Eleições Presidenciais de 2022: Predições com Dados de Pesquisas Eleitorais do Poder360 e *PollingData*

Gabriel P. Valentim[1], **Danilo B. Seufitelli**[1], **Carlos A. Silva**[1]

[1]Departamento de Informática – Instituto Federal de Minas Gerais (IFMG)
CEP 34.590-390 – Sabará, MG – Brasil

cerogabgio@gmail.com, {danilo.boechat,carlos.silva}@ifmg.edu.br

Abstract. *Election polls are a powerful tool to capture the population's aspirations. Hence, we explore the potential of combining electoral polls with prediction algorithms for the 2022 Brazilian presidential election. Specifically, we aim to determine whether integrating machine learning techniques can yield superior results compared to traditional electoral polls alone. Our results show that prediction models exhibited promising performance, outperforming the research institutes' results, particularly in the second round. This approach unveils a promising option for predicting future elections, shedding light on forecasting electoral outcomes.*

Resumo. *As pesquisas eleitorais são uma ferramenta poderosa para captar as aspirações da população. Portanto, exploramos o potencial de combinar pesquisas eleitorais com algoritmos de previsão para a eleição presidencial brasileira de 2022. Especificamente, foi preterido determinar se a integração de técnicas de aprendizado de máquina pode produzir resultados superiores em comparação com as pesquisas eleitorais tradicionais de forma isolada. Nossos resultados mostram que os modelos de previsão exibiram desempenho promissor, superando os resultados dos institutos de pesquisa, principalmente no segundo turno. Essa abordagem revela uma opção promissora para prever futuras eleições, lançando luz sobre a previsão de resultados eleitorais.*

1. Introdução

A escolha popular de representantes políticos configura-se como um dos principais processos da democracia. No entanto, tal escolha não se limita apenas a selecionar o representante da população junto ao poder. Afinal, com essa decisão, devemos incluir as diversas estratégias de condução do governo eleito para lidar com os rumos da economia, do emprego, da saúde e da educação, que afetará a vida das pessoas durante o período do mandato. Por isso, cada cidadão deve analisar com cautela os seus candidatos, e então se decidir considerando que as consequências de suas escolhas (nas urnas) exercerão uma forte influência sobre os rumos políticos nos anos vindouros. Desta forma, é preciso considerar a importância do que se consome em termos de conteúdos pelos eleitores, visto que as informações podem influenciar na definição do voto em um determinado candidato. Na atual era tecnológica e conectada, é notório o poder que as mídias sociais têm tido sobre as eleições. Trabalhos como [Sani and Azizuddin 2014], [Aiyappa et al. 2023] e [Hagemann and Abramova 2022] buscam mostrar a grande influência exercida por tais ferramentas nos processos eleitorais.

No Brasil, a cada dois anos, para diferentes cargos, a população precisa ir às urnas para escolher seus representantes do próximo ciclo para que eles possam tomar as importantes decisões dos rumos que serão seguidos. Similar ao que ocorre no cenário global, as redes sociais têm causado uma grande mudança no cenário eleitoral brasileiro, conforme demonstra o trabalho de [de Sousa 2021], que ressalta como a utilização e a interação através da internet impactou o resultado final da eleição de 2018.

No período eleitoral ocorrem as pesquisas eleitorais, que são as indagações feitas aos eleitores em um determinado momento, sobre a sua opção a respeito dos candidatos que concorrem em uma eleição.[1] Tais pesquisas são cruciais no período pré-eleitoral brasileiro, pois subsidiam as preferências da população, e consequentemente (assim como as redes sociais e a internet) também podem influenciar o voto de um indivíduo e impactar o resultado final das eleições [Venturi 1995]. Afinal, os resultados das pesquisas podem ser utilizados em aplicações como a identificação e planejamento de estratégias por parte de partidos políticos e candidatos para ampliar sua base eleitoral. No entanto, a confiança das pesquisas tem sido questionada, devido às divergências nos resultados divulgados pelos institutos.[2] Por sua vez, os institutos têm reconhecido seus erros e praticado novas abordagens para coletar dados e explorado outros métodos visando resultados mais precisos.[3]

Diante do cenário onde as pesquisas eleitorais se mostram como uma importante ferramenta de influência do voto individual, detectamos que a literatura não explora o potencial dos algoritmos de aprendizado de máquina ao combinar resultados de pesquisas eleitorais de votações anteriores para melhorar o resultado das predições das eleições presidenciais. Semelhantemente à pesquisa realizada por [Silva 2018] para as eleições presidenciais brasileiras de 2014, neste trabalho levantamos a seguinte questão de pesquisa (QP):

> ***QP - Os algoritmos de aprendizado de máquina somados às pesquisas eleitorais podem prever melhores resultados das eleições presidenciais?***

Dessa forma, este trabalho propõe uma abordagem que combina os resultados das pesquisas eleitorais com técnicas de aprendizado de máquina, com o objetivo de obter resultados mais próximos do que os alcançados pelos institutos para as eleições presidenciais brasileiras de 2022. O objetivo é empregar uma abordagem de previsão distinta das utilizadas em trabalhos prévios, fundamentando-se na análise de eleições passadas para o treinamento dos modelos preditivos.

