UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
UNIDADE ACADÊMICA DE SISTEMAS E COMPUTAÇÃO
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

DANIYEL NEGROMONTE NASCIMENTO ROCHA

AVALIAÇÃO DE MODELOS PREDITIVOS PARA A EXTRAÇÃO DE CARACTERÍSTICAS SIGNIFICATIVAS NAS ELEIÇÕES BRASILEIRAS

CAMPINA GRANDE - PB

2019

DANIYEL NEGROMONTE NASCIMENTO ROCHA

AVALIAÇÃO DE MODELOS PREDITIVOS PARA A EXTRAÇÃO DE CARACTERÍSTICAS SIGNIFICATIVAS NAS ELEIÇÕES BRASILEIRAS

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Professor Dr. Leandro Balby Marinho.

CAMPINA GRANDE - PB

2019

R672a Rocha, Daniyel Negromonte Nascimento.
Avaliação de modelos preditivos para a extração de características significativas nas eleições brasileiras. / Daniyel Negromonte Nascimento Rocha. – 2019.

14 f.

Orientador: Prof. Dr. Leandro Balby Marinho.
Trabalho de Conclusão de Curso - Artigo (Curso de Bacharelado em Ciência da Computação) - Universidade Federal de Campina Grande; Centro de Engenharia Elétrica e Informática.

1. Aprendizagem de máquina. 2. Predição. 3. Eleições. 4. Dados desbalanceados. 5. Desenvolvimento de modelos preditivos. 6. Algoritmos de aprendizagem de máquina. 7. Modelo Random Forest. 8. Regressão logística. 9. Redes neurais. I. Marinho, Leandro Balby. II. Título.

CDU:004(045)

Elaboração da Ficha Catalográfica:

Johnny Rodrigues Barbosa
Bibliotecário-Documentalista
CRB-15/626

DANIYEL NEGROMONTE NASCIMENTO ROCHA

AVALIAÇÃO DE MODELOS PREDITIVOS PARA A EXTRAÇÃO DE CARACTERÍSTICAS SIGNIFICATIVAS NAS ELEIÇÕES BRASILEIRAS

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA:

Professor Dr. Leandro Balby Marinho
Orientador – UASC/CEEI/UFCG

Professor Dr. Rohit Gheyi
Examinador – UASC/CEEI/UFCG

Professor Dr. Tiago Lima Massoni
Disciplina TCC – UASC/CEEI/UFCG

Trabalho aprovado em: 02 de julho de 2019.

CAMPINA GRANDE - PB

Avaliação de modelos preditivos para a extração de características significativas nas eleições brasileiras

Negromonte N. R., Daniyel

Departamento de Sistemas e Computação, Universidade Federal de Campina Grande (UFCG), Campina Grande, Paraíba, Brasil, daniyel.rocha@ccc.ufcg.edu.br

Balby M., Leandro

Departamento de Sistemas e Computação, Universidade Federal de Campina Grande (UFCG), Campina Grande, Paraíba, Brasil, lbmarinho@computacao.ufcg.edu.br

RESUMO

A dinâmica do sistema político brasileiro dificulta a compreensão do processo eleitoral. Essa situação pode fazer com que a sociedade seja menos participativa, sendo algo prejudicial para a democracia. Nesse sentido, foi feita uma avaliação acerca da utilização de dados públicos e técnicas de aprendizagem de máquina, com o objetivo de observar padrões a partir de eleições passadas e inferir resultados por meio do desenvolvimento de modelos preditivos. Além disso, procurou-se entender quais variáveis foram mais úteis para os modelos gerados, na tentativa de alcançar uma melhor compreensão acerca das características das eleições. Utilizando os dados coletados, os algoritmos de aprendizagem de máquina produziram modelos que foram avaliados e tiveram seus atributos mais importantes selecionados. Dessa forma, novos modelos foram obtidos a partir das características escolhidas e eles também foram examinados. Após observação dos resultados, o modelo *random forest* apresentou melhores desempenhos com relação à métrica definida e, em contrapartida, a seleção das variáveis mais importantes para a construção dos modelos não produziu melhores resultados.

PALAVRAS-CHAVE

Aprendizagem de máquina, predição, eleições, dados desbalanceados

1 Introdução

O processo eleitoral representa um dos traços mais significativos da democracia brasileira. Por meio das eleições serão escolhidas as pessoas que representarão os cidadãos pelos próximos anos. Os políticos irão trabalhar em prol da sociedade, regulando o estado através da criação de leis e fiscalização de outros poderes. Esse sistema de escolha exige que os eleitores examinem os candidatos, suas propostas e as alianças a que estão integrados. Além disso, é importante analisar o panorama político no qual a sociedade está inserida desde o período pré eleitoral.

