Análise do perfil de voto das mesorregiões de Minas Gerais no processo eleitoral de 2022

Uciolli B. F. Lemos, Carlos A. Silva, Renato M. Filho

[1]Instituto Federal de Minas Gerais - Sabará
Caixa Postal 34.590-390 – Sabará – MG – Brasil

uciollilemos@gmail.com,{carlos.silva,renato.miranda}@ifmg.edu.br

Abstract. *Elections represent crucial moments in a democratic system, where candidates face challenges of limited financial resources and time to conduct their campaigns. Understanding voter behavior is key to optimizing the costs and effectiveness of electoral campaigns. In this study, we propose the identification and analysis of the most influential attributes in candidates for state and federal deputies in the 2022 elections in each of the twelve mesoregions of the state of Minas Gerais, using data mining techniques, including the Regression Tree algorithm. The results indicate that factors such as race, gender, declared asset value, and marital status play a significant role in determining the amount of votes received by the candidate.*

Resumo. *Eleições representam momentos cruciais em um sistema democrático, no qual os candidatos enfrentam desafios de recursos financeiros e de tempo limitados para conduzir suas campanhas. Compreender o comportamento do eleitorado é fundamental para otimizar os custos e a eficácia das campanhas eleitorais. Neste estudo, propomos a identificação e análise dos atributos mais influentes nos candidatos a deputado estadual e federal nas eleições de 2022 em cada uma das doze mesorregiões do estado de Minas Gerais, utilizando técnicas de mineração de dados, incluindo o algoritmo de Árvore de Regressão. Os resultados indicam que fatores como raça, gênero, valor dos bens declarados e o estado civil desempenham um papel significativo na determinação da quantidade de votos recebidos pelos candidatos.*

1. Introdução

O período eleitoral é marcado pela escolha dos representantes da população, e compreender os fatores que influenciam o voto é crucial nesse contexto. Esse conhecimento não apenas pode orientar estudos subsequentes, mas também auxiliar os candidatos a direcionar seus esforços de campanha de maneira estratégica, contribuindo para a redução de custos. Um exemplo prático desse cenário é observado nas eleições de 2022, em que foi declarado um montante total de R$ 12.753.498.295,20[1] em despesas eleitorais, sendo 90,22% provenientes de recursos públicos como o Fundo Partidário e o Fundo Especial de Financiamento de Campanha. Dessa forma, torna-se imperativo que as campanhas sejam eficientes, uma vez que a maioria dos recursos é originária do orçamento público.

Os conjuntos de dados utilizados neste trabalho foram coletados no portal de dados abertos do TSE[2] e consistem de três arquivos de texto. O primeiro arquivo, composto por

[1]https://sig.tse.jus.br/ords/dwapr/seai/r/sig-prestacao-contas/receitas-despesas
[2]https://dadosabertos.tse.jus.br/dataset/

2.562 registros, contém as informações gerais do candidato, tais como: nome, raça, sexo e ocupação. O segundo arquivo contém informações referentes aos bens declarados pelos candidatos e possui 8.548 registros. O terceiro arquivo contém os votos por seção do estado de Minas Gerais. Os conjuntos passaram por uma etapa de pré-processamento, onde foram retirados registros que fogem do escopo deste trabalho, por exemplo, instâncias relacionadas à cargos diferentes aos de deputado estadual e deputado federal.

O conjunto de dados contendo os dados demográficos dos candidatos a deputado estadual e deputado federal, já processados, foi submetido a um modelo de aprendizado de máquina, para que fosse possível prever a quantidade de votos recebidos. O modelo foi construído utilizando um algoritmo de Árvore de Regressão. Este modelo foi avaliado por meio de uma validação cruzada e os resultados foram mensurados pelas métricas *Root Mean Squared Error* (RMSE) e *Mean Absolute Error* (MAE), pelas quais foi possível realizar um comparativo sobre a assertividade. Estas métricas foram utilizadas visando medir os impactos de possíveis *outliers* presentes na base.

Ao fim da construção do modelo foi analisada a árvore gerada e extraídas as regras que guiam a votação da população mineira, em suas diferentes mesorregiões. Os resultados obtidos mostram que as características mais relevantes, de modo geral, são a raça e o gênero do candidato.

Este trabalho está organizado da seguinte forma: Na seção 2 é feito um estudo sobre o estado da arte do problema abordado no cenário nacional e na América do Sul. Na seção 3 é apresentada a forma como os dados utilizados neste trabalho foram obtidos e tratados. A caracterização do conjunto de dados utilizado e a discussão dos resultados são apresentados na seção 4. A seção 5 descreve o processo de construção da Árvore de Regressão e analisa os resultados obtidos. Por fim, a seção 6 realiza à conclusão e apresenta propostas de trabalhos futuros.

2. Trabalhos Relacionados

Eleições são amplamente empregadas em pesquisas de mineração de dados e encontram diversas aplicações. Por exemplo, são usadas para monitorar atividades anômalas em plataformas de redes sociais [Silva 2021], ou gerar *insights* sobre comportamentos de candidatos e características de campanhas [Martins et al. 2019]. Além disso, são relevantes na detecção de violações do código eleitoral e na análise de sentimentos [Pereira et al. 2023] e na predição de alocação de recursos [Guedes 2018, Barros 2018]. No contexto de *fake news*, a aplicação de eleições é destacada em pesquisas como [Cabral et al. 2021, Leal 2018], entre outros estudos.