O restante deste artigo está organizado da seguinte forma. Inicialmente discutimos os trabalhos que abordam o problema da predição das eleições e o uso de pesquisas eleitorais na Seção 2. Em seguida, descrevemos a metodologia para a obtenção e o tratamento de dados, obtidos do portal Poder360 e do portal *PollingData*, bem como exploramos os

[1]TSE - O que é pesquisa eleitoral? Disponível em: `https://www.tse.jus.br/comunicacao/noticias/2019/Maio/o-que-e-pesquisa-eleitoral-o-glossario-do-tse-responde`. Acesso em 17/08/2023.

[2]Resultados das urnas divergem das pesquisas eleitorais em 21 estados e no DF. Disponível em: `https://bit.ly/resultado-pesquisas`. Acesso em 11/07/2023.

[3]Em Busca dos Prováveis Eleitores como Funcionam os Modelos para Medir Abstenção nas Pesquisas. Disponível em: `https://bit.ly/eleitores-pesquisas`. Acesso em 11/07/2023.

algoritmos utilizados para a tarefa de predição na Seção 3. Posteriormente, apresentamos os resultados das análises na Seção 4. As dificuldades encontradas durante o desenvolvimento são relatadas na Seção 5. Discutimos as principais conclusões na Seção 6. Por fim, realizamos a recomendações para trabalhos futuros, na Seção 7.

2. Revisão Bibliográfica

Vários institutos de pesquisa eleitoral têm enfrentado desafios em prever resultados precisos nas eleições presidenciais. Isso pode ser atribuído a problemas como a inadequada representação da população ou à relutância de alguns entrevistados em admitir apoio a figuras controversas, como apontado por [Zhou et al. 2021]. Para superar essas dificuldades, foi proposta uma abordagem baseada nas eleições presidenciais argentinas de 2019, empregando técnicas de aprendizado de máquina e análise de *big data* das redes sociais. Eles coletaram *tweets* relevantes por meio de consultas específicas, incluindo os nomes dos candidatos. Cinco modelos de classificação foram aplicados: *Logistic Regression, Support Vector Machine, Naive Bayes, Random Forest* e *Decision Tree*. O modelo *Logistic Regression* demonstrou o melhor desempenho na classificação de *tweets* relacionados aos candidatos daquela eleição. Os autores obtiveram resultados mais precisos do que os tradicionalmente alcançados por institutos convencionais de pesquisa.

Novas abordagens de previsão de resultados eleitorais têm sido propostas, como evidenciado por [Brito and Adeodato 2023]. Nesse estudo, a análise concentrou-se nas eleições presidenciais em países da América Latina (como a Argentina em 2019, Brasil, Colômbia e México em 2018). A atenção que os candidatos presidenciais receberam nas redes sociais foi examinada a partir de *posts* e interações coletadas dos perfis oficiais dos candidatos em três plataformas (*Facebook, Twitter* e *Instagram*). Além disso, dados de pesquisas eleitorais tradicionais também foram coletados e usad +os como referência para comparar os resultados obtidos pela nova metodologia. As métricas extraídas dos *posts*, como total de *posts*, curtidas, compartilhamentos e comentários, foram empregadas nos métodos MLP-BP e GRNN, junto com a utilização da Regressão Linear. A comparação entre os resultados obtidos pelo método proposto e os resultados das pesquisas eleitorais prévias revelou que os valores obtidos se assemelharam aos das pesquisas. Em alguns países, os números até se aproximaram mais do resultado final das eleições.

Em uma linha de investigação semelhante, [Tsakalidis et al. 2015] empregou dados coletados do *Twitter* e informações de pesquisas eleitorais para prever os resultados de eleições na Grécia, Alemanha e Holanda. O estudo combinou pesquisas eleitorais e *tweets*, incluindo uma análise de sentimento dos *tweets*. Três algoritmos distintos foram usados para previsão: Regressão Linear, *Gaussian Process* e *Sequential Minimal Optimization*. Os resultados excederam aqueles obtidos apenas por pesquisas eleitorais, plataformas de previsão e outros trabalhos anteriores individualmente, indicando que a combinação de abordagens traz resultados positivos.

Considerando o exposto, é evidente que os métodos tradicionais de análise não devem ser descartados, mas sim explorados de maneiras que se integrem à tecnologia, como também demonstrado por [Silva 2018]. Nesse estudo, o autor buscou reduzir o erro na estimativa da intenção de voto, combinando dados de pesquisas eleitorais de diferentes institutos para a eleição presidencial brasileira de 2014. Técnicas de aprendizado de máquina foram aplicadas, como *Local Regression, Random Forest* e *Support Vector*

Machine, resultando em melhorias notáveis em relação a cenários específicos.

Apesar da ênfase da literatura em utilizar dados recentes de pesquisas eleitorais em modelos preditivos, a exploração de dados históricos permanece ausente. Para abordar essa lacuna, o presente estudo visa investigar se a inclusão de dados de eleições presidenciais anteriores produzirá previsões mais precisas dos resultados eleitorais. Assim, este trabalho avança ao propor uma abordagem inovadora para a predição dos resultados eleitorais.

3. Metodologia

A metodologia deste trabalho é dividida em três etapas: aquisição e tratamento de dados da *web*, modelagem do problema, e análise preditiva, conforme descrito a seguir.

