Análise de dados para determinar a influência dos fatores psico-socioambientais no posicionamento ideológico

Autor: Tailan João da Silva Oliveira
Orientador: Diego Stéfano

Resumo

A presente pesquisa tem por objetivo estudar como os fatores psico-socioambientais se relacionam com o alinhamento ideológico individual, avaliando o quão relevantes podem ser estes fatores para determinar o posicionamento e se é possível criar uma regra de associação entre esses dois conjuntos de variáveis. Para isto, realizou-se uma enquete que colheu dados relativos a posicionamento ideológico e características psicológicas e sociais de 87 indivíduos. Os dados foram analisados através de técnicas de estatística e por meio de algoritmos de inteligência artificial. Ao final da pesquisa concluiu-se que os elementos estudados de fato afetam as convicções dos indivíduos, todavia, essa correlação não parece ser determinante o suficiente para que seja criada uma regra absoluta de associação entre os dois grupos de variáveis.

Palavras-Chave: Alinhamento Ideológico, fatores psico-socioambientais, estatística, inteligência artificial.

Abstract

This research aims to study how psycho-socio-environmental factors are related to individual ideological alignment, assessing how relevant these factors can be to determine positioning and whether it is possible to create an association rule between these two sets of variables. For this, a survey was carried out that collected data on the ideological position and psychological and social characteristics of 87 individuals. Data were analyzed using statistical techniques and artificial intelligence algorithms. At the end of the research, it was concluded that the elements studied affect the convictions of individuals, however, this correlation does not seem to be decisive enough to create an absolute rule of association between the two groups of variables.

Keywords: Ideological Alignment, psycho-socio-environmental factors, statistics, artificial intelligence.

1 INTRODUÇÃO

Originado na Assembleia Nacional durante a revolução Francesa, por volta de 1789, o uso político dos termos “direita” e “esquerda” servia para classificar os dois grandes grupos partidários presentes nas discussões do movimento quanto aos seus posicionamentos diante do sistema de governo vigente, divididos em conservadores e revolucionários, respectivamente. Muitos anos se passaram, desde então, os termos foram ressignificados e atualmente englobam um conjunto muito vasto de ideias, sendo comumente usados em discussões sociais e políticas com diferentes sentidos.

No cenário atual de intensificação da polarização política, os termos supracitados têm sido comumente utilizados em sentido pejorativo, quando a escolha por um ou outro é associada a fatores como escolaridade ou renda, por exemplo. Diante do fenômeno exposto, torna-se necessário verificar a premissa, isto é, os fatores ambientais realmente definem o posicionamento político-social de uma pessoa ou grupo? E para responder este questionamento, pretende-se utilizar da tecnologia da informação aplicada à análise estatística.

Como definiu Joaquim Filipe (2012), os processos de tratamento e análise de dados são fundamentais a qualquer processo investigativo, e com base nisto, o presente trabalho objetiva corroborar ou refutar a supramencionada premissa analisando um determinado público quanto ao seu posicionamento no espectro político-econômico, avaliando como se classificam, quais fatores sociais influenciam mais fortemente este posicionamento e se é possível aplicar a correlação estudada em um algoritmo preditivo.

Definida a problemática e os dados a serem avaliados, as etapas subsequentes do processo de análise estatística consistem em extração, sintetização, análise dos dados e elaboração do relatório (Silvestre, 2007). É especialmente nestas primeiras etapas apresentadas que a tecnologia da informação se tornará essencial para otimizar os procedimentos, iniciando com os formulários digitais que possuem fácil acesso, criação e amplo alcance, para a etapa de coleta dos dados, posteriormente, com as ferramentas e bibliotecas da linguagem de programação Python, que facilitarão o tratamento, avaliação dos dados e geração dos algoritmos para teste da hipótese.

Ao final das fases analíticas do processo, as informações geradas, bem como as conclusões obtidas a partir delas foram condensadas e apresentadas na conclusão deste artigo. O objetivo deste trabalho é aplicar a computação na análise de dados

para investigar a hipótese de que há uma correlação entre os fatores psicosocio-ambientais e o posicionamento ideológico, contribuindo para o campo de estudo do aprendizado de máquina e ciência de dados, ao sintetizar o conhecimento técnico sobre o funcionamento dos algoritmos utilizados, e para o campo das ciências políticas, ao investigar as premissas que são objeto de estudo na área.

2 FUNDAMENTAÇÃO TEÓRICA

Para compreender melhor a proposta do trabalho e organizar as questões que foram abordadas no projeto, o referencial das pesquisas foi desenvolvido em três frentes: A delimitação dos conceitos de posicionamento político que serão utilizados, a análise da interferência dos fatores psicossociais no alinhamento ideológico e, por fim, a fundamentação matemática e computacional dos procedimentos de análise.

2.1Posicionamento político

Em processos relacionados a gestão do conhecimento na área da tecnologia, para além das habilidades técnicas, o conhecimento do ambiente de negócio das informações pode ser elencado como a principal qualidade buscada nos profissionais que trabalham com dados (FERREIRA,2003). Por esta razão, torna-se imprescindível para o projeto proposto, aprofundar-se no campo das ciências políticas e compreender os conceitos essenciais para a formulação de uma boa metodologia investigativa.