A análise de processos eleitorais tem crescido no Brasil, porém um dos empecilhos tem sido a falta de padronização e unificação dos conjuntos de dados [Vasconcelos et al. 2021], bem como a presença de dados faltantes e inconsistentes. No trabalho supracitado produziu-se o conjunto de dados CandiData que busca solucionar estes problemas, padronizando e tratando as inconsistências presentes como, por exemplo, datas em formatos diferentes, arquivos em formatos distintos (.txt, .csv) e nomes dos atributos diferindo de base para base. Complementarmente, buscou-se agrupar dados eleitorais desde 1945 para que fosse possível a realização de análises históricas. A título de demonstração foram realizados dois experimentos. O primeiro buscou verificar a participação feminina nas eleições, observando um crescimento significativo a partir de

2008. O segundo experimento observou as profissões com maior participação no período eleitoral de 2020, e foi demonstrado que o perfil de agricultor foi o mais presente entre os candidatos.

O trabalho de [Camargo et al. 2016], buscou realizar uma análise para identificar os fatores mais relevantes para a eleição de um candidato a vereador na região do Rio Grande do Sul. Para isso utilizou-se a ferramenta WEKA (*Waikato Environment for Knowledge Analysis*) com o algoritmo J48. Assim, por meio da mineração dos dados, foram identificados que os atributos mais relevantes incidiam na verificação se o candidato era político de carreira, o grau de instrução, o gênero e a idade.

A análise proposta por [Nicolau 2014] parte da observação do eleitorado e avalia quais características mais influenciaram na votação dos candidatos à presidência da república nas eleições de 2010, sendo os candidatos deste período: Dilma Rousseff, José Serra e Marina Silva. Para o desenvolvimento do trabalho foram utilizados dados do Eseb-2010[3] (Estudo Eleitoral Brasileiro). Por meio de um algoritmo de Regressão Logística Multinomial, observou-se que a escolaridade, a região, a religião, a autoidentificação no espectro esquerda-direita e a avaliação do governo foram as características que representaram maior diferença na votação obtida pelos candidatos.

O trabalho de [De Albuquerque Filho et al. 2020] buscou encontrar candidatos "laranjas", ou seja, pessoas que cedem seu nome, com ou sem consentimento, para uso de outra pessoa, em processos eleitorais. Para isso, o trabalho recorreu ao algoritmo *Isolation Forest* em conjunto com uma Árvore de Decisão. O trabalho utilizou as bases do TSE relacionando características demográficas e despesas de campanha dos candidatos. Ao final observou-se que as candidaturas "laranjas" tendem a ter altos gastos, porém o perfil e a votação do candidato não condiz com as despesas declaradas.

[Campos-Valdés et al. 2021] buscou analisar o impacto dos perfis e decisões de campanha dos candidatos do Chile e seu desempenho nas eleições após a mudança do sistema eleitoral. Para o desenvolvimento foram utilizados os algoritmos *Regressor Random Forest* e Regressão Logística Multinomial. Foram levados em conta atributos como carreira política, gastos na campanha e orientação política. Ao fim, o atributo mais relevante encontrado foi a carreira política do candidato, já os gastos de campanha variavam em sua relevância de acordo com a coalizão política do candidato.

Em resumo, os trabalhos relacionados abordados nesta seção destacam a crescente importância da análise de processos eleitorais e apresentam diversas abordagens para compreender os fatores que influenciam as eleições. Enquanto alguns focam na padronização e tratamento de bases de dados, como o trabalho de [Vasconcelos et al. 2021], outros exploram técnicas de mineração de dados para identificar atributos relevantes na escolha de candidatos, como [Camargo et al. 2016]. O estudo de [Campos-Valdés et al. 2021] destaca a importância da carreira política na performance dos candidatos, bem como a influência da coalizão política e os gastos de campanha. Na Tabela 1 é possível identificar os atributos mais relevantes encontrados pelos trabalhos analisados nesta seção. Este trabalho analisa um grupo pouco explorado na literatura até então, os candidatos a deputado federal e estadual.

[3]https://www.cesop.unicamp.br/por/eseb/ondas/7

Tabela 1. Atributos mais relevantes segundo os trabalhos relacionados.

Trabalho	Características Observadas
[Camargo et al. 2016]	Carreira política, grau de instrução, gênero e idade.
[Nicolau 2014]	Escolaridade, região, religião, autoidentificação no espectro esquerda-direita e avaliação do governo.
[Campos-Valdés et al. 2021]	Carreira política, gastos de campanha, coalizão.

3. Construção da Base de Dados

Nesta etapa foi realizada a aquisição e tratamento dos dados que estavam disponíveis no portal de dados abertos do Tribunal Superior Eleitoral (TSE)[4]. Foram utilizados dois conjuntos de dados que continham informações relativas às candidaturas para as eleições gerais do ano de 2022 e um terceiro conjunto de dados com os votos por seção do estado de Minas Gerais.

3.1. Conjunto de Dados Originais

Foram utilizados os seguintes conjuntos de dados: a base de informações gerais dos candidatos que será chamada de CC[5] e a base de bens declarados pelos candidatos que será chamada BC[6]. A base CC possui 2.562 registros e contém as informações demográficas do candidato (gênero, estado civil, raça e idade), bem como informações relacionadas à candidatura (número do candidato, partido político, se declarou bens ou não). A base BC contém 8.548 registros e contém informações relacionadas aos bens declarados pelo candidato como o tipo do bem e o seu valor em reais (R$).