3.1. Coleta e Tratamento de Dados

Neste trabalho, foram utilizadas três bases de dados com informações sobre pesquisas eleitorais que possuem acesso público. Uma das bases é proveniente do portal Poder360, um jornal brasileiro independente que acompanha assuntos do poder e da política. Tal base está está no formato de um banco de dados relacional.[4] As outras duas bases de dados foram extraídas do portal *PollingData*, especialista em pesquisas de opinião pública e amostragem. Ambas as bases foram disponibilizadas no formato de arquivos csv.[5]

A primeira base de dados proveniente do Poder360, chamada *microdados*, possui informações de milhares de pesquisas eleitorais brasileiras, feitas no período de 2000 até 2022. Ela contém dados desde eleições municipais até nacionais realizadas pelos mais diversos institutos (tais como Data Folha, Paraná Pesquisas, Quaest, dentre outros). Esta base dispõe de 24 colunas que contém informações de pesquisas realizadas neste século – ano da eleição, nome do candidato, partido, nome do instituto que realizou a pesquisa, etc. Especificamente para a eleição presidencial de 2018, a última pesquisa registrada nesta base foi divulgada em dezembro de 2017. Portanto, outras bases de dados foram necessárias para ampliar a quantidade de pesquisas para o pleito de 2018.

Desta forma, integramos a segunda base de dados considerada neste trabalho (*PollingData*) intitulada de *2018-T1-Brasil-BR-Presidente*, que possui apenas oito colunas: *Data, Instituto, Ciro Gomes (PDT), Geraldo Alckmin (PSDB), Jair Bolsonaro (PSL), Fernando Haddad (PT), Nao Validos* e *Outros*. Esta base de dados contém apenas dados do primeiro turno das eleições presidenciais de 2018, limitada aos quatro candidatos mais bem posicionados na votação daquele ano. A terceira base de dados utilizada, nomeada de *2018-T2-Brasil-BR-President* possui as mesmas características que a anterior, no entanto ela se refere ao 2º turno daquele pleito. Nela há cinco colunas: *Data, Instituto, Jair Bolsonaro (PSL), Fernando Haddad (PT)* e *Nao Validos*.

Após a coleta das três bases de dados, o próximo passo foi realizar um tratamento para a junção e limpeza dos dados. Na base *microdados* foram excluídas as colunas que não eram necessárias para a realização das predições, permanecendo apenas *ano, cargo, data, sigla_uf, turno, tipo, nome_candidato, tipo_voto* e *percentual*. Candidatos

[4]Disponível em: `https://bit.ly/poder360-pesquisasEleitorais`. Acesso em 11/05/2023.

[5]Disponível em: `https://bit.ly/polling-dataset`. Acesso em 08/06/2023.

Tabela 1. Candidatos com pesquisas eleitorais relacionadas nos anos em que concorreram à presidência da república.

Ano	Candidatos considerados
2002	Lula, Serra, Anthony Garotinho e Ciro Gomes
2006	Lula, Geraldo Alckmin, Heloísa Helena, Cristovam Buarque, Ana Maria Rangel, Eymael e Luciano Bivar
2010	Dilma Rousseff, Serra, Marina, Plínio de Arruda Sampaio, Eymael, Zé Maria, Levy Fidelix, Ivan Pinheiro e Rui Costa
2014	Dilma Rouseff, Aécio Neves, Marina Silva, Luciana Genro, Pastor Everaldo, Eduardo Jorge, Levy Fidelix, José Maria, Eymael, Mauro Iasi e Rui Costa
2018	Jair Bolsonaro, Fernando Haddad, Ciro Gomes e Geraldo Alckmin

Fonte: elaborado pelos autores.

Tabela 2. Descrição dos metadados das pesquisas armazenadas no conjunto de dados *Dados Unidos*.

Coluna	Armazena
ano	Ano eleitoral referência
sigla_uf	Sigla do estado alvo (nulo quando nacional)
cargo	Cargo para o qual a pesquisa diz respeito
data	Data de publicação da pesquisa
tipo	Tipo da pesquisa
turno	Turno considerado
tipo_voto	Tipo do voto contabilizado
nome_candidato	Nome do candidato
percentual	Percentual que o candidato obteve na pesquisa
resultado	Resultado final que o candidato obteve na eleição

Fonte: elaborado pelos autores.

com pesquisas associadas foram identificados com nome e ano de disputa eleitoral. Isso foi feito para evitar confusão quanto aos resultados obtidos por um candidato em eleições diferentes. Por exemplo, o candidato Serra concorreu em 2002 e 2010. Desta forma, *Serra 2002* se refere às pesquisas associadas à eleição presidencial de 2002, enquanto *Serra 2010* se refere à eleição presidencial de 2010. Assim, evita-se problemas durante o treinamento dos algoritmos ao adicionar resultados obtidos em cada uma das votações.

Em seguida, filtramos a base de dados para considerar apenas as eleições nacionais (i.e., onde *sigla_uf* seja *null*). Além disso, selecionamos as pesquisas cujo *tipo* fosse *estimulada*, e *tipo_voto* fosse *votos totais*, que considera votos nulos e brancos. Para analisar os dois cenários das eleições (1º e 2º turno), separamos os dados em dois conjuntos. O primeiro com apenas pesquisas do 1º turno, e o segundo com pesquisas do 2º turno. Por fim, selecionamos os anos considerados no estudo (2002, 2006, 2010, 2014 e 2018), bem como os candidatos que possuíam pesquisas, conforme mostra a Tabela 1.