Antes de iniciar o processo de pesquisa, visando abrandar a subjetividade e delimitar a definição do que seria considerado posicionamento político, propõe-se a redução da vastidão de abrangência deste termo a uma representação mais simples, seguindo o pensamento de Bobbio: “Não faz mal colocar um pouco de ordem em nossos raciocínios [..]. O esquematismo está inerente, no caso, à simplificação de uma realidade complexa a que nos induz qualquer raciocínio através de díades ou tríades” (BOBBIO,2005).

Prosseguindo nesta reflexão, Bobbio compreende a pluralidade das linhas de pensamento da modernidade, mas busca preservar a validade da tradicional díade direita e esquerda, sob uma nova perspectiva. Em sua obra “Direita e Esquerda: razões e significados de uma distinção política”, o filósofo classifica os dois grupos como almejantes por reformas sociais, todavia, enquanto a esquerda busca a promoção da igualdade social, a direita visa a liberdade individual (BOBBIO,2012).

A definição de Bobbio, insere uma classificação dicotômica ao tema e introduz uma distinção clara entre os lados, entretanto, para este projeto, faz-se necessário objetificar ainda mais os conceitos, e um modo de se fazer isto, é desenvolver a conceituação do autor, expondo sob quais aspectos da sociedade recaem as referidas noções de igualdade e liberdade, assim como faz o economista Bresser-Pereira, ao afirmar:

> A Direita é o conjunto de forças políticas que, em um país capitalista e democrático, luta sobretudo para assegurar a ordem, dando prioridade a esse objetivo, enquanto a esquerda reúne aqueles que estão dispostos, até certo ponto, a arriscar a ordem em nome da justiça[...]. Adicionalmente, a esquerda se caracteriza por atribuir papel ativo na redução da injustiça social, enquanto a direita, percebendo que o Estado, ao se democratizar, foi saindo do controle, defende o papel do Estado mínimo, limitado à ordem pública, dando preponderância para o mercado na coordenação da vida social. (BRESSER-PEREIRA, 2006).

Como observado, a visão do autor parece complementar a menção anterior de Bobbio, explicitando desta vez, o modo como se manifestam materialmente as correntes pensamento de direita e esquerda, destrinchando a definição em dois eixos, o social e o econômico.

No plano social, a Esquerda é caracterizada por priorizar a redução das desigualdades sociais, normalmente atribuindo ao Estado o dever de fazê-lo. A Direita, por sua vez, enxerga tal igualdade como inalcançável em sua plenitude, e as intervenções estatais para este fim, limitantes das liberdades e perigosas à ordem social. Já no plano econômico, a Direita enxerga no livre mercado a capacidade de autorregulação e fomento ao desenvolvimento, enquanto a Esquerda acredita que o desequilíbrio de poder entre os que concorrem no livre comércio, torna a competição desleal e sufoca os produtores menores, à proporção que reafirma as desigualdades.

Definidas as bases que fundamentam a ideologia política de ambos os lados, segundo os referidos cientistas, o outro fator a ser analisado para a formulação da hipótese é identificar como os fatores sociais podem influenciar no posicionamento político-ideológico individual.

2.2Comportamento

Segundo Lipset (1967), o alinhamento de grupos a determinadas vertentes políticas pode ser compreendido através do estudo das clivagens históricas. Segundo ele, determinados conflitos históricos teriam culminado nas cisões entre os grupos sociais, e por conseguinte, na emergência de organizações destinados a representar politicamente os interesses de cada lado. Das ideias do sociólogo, é possível extrair que a autoidentificação com determinado grupo dentro do sistema de clivagem, é um determinante para a inclinação de uma pessoa a um espectro político partidário.

Inglehart (1977), por sua vez, condiciona os valores individuais ao ambiente onde ocorre a socialização do indivíduo, dividindo esses valores em materialistas e pós materialistas. Para o autor, cidadãos socializados em ambientes de escassez material e baixa proteção social, tendem a apoiar a segurança física e o crescimento econômico, enquanto aqueles que crescem com maior proteção social e com suas necessidades materiais supridas, tendem a valorizar a questões relativas à expressão e realização pessoal, qualidade de vida e participação política social ativa.

Consoante Freire (1997), as bases sociais que fundamentam as ideologias, tanto de Direita quanto de Esquerda, sofreram alterações no decorrer do tempo, mas ainda e possível identificar grupos específicos que tendem a compor cada um dos lados, cabendo aqui citar, a título de exemplificação, os jovens, minorias étnicas, classe média e aqueles com maior nível de instrução como componentes das bases da Esquerda, e as maiorias étnicas, pequenas burguesias e os mais religiosos, como componentes das bases de Direita.

Em todas as literaturas supracitadas é possível observar uma argumentação em torno da forte presença de diferentes fatores na percepção política, como grau de instrução ou poder econômico, o que nos inclina a uma resposta positiva para o questionamento levantado nesta pesquisa, todavia, um resultado mais conclusivo dependerá da análise e processamento de dados.

2.3 Análise Estatística

Consoante Bernoville (1939), “A estatística é um conjunto de métodos e processos quantitativos usados para medir os fenômenos coletivos”. Em outras palavras, a estatística é uma ciência que visa a compreender um evento por meio da interpretação matemática de dados extraídos de uma população.

