ISSN: 2237-826X
ARTIGO | PAPER

Classificação das percepções de stakeholders sobre o futuro do Brasil utilizando aprendizado de máquina

Classifying perceptions from Stakeholders about Brazil's future using machine learning

Amauri Ornellas da Silva[1], Daniele Gonçalves de Toledo Luchetta Raminelli[2], Bruno Samways dos Santos[3], Rafael Henrique Palma Lima[4]

[1] *Universidade Tecnológica Federal do Paraná (UTFPR), Londrina, Paraná, Brasil. ORCID: http://orcid.org/0000-0001-5572-6341*

[2] *Universidade Tecnológica Federal do Paraná (UTFPR), Londrina, Paraná, Brasil. ORCID: http://orcid.org/0000-0003-1927-0011*

[3] *Universidade Tecnológica Federal do Paraná (UTFPR), Londrina, Paraná, Brasil. ORCID: http://orcid.org/0000-0001-7919-1724*

[4] *Universidade Tecnológica Federal do Paraná (UTFPR), Londrina, Paraná, Brasil. ORCID: http://orcid.org/0000-0002-9098-3025*

Autor para correspondência/*Mail to:* Natalia Pimenta Alves Lage, nataliap.alves@gmail.com

Recebido/*Submitted*: 17 de dezembro de 2021; **Aceito/*Approved*:** 15 de julho de 2022

Resumo

Este artigo compara cinco técnicas de aprendizado de máquina (AM) para classificar as percepções dos *stakeholders* quanto ao futuro do Brasil. As técnicas de ML utilizadas foram redes neurais artificiais, *k*-vizinhos mais próximos, *naïve bayes*, floresta aleatória e máquinas de vetores de suporte. Eles foram aplicados a um conjunto de dados do Banco Mundial sobre o desenvolvimento do Brasil. O conjunto de dados foi pré-processado e configurado em duas versões diferentes: a primeira continha um subconjunto de atributos selecionados manualmente pelos autores, enquanto a segunda era composta por atributos selecionados usando a abordagem de ganho de informação. Verificou-se que todas as técnicas de ML tiveram melhor desempenho com a segunda versão do conjunto de dados, em que os atributos foram classificados com base no ganho de informação. No entanto, dentro de cada versão do conjunto de dados, todas as técnicas tiveram desempenhos semelhantes. Esta pesquisa também constatou que os atributos mais relevantes estão relacionados às oportunidades de negócios, índices de desenvolvimento associados a temas críticos e confiança nas instituições e organizações.

Palavras-chave: Aprendizado de máquina; Classificação; Brasil; *Stakeholders*.

Abstract

This paper compares five machine learning (ML) techniques to classify the perceptions of stakeholders as to Brazil's future. The ML techniques used were artificial neural networks, k-nearest neighbors, naïve bayes, random forest and support vector machines. They were applied to a dataset retrieved from the World Bank about Brazil's development. The dataset was preprocessed and configured in two different versions: the first contained a subset of attributes manually selected by the authors, whereas the second was composed of attributes selected using the information gain approach. It was found that all ML techniques performed better using the second version of the dataset, where attributes were ranked based on information gain. However, within each version of the dataset all techniques had similar performance. This research also found that the most relevant attributes are related to business opportunities, development indexes associated with critical subjects and trust in institutions and organizations.

Keywords: *Machine learning; Classification; Brazil; Stakeholders.*

INTRODUÇÃO

O Produto Interno Bruno (PIB) é um dos indicadores que reflete o crescimento econômico de uma nação, representando a soma, em valores com seleção automática de atributos monetários, de todos os bens e serviços finais produzidos em uma cidade, estado ou país em um período. O Instituto Brasileiro de Geografia e Estatística (IBGE) calcula este indicador a partir de dados como o Balanço de Pagamentos do Banco Central, Declarações de Informações Econômicas de Pessoas Jurídicas, bem como os Índices de Preços ao Consumidor (IPA), o Índice Nacional de Preços ao Consumidor Amplo (IPCA), entre outros (Instituto Brasileiro de Geografia e Estatística, 2020).