Incluímos uma coluna *resultado*, para armazenar o resultado final (em votos totais) que cada candidato obteve em cada eleição considerada. Para chegar aos valores em votos totais (pois o resultado é divulgado em votos válidos) foram realizados cálculos de todos os candidatos para cada turno das eleições presidenciais. Especificamente, considerou-se os votos válidos de cada eleição com o valor final recebido por cada candidato. O mesmo cálculo foi feito para os resultados após a predição das eleições de 2022.

Tabela 3. Exemplos dos metadados das pesquisas armazenadas no conjunto de dados *Dados Unidos*.

Coluna	Exemplo
ano	2022
sigla_uf	*AL*
cargo	*presidente*
data	*2022-10-01*
tipo	*estimulada*
turno	*1*
tipo_voto	votos totais
nome_candidato	*Ciro*
percentual	*7.8*
resultado	*11.37*

Fonte: elaborado pelos autores.

As outras duas bases de dados (eleições de 2018) foram tratadas e padronizadas para serem mescladas com a base *microdados*. Para isso, colunas que continham dados sobre cada candidato separadamente foram excluídas. Seus nomes foram armazenados na coluna *nome_candidato*, e seus respectivos valores foram para a coluna *percentual*. Desta forma, cada linha contém dados da pesquisa de um único candidato. Por fim, foram criadas novas colunas: *ano, cargo, sigla_uf, turno, tipo* e *tipo_voto*. Apenas a coluna *Data* permaneceu em seu estado original, e as demais colunas foram excluídas.

Após a remoção das colunas desnecessárias e junção das tabelas, a base de dados final foi construída. A Tabela 2 resume nosso conjunto de dados, e mostra todas as colunas, um resumo sobre o que cada coluna armazena, e, na Tabela 3, há um exemplo dos dados armazenados em cada coluna.

Especificamente para o 2º turno houve uma outra verificação, além das mencionadas. As pesquisas utilizadas para o treinamento dos algoritmos foram as divulgadas após a votação para o 1º turno, de forma que já havia uma definição oficial de quais candidatos estariam no embate final.

3.2. Modelagem do Problema

Nesta seção, modelamos nosso problema de predição das eleições presidenciais brasileiras de 2022 a partir de pesquisas eleitorais. Nosso objetivo é tentar obter resultados mais próximos do ocorrido nas urnas. Para isso, consideramos e modelamos o problema para os seguintes algoritmos de aprendizado de máquina: *KNeighbors Regressor*, *Linear Regression* e *Random Forest*. Tais algoritmos foram selecionados para a realização da presente pesquisa, por atenderem as especificidades do trabalho, adicionalmente à justificativa de escolha destes métodos, desejávamos explorar combinações de distintas técnicas utilizadas em trabalhos previamente desenvolvidos, com o intuito de averiguar a adequação dos resultados proveniente dos três algoritmos escolhidos. Eles são apresentados a seguir.

***KNeighbors Regressor* (KNR).** É uma regressão baseada em k vizinhos mais próximos. No caso da regressão, ao invés de atribuir classes, o KNR atribui valores numéricos às amostras. A previsão é feita calculando a média (ou mediana) dos valores das amostras vizinhas mais próximas. Aqui, o objetivo é prever o valor do resultado da eleição, utilizando as variáveis do nosso conjunto de dados. Neste estudo, foram realizados testes

empíricos para determinar o número ideal de vizinhos no KNR. Assim, foi observado que a quantidade ideal para este problema é igual a 3 ($k = 3$), já que ao utilizar $k = 4$, não houve diferença significativa nos resultados e quanto maior for o valor de k, mais lenta se torna a execução da predição.

***Linear Regression* (LR).** É um algoritmo de aprendizado supervisionado utilizado para modelar a relação entre variáveis independentes (as colunas de nosso *dataset*) e um valor de previsão (o resultado das eleições). Optamos por utilizar a regressão linear devido à sua capacidade de prever valores futuros com base em desempenhos passados, que é o objetivo deste estudo. Uma das vantagens da regressão linear é que não requer a configuração de parâmetros adicionais, simplificando o processo de implementação e uso.

***Random Forest* (RF).** É um algoritmo que constrói múltiplas Árvores de Decisão de forma aleatória e combina os resultados dessas árvores para chegar a um resultado final (i.e., a predição do resultado das eleições). Os ramos das árvores representam as condições que levam a diferentes valores previstos. A RF é capaz de prever valores na previsão de resultados em diversos cenários (tais como as eleições). RF requer a definição de um parâmetro: número de árvores utilizadas nas estimativas. Após testes empíricos, foi determinado que 10 é a quantidade ideal de árvores para este trabalho e, assim como no caso do KNR, uma quantidade maior não foi utilizada, pois na realização dos testes não foi observada melhora nas predições, além de, ao utilizar um maior número de árvores, a execução do código se torna mais lenta.