Em “Estatística Aplicada”, o professor Falco (2008) afirma que a teoria estatística atual pode ser dividida em duas áreas, a estatística descritiva e a indutiva, sendo esta primeira focada na interpretação de dados com o objetivo de compreender o fenômeno dentro da amostra observada, enquanto a segunda extrapola o espaço da amostra, e visa a testar hipóteses, e realizar predições. Nesta pesquisa, realizaram-se estudos em ambas as frentes, porém, para a análise indutiva, foi feito uso do aprendizado de máquina, em lugar de métodos estatísticos convencionais.

Segundo Alpaydin (2009), a resolução de problemas em um computador prescinde a construção de um algoritmo, isto é, um conjunto sequencial de instruções capaz de transformar dadas entradas em saídas esperadas, todavia, em determinadas ocasiões, não há como definir um algoritmo de aplicação global para todas as circunstâncias de um mesmo problema, ou talvez não seja viável fazê-lo, e para situações como esta, a solução computacional proposta é o aprendizado de máquina.

Aprendizado de máquina, ou *Machine Learning*, designa a técnica que consiste basicamente no fornecimento de grandes conjuntos de dados estruturados em entradas e saídas conhecidas para um computador que através de um algoritmo de aprendizado pré-definido é capaz de identificar automaticamente os padrões e extrair um modelo capaz de explicar o fenômeno observado em cada instância do problema com uma boa taxa de precisão. Parafraseando o autor: "Podemos não ser capazes de identificar o processo completamente, mas acreditamos ser capazes de criar uma aproximação boa e utilizável".

Alpaydin (2009) enumera, em rol exemplificativo, os principais tipos de aplicações do aprendizado de máquina segundo seus objetivos. No contexto desta pesquisa, utilizaram-se modelos do campo da análise preditiva, mais especificamente, o aprendizado por classificação, que, segundo o autor, é definido como uma técnica para a criação de uma regra de associação para classificar dados futuros baseado em padrões extraídos de análises de amostras históricas.

Dentre os algoritmos de classificação existentes, o escolhido para a análise preditiva dos dados foi o aprendizado por árvore de decisão. A técnica caracteriza-se por categorizar os dados por meio do uso de uma estrutura bastante similar às árvores de busca, nela, testes lógicos com as variáveis independentes do problema compõem os nós raiz e intermediários da árvore, enquanto as categorias das variáveis de

respostas são representadas pelos nós folhas. Uma visualização do arranjo de árvore pode ser obtida em: https://www.scielo.br/j/cr/a/CRMPjkLfN4jppFwB5qzdpYk/.

Ao ser submetido ao algoritmo classificador, uma das características do dado de entrada passa pelo teste do nó raiz, que, dependendo da resposta, encaminha a amostra para o próximo teste representado por um de seus nós filhos. O algoritmo realiza este processo repetidamente até que os testes terminem por apontar para um nó folha, que representa a classificação dada pelo programa à amostra recebida (MAIMON e ROKACH, 2014).

3 Metodologia

O processo de pesquisa e teste da hipótese seguiu um conjunto sequencial de passos que, como supramencionado, pode ser agrupado em 4 etapas: a extração de dados, análise exploratória de dados, análise preditiva, teste de eficiência dos algoritmos e, por fim, avaliação dos resultados.

Para a etapa de coleta de dados desenvolveu-se um formulário no Google com perguntas sobre as características de interesse para o estudo. O formulário foi distribuído para o máximo de pessoas possível, sem exigência de características específicas para os participantes, uma vez que, a heterogeneidade do espaço amostral seria essencial para a pesquisa.

Após a coleta, os dados foram analisados estatisticamente em duas frentes, primeiro em um conjunto de análises bivariadas, para determinar como cada variável influenciava individualmente o fenômeno de interesse, e posteriormente em uma análise multivariada utilizando a inteligência artificial, para determinar o efeito conjunto dos fatores situacionais no posicionamento ideológico.

Após cada fase analítica, os resultados foram avaliados matematicamente, e as métricas obtidas foram interpretadas dentro do contexto do problema, de modo a determinar por rejeitar ou confirmar a tese levantada no início deste artigo.

4 Extração de Dados

Na etapa da coleta das informações, os principais desafios envolviam ampliar o alcance da pesquisa e padronizar as respostas de modo que fossem objetivas e condizentes com o propósito da análise.

O formulário utilizado possuía 39 perguntas, agrupadas em duas subseções, a primeira focada na obtenção das informações psico-socioambientais, e a segunda na compreensão do posicionamento político do respondente.

Para aumentar a abrangência e o engajamento, o formulário não exigiu a identificação dos entrevistados e contou apenas com perguntas objetivas, de modo a evitar a exposição de dados sensíveis e reduzir a possibilidade de interpretação ambígua.

Ao final da fase de respostas, foram obtidas 87 amostras em que foram questionados do entrevistado, na primeira etapa, seu sexo, etnia, idade, renda média domiciliar, nível de escolaridade, religião, tipo de região que habita (zona rural, cidade pequena ou metrópole, por exemplo), principal meio de informação consumido, frequência com que dialoga sobre política, e nível de religiosidade, isto é, a relevância que os ideais religiosos influenciam em suas decisões individuais. As duas últimas variáveis são numéricas, e variam em uma escala de 0 a 10, enquanto as demais variáveis são categóricas.