Observa-se que o cenário apresentado é composto por vários fatores, constatando que essa é uma escolha complexa para todos os votantes e candidatos. Isso se torna mais problemático quando se percebe que mudanças são frequentemente incorporadas a esse contexto: proibições de candidaturas avulsas (BRASIL, 2017a), alterações relativas ao financiamento eleitoral por meio da criação de um Fundo Especial de Financiamento de Campanha (FEFC) (BRASIL, 2017b), além do fim de coligações e a aprovação da cláusula de desempenho (BRASIL, 2017c). Esses são alguns exemplos de modificações nas diretrizes eleitorais realizadas nos últimos anos e que também influenciarão as próximas eleições. Tais circunstâncias podem fazer com que esse processo seja ainda menos compreensível para a população e candidatos. Isso pode abrir espaço para uma diminuição da participação da sociedade na conjuntura política, podendo acontecer em decorrência das dificuldades apontadas anteriormente.

Outro aspecto que corrobora com essa situação de desarranjo verifica-se por meio da dificuldade

de obtenção e compreensão dos dados públicos em nosso país. Apesar de órgãos como o Tribunal Superior Eleitoral (TSE) disponibilizarem de forma facilitada dados de campanhas eleitorais dos últimos anos, isso não garante que o consumo dessas informações será útil para a população. Do mesmo modo, é importante frisar que cada novo pleito eleitoral resultará na obtenção de mais dados e a interpretação desse conteúdo não é algo trivial. Atingir um melhor entendimento dessas informações pode promover benefícios para a sociedade.

Diante do exposto, constata-se a necessidade de alcançar soluções que contribuam para uma melhor compreensão acerca do contexto político brasileiro. Um método que permita investigar e filtrar esses atributos utilizando os dados de eleições anteriores é algo que faz sentido, em decorrência do grande volume de dados já produzido. A disposição desses materiais e o uso dessa metodologia pode resultar em uma síntese de informações relevantes para a população, que pode aproveitá-las em prol de uma sociedade mais participativa dentro do processo eleitoral. Além disso, candidatos e partidos políticos de menor expressão também podem ser beneficiados, visto que isso pode permitir uma melhor condução dos comitês organizadores de campanhas, direcionando recursos para os aspectos mais relevantes.

Portanto, a ciência da computação pode ser utilizada para auxiliar nesse problema, produzindo resultados que gerem os benefícios esperados. A grande quantidade de dados disponibilizados é algo que tem sido muito explorado pelo campo da aprendizagem de máquina (*machine learning*), que pode manipular grandes volumes de informações para a realização de predições e análise de tendências presentes nos dados.

Assim, o presente trabalho utilizou os dados das eleições brasileiras para a avaliação de técnicas de aprendizagem de máquina e obtenção das variáveis mais importantes para a eleição de um candidato. Os dados viabilizados pelo TSE referentes aos pleitos de 2006, 2010, 2014 e 2018 para cargos de deputado federal foram empregadas para a construção de modelos preditivos que utilizaram essas informações e as compreensões obtidas acerca de quais variáveis foram mais relevantes para as eleições e com isso, a aplicação dessas técnicas que permitiram inferir resultados sobre as eleições futuras.

Observou-se ainda quais variáveis possuem maior importância em cada modelo. Dessa maneira, foram obtidas combinações de características para que novos modelos fossem gerados a partir das mesmas. Ao final, os novos modelos foram testados com os dados de 2018, avaliando quais deles obtiveram melhores resultados. Então, foi possível identificar a melhor combinação entre conjunto de variáveis e modelo. Como resultado, verificou-se que o algoritmo *random forest* é capaz de gerar modelos que têm melhor desempenho quando lidam com dados desbalanceados, apresentando um F-Score de 0,55842. Além disso, foi constatado que a utilização de variáveis de maior importância para a produção de modelos não forneceu bons resultados.

A seção 2 inclui detalhes da metodologia proposta, aplicada para a obtenção dos resultados descritos na seção 3. Ao final, foram expostas as conclusões e a discussão de trabalhos futuros.

2 Metodologia

O trabalho consiste na realização de experimentos que avaliam a qualidade de três técnicas de *machine learning* e a descoberta de quais variáveis são mais significativas para a predição do êxito de uma candidatura. Isso foi feito por meio dos dados das últimas eleições para elaborar modelos de classificação que tentaram predizer os resultados do pleito de 2018. Essa seção aborda as etapas necessárias para a realização dos experimentos e das avaliações, bem como as decisões relevantes.

