



Tudo é Ciência: do Big Bang ao Metaverso

1º Congresso Brasileiro de Ciência
e Saberes Multidisciplinares

Percepção de adolescentes quanto à identificação de notícias falsas na internet: um algoritmo preditivo baseado em aprendizado de máquina

Emanuel Santos Junior¹; 0000-0002-2173-7285

Myke Moraes de Oliveira²; 0000-0001-8426-9127

1 – UniFOA, Centro Universitário de Volta Redonda, Volta Redonda, RJ.

emanuelsantosjr@gmail.com

2 – ESALQ/USP, Universidade de São Paulo, Escola Superior de Agricultura "Luiz de Queiroz", Piracicaba, SP.

mykeoliveira@usp.br

Resumo: Atualmente, o fenômeno de disseminação de notícias falsas na internet é notável e requer cautela de usuários da internet, da sociedade e de governos para evitá-lo. O objetivo principal do estudo é propor um modelo preditivo a partir da identificação de características de indivíduos que, em pesquisas/entrevistas conduzidas pelo Cetic.br, declararam saber ver se a informação que encontrou na internet está certa. A elaboração do modelo foi efetuada por meio das técnicas de árvore de decisão e regressão logística, utilizando-se o total de 324 registros minerados de pesquisas TIC Domicílios e TIC Indivíduos de 2020. A modelagem foi realizada no programa RStudio. Inicialmente, o método de árvore de decisão foi aplicado para classificar os entrevistados e definir as variáveis explicativas mais relevantes, de acordo com o critério de partição e interrupção de crescimento da árvore. Em seguida, a regressão logística foi empregada para o desenvolvimento do algoritmo preditivo com as variáveis selecionadas pelo modelo de árvore de decisão. O objetivo da modelagem é prever a probabilidade do usuário da internet declarar saber identificar se uma notícia encontrada na internet é falsa. O modelo preditivo contém 8 características dos indivíduos (variáveis explicativas categóricas e métricas), possui acurácia de 81% e é aplicável para usuários adolescentes da faixa etária entre 11 e 17 anos.

Palavras-chave: Notícias falsas. *Fake news*. Árvore de decisão. Regressão logística.



Tudo é Ciência: do Big Bang ao Metaverso

1º Congresso Brasileiro de Ciência
e Saberes Multidisciplinares

INTRODUÇÃO

A produção de *fake news* (notícias falsas, em português) tem sido uma prática usual com o principal objetivo de desinformar ou confundir os leitores. De acordo com a intenção do autor e o grau de embasamento nos fatos verídicos, as *fake news* podem ser basicamente compreendidas como a fabricação de notícias falsas, manipulação de informações ou imagens ou, ainda, a publicidade enganosa para fins lucrativos (MOLINA *et al.*, 2021). Certamente, o fenômeno da propagação acelerada de *fake news* em mídias digitais assume uma particular importância devido às suas motivações e, sobretudo, às suas consequências potencialmente nocivas à sociedade (BEAUVAIS, 2022). Neste contexto, o letramento digital deve ocupar espaço nas escolas no que tange a formação dos estudantes na Era da Informação. Além disso, a educação digital também promove o desenvolvimento do pensamento crítico quanto ao uso da internet como fonte de informação (DUMITRU, 2020; EINAV *et al.*, 2020). Portanto, o reconhecimento de notícias ou informações falsas é uma competência essencial aos indivíduos da sociedade atual.

Nesta direção, a aplicação de técnicas de análise de dados, como os modelos de aprendizado de máquina, para checagem da informação e identificação de *fake news* é oportuna e atual. Dentre os diversos modelos de aprendizado de máquina, a técnica de árvore de decisão é amplamente empregada para tal finalidade (BLANCKE e VENTURINI, 2022; SUDHAKAR e KALIYAMURTHIE, 2023). Além disso, outras técnicas estatísticas, como a regressão logística, também podem ser aplicadas com o propósito de complementação aos modelos baseados em aprendizado de máquina. O método de árvore de decisão é uma técnica estatística preditiva supervisionada empregada em problemas de classificação, possibilitando a predição de variáveis categóricas ou numéricas. A estrutura do modelo assemelha-se à uma árvore constituída pelo nó-raiz e nós-folhas, os quais são determinados pelo algoritmo e hierarquicamente relacionados entre si. O crescimento da árvore ocorre a partir do nó-raiz em partições recursivas, de acordo com critérios de partição do algoritmo. Um dos



Tudo é Ciência: do Big Bang ao Metaverso

1º Congresso Brasileiro de Ciência
e Saberes Multidisciplinares

critérios de partição mais comumente aplicado é o índice de Gini, o qual é associado à medida de heterogeneidade dos dados (MITCHELL, 1997).