É possível verificar uma diferença na quantidade de instâncias das bases, devido que, na base BC, um mesmo candidato pode ter mais de um bem declarado, e cada um desses bens trata-se de uma instância na base.

3.2. Tratamento dos conjuntos de dados

Um novo conjunto de dados foi criado ao combinar as duas bases mencionadas na subseção anterior. Na base BC, os valores dos bens dos candidatos foram agregados, enquanto na base CC, os registros que não se referiam a candidatos a deputado estadual ou federal foram removidos. Além disso, foram selecionadas as colunas relacionadas à data de nascimento, gênero, estado civil, raça e ocupação. Isso resultou em um conjunto de dados com 2.514 instâncias e 6 atributos.

3.3. Construção do conjunto de dados com votos por cidade

Foram utilizadas informações de votos por seção[7] disponibilizadas pelo TSE. Esta base contém todos os votos de todas as seções do estado de Minas Gerais, porém, devido ao alto número de instâncias presentes (1.978.746), foram removidos os cargos diferentes

[4]https://dadosabertos.tse.jus.br/

[5]https://dadosabertos.tse.jus.br/dataset/candidatos-2022/resource/435145fd-bc9d-446a-ac9d-273f585a0bb9

[6]https://dadosabertos.tse.jus.br/dataset/candidatos-2022/resource/fac824ef-8519-4c75-b634-378e6fcc717f

[7]https://dadosabertos.tse.jus.br/dataset/resultados-2022

de deputado federal e estadual e foram agrupados os votos por candidato por cidade, totalizando 1.577.586 instâncias. Como o número de instâncias permanecia alto foram consideradas apenas algumas cidades.

As cidades selecionadas são descritas na Tabela 2. Os critérios utilizados foram: i) o município com o maior eleitorado; ii) um município aleatoriamente escolhido com o eleitorado entre 15 mil e 50 mil; e iii) um município aleatoriamente escolhido com o eleitorado abaixo de 5 mil. Esta divisão tem como objetivo analisar cidades de pequeno, médio e grande porte visando abranger diferentes realidades e perfis de eleitorado.

Tabela 2. Descrição das doze mesorregiões com os municípios selecionados.

Município	Eleitorado	Município	Eleitorado
CAMPO DAS VERTENTES		CENTRAL MINEIRA	
Lavras	74.703	Curvelo	60.206
Nepomuceno	20.896	Três Marias	24.270
Ressaquinha	4.235	Estrela do Indaiá	2.931
JEQUITINHONHA		METROPOLITANA	
Diamantina	38.011	Belo Horizonte	2.002.303
Jequitinhonha	18.190	Caeté	34.168
Bandeira	4.538	Taquaraçu de Minas	4.741
NOROESTE DE MINAS		NORTE DE MINAS	
Paracatu	66.701	Montes Claros	287.056
Buritis	19.006	Salinas	32.477
Dom Bosco	3.552	São João da Lagoa	4.866
OESTE DE MINAS		SUL DE MINAS	
Divinópolis	169.844	Poços de Caldas	121.569
Piumhi	25.682	Extrema	37.093
Santana do Jacaré	3.977	Claraval	4.563
TRIÂNGULO MINEIRO		VALE DO MUCURI	
Uberlândia	513.847	Teófilo Otoni	106.477
São Gotardo	29.076	Águas Formosas	15.854
União de Minas	4.538	Bertópolis	4.266
VALE DO RIO DOCE		ZONA DA MATA	
Governador Valadares	215.354	Juiz de Fora	417.164
Mutum	21.335	Raul Soares	18.062
Braúnas	4.402	Santa Rita de Jacutinga	4.183

A contagem de votos por candidato por cidade foi agrupada em doze conjuntos de dados distintos referentes a cada mesorregião. Cada conjunto contendo as informações de votação de cada candidato em cada um dos municípios selecionados.

4. Caracterização da Base de Dados

Após o pré-processamento das bases inicias, foram gerados dois conjuntos de dados. Esses conjuntos de dados foram usados para o construção do modelo. O primeiro conjunto de dados contém as características dos candidatos, bem como seu código sequencial do sistema do TSE e seu número na urna, que foram utilizados para que fosse possível cruzar as bases de características, bens e de votação como podemos ver na Tabela 3. O segundo conjunto de dados contém os votos dos candidatos por município (atributos mostrados na Tabela 4).

Tabela 3. Descrição da tabela das características dos candidatos.

Nome	Descrição
DS_CARGO	Descrição do cargo (Deputado Estadual, Deputado Federal)
SQ_CANDIDATO	Número sequencial do candidato no TSE
NR_CANDIDATO	Número do candidato na urna
DT_NASCIMENTO	Data de nascimento do candidato
DS_GENERO	Descrição gênero do candidato
DS_ESTADO_CIVIL	Descrição do estado civil do candidato
DS_COR_RACA	Descrição da raça do candidato
VR_BEM_CANDIDATO	Valor total dos bens declarados do candidato

Tabela 4. Descrição da tabela dos votos por município.