2.1 Base de dados e variáveis

Para a construção dos modelos, fez-se necessário adquirir os dados e definir o escopo que seria analisado. Tendo isso em vista, a escolha de unidade de um candidato foi definida como a de um concorrente ao cargo de deputado federal. O critério de escolha ocorreu de modo a simplificar a estruturação e avaliação do que foi produzido por cada técnica, visto que a eleição para essa função acontece em um turno único - diferente de outros cargos como presidente e governador, que poderia implicar na construção de modelos para o segundo turno, dificultando a avaliação.

2.1.1 Obtenção dos dados

Os dados foram coletados no Repositório de Dados Eleitorais do TSE (TSE, 2019). As informações adquiridas correspondem às eleições de 2006, 2010, 2014 e 2018 e foram separadas em três seções principais: *candidatos*, *resultados* e *prestação de contas* eleitorais. A opção por uma amostra contendo apenas esses anos eleitorais se deu em decorrência da impossibilidade em se obter as informações anteriores ao ano de 2006. A seção referente aos *candidatos* contém dados sobre as candidaturas de cada um. Os *resultados* correspondem ao que foi obtido após o processamento dos votos em cada seção eleitoral. A seção de *prestação de contas* diz respeito a receitas e despesas de campanhas, declaradas por parte de candidatos, de partidos e de comitês.

O passo seguinte foi a realização de um trabalho de limpeza desses dados, onde cada uma das seções foi processada por meio da filtragem e agrupamento desses conteúdos. Isso foi feito através do RStudio, um software gratuito e de código aberto que provém uma interface de desenvolvimento integrado (*IDE*) para a linguagem de programação R, bastante popular para a manipulação de dados e cálculos estatísticos. Assim, cada uma das partes pode ser combinada resultando em um conjunto útil de informações.

2.1.2 Descrição dos dados

A etapa anterior produziu um conjunto de dados de 19 variáveis e 16.834 registros. Entre as variáveis selecionadas, 13 são numéricas e 6 com valores categóricos. A Figura 1 a seguir, mostra a matriz de correlação das variáveis numéricas utilizada para a construção dos modelos.

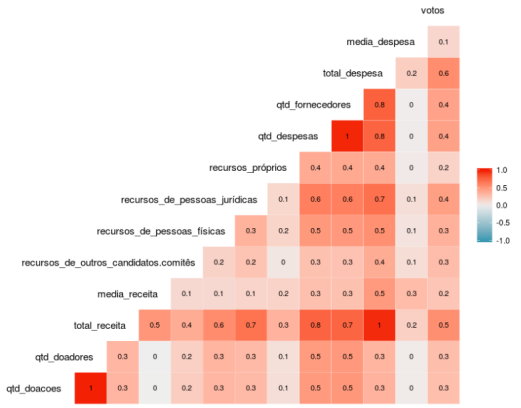

Figura 1: Matriz de correlação das variáveis numéricas

Em particular, a matriz de correlação das variáveis numéricas sumariza os coeficientes de correlação entre todos os pares de variáveis que, por sua vez, indicam quão forte elas estão linearmente relacionadas. Pela imagem, percebe-se ainda que algumas variáveis apresentam forte correlação positiva, o que pode evidenciar que esses atributos dependem uns dos outros e são candidatos a não serem incluídos no conjunto utilizado para a construção dos modelos.

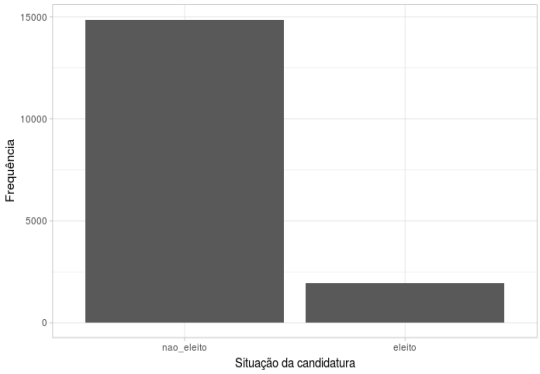

Figura 2: Distribuição da situação das candidaturas

A variável “situação_candidatura”, que indica se o candidato foi eleito ou não, foi considerada

como de maior interesse para a construção dos modelos preditivos. A Figura 2 mostra a distribuição dessa característica. É possível perceber que a classe referente à não eleição de um candidato é muito mais frequente dentro do conjunto de dados, o que caracteriza um desbalanceamento - algo já esperado, visto que existem muito mais candidatos do que vagas, o que gera esse efeito.

2.1.3 Transformação dos dados

Antes da elaboração dos modelos de classificação, os dados adquiridos ainda precisaram ser transformados novamente.

As colunas numéricas foram normalizadas e centralizadas, pois elas não estavam em uma mesma escala - assim, os conteúdos podem ser comparados de forma igualitária, com cada variável possuindo a mesma importância das demais. Além disso, as informações categóricas foram transformadas por meio da técnica de *one hot encoding*, que de maneira geral, transforma os valores dessas variáveis em colunas e atribuem 0 ou 1 para a linha correspondente. Isso facilita o trabalho dos algoritmos de predição. A Figura 03 ilustra melhor esse processo.