Resumidamente, o método de regressão logística é uma técnica estatística que difere dos demais métodos devido à sua variável dependente ser do tipo categórica. É uma técnica que permite a obtenção da probabilidade de ocorrência de um evento e a influência de cada variável independente (explicativa) sobre o evento analisado. A variável dependente é comumente binária e segue a distribuição de Bernoulli. O resultado da aplicação da técnica de regressão logística é uma combinação linear das variáveis e seus coeficientes. Os valores dos coeficientes são obtidos pelo método de máxima verossimilhança (FÁVERO e BELFIORE, 2021).

Com o propósito de monitorar o acesso, o uso e a apropriação das tecnologias de informação e comunicação (TIC) no Brasil, o Centro Regional de Estudos para o Desenvolvimento da Sociedade da Informação (Cetic.br), órgão também vinculado ao Comitê Gestor da Internet no Brasil e ao setor de Comunicação e Informação da UNESCO, realiza pesquisas contínuas no país desde 2005. As bases de dados anonimizados das pesquisas realizadas anualmente são disponibilizadas no sítio eletrônico do Cetic.br (www.cetic.br).

O objetivo deste estudo é elaborar um algoritmo preditivo a partir de características padrões de indivíduos que, na pesquisa do Cetic.br 2020 (NIC.br, 2021), declararam saber ver se a informação que encontram na internet está certa. O algoritmo determinará a probabilidade de um indivíduo acreditar saber averiguar a autenticidade de informações na internet, ou seja, declarar “Sim” para a pergunta de pesquisa. O modelo preditivo foi desenvolvido por meio das técnicas de árvore de decisão e regressão logística no programa RStudio.

MÉTODOS

Os dados utilizados neste estudo foram obtidos em duas bases de microdados da pesquisa TIC (Domicílios e Indivíduos) realizada pelo Cetic.br no ano de 2020 (NIC.br,



Tudo é Ciência: do Big Bang ao Metaverso

1º Congresso Brasileiro de Ciência
e Saberes Multidisciplinares

2021). A base de microdados TIC Domicílios contém originalmente 5590 registros, ao passo que a base de microdados TIC Indivíduos possui 4129 registros (NIC.br, 2021). A partir das bases de microdados originais, foram realizados os procedimentos de mineração e estruturação de dados. Com isso, a base de dados utilizada nesse estudo passou a conter 324 registros e 16 variáveis. As bases de microdados utilizadas possuem dados anonimizados e abertos (NIC.br, 2021), dispensando aprovação prévia de comitê de ética em pesquisa com seres humanos.

O modelo preditivo para classificação foi obtido pela técnica de árvore de decisão adotando-se o índice de Gini como critério de partição. O algoritmo de classificação utilizado na técnica de árvore de decisão foi o *Classification and Regression Trees* - CART (BREIMAN *et al.*, 1984). A variável R2_C (Você sabe ver se uma informação que você achou na internet está certa?) é a variável de pesquisa e, portanto, foi utilizada como o nó-raiz para o crescimento da árvore. A partir destes resultados, as variáveis mais importantes foram selecionadas e, em seguida, aplicou-se a técnica de regressão logística para o desenvolvimento do algoritmo preditivo. A modelagem matemática baseada em aprendizado de máquina foi elaborada no programa RStudio.