3.3. Análise Preditiva

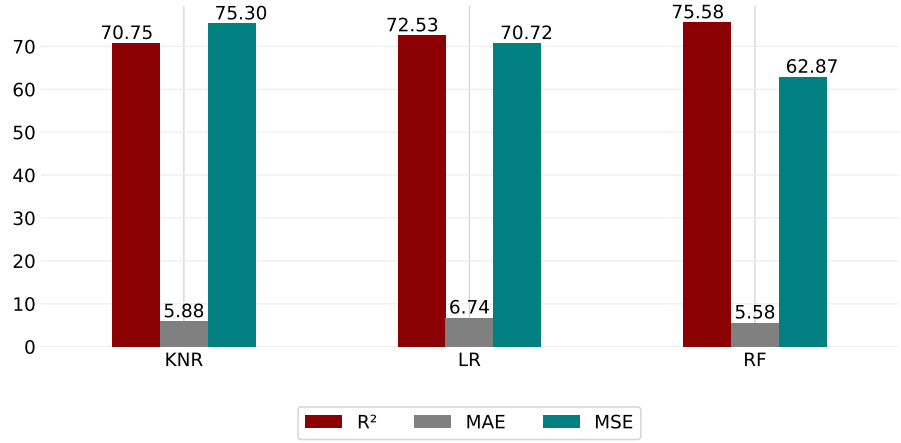
Por fim, foram executados os modelos preditivos. A execução foi dividida em treino e teste, onde 70% dos dados foram utilizados para a fase de treino e os outros 30% para testes. As métricas consideradas para avaliar o desempenho dos algoritmos foram Coeficiente de Determinação (*R2*), Erro Absoluto Médio (*MAE*) e Erro Quadrático Médio (*MSE*).

Coeficiente de Determinação. Também conhecido como R^2, é uma métrica que avalia o quão bem os dados se ajustam à regressão. Ele representa a proporção da variação na variável dependente que pode ser explicada pela regressão em relação à variação total da variável dependente. Em outras palavras, o R^2 indica o quão bem as previsões de regressão se aproximam dos valores reais. Quanto maior a pontuação de R^2, melhor o modelo se ajusta aos valores reais em comparação com a utilização de apenas a média dos valores.

Erro Absoluto Médio. É calculado como a média das diferenças absolutas entre as previsões do modelo e os valores reais. Quanto menor for o valor do *MAE*, menor será a quantidade de erros ocorridos nas previsões do modelo. Por outro lado, um *MAE* alto indica que o modelo apresenta uma maior diferença em relação aos valores reais, sugerindo uma menor precisão nas previsões. Portanto, ao avaliar um modelo, é desejável obter um valor de *MAE* o mais baixo possível.

Erro Quadrático Médio. É calculado como a média das diferenças quadráticas entre as previsões do modelo e os valores reais. Quanto menor for o valor do *MSE*, menor será a distância média dos erros em relação à média. Todavia, um *MSE* maior indica que

Figura 1. Resultados dos testes para o cenário do 1º turno.



Fonte: elaborado pelos autores.

o modelo apresenta uma maior dispersão em relação aos valores reais, sugerindo menor precisão nas previsões. Logo, é desejável obter um valor de *MSE* o mais baixo possível.

4. Discussão e Análise dos Resultados

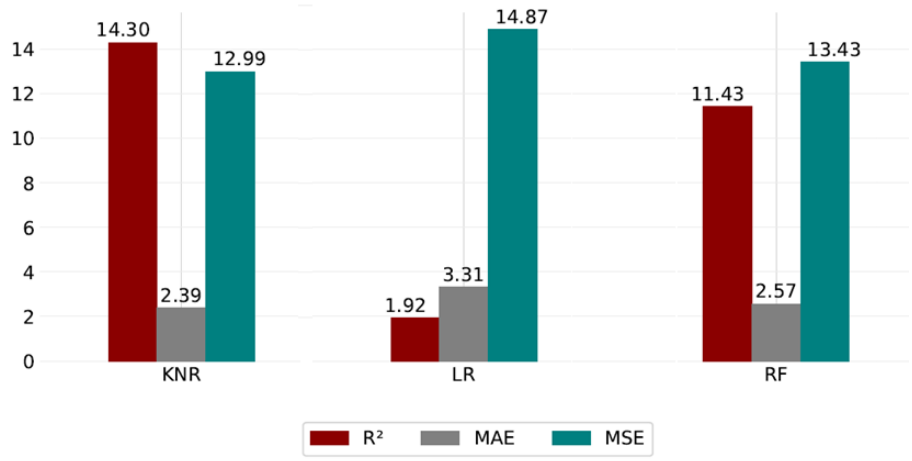
Para proceder com a análise dos resultados foi necessário realizar as fases de treino e teste dos algoritmos de predição, conforme descrito a seguir.

Fase de Treino. Primeiro, os três algoritmos foram treinados com dados das pesquisas de primeiro turno de eleições (2002-2018). Cada modelo foi executado uma vez e então calculadas as métricas de avaliação. A Figura 1 mostra o resultado da execução para o 1º turno. Note que os valores de R^2 foram multiplicados por 100. Os três algoritmos obtiveram valores de R^2 próximos, variando em poucos mais de 5%, assim como o *MAE* que variou em pouco mais de 1%. No entanto, o *MSE* obteve uma maior variação, passando dos 13%. Assim, o RF obteve o maior R^2 e os menores erros, saindo vitorioso. Já o KNR, embora não tenha o maior *MAE*, possui um maior *MSE* e menor R^2 tendo um resultado inferior.

