wiley.com/10.1111/j.1468-0491.2004.00246.x>. Citado na página 42.

YANO, T.; SMITH, N. a.; WILKERSON, J. D. Textual Predictors of Bill Survival in Congressional Committees. **NAACL HLT '12 Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, p. 793–802, 2012. Citado nas páginas 48 e 62.

YU, H.; ROBINSON, D. G. The New Ambiguity of 'Open Government'. **SSRN Electronic Journal**, v. 178, n. 2012, p. 178–208, 2012. ISSN 1556-5068. Disponível em: <http://www.ssrn.com/abstract=2012489>. Citado na página 41.

APÊNDICE

A

CARACTERÍSTICAS PRÉ-SELECIONADAS

A princípio havia 27 características pré-selecionadas para cada PLO aprovado ou arquivado na Câmara dos Deputados. Das 27, 14 são relacionadas ao projeto de lei e 13 são relacionadas ao autor do projeto lei.

As 27 características estão listadas abaixo, enquanto as características relacionadas ao projeto são inciadas com 'p_', as característica do autor principal do projeto se iniciam com 'l_'.

1. **l_autores_grupo_mesclado** (nominal): Se há mais de um autor, se todos os autores pertencerem ou à oposição ou à situação, é Falso. Se há autores tanto da oposição quanto da situação é Verdadeiro.

2. **l_estado** (ordinal): Indica o total de cadeiras ocupadas na Câmara pelos parlamentares do mesmo estado do autor principal do projeto de lei.

3. **l_governo** (nominal): Indica se o autor principal do projeto é membro de algum partido da coalizão do governo.

4. **l_lideranca** (nominal): Indica se o autor era líder de governo e de partido no momento da apresentação do projeto.

5. **l_liderancas_total** (ordinal): Indica quantas vezes o autor foi líder de partido ou de governo até o momento da apresentação do projeto.

6. **l_mandatos** (ordinal): Indica o total de mandatos que o autor possuía até o momento da apresentação do projeto.

7. **l_membro_comissao** (nominal): Indica se o autor principal do projeto é membro, vice-presidente ou presidente de uma das comissão que analisará o projeto no momento de sua apresentação.

8. **l_partido_proporcao** (ordinal): Indica a proporção de cadeiras ocupadas pelo partido do autor na Câmara dos Deputados.

9. **l_presidente_flag** (nominal): Indica se o autor é o do mesmo partido do presidente da Câmara dos Deputados no momento da apresentação do projeto de lei.

10. **l_tema_projeto_score** (ordinal): Indica o total de projetos de lei do mesmo tema do que o autor havia apresentado no momento da apresentação do projeto de lei.

11. **l_tema_relatorias_score** (ordinal): Indica quantas relatorias sobre o mesmo tema da proposição o autor principal havia apresentado no momento da apresentação do projeto de lei

12. **l_tipo_autor** (nominal): Indica se o autor principal pertence ao Poder Legislativo ou ao Poder Executivo ou ao Poder Judiciário ou ao Ministério Público

13. **l_vice_presidente_partido_flag** (nominal): Indica se o autor é o do mesmo partido do vice-presidente da Câmara dos Deputados no momento da apresentação do projeto de lei.

14. **p_apensado_flag** (nominal): Indica se o projeto foi apensado à outro durante sua tramitação.

15. **p_assunto** (nominal): Indica o tema do projeto de lei.

16. **p_casa_revisora_flag** (nominal): Indica se o projeto é proveniente do Senado.

17. **p_emendas** (ordinal): Total de propostas de emendas que o projeto recebeu.

18. **p_total_pareceres** (ordinal): Indica o total de pareceres que o um determinado projeto de lei recebeu.

19. **p_processo** (ordinal): Indica se o projeto é conclusivo pelas comissões ou se precisará passar pelo plenário. Quanto maior o valor, maior será o caminho que o projeto precisará percorrer para ser aprovado. Diversos(0), Comissão(1) e Plenário(2).

20. **p_regime** (ordinal): Indica o regime de tramitação do projeto. Quanto maior o valor, menores são os prazos do Processo Legislativo e mais rápido o projeto de lei caminha. Os seguintes regimes foram considerados: Ordinária(1), Prioridade(2) e Urgência(3).

21. **p_sessao_legislativa** (nominal): Indica em qual ano da legislatura o projeto foi apresentado. A sessão legislativa pode variar entre 1 a 4.

22. **p_total_apensados** (ordinal): Indica o total de projetos de lei foi apensado à um determinado projeto de lei.

23. **p_total_autores** (ordinal): Indica o total de autores do projeto de lei.

24. **p_total_dias** (ordinal): Total de dias que o projeto ficou em tramitação.
25. **p_total_orgaos** (ordinal): Total de comissões da Câmara dos Deputados que foram designadas para analisar o projeto.
26. **p_total_tramitacoes** (ordinal): Total de tramitações que o projeto recebeu.
27. **p_velocidade** (ordinal): Total de tramitações dividido por total de dias no últimos 6 meses de tramitação do projeto antes dele ser arquivado ou aprovado.

Por exemplo, o PLO 7410/2006 que "Altera o § 4º do art. 7º da Lei nº 11.096, de 13 de janeiro de 2005, que instituiu o Programa Universidade para Todos (Prouni), para dispor sobre a desvinculação dos cursos com desempenho insuficiente no Sistema Nacional de Avaliação da Educação Superior (Sinaes)."e que foi aprovado possui as seguintes características:

l_autores_grupo_mesclado: não.

l_estado: 8.

l_governo: sim.

l_lideranca: líder de partido.

l_liderancas_total: 5.

l_mandatos: 2.

l_membro_comissao: não.

l_partido_proporcao: 0,255.

l_presidente_flag: não.

l_tema_projeto_score: 4.

l_tema_relatorias_score: 1.

l_tipo_autor: legislador.

l_vice_presidente_partido_flag: não.

p_apensado_flag: não.

p_assunto: educação

p_casa_revisora_flag: sim.

p_emendas: 1 emenda.

p_total_pareceres: 2.

p_processo: Apreciação Conclusiva pelas Comissões.

p_regime: Prioridade;

p_sessao_legislativa: 4.

p_total_apensados: 0.

p_total_autores: 1.

p_total_dias: 336 dias.

p_total_orgaos: 2.

p_total_tramitacoes: 35 tramitações.

p_velocidade: 0,11.

aprovado: sim.

ICMC USP
SÃO CARLOS