RESULTADOS E DISCUSSÃO

Uma breve análise descritiva dos dados mostra que os registros (observações) utilizados no estudo referem-se à adolescentes da faixa etária 11–17 anos com predominância de renda familiar de até R\$ 1.908,00 (232 observações). A maior parte dos entrevistados reside em áreas urbanizadas (272) distribuídas nas cinco regiões brasileiras e, ainda, 52% dos respondentes é do gênero feminino. A Figura 1 apresenta a quantidade de respondentes em relação à idade e respostas “Sim” para a variável de pesquisa (R2_C). Verifica-se a tendência de crescimento (em %) de respostas afirmativas para a variável de pesquisa com o aumento da idade. Cerca de 51% dos entrevistados com idade de 11 anos declararam saber verificar se a informação encontrada na internet é correta. Este percentual aumenta para 91% quando os entrevistados têm 17 anos de idade. Embora 90% dos entrevistados tenham acesso à internet no domicílio, é interessante observar que 56% (182) dos respondentes informaram não possuir os equipamentos computador de mesa,

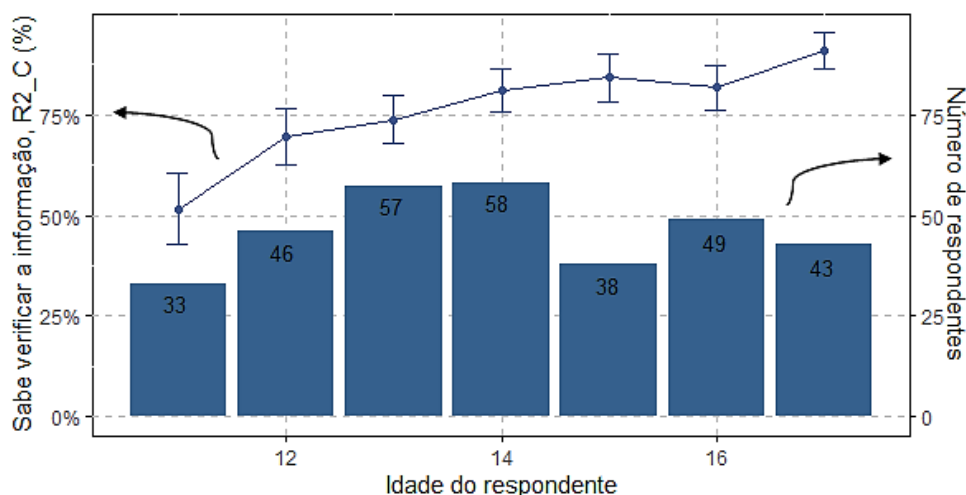


Tudo é Ciência: do Big Bang ao Metaverso

1º Congresso Brasileiro de Ciência
e Saberes Multidisciplinares

notebook e *tablet* na residência. Destes, 126 entrevistados declararam “Sim” para a variável de pesquisa. Portanto, o conjunto de dados sugere a preferência dos entrevistados pelo uso do aparelho celular para acesso à internet.

Figura 1 – Respostas “Sim” para a variável de pesquisa (R2_C) e quantidade de respondentes em relação à idade.



Fonte: os autores.

A Figura 2 mostra a árvore de decisão resultante com crescimento a partir da variável de pesquisa R2_C (nó-raiz). Este modelo de aprendizado de máquina foi aplicado como classificador para apurar as variáveis mais dominantes em relação ao atributo do nó-raiz, de acordo com o critério de partição do algoritmo. Verifica-se que 77% dos respondentes (291 observações) declararam saber checar as informações. Destes, 90% dos entrevistados declararam escolher palavras-chave para buscas na internet (variável R2_D), com probabilidade de 81% de ocorrência deste evento. Do total de 324 registros, 50% (162) também declarou ter lido jornais, revistas ou notícias na internet nos 3 meses que antecederam a entrevista (variável C9_D), com probabilidade de ocorrência de 90%. Portanto, a análise qualitativa da árvore de decisão indica que as variáveis R2_D e C9_D são as mais relevantes quanto à percepção dos entrevistados com relação à checagem das informações acessadas na internet. Por outro lado, a área (rural ou urbana) do domicílio assume importância