Já na segunda etapa foram obtidas informações sobre o posicionamento político do entrevistado. Por entender-se que a definição de posicionamento político possa ser comumente entendida de modo equivocado, foi necessário utilizar-se de ferramentas para a redução da subjetividade nas respostas, e o instrumento escolhido foi o compasso político.

O Compasso Político, ou Bússola Política, é um formulário composto por um conjunto de sentenças que sintetizam ideias típicas da esquerda ou da direita, e uma escala de aceitação desses ideais que varia entre a concordância total e a rejeição total da premissa. Conforme o entrevistado se posiciona diante das asserções, lhe são atribuídas pontuações, valores numéricos positivos para respostas com inclinações à direita, e negativos para respostas com inclinações à esquerda. Ao final, com o somatório dos valores obtidos nos resultados, o indivíduo é posicionado em um plano cartesiano, em que a reta horizontal representa a sua posição no espectro econômico, e o eixo vertical o posicionamento no espectro social.

Se desconsiderada a subjetividade da intepretação dos enunciados por parte de quem fórmula e de quem responde ao questionário, a Bússola Política consegue situar objetivamente um conjunto ideológico e, apesar de, por vezes, ter sua legitimidade questionada, ainda é a ferramenta menos enviesada e a que mais reflete os conceitos ideológicos apresentados no início deste artigo

5 Análise Descritiva.

Colhidas as amostras, a etapa subsequente consiste na análise exploratória dos dados. Nesta fase, dado o objetivo do estudo de identificar uma associação entre as informações colhidas, a investigação estatística focou na compreensão do comportamento conjunto das variáveis, por meio de ferramentas de análise bivariada.

5.1 Análise Bivariada

A análise bivariada consiste na utilização do aparato estatístico para estudar a dependência entre duas variáveis, a partir da observação do comportamento de uma em relação à outra.

Os instrumentos utilizados variam conforme a natureza das variáveis, que podem ser quantitativas ou categóricas. Dada a natureza mista dos dados colhidos neste estudo, foram utilizados dois parâmetros: O coeficiente de correlação de Pearson, para os dados numéricos, e a distribuição Qui-Quadrado para os dados nominais.

5.1.1 Correlação de Pearson

O posicionamento político, como anteriormente explicado, foi determinado nas amostras por meio de um conjunto de duas coordenadas cartesianas, que, posteriormente, foram reduzidas à duas classificações cada, “direita” e “esquerda”. Estas variáveis, portanto, podem ser representados por valores numéricos ou categóricos.

Dentre os dados psico-socioambientais, por sua vez, há duas variáveis numéricas de variação discreta, o grau de religiosidade e a frequência com que os entrevistados conversam sobre política, ambos dados em escalas de 0 a 10.

O grau de dependência entre as duas variáveis pode ser expresso pelo coeficiente de correlação de Pearson, que é dado pela Eq. 1.

$$\mathrm{p} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}}(x - \bar{x})^2}\sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}}(y_i - \bar{y})^2}} \qquad \text{Eq. 1}$$

Em que x_i e y_i são os valores obtidos nas amostras, e $\bar{x}$ e $\bar{y}$ são as médias aritméticas dos respectivos dados.

O valor obtido no cálculo é um coeficiente que varia entre -1 e 1 e mensura a força da correlação analisada, sendo mais forte quão mais próximo o valor estiver do módulo de 1, e a direção da correlação, sendo positiva para valores maiores que 0 e negativa para valores inferiores a 0. (PARANHOS, 2014)

Utilizando-se a biblioteca de análise de dados, Pandas, no Python, plotou-se a matriz de correlação, que mostra os coeficientes de Pearson para as variáveis numéricas de uma amostra, e o resultado obtido pode ser observado na Figura 1.

	Religiosidade	Frequência de Diálogo	Eixo Social	Eixo Econômico
Religiosidade	1.000000	0.053846	0.403241	0.267108
Frequência de Diálogo	0.053846	1.000000	0.040997	-0.245118
Eixo Social	0.403241	0.040997	1.000000	0.498532
Eixo Econômico	0.267108	-0.245118	0.498532	1.000000

Figura 1- Matriz de correlação dos dados Numéricos.

Fonte: Autor

Na matriz é possível observar uma relação positiva fraca entre a religiosidade e o posicionamento nos eixos social e econômico. Já em relação à frequência de diálogo, o eixo econômico tem uma dependência negativa fraca, e o eixo social apresenta um valor de dependência quase nulo.

Plotando-se os gráficos de dispersão das correlações estudadas, é possível observar uma sutil tendência a valores mais altos no eixo social (o que indica inclinação à direita), com o crescimento do nível de religiosidade. Estas duas variáveis foram as que apresentaram a maior correlação observada, todavia, a dependência ainda é bastante fraca e, portanto, não pode ser considerada para quaisquer inferências.