Nome	Descrição
NM_MUNICIPIO	Nome do município
NR_VOTAVEL	Número do candidato na urna
DS_CARGO	Descrição do cargo (Deputado Estadual, Deputado Federal)
QT_VOTOS	Quantidade de votos

As bases de características conta com 2.514 candidatos, entre eles 1.411 concorreram ao cargo de Deputado Estadual e os 1.103 restantes ao de Deputado Federal. Neste conjunto de dados, 1.673 candidatos se autodeclararam do gênero masculino e outros 841 do gênero feminino. Conforme a Figura 1, é possível observar a discrepância entre o número de candidatos em relação ao gênero. O número de candidatos do gênero masculino em ambos os cargos representa aproximadamente o dobro do numero de candidatos do gênero feminino.

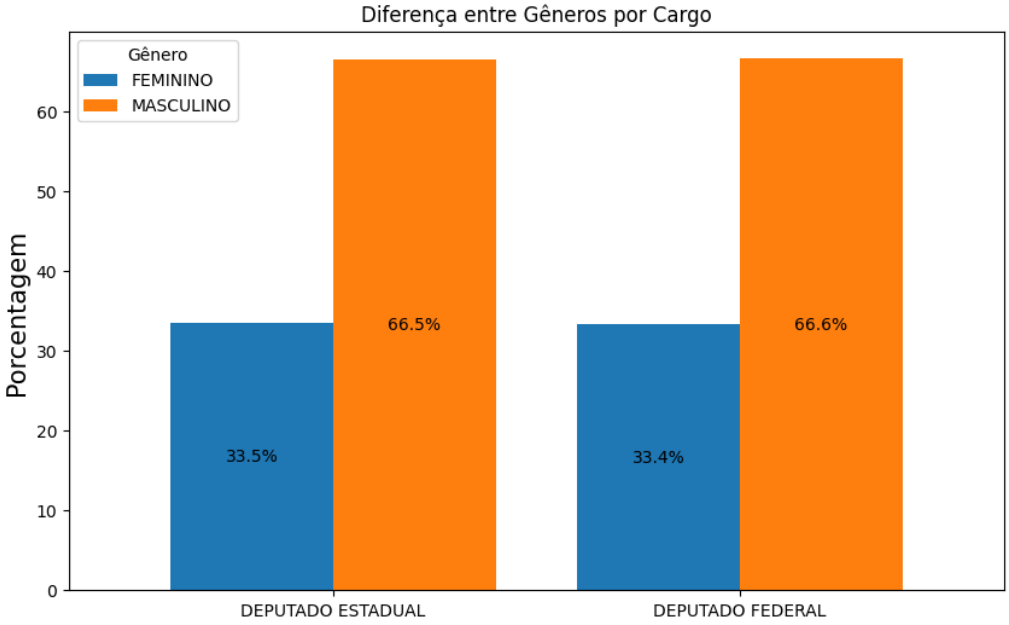

Figura 1. Relação entre o cargo almejado e o gênero do candidato.

Em relação à raça autodeclarada, 1.199 são brancos, 889 são pardos, 398 são pretos, 12 não informaram, 9 são amarelos e 7 são indígenas como mostra a Figura 2.

As raças Indígena, Amarela e Não Informado foram agrupadas na categoria “OUTROS”, pois somadas representam cerca de 1% do total de candidaturas.

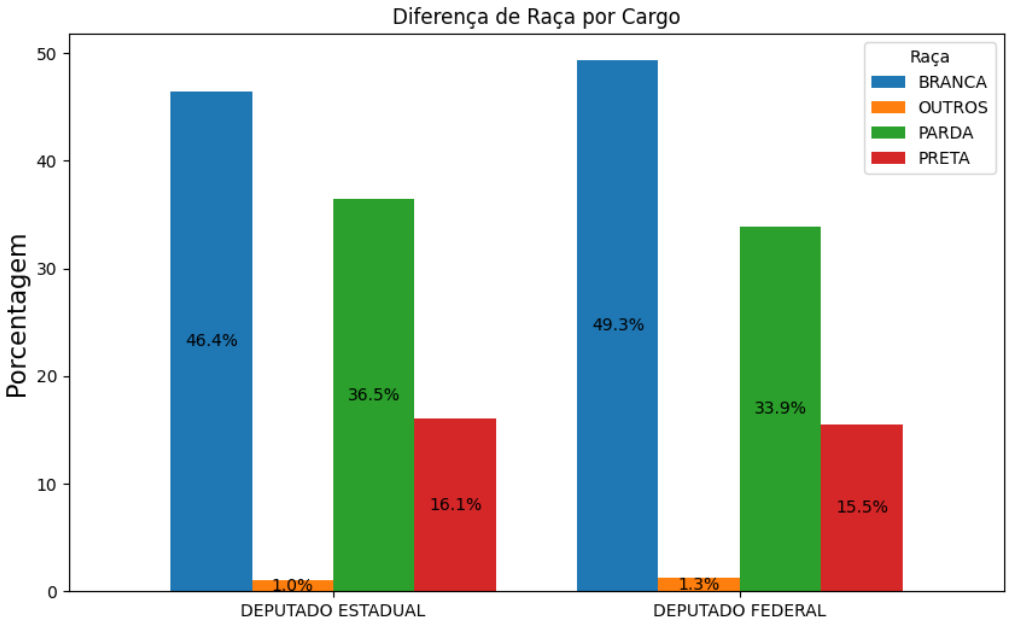

Figura 2. Relação entre o cargo almejado e a raça (autodeclarada) dos candidatos.