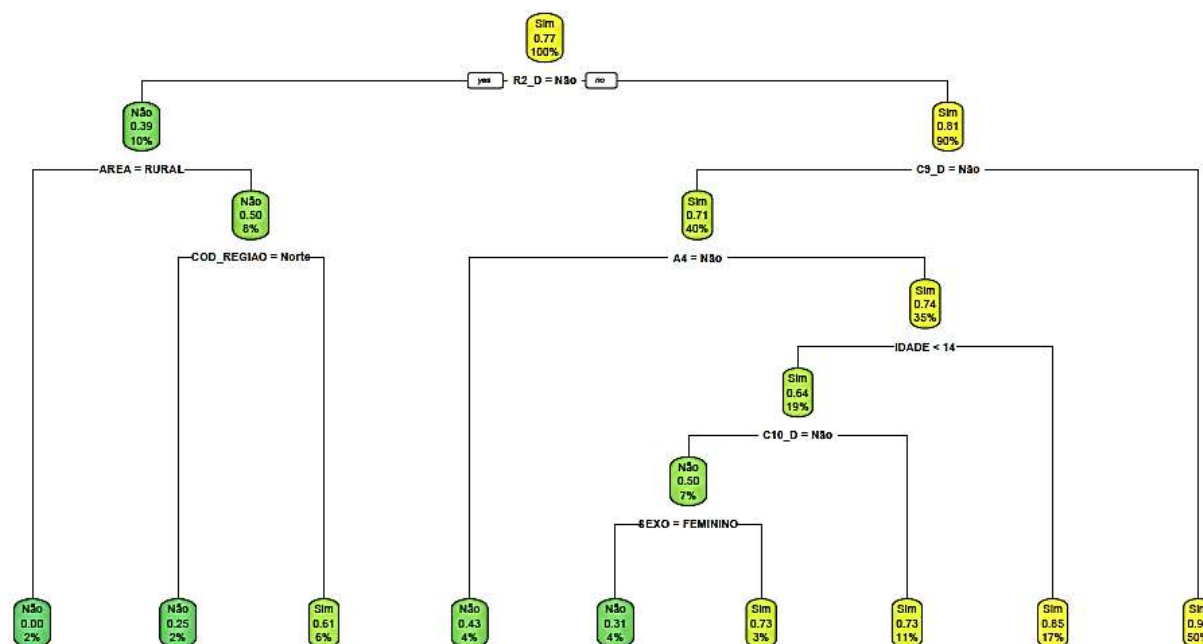


Tudo é Ciência: do Big Bang ao Metaverso

1º Congresso Brasileiro de Ciência
e Saberes Multidisciplinares

quando o entrevistado declara não saber usar palavras-chave para buscas na internet (R2_D).

Figura 2 – Árvore de decisão resultante a partir da variável R2_C (nó-raiz): “Você sabe ver se uma informação que você achou na internet está certa?”.



Fonte: os autores.

Com base na árvore de decisão mostrada acima, o Quadro 1 apresenta a descrição das variáveis e o grau de importância de cada uma calculado a partir do modelo previsto e, conseqüentemente, as variáveis selecionadas para elaboração do modelo preditivo pela técnica de regressão logística. O grau de importância é um valor absoluto e adimensional, o qual indica a intensidade da correlação entre a variável de pesquisa (R2_C) e as variáveis explicativas na modelagem. Os resultados corroboram a análise qualitativa, sugerindo que as variáveis R2_D, C9_D e IDADE estão, de fato, mais estreitamente vinculadas à variável de pesquisa. A acurácia é uma medida que indica o percentual dos registros para os quais a predição de um modelo está correta. Para a árvore de decisão obtida, a acurácia de 82% sugere que o modelo é robusto. Entretanto, este algoritmo não se aplica para obtenção de uma relação matemática

Tudo é Ciência: do Big Bang ao Metaverso

1º Congresso Brasileiro de Ciência
e Saberes Multidisciplinares

para fins preditivos. Então, a técnica de regressão logística torna-se uma alternativa adequada para tal finalidade a partir de variáveis categóricas.

Quadro 1 – Descrição das variáveis da árvore de decisão* e o grau de importância de cada variável.

