Trabalho de Conclusão de Curso apresentado para obtenção do título
de especialista em Data Science e Analytics – 2025
.

Importância de variáveis na predição conjunta de variáveis Categóricas com Floresta Aleatória

Giovana Marques Trindade[1]*; Sandro Ricardo Fuzatto[2]

[1] Cientista de Dados. Av. Edward Fru Fru Marciano da Silva, 537 – Jardim São Guilherme; 18074-621 Sorocaba, São Paulo, Brasil
[2] Engenheiro Agrônomo Especialista em Melhoramento Genético de Plantas. Rua Phenom, 35 – Portal das Araras; 79644-256, Três Lagoas, MS, Brasil
*autor correspondente: gvnmrqs@gmail.com

Importância de variáveis na predição conjunta de variáveis categóricas com Floresta Aleatória

Resumo

A mensuração de níveis de relevância de variáveis explicativas na classificação de variáveis resposta é uma importante etapa da seleção de variáveis para a modelagem estatística para evitar problemas de “overfitting” e falsas descobertas. Este estudo tem como objetivo abordar o uso do algoritmo floresta aleatória para este fim no caso em que se tem interesse em avaliar a relevância de variáveis explicativas na discriminação conjunta de múltiplas variáveis resposta. Um estudo de caso foi realizado com os dados da pesquisa perfil ideológico de 2013, conduzida pelo Datafolha. Primeiramente, o algoritmo foi aplicado de forma univariada para a modelagem de cada uma das variáveis de opinião pública de interesse com relação a variáveis socioeconômicas e de posicionamento político disponíveis. Observou-se que as variáveis a respeito da situação empregatícia do entrevistado, sua renda familiar e se se considera petista ou bolsonarista são as que mais influenciam em sua opinião a respeito de temas sociais e políticos polêmicos em detrimento das demais variáveis de cunho socioeconômico, como sexo e região, por exemplo. Além disso foram propostas técnicas de sumarização ou modelagem multivariada para avaliar o grau de relevância das variáveis explicativas na predição conjunta das 20 variáveis resposta. Procurou-se avaliar a coerência entre os resultados obtidos na análise multivariada e o que se observou na análise univariada e concluiu-se que incluir variáveis resposta pouco explicadas pelas variáveis explicativas pode levar a uma complexidade prejudicial em um modelo multivariado, gerando pouco poder preditivo e complexidade na inferência do grau de relevância das variáveis preditoras.
Palavras-chave: seleção de variáveis; análise multivariada; variáveis de opinião; variáveis socioeconômicas.

Feature importance on multivariate prediction of categorical variables with Random Forest

Abstract

The measurement of the relevance levels of explanatory variables in the classification of response variables is an important step in variable selection for statistical modeling to avoid overfitting and false discoveries. This study aims to address the use of the random forest algorithm for this purpose in cases where there is interest in evaluating the relevance of explanatory variables in the joint discrimination of multiple response variables. A case study was conducted with data from the 2013 Ideological Profile survey, conducted by Datafolha. Initially, the algorithm was applied univariately to model each of the public opinion variables of interest with respect to all available socioeconomic and political stance variables. It was observed that variables regarding the interviewee's employment status, family income, and whether they consider themselves a supporter of the Workers' Party (PT) or Bolsonaro are the most influential in their opinion on controversial social and political issues, compared to other socioeconomic variables such as gender and region. Additionally, summarization techniques or multivariate modeling were proposed to assess the relevance of explanatory variables in the joint prediction of the 20 response variables. The coherence between the results obtained in the multivariate analysis and those observed in the univariate analysis was evaluated, and it was concluded that including response variables poorly explained by the explanatory variables can lead to detrimental complexity in a multivariate model, generating low predictive power and complexity in inferring the relevance of predictor variables.
Keywords: feature selection; multivariate analysis; opinion variables; socioeconomic variables.

Introdução

A identificação de graus de importância de variáveis independentes na discriminação de variáveis resposta é crucial para inúmeras áreas, como a saúde, finanças e marketing. Este tipo de análise permite entender quais características são mais relevantes para a discriminação das informações de interesse, proporcionando perspectivas valiosas para a tomada de decisões. Para tal, métodos robustos e precisos são fundamentais para garantir a eficácia das análises e sua aplicabilidade prática.

O algoritmo floresta aleatória, ou em inglês “random forest”, introduzido por Leo Breiman (2001), revolucionou o campo da aprendizagem de máquina ao combinar múltiplas árvores de decisão, sendo portanto, um modelo *“*ensemble” (técnica de combinação de resultados de vários modelos) para criar um modelo preditivo não paramétrico com excelente desempenho e eficaz contra o problema de “overfitting” (Parmar et al., 2018). Hoje é amplamente utilizado para problemas de classificação e regressão, bem como para identificação de graus de importância de variáveis independentes na predição de variáveis resposta e seleção de variáveis “features”, como abordado nos artigos de Huljanah et al. (2019), Ibrahim et al. (2017) e Saraswat e Arya (2014).

O modelo de floresta aleatória pode ser estendido para a análise multivariada como abordado por Segal e Yuanyuan (2011). O estudo utiliza dados de um experimento ecológico para demonstrar como a floresta aleatória multivariada pode ser uma metodologia capaz de modelar simultaneamente múltiplas respostas e identificar, neste caso, padrões complexos de coocorrência entre espécies. O artigo também explora o uso do método para identificação de graus de importância das variáveis independentes envolvidas no modelo e para o entendimento de como as interações entre as variáveis podem afetar as respostas simultâneas.

Este trabalho tem como objetivo propor soluções para identificação de graus de relevância de variáveis explicativas na distinção entre categorias de variáveis dependentes, sem a intenção de construir modelos preditivos. Utilizando o algoritmo de floresta aleatória, a proposta é mensurar a importância relativa de cada variável independente com base nos dados da pesquisa Perfil Ideológico de 2023 do Datafolha. Essa abordagem é especialmente útil em contextos com grande número de variáveis, auxiliando na seleção de atributos informativos e na prevenção de “overfitting” e descobertas espúrias (Shi et al., 2019).

Material e Métodos

O método floresta aleatória é um algoritmo de aprendizado de máquina supervisionado que se baseia na construção de diversas árvores de decisão com amostras geradas via “bootstrap”, método estatístico introduzido por Efron (1979) baseado em reamostragem, em

geral utilizado para estimar a distribuição de uma estatística mas também amplamente utilizado em modelagem a fim de gerar múltiplos ensaios de um modelo para minimizar a variância das estimativas e aperfeiçoar a acurácia. As árvores de decisão do método floresta aleatória permitem a compreensão de padrões complexos existentes nos dados em termos das relações entre as variáveis explicativas e a variável resposta, que pode ser tanto categórica, o que qualifica o problema como um problema de classificação, quanto contínua, classificando o problema como regressão. O algoritmo combina os resultados das árvores de decisão aperfeiçoando sua precisão de predição e reduzindo o risco de "overfitting" e é capaz de fornecer uma medida de importância para cada variável explicativa utilizada baseado no ganho de informação promovido por sua participação no modelo.