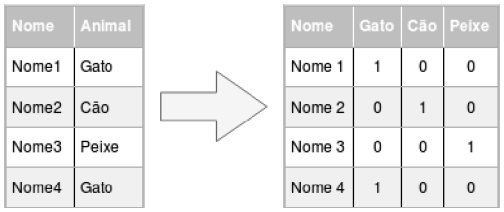

Figura 3: Resultado da transformação das informações categóricas

Com isso, variáveis como “estado_civil”, “genero”, “grau_instrucao” e “sigla_partido” foram transformadas, resultando na adição de 55 novas colunas.

Em relação ao desbalanceamento de classes, foi empregada a técnica SMOTE (CHAWLA et al., 2002), que é um método popular de reamostragem dos dados e utilizado para a correção da situação onde existem classes muito desproporcionais. Isso evita que os algoritmos adquiram um viés de predição para a classe de maior frequência. Esse mecanismo foi utilizado para os dados de treino, permitindo que os modelos que os realizassem previsões mais confiáveis. É importante ressaltar que os dados separados em validação e teste ainda apresentam o desbalanceamento, pois eles simulam os dados reais e serviram para avaliar os modelos.

Desta forma, os dados transformados foram organizados em dados de teste, validação e treino. Essa estrutura é empregada para a construção dos modelos (por meio dos dados de treino e validação) e avaliação dos mesmos (dados de teste). De modo geral, o conjunto de teste é composto pelos dados do ano de 2018, enquanto o conjunto de treino é formado pela combinação de 70% de cada um dos anos. Os conjuntos de validação correspondem aos 30% de 2006, 2010 e 2014, respectivamente.

2.2 Algoritmos de classificação

Os algoritmos de *machine learning* procuram aprender a partir dos dados fornecidos, sendo capazes de identificar padrões a partir dos dados. Com isso, eles utilizam novos dados para inferir resultados com base no aprendizado obtido anteriormente.

O termo classificação refere-se a uma classe de problema que procura identificar se um conjunto de observações pertence a determinado grupo. Nesse caso, a categoria estudada corresponde à classificação binária, uma vez que os modelos construídos levaram em conta apenas duas classes: se um candidato foi eleito ou não.

Com isso, foram escolhidos três algoritmos geralmente utilizados para problemas de classificação. Cada um deles empregou o conjunto de dados de treino definido anteriormente para a construção dos modelos preditivos, que foram examinados em duas etapas. Inicialmente observou-se a performance dos modelos em relação aos dados de validação, de modo a entender a capacidade de generalização de cada um com relação às eleições dos anos de 2006, 2010 e 2014.

Os modelos foram elaborados por meio da utilização do pacote Caret 6.0.84 (KUHN, 2019a), que fornece um conjunto de funções para a criação de modelos preditivos.

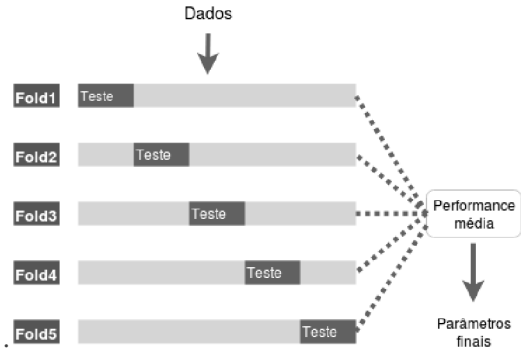

Figura 4: Representação da validação cruzada

A estruturação dos classificadores foi feita com o mesmo conjunto de treino e todos aplicam validação cruzada, que é uma técnica de validação que particiona os dados de treino em subconjuntos mutuamente exclusivos para a obtenção de parâmetros mais precisos e genéricos. A Figura 4 apresenta uma representação visual da estratégia. Todos os modelos adotaram essa estratégia, de modo a segmentar os dados de treino em 5 partes com o objetivo de obter valores que maximizam a métrica de avaliação escolhida.

2.2.1 Métricas de avaliação

Durante a etapa de construção dos modelos, os dados de treino são fornecidos para os algoritmos, eles aprendem os padrões inerentes a esse conjunto e, após isso, se um novo valor é fornecido ele será capaz de classificar a entrada para uma das classes de saída. Entretanto, o mapeamento nem sempre é correto. O modelo pode classificar um valor para uma classe que não corresponde à categoria que ela foi rotulada. Analogamente, isso acontece quando um e-mail é rotulado como spam mas na realidade ele tem um conteúdo importante e deveria estar na caixa de entrada. Dessa maneira, são necessárias métricas bem definidas para avaliar os modelos em relação aos dados que eles estão tentando predizer.