Variável	Descrição	Importância
R2_D	Você sabe escolher que palavras usar para procurar alguma coisa na Internet? (Sim, Não)	30
C9_D	Nos últimos 3 meses, você utilizou a internet para ler jornais, revistas ou notícias? (Sim, Não)	15
IDADE	Idade do respondente	11
AREA	Área do domicílio (Urbana, Rural)	9
A4	Este domicílio tem acesso à Internet? (Sim, Não)	7
C10_D	Nos últimos 3 meses, você utilizou a Internet para fazer cursos à distância? (Sim, Não)	7
SEXO	Gênero do respondente (Masculino, Feminino)	6
COD_REGIAO	Região do domicílio (Norte, Nordeste, Centro-Oeste, Sudeste, Sul)	5

* O nó-raiz da árvore de decisão é a variável R2_C (Você sabe ver se uma informação que você achou na internet está certa?).

Fonte: os autores.

A aplicação da técnica de regressão logística, utilizando apenas as variáveis apresentadas no Quadro 1, gerou o modelo preditivo mostrado no Quadro 2. A relação matemática apresenta a variável de pesquisa (R2_C) em função das variáveis explicativas.

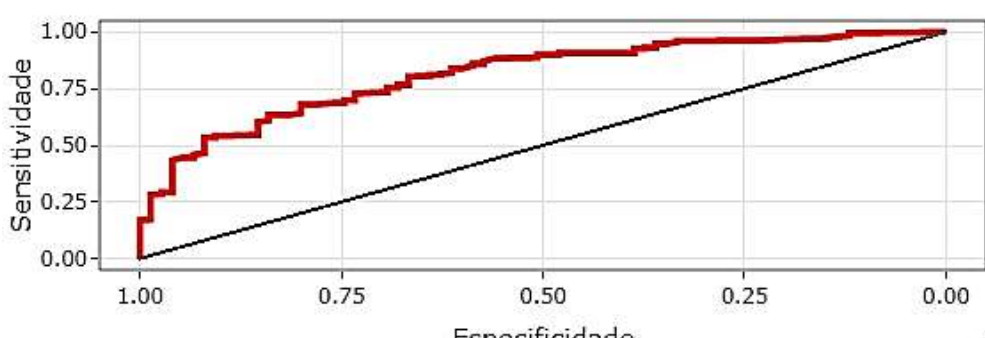
O modelo apresentado no Quadro 2 possui acurácia de 81% e coeficiente de Gini igual à 0,63, o qual corresponde a uma métrica da frequência com que um elemento escolhido aleatoriamente seja identificado incorretamente. Neste contexto, quanto menor o coeficiente de Gini, mais preciso é o modelo. A curva ROC é construída a partir da relação entre sensibilidade (acertos das previsões) e especificidade (previsão falso-positivo) do modelo. Quanto mais próximo de 1,0 a área sob a curva, mais robusto é o modelo preditivo. Isto posto, a partir do modelo preditivo proposto é possível realizar previsões com 81,3% de assertividade.

Tudo é Ciência: do Big Bang ao Metaverso

1º Congresso Brasileiro de Ciência
e Saberes Multidisciplinares

As limitações da modelagem proposta estão relacionadas ao tamanho amostral reduzido e à faixa etária restrita (11 – 17 anos) dos indivíduos da amostra.

Quadro 2 – Características do modelo preditivo obtido a partir das técnicas de árvore de decisão e regressão logística.

MODELO			
$pR2_C = 1/[1 + e^{-(\alpha + \beta_1 \cdot R2_D + \beta_2 \cdot \text{Área Rural} + \dots + \beta_{11} \cdot \text{Gênero Feminino})}]$			
COEFICIENTES			
Variável	Coeficiente (β_k)	Desvio-padrão	Nível de significância
Intercepto (α)	-5.33359	1.37919	0.000110
R2_D	1.75343	0.45273	0.000108
Área rural	-0.31196	0.40217	0.437927
C9_D	0.91678	0.32521	0.004817
Região Nordeste	-0.14346	0.41297	0.728295
Região Sudeste	0.90168	0.48428	0.062617
Região Sul	0.02663	0.50955	0.958325
Região Centro-Oeste	0.77788	0.59521	0.191250
A4	1.22633	0.45612	0.007176
IDADE	0.25122	0.08535	0.003248
C10_D	0.34994	0.32138	0.276208
Gênero Feminino	-0.58393	0.31251	0.061695
Curva ROC			
Área abaixo da curva: 0.813 Coeficiente de Gini 0.627 			

Fonte: os autores.