O algoritmo, disponível no "Comprehensive R Archive Network" (CRAN), o repositório oficial de pacotes para a linguagem R, através do pacote "randomForest", desenvolvido por Liaw e Wiener (2002), foi aplicado aos dados da questionário do Perfil Ideológico de 2013, pesquisa realizada com eleitores de 16 anos ou mais sobre temas sociais e políticos polêmicos tais como aborto, posse de arma, racismo e homossexualidade, conduzida pelo Instituto de Pesquisas Datafolha, que realiza pesquisas nas áreas de Opinião Pública e Inteligência de Mercado há mais de 30 anos (Folha de São Paulo, 2022). A pesquisa é disponibilizada pelo Centro de Estudos de Opinião Pública [CESOP], núcleo de pesquisa da Universidade Estadual de Campinas que desenvolve pesquisa científica no campo do comportamento político e social, que fornece tanto o questionário e um relatório de análise descritiva univariada das informações coletadas em formato "pdf" quanto os microdados em formato "SPSS" mediante permissão concedida a partir do fornecimento de informações sobre a pesquisa a ser desenvolvida com sua utilização através do preenchimento de um formulário obrigatório.

Os microdados são compostos de variáveis de cunho socioeconômico como idade aberta, faixa de idade, sexo, ocupação, classe social domiciliar (composta a partir de questões pertinentes ao domicílio do entrevistado), escolaridade do entrevistado e do chefe da família, renda familiar, cor e região em que reside, e variáveis de posicionamento político como quão satisfeito o entrevistado está com o atual governo e em que ponto da escala entre bolsonarista e petista o entrevistado considera estar, além de variáveis de opinião a respeito de temas sociais e políticos coletadas através do nível de concordância em escala "likert" (técnica de medição amplamente utilizada em pesquisas sociais e psicológicas para medir atitudes, opiniões e percepções (Likert, 1932)) com 20 diferentes frases. As frases são apresentadas na Tabela 1 e os entrevistados declararam seu nível de concordância através das opções de resposta "Concorda totalmente", "Concorda em parte", "Nem concorda, nem discorda", "Discorda em parte", "Discorda totalmente" e "Não sabe". As variáveis de opinião em escala

“likert” foram modeladas com relação as demais variáveis, isto é, utilizadas como variáveis alvo para os modelos univariados e multivariados.

Tabela 1: Perguntas de opinião do entrevistado no questionário, variáveis alvo na modelagem

Pergunta	Frase
p901a	"Possuir arma legalizada deveria ser um direito do cidadão para se defender"
p901b	"A homossexualidade deve ser aceita por toda a sociedade"
p901c	"O aborto deve ser um direito da mulher"
p901d	"Os negros têm mais dificuldades de encontrar um trabalho do que os brancos"
p901e	"As leis ambientais devem ser mais flexíveis para o avanço do agronegócio"
p901f	"O aquecimento global é uma realidade"
p901g	"A família deve ser formada por um homem e uma mulher"
p901h	"Os humilhados serão exaltados"
p901i	"A ditadura, que durou de 1964 a 1985, trouxe benefícios para o Brasil"
p901j	"Os governos do PT no passado trouxeram benefícios para o país"
p901k	"O governo do ex-presidente Bolsonaro trouxe benefícios para o país"
p901l	"O Brasil corre o risco de se tornar um país comunista"
p901m	"As pessoas devem ter o direito de dizer o que pensam nas redes sociais, mesmo que isso ofenda alguém"
p901n	"No Brasil tem muita mordomia para bandido"
p901o	"O PT é um partido que não respeita a família cristã"
p901p	"Cada pessoa deve decidir se deve ou não tomar vacinas ou vacinar seus filhos"
p901q	"Hoje em dia as pessoas veem racismo em tudo"
p901r	"Política e valores religiosos devem andar sempre juntos para que o Brasil possa prosperar"
p901s	"A igreja deve fazer parte do projeto de vida das pessoas"
p901t	"As mulheres deveriam ocupar mais cargos de liderança"

Fonte: Dados originais da pesquisa

Inicialmente foram modeladas todas as variáveis de opinião em escala “likert” de maneira univariada com relação a todas as demais variáveis disponíveis, isto é, as variáveis de cunho socioeconômico e duas variáveis a respeito do posicionamento político dos entrevistados, como a satisfação com o atual governo e em que escala entre bolsonarista e petista o entrevistado considera estar. Os resultados preliminares desta etapa adiantam o que se deve esperar de sua sumarização, que foi realizada em seguida.

A avaliação do grau de relevância das variáveis preditoras envolvidas nos modelos univariados foi feita através da média no decréscimo do indicador de impureza de Gini dada sua adição aos modelos de árvore de decisão. A Impureza de Gini indica a probabilidade de se classificar incorretamente a variável resposta em determinado nó da árvore. Em outras palavras, a impureza de Gini mensura o grau de aleatoriedade das categorias da variável

resposta a partir da divisão na variável explicativa e é usada para determinar a qualidade da divisão, ou seja, quão bem a divisão separa os dados em diferentes categorias. É, portanto, uma métrica do ganho de informação promovido pela inclusão das “features” no modelo. Desta forma, dada a adição de uma variável preditora, o decréscimo na impureza de Gini com relação ao nó anterior pode mensurar sua relevância para a classificação da variável resposta, Breiman et al. (1984).

Quanto à análise multivariada ou à sumarização dos resultados dos modelos univariados, diversas abordagens foram estudadas. Primeiramente, com uso dos pacotes “MultivariateRandomForest” e “MulvariateRandomForestVarImp” do software R, disponíveis no CRAN, foram executadas 50 árvores de decisão multivariadas com diferentes subamostras dos dados, geradas via “bootstrap”, a fim de se extrair a importância de cada variável preditora. O primeiro pacote avalia a importância de cada “feature” pela quantidade de vezes que cada uma delas é escolhida como a melhor variável para divisão de um nó da árvore, uma vez que a escolha da variável leva em consideração o custo do nó, método proposto por Rahman (2017). O segundo baseia a medida de importância pela diferença entre a soma dos quadrados nos nós pai e filhos, onde a “feature” foi utilizada para dividir a árvore de decisão, e calcula a média dos resultados ao longo de todas as árvores da floresta (Sharmistha e Giles, 2021).