Com isso em mente, verifica-se que classificadores binários podem produzir resultados relativos a quatro categorias: verdadeiro positivo (VP), falso positivo (FP), verdadeiro negativo (VN) e falso negativo (FN). A Figura 5 resume o exposto:

		Real	
		Positivo	Negativo
Predito	Positivo	VP	FP
	Negativo	FN	VN

Figura 5: Matriz de confusão dos dados reais e previstos

A eleição de um candidato é definida como sendo a classe positiva, enquanto a não eleição corresponde à negativa. Nesse sentido, pretende-se maximizar o número de VP que o modelo produz, não havendo grande relevância em relação à quantidade de VN que ele gerar.

Uma métrica muito comum para avaliar modelos em relação aos valores obtidos é a acurácia, que corresponde à fórmula:

$$Acurácia = \frac{VP + VN}{VP + FN + FP + VN}$$

Entretanto, os dados de validação e teste apresentam uma maior frequência de candidatos não eleitos e, por isso, os valores preditos da classe negativa tendem ser mais frequentes também - gerando alta acurácia e provocando a impressão de que o modelo tem boa performance.

Métricas como o *F-score* são alternativas para contornar o problema da acurácia com os dados desbalanceados (ALI A.; SHAMSUDDIN S. M.; RALESCU A. L., 2015). Ela é utilizada como forma de balancear outras duas medidas: precisão e *recall*, que buscam minimizar os custos de detecção de FP e de FN, respectivamente (HE, H.; GARCIA E. A., 2009). O custo um FP é alto porque o modelo indica uma falsa impressão de que o candidato pode se eleger, enquanto que o custo de um FN gera a

ideia de que o candidato não é capaz de se eleger mesmo estando numa condição já favorável. *F-score* é definido pela fórmula:

$$F\text{-}Score = \frac{(1+\beta)^2 \times precisão \times recall}{\beta^2 \times precisão + recall} ,$$

em que,

$$precisão = \frac{VP}{VP + FP} ,$$

$$recall = \frac{VP}{VP + FN} ,$$

onde β é o coeficiente que ajusta a importância relativa entre precisão e *recall*. Como as duas medidas tem a mesma importância, o coeficiente é definido por $\beta = 1$. Caso a precisão e o recall sejam iguais a 1, o F-score também será igual a 1, indicando que o modelo analisado foi muito preciso, realizando muitas classificações corretas.
Com isso, utilizou-se a métrica F-score para minimizar os custos associados a falsos positivos e falsos negativos. Dessa maneira, ela foi empregada pelos algoritmos de modo a otimizar esse valor, servindo também de medida para comparação entre os resultados que foram produzidos pelos modelos.

2.2.2 Regressão Logística

É um tipo de regressão que tem como objetivo utilizar os dados para encontrar parâmetros para as variáveis independentes de uma função linear, onde essa combinação possui o menor erro. Por vezes, o termo função denota uma linha que, após a identificação dos valores ideais, melhor representa os dados e mais se ajusta a eles. A variável dependente, que corresponde à saída do modelo, diz respeito a uma variável categórica, verificando-se como uma opção viável para a classificação binária.

De modo geral, a função definida de acordo com os parâmetros obtidos será utilizada pela função logística, que mapeia a função original para valores no intervalo entre 0 e 1. A função logística é definida pela fórmula:

$$f(x) = \frac{1}{1+e^{-x}} ,$$

onde x é função linear que foi obtida.

Com isso, ao tentar realizar a classificação de um novo dado, os valores produzidos pela função logística serão comparados com um limiar (geralmente estabelecido como 0,5): resultados acima do limiar são classificados como pertencentes à classe positiva, enquanto os abaixo concernem à classe negativa. O modelo utilizado fez uso do limiar igual a 0,5.

2.2.3 Random Forest

Essa técnica está ligada à ideia de árvores de decisão. Nela os dados são utilizados para a construção de nós de decisão, onde um atributo será testado e dividido em categorias, que representam o resultado do teste. Isso será repetido para cada nó de decisão até que todos os atributos sejam avaliados - resultando em um conjunto de folhas que caracterizam uma das categorias da variável alvo. Assim, se faz possível a utilização de árvores de decisão para resolução de problemas de classificação. A Figura 6 mostra um exemplo de árvore de decisão.

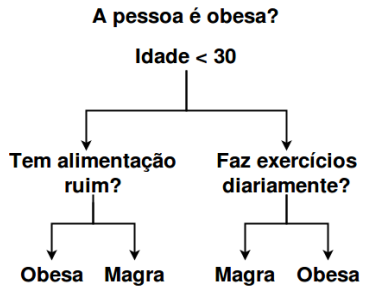

Figura 6: Exemplo de árvore de decisão

Dessa maneira, *random forest* é uma estratégia que combina diversas árvores de decisão, onde em cada uma delas as variáveis são avaliadas de forma aleatória. Com isso os resultados parciais de cada uma das árvores são combinados e o resultado da classificação consiste na seleção da classe de maior frequência.