Tudo é Ciência: do Big Bang ao Metaverso

1º Congresso Brasileiro de Ciência
e Saberes Multidisciplinares

A partir do modelo preditivo proposto pode-se determinar, por exemplo, que a probabilidade de resposta “Sim” é de 83,3% para a variável de pesquisa (R2_C) para um indivíduo com as seguintes características: gênero feminino, 15 anos de idade, residente em área urbana, na região Nordeste, com acesso à internet no domicílio, que utiliza palavras-chave para realizar buscas na internet, tem lido notícias na internet e não realizou cursos na internet por conta própria. Portanto, considerando as limitações do estudo e baseando-se nestas classes de características dos indivíduos, o modelo apresentado possibilita a predição da percepção de um adolescente quanto à sua capacidade de saber se a notícia encontrada na internet é correta.

CONSIDERAÇÕES FINAIS

A publicação de notícias falsas na internet é um evento indiscutível que exige prudência dos usuários. A identificação de notícias falsas é, de fato, uma competência requerida na sociedade contemporânea, tornando a educação digital um pilar importante à promoção da cidadania. Utilizando-se dados minerados de pesquisas realizadas pelo Cetic.br no Brasil, foi elaborado um algoritmo a partir da técnica de árvore de decisão para classificar os entrevistados que declararam saber identificar notícias falsas na internet. Em seguida, um modelo preditivo foi desenvolvido por meio do método de regressão logística para determinar a probabilidade de um usuário da internet declarar saber reconhecer se a notícia encontrada é correta. A modelagem proposta possui acurácia de 81%, sendo aplicável para adolescentes da faixa etária entre 11 e 17 anos usuários da internet.

REFERÊNCIAS

- BEAUVAIS, C. Fake news: why do we believe it?. **Joint Bone Spine** 89(4): 1-6, 2022.
- BLANCHE, T.; VENTURINI, T. A network view on reliability: using machine learning to understand how we assess news websites. **Journal Of Computational Social Science** 5(1): 69-88, 2021.
- BREIMAN, L.; FRIEDMAN, J.H.; OLSHEN, R.A.; STONE, C.J. 1984. Classification and Regression Trees. 1ed. Routledge. New York, NY, USA.



Tudo é Ciência: do Big Bang ao Metaverso

1º Congresso Brasileiro de Ciência
e Saberes Multidisciplinares

DUMITRU, E.A. Testing children and adolescents' ability to identify fake news: a combined design of quasi-experiment and group discussions. **Societies** 10(3): 71-83, 2020.

EINAV, S.; LEVEY, A.; PATEL, P.; WESTWOOD, A. Epistemic vigilance online: textual inaccuracy and children's selective trust in webpages. **British Journal of Developmental Psychology** 38(4): 566-579, 2020.

FÁVERO, L.P.; BELFIORE, P. 2021. Manual de análise de dados: estatística e modelagem multivariada com Excel, SPSS e Stata. 1ed. LTC. Rio de Janeiro, RJ, Brasil.

MITCHELL, T.M. 1997. Machine Learning. 1ed. McGraw-Hill Science. New York, NY, USA.

MOLINA, M.D.; SUNDAR, S.S.; LE, T.; LEE, D. "Fake News" is not simply false Information: a concept explication and taxonomy of online content. **American Behavioral Scientist** 65(2): 180-212, 2019.

NIC.br, Núcleo de Informação e Coordenação do Ponto BR. 2021. Pesquisa sobre o uso das tecnologias de informação e comunicação nos domicílios brasileiros: pesquisa TIC Domicílios (Edição COVID-19 - Metodologia adaptada), ano 2020. Disponível em: <https://cetic.br/pt/arquivos/domicilios/2020/domicilios/>. Acesso em: 31 julho 2022.

SUDHAKAR, M.; KALIYAMURTHIE, K.P. Efficient prediction of fake news using novel ensemble technique based on machine learning algorithm. **Lecture Notes in Networks and Systems** 401: 1-8, 2023