Quanto ao estado civil, 1.344 são casados, 754 são solteiros, 337 são divorciados, 55 são viúvos e 24 são separados judicialmente, como pode ser observado na Figura 3. Após as análises é possível perceber que os valores percentuais de gênero, raça e estado civil se mantém semelhantes entre os cargos.

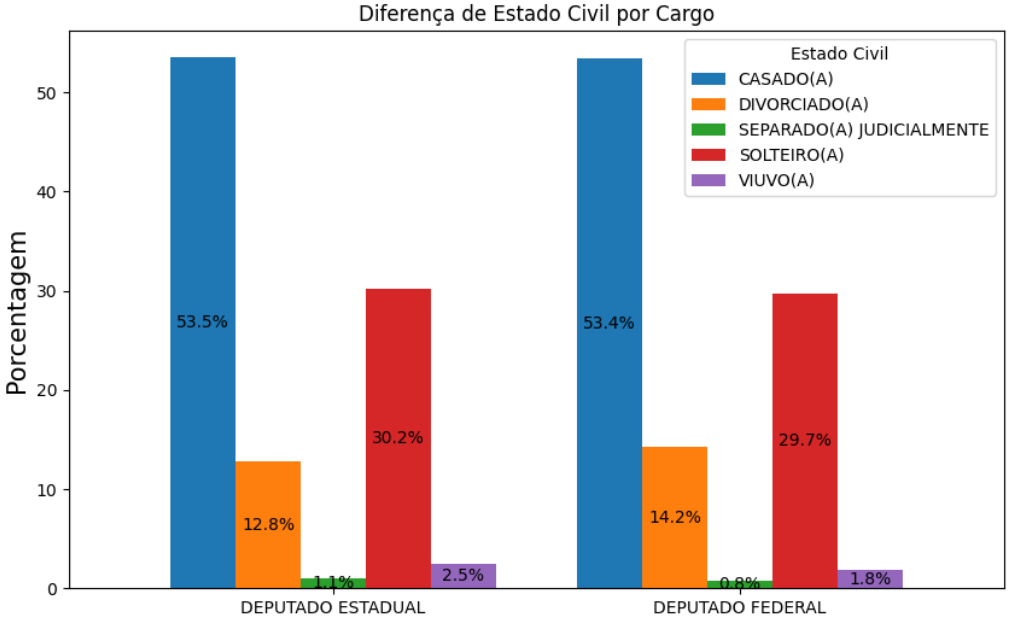

Figura 3. Relação entre o cargo almejado e o estado civil dos candidatos.

Em relação à idade dos candidatos, a maioria tem entre 40 e 60 anos, sendo a média de idade de 49 anos. O candidato mais novo possui 21 anos e o mais velho 91 anos, como podemos observar na Figura 4.

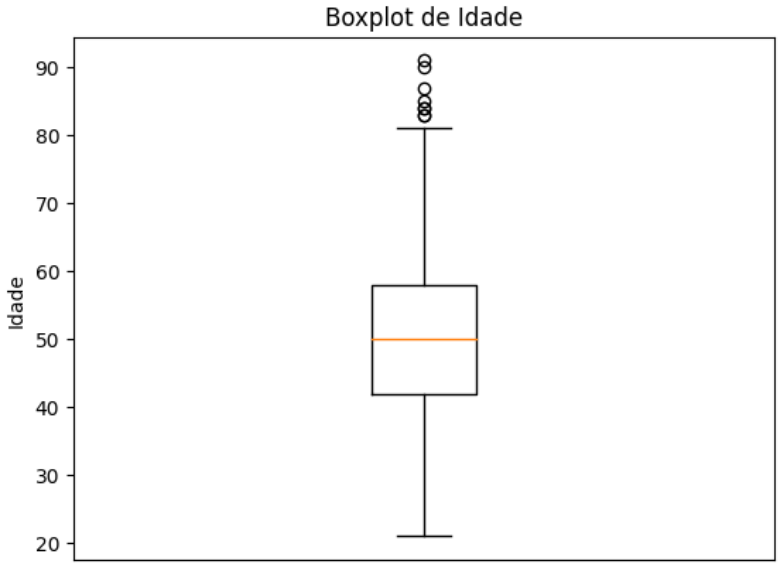

Figura 4. Distribuição das idades dos candidatos.

Por fim, em relação aos valores de bens declarados, podemos observar na Figura 5 que cerca de 35% dos candidatos não os declarou, e que 34,7% dos candidatos declarou entre R$ 100.000,00 e R$ 1.000.000,00. Podemos observar também que existem mais candidatos com alto valor de bens declarados (mais que R$ 1.000.000,00) que candidatos com valor baixo de bens declarados (menor que R$ 10.000,00).