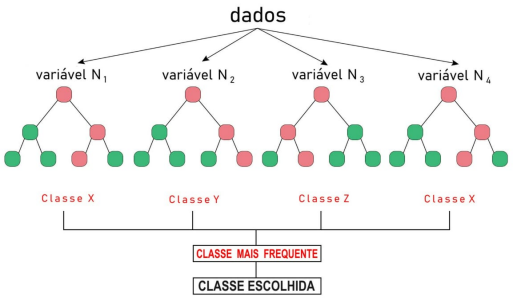

Figura 7: Exemplo de execução do *random forest* para realização de classificação

A Figura 7 exibe como ocorre o processo de classificação do algoritmo. É possível ver como são criadas diversas árvores em que os nós de decisão são distintos. Ao final, a classe mais frequente é selecionada como saída.

O modelo construído fez uso de apenas 2 variáveis para serem divididas em cada nó das árvores de decisão.

2.2.4 Redes Neurais

Esse método consiste em modelos que simulam neurônios e sinapses interligadas, sendo possível a identificação de padrões a partir dos dados. As redes neurais consistem em nós, que são dispostos em camadas e podem ser ativados ou não; e conexões, que são feitas entre as camadas de nós e servem para a ativação de nós.

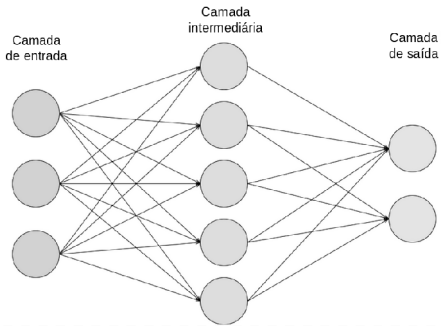

Figura 8: Representação de uma rede neural

A Figura 8 mostra a representação das camadas dentro das etapas de aprendizagem. A primeira camada contém as informações que precisam ser aprendidas, ou seja, os dados de entrada. Os níveis seguintes correspondem a camadas internas e a camada de saída, onde os nós intermediários são ativados de acordo com o que é aprendido na camada anterior, enquanto a camada final refere-se ao resultado final.

O funcionamento da rede ocorre por meio da ativação dos nós e ajuste dos pesos de cada conexão entre eles, que pode ter seu valor aumentado ou diminuído de acordo com a precisão das predições. Dessa forma é possível fornecer os dados de entrada e a rede aprende a partir deles e pode realizar classificações.

A rede utilizada foi uma rede artificial que fez uso de apenas uma camada intermediária contendo 5 nós, cuja finalidade é identificar as características dos dados.

2.3 Seleção de variáveis

A triagem das variáveis de maior importância foi necessária para analisar a relevância de cada atributo na construção do modelo e entender o impacto das variáveis na eleição dos candidatos.

Para cada um dos modelos definidos, o cálculo é realizado de maneira distinta, através da função *varImp* disponibilizada pelo Caret (KUHN, 2019b).

- A regressão logística, por ser um modelo linear, tem a importância das variáveis calculadas pela estatística de teste, normalmente utilizada no teste t de Student em testes de hipótese.
- Para o *random forest*, a contabilidade das importâncias é feita por meio de uma média ponderada das acurácias de todas as árvores construídas. Ao final, o valor obtido é normalizado.
- O cálculo definido para as variáveis do modelo de rede neural é definido por uma combinação dos valores absolutos dos pesos das conexões.

Assim, os algoritmos foram utilizados para a construção de novos modelos a partir do conjunto de variáveis de maior importância. Para isso, foram selecionadas as 15 variáveis mais significativas, de modo a facilitar a interpretação dos modelos produzidos.

3 Resultados

A partir dos dados coletados e utilização do R e do Caret para a aplicação dos métodos e técnicas apresentados, foi possível obter modelos resultantes de cada um dos algoritmos selecionados - com as ferramentas escolhidas facilitando a avaliação e construção dos mesmos. Inicialmente eles foram avaliados com os dados das eleições anteriores a 2018, de modo a observar a qualidade desses algoritmos e a capacidade de generalização dos modelos produzidos. A Tabela 1 resume os resultados do modelo de regressão logística, enquanto as Tabelas 2 e 3 apresentam os resultados para os modelos de *random forest* e rede neural, respectivamente.