Além disso, o pacote “mvpart”, desenvolvido por De'ath (2014), do software R, disponível entre os pacotes arquivados do CRAN, também foi utilizado para gerar uma floresta aleatória multivariada. Com ele foi possível observar a qualidade do modelo gerado e a composição, em termos de variáveis envolvidas e sua atuação no modelo, da melhor árvore da floresta de acordo com sua capacidade de predição.

Por último, uma quarta alternativa foi considerada baseada na sumarização dos resultados dos modelos univariados ponderada pela qualidade de predição da floresta aleatória, isto é, sua acurácia, uma vez que a modelagem multivariada se apresentou computacionalmente cara e incapaz de identificar a falta de capacidade de predição de algumas das variáveis resposta envolvidas no vetor de variáveis pelas variáveis explicativas, o que levou à complexidade na inferência de níveis de importância e na identificação de grupos distintos de observações com respeito às variáveis resposta.

Resultados e Discussão

Uma análise descritiva das penetrações de cada categoria dentre os graus de concordância disponíveis como opção de resposta às variáveis de opinião é apresentada na

Figura 1, em que são observadas as proporções de cada resposta para cada uma das variáveis p901a a p901t.

Observa-se que para a maioria das variáveis de opinião, o nível neutro (“Nem concorda, nem discorda” ou “Não sabe”) é pouco manifestado, salvo algumas exceções como para as variáveis p901i ("A ditadura, que durou de 1964 a 1985, trouxe benefícios para o Brasil") e p901o ("O PT é um partido que não respeita a família cristã"). Além disso, é possível separar as variáveis em 3 grandes categorias: Variáveis cuja concordância prevalece (I), variáveis cuja discordância prevalece (II) e variáveis equilibradas entre concordância e discordância (III).

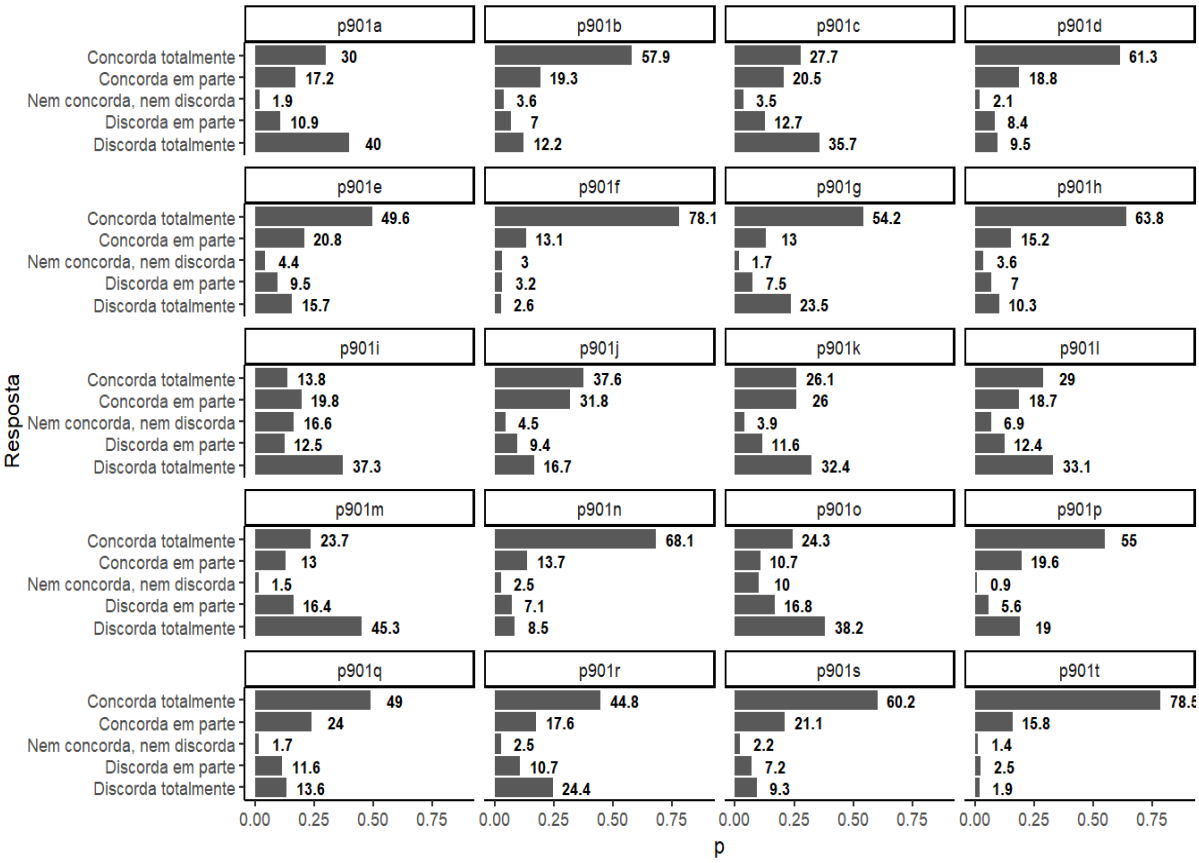

Figura 1. Análise descritiva univariada com porcentagem de cada categoria de resposta para as variáveis de opinião.

A Tabela 2 apresenta a relação entre as perguntas (e suas frases) e sua categoria dentre estas mencionadas. As categorias foram compostas segundo as seguintes regras: caso a razão entre a soma das penetrações das opções “Concorda totalmente” e “Concorda em parte” e a soma das penetrações das opções “Discorda totalmente” e “Discorda em parte” fosse superior a 1.5, a frase pertenceria à categoria I; caso a razão inversa fosse superior a 1.5, a frase pertenceria à categoria II; caso contrário, a frase pertenceria à categoria III. De acordo com a classificação, a maioria das frases têm prevalência de concordância, e a minoria

prevalência de discordância. No meio termo ficam as variáveis cuja concordância e discordância se equivalem.

Tabela 2. Categoria das perguntas segundo a distribuição de suas opções de resposta

Pergunta	Frase	Categoria
p901a	"Possuir arma legalizada deveria ser um direito do cidadão para se defender"	III
p901b	"A homossexualidade deve ser aceita por toda a sociedade"	I
p901c	"O aborto deve ser um direito da mulher"	III
p901d	"Os negros têm mais dificuldades de encontrar um trabalho do que os brancos"	I
p901e	"As leis ambientais devem ser mais flexíveis para o avanço do agronegócio"	I
p901f	"O aquecimento global é uma realidade"	I
p901g	"A família deve ser formada por um homem e uma mulher"	I
p901h	"Os humilhados serão exaltados"	I
p901i	"A ditadura, que durou de 1964 a 1985, trouxe benefícios para o Brasil"	III
p901j	"Os governos do PT no passado trouxeram benefícios para o país"	I
p901k	"O governo do ex-presidente Bolsonaro trouxe benefícios para o país"	III
p901l	"O Brasil corre o risco de se tornar um país comunista"	III
p901m	"As pessoas devem ter o direito de dizer o que pensam nas redes sociais, mesmo que isso ofenda alguém"	II
p901n	"No Brasil tem muita mordomia para bandido"	I
p901o	"O PT é um partido que não respeita a família cristã"	II
p901p	"Cada pessoa deve decidir se deve ou não tomar vacinas ou vacinar seus filhos"	I
p901q	"Hoje em dia as pessoas veem racismo em tudo"	I
p901r	"Política e valores religiosos devem andar sempre juntos para que o Brasil possa prosperar"	I
p901s	"A igreja deve fazer parte do projeto de vida das pessoas"	I
p901t	"As mulheres deveriam ocupar mais cargos de liderança"	I