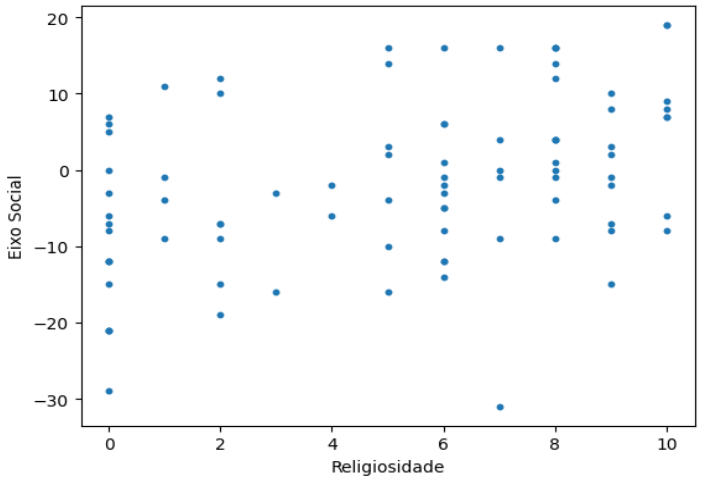

Figura 2- Gráfico de Dispersão (Religiosidade e Posicionamento Social)

Fonte: Autor

5.1.2 Distribuição Qui-Quadrado

Para a determinação da associação entre variáveis categóricas, a abordagem proposta é a determinação da distribuição Qui-quadrado, que avalia a relação entre dados a partir da comparação entre os resultados obtidos em uma distribuição e os valores esperados.

A primeira etapa para a obtenção do parâmetro é a criação de uma tabela de contingência, estrutura que organiza duas características do estudo em linhas e colunas e contabiliza a ocorrência simultânea de duas categorias dessas variáveis. Um exemplo de como se estrutura a tabela pode ser observado na figura 3.

Social Renda	Direita	Esquerda
Acima de R$22000,00	4	2
Entre 0 e R$ 2.200,00	8	27
Entre R$11.000,01 e R$22.000,00	4	1
Entre R$2.200,01 e R$4.400,00	8	14
Entre R$4.400,01 e R$11.000,00	11	8

Figura 3- Tabela de Contingência (Renda X posicionamento Social)
Fonte: Autor

Conforme visto na Figura 3, a partir da tabela de contingência já é possível notar algumas tendências, como a inclinação dos indivíduos com as rendas mais baixas à esquerda social, todavia, para reiterar esta tese, determina-se o Qui-quadrado por meio da seguinte equação:

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad \text{Eq.2}$$

Em que O_i é o valor da célula na tabela de contingência, e E_i é o valor que esperava-se encontrar na célula, dado pelo somatório dos valores da linha em que se encontra a célula, multiplicado pelo somatório dos valores da coluna e divido pelo total de amostras.

Em posse do valor de X^2 para uma tabela, é necessário observar se este valor já era estatisticamente esperado, o que indicaria uma distribuição comum dos dados e, portanto, nenhum tipo de relação entre as características analisadas.

Os valores de X^2 esperados para a tabela, dependem da estrutura dela, mais especificamente, dos graus de liberdade, dados por: $(nº\ de\ linhas - 1) * (nº\ de\ colunas - 1)$, sendo constantes já tabeladas e agrupadas segundo o percentual de certeza adotado, isto é, o percentual de mínimo de precisão estatística aceitável para que a hipótese da independência entre os dados seja aceita.

Os valores X^2 tabelados podem ser interpretados, portanto, como sendo valores limites correspondentes ao percentual de certeza adotado. Caso o valor X^2 obtido supere o tabelado, o percentual para a aceitação da hipótese nula na amostra do estudo também será inferior ao mínimo aceitável e, portanto, teremos uma margem segura para afirmar que os dados são dependentes. Os valores percentuais de aceitação da Hipótese nula podem ser observados na Tabela 1.

Variáveis Relacionadas	*P-Value*	
	Posicionamento Social	Posicionamento Econômico
Sexo	0.335	0.103
Etnia	0.008	0.011
Renda	0.017	0.627
Escolaridade	0.339	0.382
Religião	0.029	0.295
Religiosidade	0.083	0.583
Meio de Informação	0.526	0.146
Frequência de Diálogo	0.129	0.396
Tipo de Habitação	0.189	0.300

Tabela 1 - Tabela Com os *P-valores* das variáveis do estudo
Fonte: Autor

Considerando um valor de aceitação de hipótese nula de 5%, observa-se que o posicionamento social individual é influenciado pela etnia, renda e religião. O posicionamento no espectro econômico, por sua vez, tem dependência apenas com a variável etnia.

A partir da tabela, segundo os critérios definidos, pode-se inferir que poucas das variáveis analisadas influem significativamente no posicionamento político, o que poderia indicar uma independência do fenômeno estudado com as demais

características analisadas, ou uma contribuição individual destas pouco relevante. De todo modo, a obtenção do resultado conclusivo requer o estudo do efeito conjunto das propriedades observadas, aqui feito por meio do uso de algoritmos preditivos de *Machine Learning.*

6 Análise Preditiva

Para a análise preditiva, as informações colhidas alimentaram um algoritmo de classificação que, posteriormente, teve seu desempenho avaliado, de modo a determinar a validade da associação criada pela inteligência artificial.