Tabela 1: Resultados do modelo utilizando regressão logística

Ano	VP	FP	VN	FN	Precisão	Recall	F-score
2006	97	117	752	50	0,45327	0,65986	0,5374
2010	118	104	924	35	0,53153	0,77124	0,62933
2014	108	99	1229	44	0,52174	0,71053	0,60167
Média					0,50218	0,71387	0,58946

Tabela 2: Resultados do modelo utilizando *random forest*

Ano	VP	FP	VN	FN	Precisão	Recall	F-score
2006	101	54	815	46	0,65161	0,68707	0,66887
2010	125	98	930	28	0,5605	0,8170	0,6649
2014	133	102	1226	19	0,6649	0,8750	0,68734
Média					0,62567	0,79302	0,67370

Tabela 3: Resultados do modelo utilizando rede neural

Ano	VP	FP	VN	FN	Precisão	Recall	F-score
2006	115	84	785	32	0,5779	0,7823	0,6647
2010	131	124	904	22	0,5137	0,8562	0,6422
2014	125	116	1212	27	0,51867	0,82237	0,63613
Média					0,53675	0,82029	0,64767

Analisando os resultados, é possível observar que os valores do F-score dos três modelos estão entre 0,53 e 0,68, o que indica um desempenho razoável se considerarmos a métrica escolhida - demonstrando também uma capacidade satisfatória para o entendimento das eleições de 2006 a 2014.

Em relação à Tabela 2, é possível sintetizar um F-score médio de 0,67 para o modelo que empregou *random forest*, sendo o maior em comparação à média da mesma medida para as outras tabelas, indicando que ele foi o modelo com o melhor desempenho. De mesmo modo, por meio da Tabela 3 percebe-se que o modelo de rede neural foi o que produziu a maior quantidade de VP, passando uma impressão positiva sobre seu desempenho. Entretanto, a mesma tabela também mostra que o modelo possui a maior quantidade de FP, evidenciando que ele é pouco confiável visto o grande número de predições incorretas.

Analogamente, a grande frequência de FN observada na Tabela 1 indica que o modelo de regressão perde credibilidade por predizer muitos resultados como positivos e na realidade não são. O reflexo disso é um F-score médio mais baixo, que é observado como o menor entre os três modelos. Dessa maneira, salienta-se a importância da métrica escolhida, que procura encontrar um equilíbrio entre FP e FN, em decorrência do custo associado à ocorrência dessas duas medidas. Nesse sentido, verifica-se que o *random forest* possuiu o melhor desempenho médio entre os três modelos.

Na etapa seguinte, foi selecionado o conjunto de variáveis com maior importância para cada modelo e novas versões foram construídas considerando apenas as variáveis escolhidas. Elas foram escolhidas por meio da função varImp do Caret. Feito isso, os novos modelos foram testados com os dados de 2018, que simula dados reais e permite analisar a capacidade preditiva de cada um.

Tabela 4: Resultados dos novos modelos para os dados de teste

Modelo	VP	FP	VN	FN	Precisão	Recall	F-score
Regressão Logística	352	483	3637	96	0,42156	0,78571	0,54871
Random Forest	368	502	3618	80	0,42299	0,82143	0,55842
Rede Neural	412	752	3368	36	0,35395	0,91964	0,51117

A tabela 4 expõe os resultados dos novos modelos, que foram testados por meio dos dados das eleições de 2018. Como consequência da seleção das informações, os algoritmos precisaram aprender os padrões dos dados com apenas 15 variáveis entre as 72 utilizadas pelos modelos originais.

A partir da triagem realizada, esperava-se que os modelos atualizados tivessem um melhor desempenho, visto que as variáveis pouco significativas foram descartadas e os algoritmos aprenderiam os padrões a partir das informações mais úteis. No entanto, de acordo com a Tabela 4, verifica-se que o resultado obtido foi pior em relação às médias dos *F-scores* dos modelos anteriores - o que pode indicar que na prática os modelos não apresentam desempenhos muito bons. Ainda assim, é possível observar que o modelo produzido por meio do *random forest* obteve o melhor resultado para esse conjunto de variáveis, evidenciando sua boa capacidade preditiva para classificação de dados desbalanceados - caracterizando-se como uma opção viável para eleições futuras.

Mediante o exposto, a queda de desempenho constatada pode ser um indicativo de que a redução na quantidade de dados é prejudicial para o treinamento de modelos. Além disso, também foi possível verificar que o método utilizado para a seleção de variáveis foi pouco eficaz. Dessa forma, os modelos resultantes podem apresentar desempenhos abaixo do esperado.