Fonte: Dados originais da pesquisa

A Figura 2 apresenta os resultados obtidos da análise de correlações de Spearman, medida introduzida por Charles Spearman (1904) que de forma não paramétrica mensura o grau de associação monotônica, isto é, não linear, entre duas variáveis contínuas, discretas nominais ou discretas ordinais. Não é possível identificar nenhum par de variáveis com grau de associação muito forte, seja negativo ou positivo, e sim apenas alguns poucos pares com uma leve indicação de associação. São os casos dos pares de afirmações "O PT é um partido que não respeita a família cristã" e "Os governos do PT no passado trouxeram benefícios para o país" (-42,05%); e "A família deve ser formada por um homem e uma mulher" e "A homossexualidade deve ser aceita por toda a sociedade" (-39,44%), com associações

negativas, isto é, a medida que em uma das variáveis os respondentes concordam com a afirmação, na outra discordam; e também dos pares "O PT é um partido que não respeita a família cristã" e "O governo do ex-presidente Bolsonaro trouxe benefícios para o país" (48,2%); "O PT é um partido que não respeita a família cristã" e "O Brasil corre o risco de se tornar um país comunista" (47,02%); e "O Brasil corre o risco de se tornar um país comunista" e "O governo do ex-presidente Bolsonaro trouxe benefícios para o país" (40,19%), com associações positivas, isto é, a medida que se concorda com a primeira frase, também se concorda com a segunda.

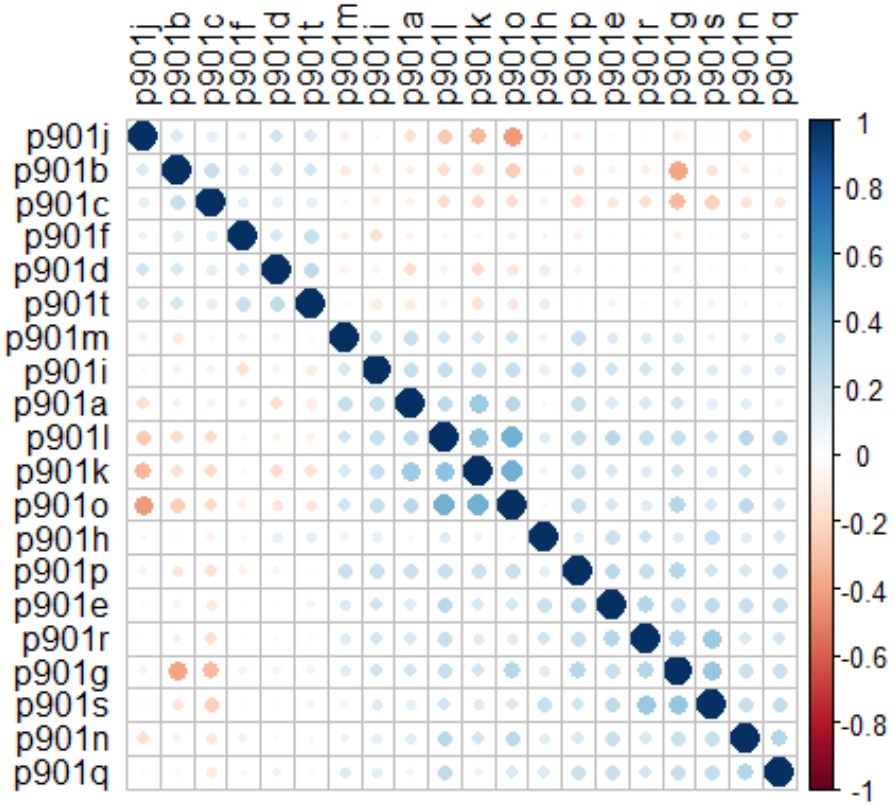

Figura 2. Correlograma de Spearman das variáveis de opinião.

O resultado do correlograma antecipa uma possível similaridade no grau de relevância das variáveis preditoras na classificação das variáveis pertencentes aos pares com forte correlação (positiva ou negativa), isto é, é possível que a ordem de importância das variáveis preditoras em suas classificações seja muito similar ou a mesma entre os pares "O PT é um partido que não respeita a família cristã" e "O governo do ex-presidente Bolsonaro trouxe benefícios para o país", por exemplo, variáveis altamente e positivamente correlacionadas.

As 20 variáveis de opinião foram modeladas de forma univariada com relação às 10 variáveis de cunho socioeconômico e de posicionamento político listadas na Tabela 3. As taxas de erro de classificação são apresentados na Tabela 4.

Tabela 3. Variáveis explicativas consideradas na modelagem das variáveis de opinião

Variável	Conteúdo
FX_IDADE	“16 a 24”, “25 a 34”, “35 a 44”, “45 a 59”, “60 anos ou mais”
SEXO	“Feminino”, “Masculino”
REGIAO	“Sudeste”, “Sul”, “Nordeste”, “Centro-Oeste”, “Norte
METROP	“Capital”, “Interior”, “Outros municípios da Região Metropolitana”
RCLASSE2	“A”, “B1”, “B2”, “C1”, “C2”, “D/E”
cor	“Branca”, “Preta”, “Parda”, “Amarela”, “Indígena”, “Outras”
rendaf	"Até R$ 1.302,00", "De R$ 1.303,00 até R$ 2.604,00", "De R$ 2.605,00 até R$ 3.906,00", "De R$ 3.907,00 até R$ 6.510,00", "De R$ 6.511,00 até R$ 13.020,00", "De R$ 13.021,00 até R$ 26.040,00", "De R$ 26.041,00 até R$ 65.100,00", "R$ 65.101,00 ou mais", "Recusa", "Não sabe"
pea	"Assalariado registrado", "Assalariado sem registro", "Funcionário público", "Autônomo regular", "Profissional liberal", "Empresário", "Free-lance/ bico", "Estagiário/ aprendiz", "Outros", "Desempregado (Procura emprego)", "Desempregado (Não procura emprego)", "Dona de casa", "Aposentado", "Estudante", "Vive de rendas"
p1	Como o entrevistado avalia o atual governo: “Ótimo”, “Bom”, “Regular”, “Ruim”, “Péssimo”, “Não sabe”
p906	Em que escala entre “Bolsonarista” e “Petista” o entrevistado considera estar: "Bolsonarista", "2", "3", "4", "Petista", "Nenhum", "Não sabe"

Fonte: Dados originais da pesquisa

Tabela 4. Taxa de erro nos dados de treino dos modelos de classificação das variáveis de opinião.