Para o segundo turno houve oscilação no R^2 do RF em mais de 10% entre as execuções. Uma possível explicação pode ser atribuída ao fato de a cada vez serem geradas novas árvores aleatórias (o que não ocorre com KNR e LR). Além disso, há uma quantidade menor de dados em comparação com o 1º turno. Obtivemos valores inferiores para R^2, embora os resultados de erros foram bem abaixo comparado aos resultados do 1º turno. Tal resultado pode ser consequência direta do menor número de pesquisas eleitorais, uma vez que o segundo turno compreende um período de apenas três semanas. Com menos dados, notamos uma diminuição nos erros em todos os modelos. A Figura 2 mostra os resultados obtidos para as previsões do 2º turno. Similar ao resultado anterior, os valores de R^2 foram multiplicados por 100.

Fase de Teste. Em seguida, executamos os algoritmos para prever o resultado da eleição presidencial de 2022. Os modelos de aprendizado de máquina não obtiveram resultados tão assertivos quanto uma pesquisa eleitoral, visto que ao somar os valores de todos os candidatos o resultado encontrado é superior a 100. Isso ocorreu, pois não houve uma

Figura 2. Resultados dos testes para o cenário do 2º turno.



Fonte: elaborado pelos autores.

limitação para os resultados, a predição foi realizada exclusivamente com base nos resultados dos candidatos nas pesquisas eleitorais para a eleição de 2022 e não foi feita uma padronização para que o resultado do somatório fosse inferior a 100. Apesar disso, os resultados encontrados foram factíveis com a realidade, tendo em vista o ocorrido nas eleições, sendo o *Random Forest* o modelo que chegou mais próximo da diferença final (votos válidos) entre os dois primeiros candidatos (um dos grandes erros das pesquisas em 2022).[6] Ele registrou uma diferença de 4.01% entre eles, enquanto nas urnas a diferença foi de 5%. O algoritmo previu Lula à frente (como de fato ocorreu), além de acertar a colocação final de seis dos onze candidatos. O KNR apontou uma diferença de 0.52% entre os dois primeiros e colocou Bolsonaro em primeiro lugar, e acertou a posição de apenas dois dos onze candidatos. Por fim, LR apontou uma diferença de 10%, entre os dois primeiros candidatos, tendo Lula à frente, além de que teve acerto da posição final de quatro dos onze candidatos, sendo superior ao KNR e inferior ao RF.

Em outra análise, comparamos os resultados obtidos e as últimas pesquisas eleitorais. Foi considerado apenas o resultado do *Random Forest* pois ele apresentou melhores resultados. O RF se aproximou do resultado final apenas do candidato *Léo Péricles* (0.06%), enquanto o resultado real das pesquisas foi de 0.05% de votos. No entanto, acertou a ordem do maior número de candidatos, tais como o *Instituto Futura* e *Paraná Pesquisas*. Por fim, o RF se aproximou da diferença entre os dois primeiros candidatos, conforme Tabela 4.

Na Tabela 5, pode ser visto que o algoritmo que obteve melhores resultados para a predição do 1º Turno das eleições foi o *Random Forest*. Ainda que ele não tenha sido o algoritmo que mais se aproximou dos resultados finais de todos os candidatos, foi o que possuiu a menor quantidade de erros médios absoluto e quadrático, em comparação aos outros algoritmos. Isso pode ter ocorrido, pelo fato de o RF ser um modelo de conjunto que combina várias árvores de decisão para fazer previsões, onde cada uma delas é treinada em uma amostra aleatória dos dados e as previsões são combinadas para ob-

[6]CNN: Pesquisas erram e divergem dos resultados das urnas. Disponível em: `https://bit.ly/cnn-diverge-pesq`. Acesso em 02/07/2023.

Tabela 4. Diferença (em votos totais) entre Bolsonaro e Lula nas pesquisas, predição e urnas.

Preditor	% Votos Totais	Preditor	% Votos Totais
Brasmarket	14.5	MDA	7.9
Data Folha	14.0	Paraná Pesquisas	6.6
Ipec	13.0	***Random Forest***	***4.01***
Ipesp	13.0	Veritá	3.0
Atlas Intel	9.1	Futura	2.9
Resultado das Urnas (1º Turno)			**5.0**

Fonte: elaborado pelos autores.

ter uma previsão final. Isso permite que ele capture relacionamentos complexos entre as variáveis, o que é positivo em um cenário de muita oscilação dos dados (tais como em pesquisas eleitorais).