4 Conclusões

A aplicação de algoritmos de *machine learning* para a construção de modelos preditivos possibilita a resolução de problemas em diferentes cenários, entre eles a predição da eleição de candidatos. Os resultados produzidos mostram que os modelos selecionados apresentam performances razoáveis, de acordo com a métrica definida. Com isso, é possível perceber quão bem cada modelo é capaz de compreender e generalizar os dados das eleições brasileiras. O algoritmo *random forest* gerou modelos que obtiveram os melhores desempenhos, demonstrando que essa estratégia produz resultados satisfatórios dentro de um cenário eleitoral. Ainda verificou-se que a seleção das variáveis importantes para um modelo e a consequente utilização para o desenvolvimento de novos modelos não produziu melhores modelos.

Como sugestão para trabalhos futuros, outros modelos podem ser testados dentro do escopo deste trabalho. Além disso, desenvolver os modelos de forma mais específica, procurando os parâmetros ideais para cada um deles, é uma opção que pode gerar melhores resultados. Com relação à seleção das variáveis mais significativas, outras abordagens também podem ser testadas. Uma opção é a utilização do LIME (RIBEIRO, M.T.; SINGH S.; GUESTRIN C., 2016), uma técnica conhecida que tenta interpretar os preditores de um dado modelo, analisando seus componentes e a interação entre eles.

REFERÊNCIAS

ALI, A.; SHAMSUDDIN, S. M.; RALESCU, A. L. **Classification with class imbalance problem**: **A Review**. v 7, p. 176–204, 2015. Disponível em https://pdfs.semanticscholar.org/1e48/70524f8de44d4f18c8f9f80eb797dfd25c89.pdf. Acesso em 18 de jun. de 2019.

BRASIL. Lei nº 13.488, de 06 de outubro de 2017. Altera as Leis n º 9.504, de 30 de setembro de 1997 (Lei das Eleições), 9.096, de 19 de setembro de 1995, e 4.737, de 15 de julho de 1965 (Código Eleitoral), e revoga dispositivos da Lei nº 13.165, de 29 de setembro de 2015 (Minirreforma Eleitoral de 2015), com o fim de promover reforma no ordenamento político-eleitoral. **Diário Oficial da União**, Brasília, DF, 06 out. 2017. p. 1. Disponível em http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2017/lei/L13488.htm. Acesso em 18 jun. 2019.

BRASIL. Lei nº 13.487, de 06 de outubro de 2017. Altera as Leis n º 9.504, de 30 de setembro de 1997, e 9.096, de 19 de setembro de 1995, para instituir o Fundo Especial de Financiamento de Campanha (FEFC) e extinguir a propaganda partidária no rádio e na televisão. **Diário Oficial da União**, Brasília, DF, 06 out. 2017. p. 1. Disponível em http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2017/Lei/L13487.htm. Acesso em 19 jun. 2019.

BRASIL. Emenda Constitucional nº 97, de 04 de outubro de 2017. Altera a Constituição Federal para vedar as coligações partidárias nas eleições proporcionais, estabelecer normas sobre acesso dos partidos políticos aos recursos do fundo partidário e ao tempo de propaganda gratuito no rádio e na televisão e dispor sobre regras de transição. **Diário Oficial da União**, Brasília, DF, 05 out. 2017. p. 1. Disponível em https://legis.senado.leg.br/norma/26247394/publicacao/26247403. Acesso em: 19 jun. 2019.

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. Disponível em https://arxiv.org/pdf/1106.1813.pdf. Acesso em 18 de jun. de 2019.

KUHN, M. **The Caret Package: Introduction**, 2019a. Disponível em http://topepo.github.io/caret/index.html. Acesso em 18 de jun. de 2019.

KUHN, M. **The Caret Package: Variable Importance**, 2019b. Disponível em https://topepo.github.io/caret/variable-importance.html . Acesso em 19 de jun. de 2019.

HE, H.; GARCIA, E. Learning from Imbalanced Data. **IEEE Transactions on Knowledge and Data Engineering**, v. 21, n. 9, p. 1263–1284, 2009. Disponível em https://www.ele.uri.edu/faculty/he/PDFfiles/ImbalancedLearning.pdf. Acesso em 19 jun. 2019.

Caret, 2019. Disponível em https://www.rdocumentation.org/packages/caret/versions/6.0-84/topics/varImp. Acesso em 19 de jun. de 2019.

TRIBUNAL SUPERIOR ELEITORAL, TSE: **Repositório de dados eleitorais**, 2019, Disponível em http://www.tse.jus.br/eleicoes/estatisticas/repositorio-de-dados-eleitorais-1/repositorio-de-dados-eleitorais. Acesso em 18 de jun. de 2019.

RIBEIRO, M.T.; SINGH S.; GUESTRIN C. **"Why Should I Trust You?": Explaining the Predictions of Any Classifier**, 2016. DIsponível em https://arxiv.org/pdf/1602.04938.pdf. Acesso em 19 de jun. de 2019