Variável	Taxa de Erro
p901t	21,18%
p901f	21,75%
p901n	33,24%
p901h	36,31%
p901d	39,66%
p901g	40,58%
p901s	42,37%
p901b	43,30%
p901p	46,01%
p901k	47,29%
p901j	49,36%
p901o	49,36%
p901a	50,93%
p901e	51,50%
p901q	53,21%
p901r	53,85%
p901l	55,35%
p901m	57,56%
p901c	60,70%
p901i	64,34%

Fonte: Dados originais da pesquisa

A Tabela 4 indica que a opinião sobre a frase “O aquecimento global é uma realidade” (p901t) é melhor modelada com relação às variáveis independentes do que a opinião sobre a frase “A ditadura, que durou de 1964 a 1985, trouxe benefícios para o Brasil” (p901i) por exemplo, que apresentou o modelo com maior taxa de erro.

Conforme elucidado por Esteves (2015) em “Sobre a Opinião Pública que já não o é – ao ter deixado de ser propriamente pública e também uma opinião” e Manstead (2018) em “The psychology of social class: How socioeconomic status impacts thought, feelings, and behaviour” a relação entre variáveis socioeconômicas e variáveis de opinião pública pode ser complexa e influenciada por múltiplos fatores. Isto leva a uma modelagem complexa e com possivelmente baixa acurácia, uma vez que a opinião expressa através de graus de concordâncias com as frases pode ser subjetiva e relativa.

Independentemente da acurácia dos modelos, é possível obter a importância das variáveis independentes na classificação através da impureza de Gini. A distribuição das importâncias das variáveis independentes ao longo das variáveis resposta é apresentada na Figura 3.

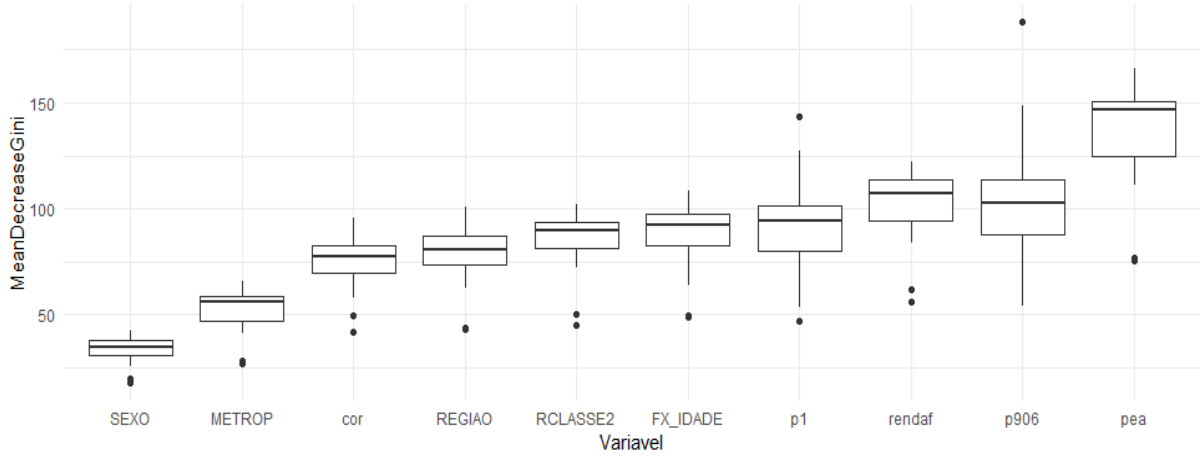

Figura 3. Distribuição da média no decréscimo da impureza de Gini dada a adição das variáveis independentes às árvores de decisão nos modelos univariados de classificação.

Nota-se que de maneira geral a variável “SEXO” é a variável que menos contribui para a classificação das variáveis resposta, enquanto que as variáveis “pea” (População Economicamente Ativa), “p906” (em que ponto da escala entre bolsonarista e petista o entrevistado considerada estar) e “rendaf” (faixa de renda familiar) se mostraram as mais importantes no processo de classificação, sendo portando responsáveis por maior ganho de informação para as árvores de decisão. Nota-se também que há um caso em que “pea” desempenhou um papel muito abaixo de sua média e que para uma das variáveis resposta, “p906” se destacou mais. Além disso, o segundo lugar é bem disputado entre as variáveis

"rendaf" e "p906", sendo possível afirmar que há poucas diferenças entre seus níveis de relevância na classificação univariada das variáveis resposta.

Os gráficos de barras das importâncias das variáveis independentes na classificação univariada de cada variável resposta são apresentados nas Figuras 4 e 5. Como a variável "pea" desempenha um papel importante na descriminação das variáveis de opinião como um todo, assim como a variável "SEXO" tem pouca importância de modo geral, é esperado que em uma sumarização dos resultados este resultado se replique.