O algoritmo escolhido para o propósito foi a árvore de decisão, que, por meio de cálculos relativamente simples, fornecem resultados de precisão satisfatória, ao focar-se nas variáveis mais relevantes para o problema, ideal para a hipótese estudada, em que alguns atributos são significativamente mais influentes no fenômeno de interesse.

6.1 Pré-Processamento

Antes do processo de aprendizado de fato, as informações colhidas precisaram ser pré-processadas de modo que ficassem no formato adequado para o algoritmo.

Algoritmos de classificação são um tipo preditivo específico que utiliza os padrões identificados nos dados históricos para analisar novos conjuntos de dados e atribuir-lhes uma classificação. A primeira etapa do pré-processamento, portanto, é a separação da base em dados de entrada e dados de resposta.

A saída esperada para um algoritmo classificativo é apenas uma classe baseada em um vetor de possibilidades. Os dados de saída deste estudo, entretanto, estão estruturados em duas variáveis, posicionamento social e econômico. Diante disto, há duas maneiras de proceder, criando dois algoritmos preditivos, um para cada saída esperada, ou agrupando as saídas em um conjunto único de classes.

A primeira opção proposta acima, restringiria o estudo à análise do efeito conjunto dos fatores psico-socioambientais nos eixos separadamente, porém, poderia aumentar a taxa de precisão por acertos acidentais, devido à quantidade reduzida de classes, o que prejudicaria a análise de eficiência do algoritmo, além disso, valores de precisão muito discrepantes entre os dois algoritmos resultaria em uma previsão de posicionamento pouco útil. Diante disto, optou-se por agrupar as classes em uma

única variável, o que, em termos práticos, treina o algoritmo para prever em qual quadrante do plano cartesiano uma nova entrada se localizaria.

Separadas as variáveis adequadamente, os valores categóricos precisaram ser convertidos em numéricos, uma vez que os algoritmos trabalham com operações matemáticas. Para isto, cada coluna nominal da tabela foi separada em um conjunto de colunas que representam cada uma das categorias das variáveis. Os valores das células passaram a ser binários, assumindo 1 caso o indivíduo representado na linha esteja na categoria indicada pela respectiva coluna, e 0 para as demais categorias.

Após isto, todos os dados foram normalizados, para que variassem em uma mesma escala, de modo que os valores originalmente numéricos, não fossem erroneamente interpretados como mais relevantes durante os cálculos. Por fim, os dados foram novamente subdivididos em dois grupos, 80% da base foi reservada para o aprendizado do algoritmo, de onde ele irá identificar os padrões, e o restante foi utilizado no teste de desempenho, para averiguar a eficiência da regra de associação encontrada.

6.2 Árvore de Decisão

O trabalho da inteligência artificial, neste procedimento, é identificar dentre as variáveis independentes àquelas que são mais relevantes para processo de classificação. Os parâmetros mais comumente utilizados para mensurar essa relevância são a entropia e o ganho de informação.

Em teoria da informação, entropia é definida como uma medida de desordem nos dados observados e nos indica um grau de incerteza probabilística. Em tese, uma maior entropia significa menor certeza de ocorrência de um fenômeno diante de um evento aleatório e, por conseguinte, uma maior possibilidade de obtenção de informações relevantes sobre as amostras por meio da observação de tais fenômenos.

Na análise de dados categóricos, a entropia é proporcional à quantidade de categorias na variável e à distribuição das amostras entre as categorias, sendo obtida pela Eq. 3.

$$E(S) = \sum_{i=1}^{n} -p_i \log_2 p_i \qquad \text{Eq. 3}$$

Em que "p_i" é a probabilidade de ocorrência de uma das categorias da variável, dada pela razão da quantidade de ocorrências da categoria pela quantidade total de amostras, e "n" a quantidade total de categorias na variável.

Nota-se que quão mais próximo de 1 for a probabilidade do evento, mais a entropia tenderá a 0.

No procedimento de montagem da árvore, após definir a entropia da variável de resposta do problema, é necessário percorrer o caminho inverso, isto é, determinar quais variáveis mais contribuem para reduzir o grau de incerteza de ocorrência dos fenômenos estudados, e para isto, a métrica utilizada é o ganho de informação.

O Ganho de informação é a medida do potencial de um atributo de redução de entropia em um fenômeno pela sua partição em classes. É o parâmetro utilizada para determinar quais variáveis devem ser consultadas, e em que ordem isso deve ser feito, para que o algoritmo gere uma classificação mais precisa no menor tempo de execução. O ganho de informação é dado pela Eq. 4.

$$G(S, A) = E(S) - \sum \frac{|S_v|}{|S|} E(S_v) \qquad \text{Eq. 4}$$

Em que "S_v" é o subconjunto composto pelas amostras do fenômeno de interesse "S" que receberam a classificação "v" na variável de partição.

Definidos os valores de ganho de informação de cada uma das variáveis, o algoritmo monta a árvore posicionando os atributos mais relevantes nos nós superiores, não necessariamente utilizando todas as variáveis do dataset.

Neste estudo, para aplicação da árvore de decisão ao problema, foi feito uso da biblioteca de aprendizado de máquina Sklearn do Python. O algoritmo foi treinado com 70 amostras e, após criada a regra de associação, teve a eficiência mensurada com uma base de testes de 17 amostras.