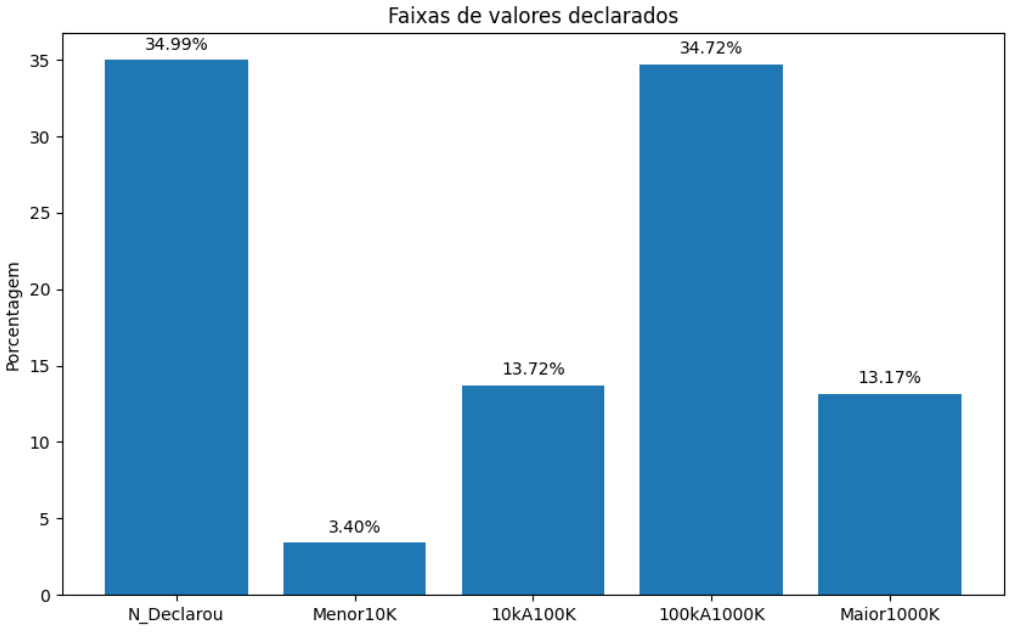

Figura 5. Proporção do valor de bens declarados.

5. Construção do Modelo

O desenvolvimento do modelo foi realizado em Python, fazendo uso da biblioteca *scikit-learn*. Os conjuntos de dados gerados, conforme mencionado na Seção 3, foram empregados como entrada para a construção de uma Árvore de Regressão. Os resultados foram avaliados por meio das métricas de Raiz do Erro Quadrático Médio (RMSE) e Erro Médio Absoluto (MAE). Foram geradas árvores para cada uma das doze mesorregiões do estado de Minas Gerais e os atributos mais relevantes foram analisados.

5.1. Árvore de Regressão

Uma Árvore de Regressão funciona de forma a dividir um ponto em diferentes nós, por meio de critérios que serão aprendidos durante o treinamento. A divisão do nó é semelhante à arvore de decisão, porém ao invés de usar valores discretos (0/1, Verdadeiro/Falso) são usados limites numéricos. Os nós são divididos em novos limites numéricos até que o erro seja minimizado, quando um valor mínimo de erro é atingido, temos uma folha que corresponde ao resultado.

Para o problema foi utilizado como critério de divisão a medida de *Friedman*, que utiliza o erro quadrático médio para melhorar o potencial das divisões da árvore. A quantidade máxima de níveis da árvore foi limitada a 4. O parâmetro de divisão foi o *random* que escolhe o melhor divisor aleatório. Para definir os parâmetros descritos acima, foi utilizado o método *GridSearchCV* que retornou a melhor combinação de parâmetros através de um processo de validação cruzada com 5 K-folds [Zahedi et al. 2021].

5.2. Pré-processamento

Para utilizar a Árvore de Decisão foi necessário realizar uma discretização dos atributos. Colunas que tinham valor categórico, como cargo, raça, estado civil e gênero, viraram múltiplas colunas representando os possíveis valores presentes nas mesmas por meio do método *One-Hot Encoding*. Os itens listados abaixo são referentes as novas colunas geradas para cada atributo que o método foi aplicado.

- Cargo:
  - Deputado Estadual
  - Deputado Federal
- Raça:
  - Amarela
  - Indígena
  - Não Informada
  - Parda
  - Preta
  - Branca
- Gênero:
  - Feminino
  - Masculino
- Estado Civil:
  - Casado
  - Divorciado(a)
  - Separado(a) Judicialmente

  - Solteiro(a)
  - Viuvo(a)

A coluna do valor dos bens do candidato também foi transformada em múltiplas colunas com valores 0/1, representando respectivamente Verdadeiro/Falso que são referentes a faixas de valor de bens declarados como mostrado nos itens abaixo.

- Valor dos Bens do Candidato:
  - N_Declarou: Não declarou nenhum bem
  - Menor10k: Soma dos bens menor que R$ 10.000,00
  - 10kA100k: Soma dos bens entre R$ 10.000,00 e R$ 100.000,00
  - 100kA1000k: Soma dos bens entre R$ 100.000,00 e R$ 1.000.000,00
  - Maior1000k: Soma dos bens maiores que R$ 1.000.000,00

A base de votos por cidade foi separada por cada uma das mesorregiões, deixando somente as colunas de quantidade de votos e número dos candidatos. Cada uma das bases de voto por região foi combinada com a base de características do candidato, sendo essas novas bases geradas as utilizadas no modelo.

5.3. Treinamento do Modelo e Visualização das Árvores

Para a construção do modelo, separamos a base em atributos (X) e saída do modelo (Y). Em X deixamos todas as colunas, com exceção do número do candidato e quantidade de votos, e Y representa a quantidade de votos. Foi utilizado o método de validação cruzada com 5 K-folds para avaliar o desempenho do modelo.