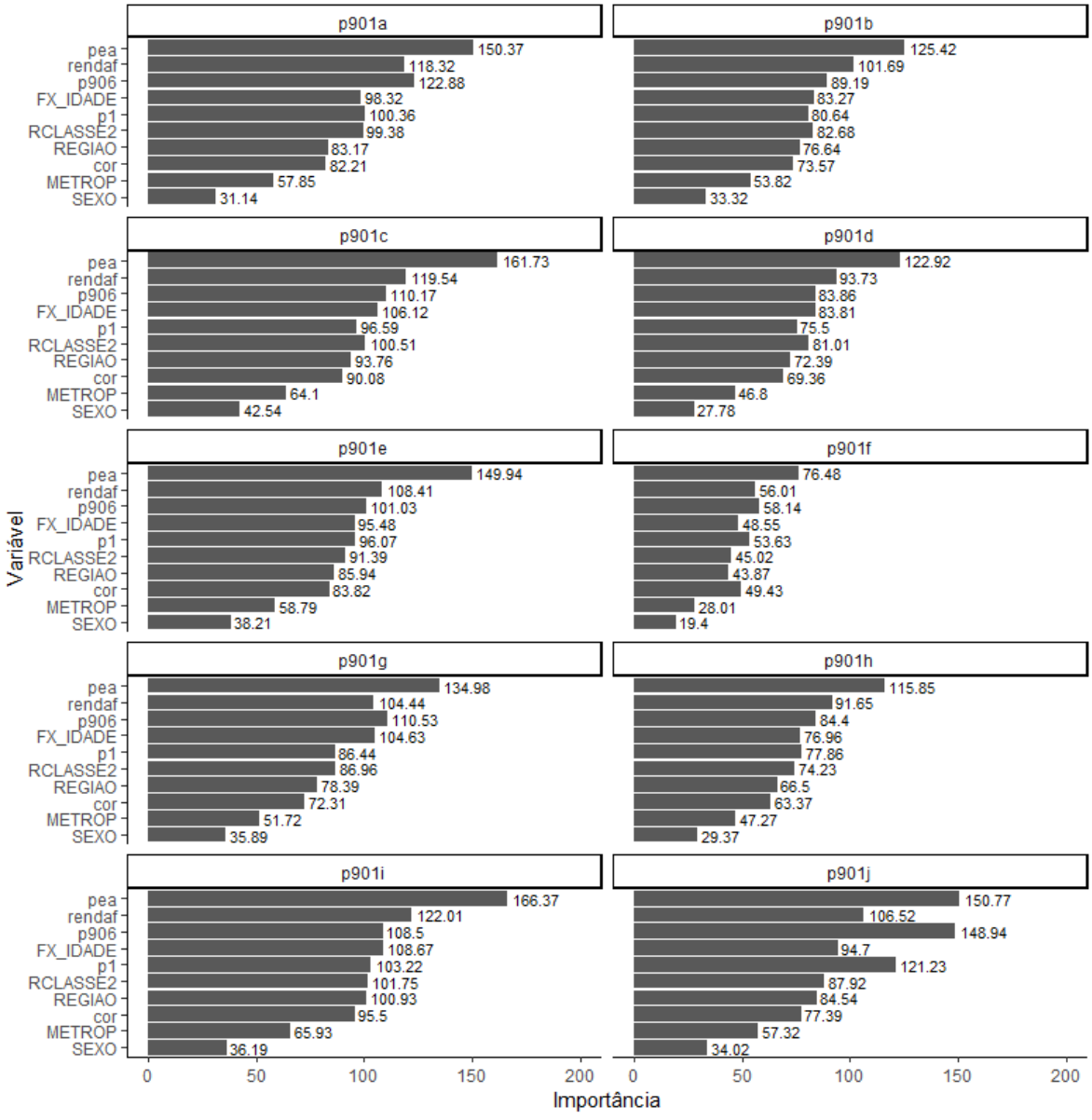

Figura 4. Gráficos de barras dos graus de importâncias das variáveis independentes na classificação das variáveis resposta p901a a p901j.

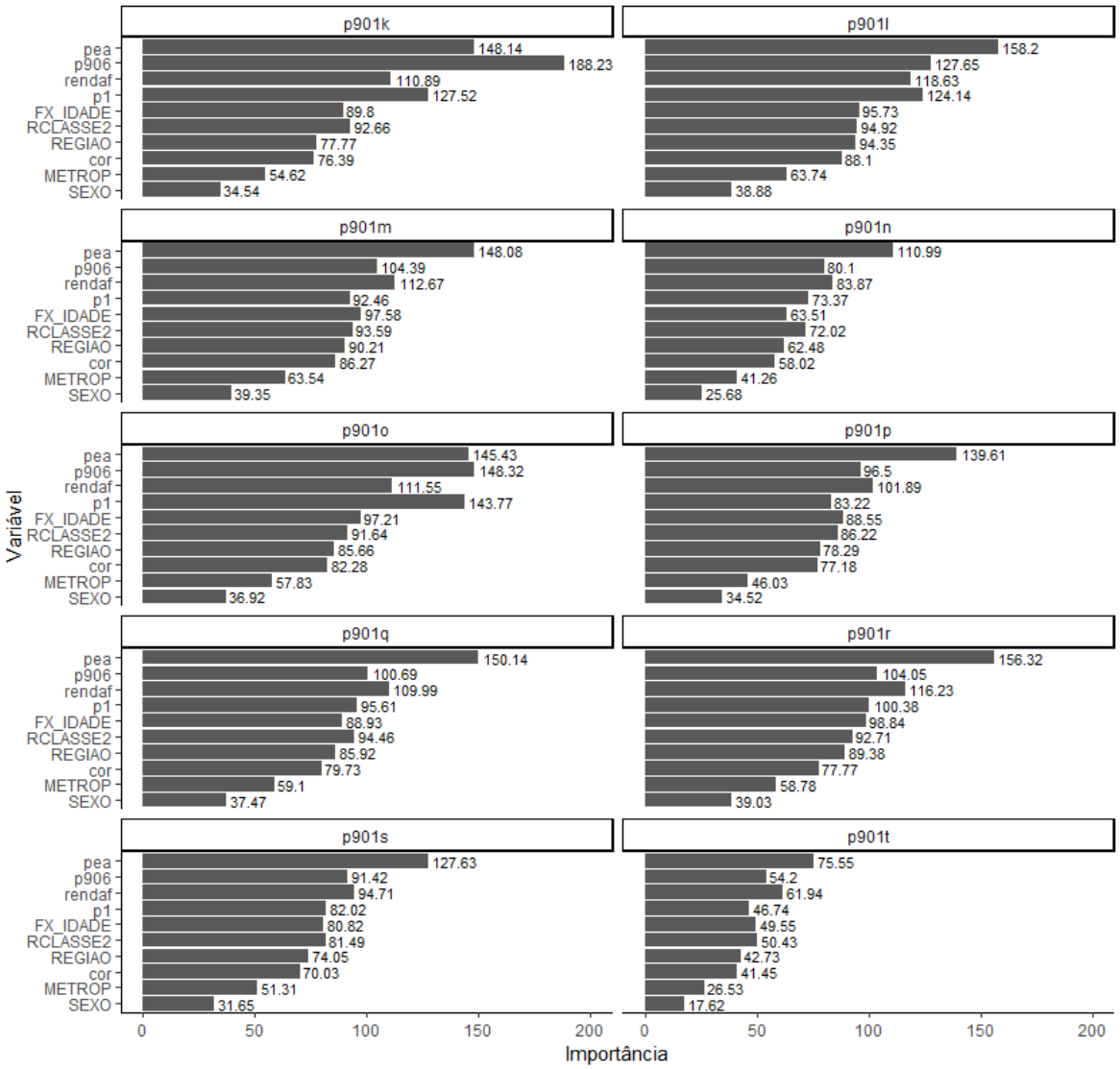

Figura 5. Gráficos de barras dos graus de importâncias das variáveis independentes na classificação das variáveis resposta p901k a p901t.