No teste inicial com a regra de associação gerada, a acurácia obtida para o algoritmo foi de 42%, valor consideravelmente baixo, mesmo se considerado o nível de correlação obtido entre as variáveis. Diante disto, visando a um aumento no desempenho, propôs-se uma abordagem diferente com a árvore de decisão, a aplicação em uma *Random Forest.*

6.3 Random Forest

Árvores, como anteriormente demonstrado, são estruturas similares a fluxogramas que fazem a classificação baseada no caminho percorrido pelo dado

dentro do fluxo. Um mesmo problema pode ser resolvido com a utilização de diferentes configurações de árvores, todavia, algumas estruturas podem ser pouco eficientes para o tipo de análise desejada, isso ocorre, geralmente, quando a profundidade das árvores cresce além do necessário.

Quando uma árvore de decisão se torna profunda demais, é sinal de que houve um overfitting, isto é, a inteligência apenas decorou as informações e se adaptou demais aos dados históricos recebido, o que é um problema quando se deseja uma generalização classificadora.

Uma solução para o overfitting é a *Random Forest*. Neste algoritmo, várias árvores de decisão são geradas utilizando os mesmos cálculos acima explicados, todavia, cada árvore utiliza um subconjunto aleatório das variáveis e das amostras do dataset passado, o que cria um conjunto diversificado de estruturas de classificação. O valor de saída de uma floresta randômica é o resultado da média simples das classificações feitas por todas as árvores à entrada recebida (MAIMON e ROKACH, 2014).

Novamente utilizando o SkLearn, os dados do problema foram submetidos ao algoritmo de *Random Forest* em um arranjo composto por 80 árvores. Nos novos testes, a acurácia média das classificações subiu para 67%, um valor bem melhor se comparado ao anterior, entretanto, uma avaliação de desempenho de um algoritmo de aprendizado requer a análise de mais métricas para além da acurácia.

6.4 Teste de Desempenho

Até o momento, os algoritmos foram avaliados com base em sua acurácia, uma métrica que fornece um panorama geral sobre a assertividade das classificações do programa, obtida pela razão da quantidade de predições feitas corretamente pelo total de predições realizadas.

No teste da árvore de decisão, o valor de acurácia obtido demonstrou-se insatisfatório, o que foi interpretado como ineficácia da regra de associação criada pelo algoritmo e motivou a adoção de uma técnica diferente. Na *Random Forest*, o valor de acurácia teve significativa melhora, o que pode ser indicativo de uma generalização aceitável, porém, para afirmar de forma mais conclusiva, foi necessária uma avaliação mais detalhada, e a ferramenta usada para tal foi a matriz de confusão.

A matriz de confusão avalia a performance do algoritmo individualmente nas classes por meio de uma matriz que compara as predições com as classes reais dos dados em uma estrutura como a apresentada na Figura 4.

		Valor Predito	
		Sim	**Não**
Real	**Sim**	Verdadeiro Positivo (TP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (TN)

Figura 4 – Estrutura de uma matriz de confusão

Fonte: Blog de Inteligência artificial Diego Nogare.

Disponível em: https://diegonogare.net/2020/04/performance-de-machine-learning-matriz-de-confusao

Na diagonal principal são representados os acertos da classificação, com os verdadeiros positivos e verdadeiros negativos quantificando os valores que foram categorizados corretamente, isto é, pertencentes e não pertencentes, respectivamente, à referida classe. Já na diagonal secundária são representados os erros do algoritmo, em que os falsos positivos representam os valores que foram categorizados como pertencentes à classe, sem de fato sê-lo, e os falsos negativos representando o inverso, isto é, os valores que pertenciam à classe analisada, mas não foram reconhecidas de tal modo (FRANCESCHI, 2019) .

Como anteriormente mencionado, a acurácia é a razão de acertos (diagonal principal) pelas classificações totais, um parâmetro que, por determinar o desempenho médio das classificações, pode mascarar tendências da inteligência artificial em ser assertiva em apenas uma das classes, por isso, outros parâmetros relevantes para a análise são a precisão e o recall, dados pelas Eq. 5 e 6, respectivamente.

$$Precisão = \frac{TP\ (True\ Positive)}{TP(True\ Positive) + FP(False\ Positive)} \quad \text{Eq. 5}$$

$$Recall = \frac{TP(\ True\ Positive)}{TP(True\ Positive) + FN\ (False\ Negative)} \quad \text{Eq. 6}$$

O valor de precisão é a métrica que avalia dentre os valores categorizados como pertencentes à uma classe, quantos o são de fato. O recall, por sua vez, é o

parâmetro que determina qual o percentual das amostras pertencentes a uma classe de fato foi reconhecido desse modo pelo algoritmo.