A avaliação dos resultados foi feita utilizando dois métodos: *Root Mean Square Error* (RMSE) e *Mean Absolute Error*. O RMSE calcula a raiz do erro quadrático médio e é utilizada quando se busca precisão em acertar valores extremos, o que a torna também muito sensível aos erros nestes valores. Tomando conhecimento da disparidade da votação e antecipando o comportamento da métrica RMSE foi utilizada uma outra métrica para garantir a assertividade do método, a métrica *Mean Absolute Error* que representa a média absoluta do erro do modelo.

Utilizamos as duas métricas para avaliar o desempenho dos modelos, pois a RMSE é sensível a *outliers* e a MAE é mais tolerante aos mesmos. Na Tabela 5 é possível verificar esses resultados.

Tabela 5. Resultados das métricas RMSE e MAE.

Região	RMSE	MAE
CAMPO DAS VERTENTES	458,21	83,68
CENTRAL	361,05	77,26
JEQUITINHONHA	197,49	50,71
METROPOLITANA	6753,12	1648,88
NOROESTE DE MINAS	470,49	82,33
NORTE DE MINAS	1517,05	314,79
OESTE DE MINAS	1031,95	190,56
SUL DE MINAS	609,45	132,74
TRIÂNGULO MINEIRO	2460,94	485,60
VALE DO MUCURI	460,48	87,08
VALE DO RIO DOCE	1168,05	211,08
ZONA DA MATA	1753,00	374,29

Vistas as métricas da etapa de construção do modelo, foi observado a diferença dos valores entre a métrica RMSE e a métrica MAE, isto se dá devido ao fato da votação nestas regiões ser heterogênea, tendo diversos perfis de candidatos recebendo uma variação alta de votos. A presença destes *outliers* e dos valores das métricas já era prevista devido ao fato da análise não levar em conta a carreira política do candidato, e nem sua fama na região, o que pode fazer um indivíduo fora dos padrões definidos pelo modelo ter muitos votos.

É possível observar também que o grau de erro das métricas apresenta discrepância entre cada região. Uma das razões é o número de votantes que varia em cada região, e, nas regiões mais populosas apresenta maior erro que nas regiões menos populosas.

Vale ressaltar que, embora o modelo busque prever a quantidade de votos obtidos pelos candidatos, a precisão na predição dos votos não era o principal fator de interesse. O foco do trabalho estava em compreender as características que influenciam a votação, em vez de apenas prever o número exato de votos. Também, optou-se por utilizar o número de votos em vez da informação sobre a eleição ou não do candidato devido ao impacto da legenda partidária, que pode influenciar significativamente os resultados eleitorais.

Ao final do treinamento, foi gerado uma Árvore de Regressão para cada uma das mesorregiões analisadas. Os atributos dos nós raiz (1° Nível) e do 2° Nível de cada árvore estão descritos na Tabela 6.

Tabela 6. Atributos encontrados nas árvores das Mesorregiões do Estado

Região	**1º Nível**	**2º Nível**
Campo das Vertentes	ESTADO_CIVIL_CASADO	RACA_BRANCA, Maior10000K
Central	RACA_BRANCA	CARGO_DEPUTADO ESTADUAL, GENERO_MASCULINO
Jequitinhonha	ESTADO_CIVIL_DIVORCIADO(A)	RACA_BRANCA, GENERO_FEMININO
Metropolitana	RACA_PARDA	IDADE, CARGO_DEPUTADO FEDERAL
Noroeste de Minas	RACA_BRANCA	ESTADO_CIVIL_DIVORCIADO, GENERO_MASCULINO
Norte de Minas	RACA_PARDA	RACA_PRETA, Menor10K
Oeste de Minas	RACA_PRETA	N_Declarou,
Sul de Minas	GENERO_FEMININO	RACA_BRANCA, RACA_PARDA
Triângulo Mineiro	GENERO_FEMININO	N_Declarou, 100kA1000K
Vale do Mucuri	N_Declarou	RACA_BRANCA, ESTADO_CIVIL_SOLTEIRO
Vale do Rio Doce	IDADE	CARGO_DEPUTADO FEDERAL, 10kA100K
Zona da Mata	RACA_PRETA	GENERO_FEMININO, IDADE

A análise dos resultados indica que os atributos referentes à raça e gênero são os principais fatores encontrados pelo modelo. Valor de bens declarados, estado civil e

idade apresentam relevância um pouco menor. É possível observar também que características da população regional tem forte influência nos atributos que foram encontrados pelo modelo.

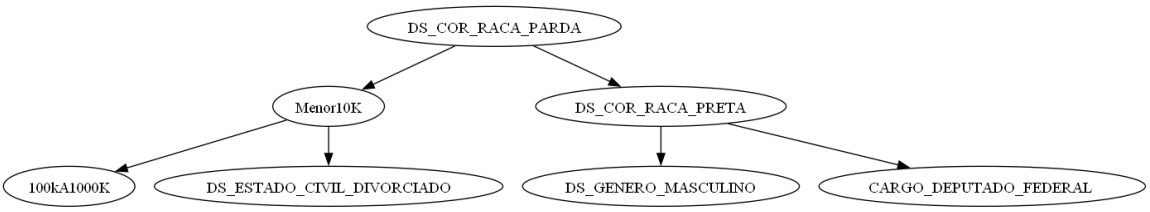

Figura 6. Nós iniciais da árvore gerada para a região norte

Observando os resultados referentes à região norte Figura 6, é possível observar que a característica presente em seu nó raiz, é a de cor parda, e, segundo o censo demográfico de 2010, 61% da população da região é parda. A princípio, é intuitivo dizer que os candidatos com essa característica obteriam mais votos. Porém isso não acontece, na realidade o modelo aponta que candidatos pardos tem tendencia a ter uma média menor de votação (36 caso o candidato seja pardo contra 185 caso não seja) que candidatos não pardos na região, demonstrando que as características do candidato podem influenciar tando positivamente quanto negativamente a votação.