Por outro lado, o *KNeighbors Regressor* obteve uma performance mais inferior, com o menor Coeficiente de Determinação e também maior Erro Médio Quadrático. O KNR acertou a posição de apenas dois candidatos e errou a ordem final dos dois primeiros. O KNR é um modelo simples que faz previsões com base nas médias dos k vizinhos mais próximos e pode não ser capaz de capturar relacionamentos não lineares nos dados, podendo prejudicar seu desempenho no cenário de primeiro turno, tendo em vista a enorme variação da pontuação de cada candidato no decorrer do ciclo de pesquisas eleitorais. Além disso, as pesquisas utilizadas no estudo começaram a serem divulgadas em 2019 (três anos antes da eleição), quando alguns dos candidatos ainda não eram conhecidos pela maioria da população. Isso pode influenciar no desempenho dos candidatos, iniciando com baixa intenção de voto, e então oscilando positivamente. Por exemplo, a candidata Soraya Thronicke passou a ter mais visibilidade nas pesquisas após suas participações em debates televisionados. Por outro lado, há o caso em que candidatos conhecidos iniciavam a corrida eleitoral melhores pontuados, e então oscilaram negativamente à medida que a votação foi se aproximando (por exemplo, Ciro Gomes).[7]

Para o segundo turno, os resultados foram bem próximos aos reais. A LR foi a que mais se aproximou do resultado final (votos válidos). Ela apontou que Lula teria 48.04%, enquanto Bolsonaro teria 46.24% dos votos. Para o RF, os resultados foram 51.56% e 46.32%, respectivamente para Lula e Bolsonaro. Por fim, o KNR inferiu que Lula receberia 49.48%, enquanto Bolsonaro teria 46.01% dos votos. Também foi comparado o resultado dos algoritmos com o resultados obtido nas urnas, conforme mostra a Tabela 6.

Nota-se que para o segundo turno, a regressão linear foi superior às pesquisas eleitorais tradicionais, com uma diferença de 1.8% entre os candidatos, enquanto a diferença real foi de aproximadamente 1.72% (votos totais). Dentre as pesquisas, a melhor diferença entre os candidatos foi na pesquisa MDA, que sugeriu uma diferença de 2%, como pode ser visto na Tabela 7.

Como há menor variação dos resultados nas pesquisas eleitorais em segundo

[7]Pesquisa eleitoral: Lula sobe 3 pontos e Ciro e Tebet caem 2, diz BTG/FSB. Disponível em: `https://exame.com/brasil/pesquisa-eleitoral-lula-sobe-3-pontos-e-ciro-e-tebet-caem-2-diz-btg-fsb/`. Acesso em 17/08/2023.

Tabela 5. Resultados (em votos totais) obtidos pelos preditores RF, KNR e LR comparados ao resultado real das urnas para o 1º turno.

Candidato	RF	KNR	LR	Urnas
Lula	40.10	41.09	45.21	46.29
Bolsonaro	36.00	41.61	34.32	41.29
Simone Tebet	9.60	7.49	7.98	3.98
Ciro	10.21	16.75	11.62	2.90
Soraya Thronicke	8.66	6.002	5.99	0.49
Luiz Felipe d'Avila	1.18	2.42	6.17	0.45
Padre Kelmon	0.26	0.058	5.53	0.07
Leonardo Péricles	0.06	0.036	5.35	0.05
Sofia Manzano	0.15	0.06	5.26	0.04
Vera Lúcia Salgado	0.14	0.06	5.26	0.02
Eymael	0.15	0.06	5.26	0.01

Fonte: elaborado pelos autores.

Tabela 6. Resultados obtidos pelos preditores RF, KNR e LR comparados ao resultado real das urnas para o 2º turno.

	% Votos Totais	
Preditor	**Lula**	**Bolsonaro**
Random Forest Regressor	51.56	46.32
K Neighbors Regressor	49.48	46.01
Linear Regressor	48.04	46.24
Resultado das Urnas (2º Turno)	**48.56**	**46.85**

Fonte: elaborado pelos autores.

turno, a predição do resultado é mais assertiva. Isso ocorre, pois com a menor variação, não existe grande dificuldade para os algoritmos chegarem na previsão mais correta, a exemplo do que foi mencionado anteriormente em relação ao que ocorreu com KNR no primeiro turno.

5. Limitações e Ameaças à Validade

Durante o desenvolvimento do projeto, deparou-se com um desafio crucial relacionado ao volume de pesquisas por candidato. Alguns candidatos que participaram da eleição presidencial de 2022 foram objeto de mais de 400 pesquisas, notavelmente Lula, Bolsonaro e Ciro. Em contraste, candidatos como Leonardo Péricles, Sofia Manzano e Eymael apareceram em menos de 150 pesquisas, enquanto o candidato Padre Kelmon teve sua presença registrada em apenas 40 pesquisas. Isso criou uma complexidade adicional para os algoritmos de previsão, já que a escassez de dados tende a dificultar a obtenção de resultados confiáveis.

Outra complicação surgiu da variação nos períodos de pesquisa para diferentes candidatos. Por exemplo, em 2019 já havia pesquisas eleitorais para a eleição presidencial de 2022, incluindo os nomes de Lula, Ciro e Bolsonaro. Em contraste, Simone Tebet, que superou Ciro nas urnas, só foi incluída em pesquisas a partir do meio de 2021. Ela era menos conhecida pelo público em geral e ganhou visibilidade quando os debates televisivos começaram. Isso resultou em Ciro mantendo uma visibilidade prolongada, o que se

Tabela 7. Comparação (em votos totais) de resultados obtidos pelas pesquisas, predição e urnas.