A partir dos valores, é possível ainda ter uma dimensão geral de qualidade para cada classe, calculando-se a média harmônica dos parâmetros, em um indicador conhecido como f1-score, denotado pela Eq. 7.

$$F1 = \frac{2 * precisão * recall}{precisão + recall} \quad \text{Eq. 7}$$

No Python, com uso do pacote metrics, do Sklearn, foram obtidas cada uma das métricas apresentadas para o algoritmo criado, e o resultado exibido pode ser visto na figura 5.

	precision	recall	f1-score
Direita Direita	0.50	1.00	0.67
Direita Esquerda	1.00	0.33	0.50
Esquerda Direita	0.60	0.60	0.60
Esquerda Esquerda	0.75	0.75	0.75
accuracy			0.67
macro avg	0.71	0.67	0.63
weighted avg	0.72	0.67	0.66

Figura 5 – Métricas de desempenho do algoritmo de *Random Forest*
Fonte: Autor

À esquerda da tabela são mostradas as classes, organizadas em posicionamento social e econômico, nesta ordem, e à direita, seus respectivos parâmetros.

Ao observar os dados, notam-se alguns pontos críticos, a começar pela discrepância na qualidade das classificações. As duas últimas categorias, esquerda social com diferentes posicionamentos econômicos, têm valores de precisão e recall razoáveis, e um equilíbrio entre essas métricas, o que aumenta a confiabilidade no seu f1-score.

As classes de direita econômica, em contrapartida, demonstram considerável desequilíbrio na assertividade, com um valor de precisão muito baixo para a primeira classe, o que indica grande quantidade de classificações falsamente positivas, e valor de recall mais baixo ainda para a segunda classe, o que indica a incapacidade do algoritmo de reconhecer muitos dos dados como pertencentes àquela categoria (muitos falsos negativos).

7 Conclusão

O presente estudo permitiu observar do ponto de vista estatístico algumas premissas formuladas nas ciências sociais, ao avaliar matematicamente a influência dos fatores psico-socioambientais no posicionamento político ideológico individual.

A análise exploratória dos dados obtidos comprovou que de fato há uma relação entre as variáveis do estudo, e que o conhecimento de algumas características demográficas de um indivíduo pode fornecer indícios de seu alinhamento ideológico.

O estudo do efeito conjunto das variáveis pela generalização criada pela inteligência artificial, entretanto, revelou que a associação das variáveis não é determinante o suficiente para predizer objetivamente as concepções ideológicas de uma pessoa.

No contexto do ambiente da pesquisa, concluiu-se que os fatores demográficos exercem certa influência sobre o posicionamento pessoal, entretanto, não é possível determinar de modo claro uma generalização que explica como essa associação ocorre em todas as instâncias do problema.

É importante frisar aqui que o tamanho da base de dados utilizada, relativamente pequena para representar todas as possíveis combinações das características avaliadas, pode ter impactado na identificação dos padrões pelo algoritmo, e por conseguinte, nos resultados desse estudo.

Além disso, há imprecisões inerentes às ferramentas de pesquisa adotadas, uma vez que, dentro das ciências sociais, os conceitos não são tão estáticos.

Diante do exposto, depreende-se, portanto, que seria imprudente descartar totalmente a hipótese levantada no início deste artigo, mas, com base nos indícios do estudo, e considerados os limites dos métodos da pesquisa, os indícios apontam para a rejeição da premissa levantada, isto é, os fatores ambientais não são determinantes do alinhamento ideológico individual.

8 Referências

ALPAYDIN, Enthen. ***Introduction to Machine Learning***.2ª Edição. *Massachussets: MIT Press.* 2009.

ARAÚJO, Joaquim Filipe; SILVESTRE, Hugo Consciência. **Metodologia para a Investigação Social**. São Paulo: Escolar Editora, 2012.

BERNOVILLE, Dugé de. **Introdução à Análise Estatística**. *Libr. Générale de Droit.* 1939.

BOBBIO, Norberto**. As ideologias e o poder em crise**. Brasília: UNB, 2005.

BOBBIO, Norberto**. Direita e Esquerda: Razões e significados de uma distinção política**. 3ª edição. São Paulo: Editora Unesp, 2012.

BRESSER-PEREIRA, Luiz Carlos. **O paradoxo da esquerda no Brasil.** SciELO, 2006, disponível em: https://doi.org/10.1590/S0101-33002006000100003.

FALCO, Javert Guimarães. **Estatística Aplicada**. Cuiabá: EdUFMT.2008.

FRANCESCHI, Pietro Reinheimer. **Modelos Preditivos de Churn: O caso do Banco do Brasil.** 2019. p. 121. Dissertação de Mestrado – UNISINOS, Porto Alegre, 2019.

FREIRE, Paulo. **Pedagogia do Oprimido**. 11ª edição. São Paulo: Terra e Paz, 1998.

INGLEHART, Ronald. ***The Silent Revolution****. New Jersey: Princeton University.1977.*

LIPSET, Seymour Martin. ***Cleavage Structures, Party Systems and Votter Alignmments***.1965.

MAIMON, Oded; ROKACH, Lior. ***Data Mining with Decision Trees: Theory and Applications****.* 2ª edição. USA: World Scientific Public Co, 2014.

PARANHOS, R.; FIGUEIREDO FILHO, D. B.; ROCHA, E. C. da; SILVA JÚNIOR, J. A. da; NEVES, J. A. B.; SANTOS, M. L. W. D. **Desvendando os Mistérios do Coeficiente de Correlação de Pearson: o Retorno**. Leviathan (São Paulo**)**, n. 8, p. 66-95, 2014. Disponível em:https://www.revistas.usp.br/leviathan/article/view/132346