Os gráficos apresentados nas Figuras 4 e 5 apresentam uma ordenação padrão das variáveis explicativas segundo seus níveis de relevância. A variável “pea” por exemplo é a variável mais importante para a classificação de 18 das 20 variáveis resposta. Em alguns dos casos, como para as variáveis resposta “p901c”, “p901e”, “p901i”, “p901p”, “p901q” e “p901r”, ela está muito à frente da segunda variável mais relevante e em outros, como para a variável resposta “p901j”, “p906” (se o entrevistado se considera bolsonarista ou petista, e quanto) é tão relevante quanto. A variável “p906” é a variável mais importante para a classificação das opiniões “p901k” ("O governo do ex-presidente Bolsonaro trouxe benefícios para o país") e “p901o” ("O PT é um partido que não respeita a família cristã"), que são, junto da variável “p901j”, as únicas variáveis de opinião que citam explicitamente os governos petista e bolsonarista.

Através da modelagem multivariada via floresta aleatória das 20 variáveis de opinião com relação às 10 variáveis explicativas do pacote “MultivariateRandomForest” obteve-se a importância de cada variável explicativa para o ganho de informação nas árvores de decisão dado o custo dos nós das árvores. Este resultado é apresentado na Figura 6. Curiosamente, a variável “SEXO” que na modelagem univariada, num geral, se apresentou como a variável com pior capacidade preditiva das variáveis resposta, na análise multivariada se apresentou como a segunda variável mais importante. Enquanto que “pea”, que havia se apresentado como a variável mais relevante, consta em quinto lugar. As variáveis “p906”, “p1” e “FX_IDADE” se mantiveram entre as 5 mais relevantes.

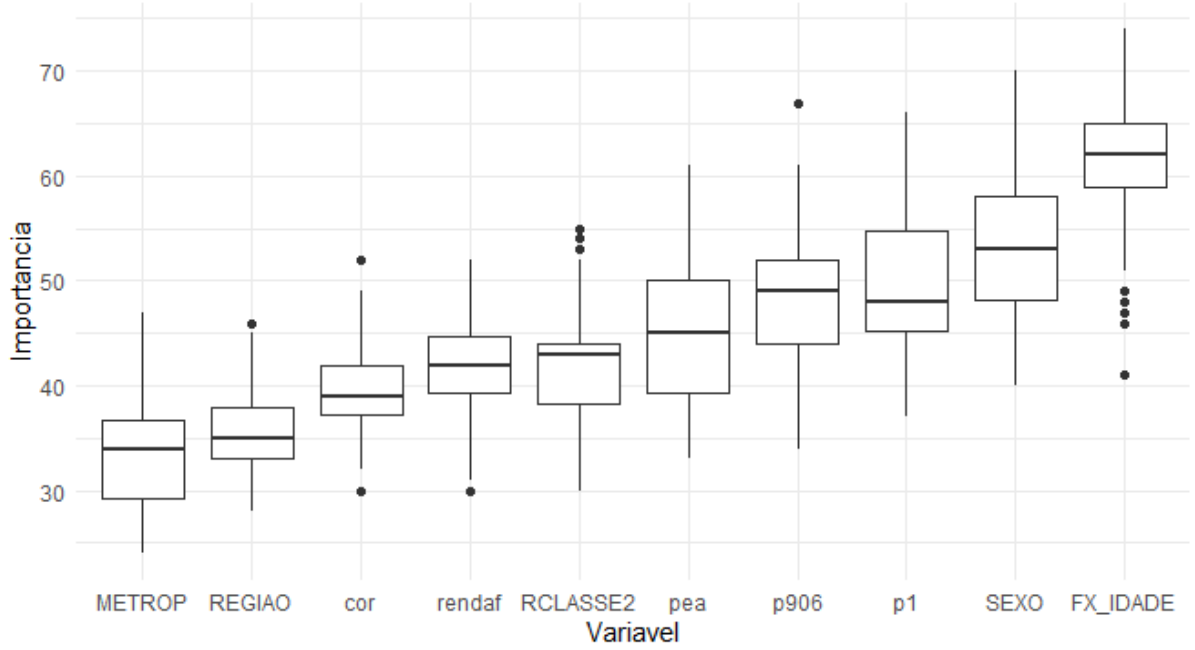

Figura 6. Distribuição do número de vezes em que cada variável se apresentou como a mais relevante para divisão de ramos das árvores de decisão multivariadas.

O pacote “MulvariateRandomForestVarImp” também foi utilizado para geração de uma floresta aleatória multivariada e propôs resultados de graus de relevância das variáveis explicativas diferentes dos propostos pelo pacote “MultivariateRandomForest”. De forma incoerente com relação aos resultados univariados, a variável “pea” consta como uma das menos relevantes enquanto que “REGIAO”, da qual se esperava pouca relevância, consta em 4º posição. Os resultados de todas as árvores de decisão são concatenados em um único valor para cada variável, desta forma, o nível de relevância inferido de cada uma das variáveis preditoras é apresentado na Tabela 5. É possível que os diferentes métodos de cálculo de nível de importância utilizados pelos pacotes tenha influenciado esta discrepância de

resultados especialmente neste caso em que muitas das variáveis resposta modeladas são pouco explicadas pelas variáveis “feature” conforme indica a Tabela 4.

Tabela 5. Nível de importância de cada variável preditora na classificação conjunta das variáveis resposta pelo pacote “MulvariateRandomForestVarImp”

Variável	Nível de importância
FX_IDADE	287.1725
p906	252.2325
p1	185.5881
REGIAO	184.1406
RCLASSE2	177.9197
cor	177.7300
METROP	119.5594
SEXO	117.6925
pea	105.9501
rendaf	104.6627

Fonte: Dados originais da pesquisa

O pacote “mvpart” também foi utilizado para modelar as 20 variáveis resposta de forma multivariada e avaliar a qualidade da melhor árvore de decisão obtida e sua composição em termos de variáveis “feature”. A árvore com melhor poder de predição envolve apenas uma variável explicativa e possui alta taxa de erro, como apresentado na Figura 7.

A variável “p906”, que contém os valores “1- Bolsonarista”, “2”, “3”, “4”, “5- Petista”, “Nenhum” e “Não sabe”, é a única variável utilizada na árvore de decisão, e divide as observações da seguinte maneira: caso a resposta seja “1- Bolsonarista” ou “2”, a observação vai para a direita, caso contrário, esquerda. A taxa de erro do modelo é muito alta, 92.9%, o que significa que apesar de esta ser a melhor árvore multivariada do modelo, não é uma boa árvore para classificação, o que indica que as variáveis resposta têm sua variabilidade pouco explicada pelas “features” envolvidas.