	% Votos Totais	
Preditor	**Lula**	**Bolsonaro**
Brasmarket	41.50	48.00
Ipespe	50.00	44.00
Data Folha	49.00	45.00
Futura	46.60	47.20
Veritá	43.60	45.80
Ipec	50.00	43.00
Paraná Pesquisas	47.10	46.30
MDA	46.90	44.90
Atlas Intel	52.40	45.30
Linear Regressor	**48.04**	**46.24**
Resultado das Urnas (2º Turno)	**48.56**	**46.85**

Fonte: elaborado pelos autores.

refletiu nas pesquisas. Ele acumulou uma porcentagem maior de votos devido à sua longa trajetória, enquanto Tebet, estreante na eleição presidencial de 2022, teve números inferiores aos de Ciro na maioria das pesquisas. Esses padrões tiveram impacto nos resultados dos algoritmos de previsão.

Finalmente, enfrentou-se o desafio da escassez de pesquisas de segundo turno após a definição dos candidatos. Com uma diferença de apenas quatro semanas entre os dois turnos, a base de dados continha apenas 64 pesquisas para o segundo turno, em comparação com as mais de 400 pesquisas do primeiro turno. A falta de informações dificultou a previsão dos algoritmos, resultando em dificuldades semelhantes às observadas no primeiro turno para alguns candidatos, como mencionado anteriormente.

6. Conclusão

Neste trabalho, o objetivo foi se aproximar dos resultados eleitorais obtidos nas pesquisas prévias às eleições. Essa abordagem foi concretizada ao empregar dados provenientes dos próprios institutos de pesquisa, juntamente com técnicas de aprendizado de máquina. Especificamente, os modelos utilizados incluíram o *Random Forest Regressor, KNeighbors Regressor* e *Linear Regression*. Para a compilação das informações de pesquisa, uma ampla base de dados eleitorais foi empregada, abrangendo todas as eleições presidenciais brasileiras do século 21. A análise dos resultados destacou que, ao lidar com um conjunto extenso de dados para a previsão de uma ampla gama de valores, o modelo mais eficaz dentre os adotados foi o *Random Forest Regressor*. Por outro lado, quando há uma disponibilidade limitada de dados e um número reduzido de valores a serem previstos, o *Linear Regression* se destaca como a escolha mais apropriada.

Chega-se à conclusão de que a aplicação do aprendizado de máquina pode desempenhar um papel significativo na busca por resultados eleitorais mais aproximados às decisões das urnas em futuras eleições no Brasil. Ainda que nem todos os resultados tenham superado as previsões das tradicionais pesquisas eleitorais, em grande parte das situações analisadas, notou-se uma maior fidelidade em relação aos desfechos da eleição presidencial brasileira de 2022. Por fim, é relevante destacar que a adoção do apren-

dizado de máquina não deve substituir as pesquisas eleitorais convencionais. Conforme evidenciado neste estudo, ambos podem e devem coexistir, colaborando mutuamente para alcançar os melhores resultados possíveis.

7. Recomendações para Trabalhos Futuros

Recomenda-se para o desenvolvimento de trabalhos futuros, a realização de predição de resultados de votações estaduais e/ou municipais, com a utilização de dados de eleições anteriores para o mesmo estado e/ou município para a qual a pesquisa for realizada. Desta forma, pode ser realizada uma análise da possibilidade de obtenção de resultados superiores às de pesquisas eleitorais para governador e prefeito com algoritmos de predição, tendo em vista que a quantidade de pesquisas para esses cargos é inferior ao número de pesquisas divulgadas para presidente.

Além disso, sugere-se a utilização de outros modelos de predição, ampliando a análise exploratória na busca de melhores resultados na predição de eleições adaptadas ao contexto e tamanho de cada votação.

Referências

Aiyappa, R., DeVerna, M. R., Pote, M., Truong, B. T., Zhao, W., Axelrod, D., Pessianzadeh, A., Kachwala, Z., Kim, M., Seckin, O. C., et al. (2023). A multi-platform collection of social media posts about the 2022 us midterm elections. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 981–989.

Brito, K. and Adeodato, P. J. L. (2023). Machine learning for predicting elections in latin america based on social media engagement and polls. *Government Information Quarterly*, 40(1).

de Sousa, L. S. B. (2021). As redes sociais na tomada de decisão do voto e as eleições para presidente do brasil em 2018. *e-Com*, 14:63–76.

Hagemann, L. and Abramova, O. (2022). Crafting audience engagement in social media conversations: Evidence from the us 2020 presidential elections.

Sani, M. and Azizuddin, M. (2014). The social media election in malaysia: The 13th general election in 2013. *Kajian Malaysia: Journal of Malaysian Studies*, 32.

Silva, J. B. d. (2018). Aplicação de técnicas de machine learning na combinação de pesquisas eleitorais. B.S. thesis, UFRN.

Tsakalidis et al. (2015). Predicting elections for multiple countries using twitter and polls. *IEEE Intelligent Systems*, 30(2):10–17.

Venturi, G. (1995). Pesquisas pré-eleitorais: legitimidade, influência e contribuições à cidadania. *Opinião Pública*, 3(2):129–145.

Zhou, Z. et al. (2021). Why polls fail to predict elections. *Journal of Big Data*, 8(1):1–28.