6. Conclusão e Trabalhos Futuros

Neste trabalho foram analisados os padrões de votação das mesorregiões de Minas Gerais utilizando bases de dados disponibilizadas pelo Tribunal Superior Eleitoral (TSE) referentes ao processo eleitoral do ano de 2022.

A presença de grandes discrepâncias dentre as votações apresentadas pelas métricas de avaliação do modelo mostram uma heterogeneidade da votação. Os fatores mais relevantes para a votação dos candidatos, segundo o modelo, foram de forma geral: a **raça** do candidato e o **gênero** do mesmo. Foi verificado que o valor dos bens declarados, estado civil e idade também afetam a votação, porém em menor intensidade. Aspectos populacionais da região analisada demonstraram ter impacto na construção da árvore, um exemplo é a região Norte, que tem o atributo RACA_PARDA em seu nó raiz e, segundo o censo de 2010 tem 61% de sua população autodeclarada como parda. É possível perceber também que as características afetam tando positivamente quanto negativamente os candidatos.

Para trabalhos futuros vê-se a necessidade de uma análise aprofundada levando em consideração todos os municípios do estado, bem como a carreira política do candidato, a orientação política do mesmo e as propostas, para que o modelo possa lidar com uma perspectiva mais orgânica e real do problema. Também é possível utilizar mais bases de dados para análises aprofundadas, como o PIB e o IDH da região.

Referências

Barros, J. C. d. (2018). Concentração de recursos públicos no financiamento de campanhas eleitorais: análise dos padrões distributivos em 2016 e 2018. Trabalho de Conclusão de Curso, Escola de Direito, Fundação Getúlio Varga, Rio de Janeiro.

Cabral, L., Monteiro, J. M., da Silva, J. W. F., Mattos, C. L. C., and Mourao, P. J. C. (2021). Fakewhastapp. br: Nlp and machine learning techniques for misinformation detection in brazilian portuguese whatsapp messages. In ICEIS (1), pages 63–74.

Camargo, A., Silva, R., Amaral, E., Heinen, M., and Pereira, F. (2016). Mineração de dados eleitorais: descoberta de padrões de candidatos a vereador na região da campanha do rio grande do sul. Revista Brasileira de Computação Aplicada, 8(1):64–73.

Campos-Valdés, C., Álvarez Miranda, E., Morales Quiroga, M., Pereira, J., and Liberona Durán, F. (2021). The impact of candidates' profile and campaign decisions in electoral results: A data analytics approach. Mathematics, 9(8).

De Albuquerque Filho, J. E., LIMA, C., Perboire, L., and Pithon, P. (2020). Identificação de indícios de candidaturas de fachada nas eleições de 2018. Revista de Engenharia e Pesquisa Aplicada, 5(1):104–109.

Guedes, P. C. (2018). Aplicação de técnicas de data mining para previsibilidade eleitoral. Technical report, Universidade Federal de Santa Catarina.

Leal, I. H. d. S. (2018). O uso de aprendizagem de máquina para identificação e classificação de fake news no Twitter referentes a eleição presidencial de 2018. Trabalho de Conclusão de Curso, Instituto Ensinar Brasil, Faculdades Doctum de Caratinga, Caratinga.

Martins, E., Gonçalves, K., and Filho, R. M. (2019). Caracterizando a campanha presidencial brasileira em 2018 usando dados do twitter. In Anais do VIII Brazilian Workshop on Social Network Analysis and Mining, pages 131–142, Porto Alegre, RS, Brasil. SBC.

Nicolau, J. (2014). Determinantes do voto no primeiro turno das eleições presidenciais brasileiras de 2010: uma análise exploratória. Opinião Pública, 20(3).

Pereira, R., Alves, A., Vidal, D., Moura, F., Cabral, L., Paulino, R., Serrufo, M., and Figueiredo, K. (2023). Análise de sentimento de postagens de usuários no twitter combinando gpt-3 e aprendizado de máquina: Um estudo de caso sobre o 2º turno das eleições presidências brasileiras. In Anais do XIV Workshop sobre Aspectos da Interação Humano-Computador para a Web Social, pages 20–27, Porto Alegre, RS, Brasil. SBC.

Silva, A. V. V. d. (2021). Detecção de contas anômalas e influência em redes sociais na propaganda política e em eleições. Dissertação de Mestrado, Faculdade de Ciências e Tecnologia, Universidade Nova Lisboa, Lisboa.

Vasconcelos, F., Tavares, J., Ribeiro, M., Coutinho, F. J., and Clarindo, J. P. (2021). Candidata: um dataset para análise das eleições no brasil. In Anais do III Dataset Showcase Workshop, pages 160–168, Porto Alegre, RS, Brasil. SBC.

Zahedi, L., Mohammadi, F. G., Rezapour, S., Ohland, M. W., and Amini, M. H. (2021). Search algorithms for automated hyper-parameter tuning. arXiv preprint arXiv:2104.14677.