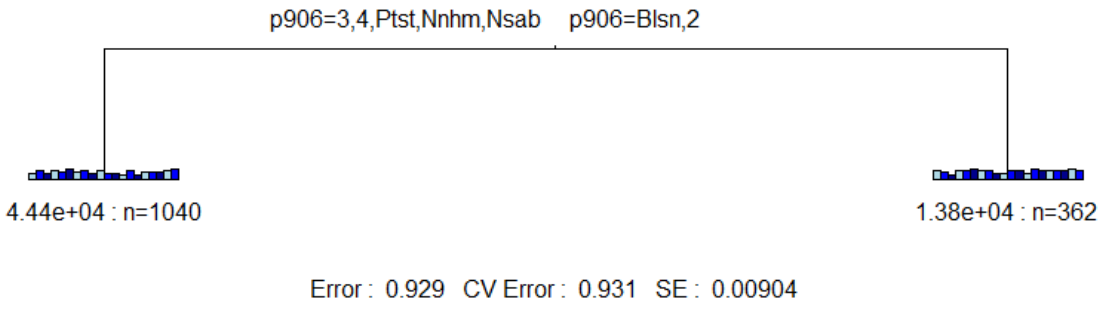

Figura 7. Melhor árvore de decisão multivariada obtida pelo pacote “mvpart”.

Uma sumarização dos resultados dos graus de relevância das variáveis preditoras na classificação das variáveis explicativas pode ser obtida através da média ponderada pela acurácia do modelo. Desta forma, modelos com maior poder preditivo têm seus resultados sobre a relevância das variáveis com maior participação na sumarização em detrimento de modelos com menor poder. Esta abordagem foi considerada uma vez que a modelagem multivariada se mostrou computacionalmente cara, levando muito tempo de execução para geração de florestas aleatórias com baixo número de árvores de decisão, e também apresentou resultados incoerentes, entre os dois diferentes métodos testados, e não esperados quando comparados com os resultados individuais das modelagem univariadas, o que pode sugerir que na análise multivariada, a presença de variáveis cuja variabilidade é pouco explicada pelas variáveis preditoras no conjunto de variáveis a serem modeladas conjuntamente pode adicionar uma complexidade prejudicial ao modelo. Os resultados da sumarização são apresentados na Tabela 6.

Tabela 6. Média ponderada das importâncias das variáveis preditoras pela acurácia dos modelos univariados

Variável	Nível de importância
pea	142.62
p906	109.89
rendaf	106.32
p1	97.05
FX_IDADE	91.22
RCLASSE2	88.46
REGIAO	81.91
cor	77.84
METROP	55.24
SEXO	34.60

Fonte: Dados originais da pesquisa

Considerações Finais

Os algoritmos de florestas aleatórias multivariadas apresentaram uma ordenação de importância das variáveis explicativas diferente do esperado em comparação com os modelos univariados e entre si, além de serem computacionalmente caros e demorados. A complexidade da modelagem multivariada pode levar a conclusões incertas, especialmente quando as variáveis preditoras explicam pouco a variabilidade das variáveis resposta. Por isso, optou-se por sumarizar os resultados dos modelos univariados ponderados por suas acurácias, atribuindo maior peso aos modelos com maior capacidade preditiva. Essa abordagem simplificou identificação de níveis de relevância para as variáveis explicativas e

sua ordenação, destacando "pea" (População Economicamente Ativa), "p906" (posição política) e "rendaf" (faixa de renda familiar) como as mais relevantes na classificação das variáveis de opinião modeladas.

Referências

Breiman, L.; Friedman, J.; Olshen, R.A.; Stone, C.J. 1984. Classification and Regression Trees. 1ed. Routledge. Belmont. California. US..

Breiman, L. 2001. Random Forests. Machine Learning 45: 5–32.

Centro de Estudos de Opinião Pública [CESOP]. 2023. PERFIL IDEOLÓGICO. Disponível em: <https://www.cesop.unicamp.br/por/banco_de_dados/v/4712>. Acesso em 11 jan. 2025.

De'ath, G. 2014. mvpart: Multivariate partitioning. R package version 1.6-2. Disponível em: <https://CRAN.R-project.org/package=mvpart>. Acesso em 04 abr. 2025.

Efron, B. 1979. Bootstrap Methods: Another Look at the Jackknife. The Annals of Statistics, 7(1), 1–26.

Esteves, J. P. 2015. Sobre a Opinião Pública que já não o é – ao ter deixado de ser propriamente pública e também uma opinião. Intexto 0(34): 276-293.

Folha de São Paulo. 2022. Estrutura e Missão do Datafolha. Disponível em: <https://datafolha.folha.uol.com.br/sobre/2022/07/estrutura-e-missao-do-datafolha.shtml>. Acesso em 11 jan. 2025.

Huljanah, M.; Rustam, Z.; Utama, S.; Siswantining, T. 2019. Feature Selection using Random Forest Classifier for Predicting Prostate Cancer. IOP Conference Series: Materials Science and Engineering, 546, 052031.

Ibrahim, D. R.; Ghnemat, R.; Hudaib, A. 2017. Software Defect Prediction using Feature Selection and Random Forest Algorithm. 2017 International Conference on New Trends in Computing Sciences (ICTCS).

Liaw, A.; Wiener, M. 2002. Classification and Regression by randomForest. R News, 2(3), 18-22. Disponível em: <https://CRAN.R-project.org/doc/Rnews/>. Acesso em 04 abr. 2025.

Likert, R. 1932. A Technique for the Measurement of Attitudes. Archives of Psychology, 140, 1-55.

Manstead, A. S. R. 2018. The psychology of social class: How socioeconomic status impacts thought, feelings, and behaviour. British Journal of Social Psychology 57(2): 267–291.

Parmar, A.; Katariya, R.; Patel, V. 2018. A Review on Random Forest: An Ensemble Classifier. Lecture Notes on Data Engineering and Communications Technologies: 758–763.

Rahman, R. 2017. MultivariateRandomForest: Models Multivariate Cases Using Random Forests. R package version 1.1.5. Disponível em: <https://CRAN.R-project.org/package=MultivariateRandomForest>. Acesso em 04 abr. 2025.

Saraswat, M.; Arya, K. V. 2014. Feature selection and classification of leukocytes using random forest. Medical & Biological Engineering & Computing 52(12): 1041–1052.

Segal, M; Yuanyuan, X. 2011. Multivariate random forests. Wiley interdisciplinary reviews: Data mining and knowledge discovery 1(1): 80-87.

Sharmistha, S.; Giles, H.; 2021. MulvariateRandomForestVarImp: Variable Importance Measures for Multivariate Random Forests. R package version 0.0.2. Disponível em: <https://CRAN.R-project.org/package=MulvariateRandomForestVarImp>. Acesso em 04 abr. 2025.

Shi, L.; Westerhuis, J. A.; Rosén, J.; Landberg, R.; Brunius, C. 2019. Variable selection and validation in multivariate modelling. Bioinformatics 35(6): 972-980.

Spearman, C. 1904. The Proof and Measurement of Association between Two Things. American Journal of Psychology 15(1): 72-101.