As variações positivas no PIB são fundamentais para a manifestação do desenvolvimento econômico, mas deve-se levar em consideração também aspectos qualitativos e subjetivos. O desenvolvimento econômico é um conceito complexo que pode ser visto tanto como um meio quanto como um fim. Analisando-se este desenvolvimento como um meio, então este envolve a riqueza, prosperidade, progresso técnico, bem-estar, sustentabilidade e liberdade, focando-se na sociedade (Viana & Lima, 2010). Já como um fim, o desenvolvimento transforma-se no objetivo final do planejamento e no desenvolvimento de estratégias e políticas para alcançá-lo (Souza & Spinola, 2017).

Os ambientes impactam diretamente as decisões tomadas por instituições e organizações e suas expectativas futuras sobre um determinado objeto podem aumentar e diminuir a partir destes aspectos. Pesquisas já mostraram que percepções otimistas ou pessimistas podem influenciar a tomada de decisões de indivíduos, empresas e órgãos

governamentais, impactando indicadores como as taxas de câmbio, preços de commodities, valores de ações e prêmios em seguros (Puri & Robinson, 2007).

Pesquisas são realizadas periodicamente no Brasil para avaliar as expectativas de mercado, sendo a principal delas o Relatório Focus, publicado semanalmente pelo Banco Central do Brasil (BCB) com expectativas de mercado quanto ao crescimento da atividade econômica e variações nos índices de preços, taxas de juros e outros indicadores relevantes (Banco Central do Brasil, 2020). O Banco Mundial (WBG – *World Bank Group*) também realiza pesquisas de opinião quanto ao cenário econômico brasileiro, englobando diversos *stakeholders* como agências governamentais, empresas, mídia, sociedade civil e academia. Em 2019, o WBG realizou uma pesquisa sobre a visão dos entrevistados quanto ao ambiente econômico no Brasil, cujo conjunto de dados está disponível publicamente no website do Banco Mundial (WBG, 2020).

Técnicas de *machine learning* (ML) têm sido amplamente utilizadas no contexto de opiniões nas mais diversas áreas, como por exemplo em acidentes (Wang, Chi, Liu, & Wang, 2019), vacinação e epidemias (Featherstone, Ruiz, Barnett, e Millam, 2020, Haihong et al., 2019, Haupt, Jinich-Diamant, Li, Nali, e Mackey, 2021), alimentação infantil controlada (Kang, Wang, Zhang, & Zhou, 2017) e em finanças (Liu, Ergu, Cai, Gong, & Sheng, 2019). Porém, esses estudos usam a mineração de textos para construir suas bases de dados, logo sendo originadas a partir de uma fonte não estruturada de coleta de dados. Entretanto, nas áreas de economia e finanças há poucos estudos que utilizam questionários estruturados para construir bases de dados para alimentar sistemas de ML, o que traria a possibilidade de analisar outras perspectivas a respeito da opinião dos entrevistados.

Neste contexto, utilizando o conjunto de dados do WBG de 2019, esta pesquisa teve como objetivo comparar cinco técnicas de ML na classificação de *stakeholders* quanto às suas expectativas em relação à economia. Para isto, cinco técnicas de ML foram aplicadas, sendo elas Redes Neurais Artificiais (*Artificial Neural Networks*, ANN), k-Vizinhos Mais Próximos (*k-Nearest Neighbors*, KNN), Naïve Bayes (NB), Floresta Aleatória (*Random Forest*, RF) e Máquina de Vetores de Suporte (*Support Vector Machine*, SVM), classificando os entrevistados como "Otimistas" ou "Pessimistas" em relação ao desenvolvimento futuro do Brasil. As comparações foram realizadas por três métricas diferentes e também em dois conjuntos de dados: (i) conjunto completo, e; (ii) conjunto reduzido com seleção automática de atributos utilizando a técnica de Ganho de Informação (*Information Gain*, IG).

Primeiramente, este trabalho se justifica pela aplicação, pois técnicas computacionais na análise de dados econômicos é um tema recente de pesquisa que ainda requer a realização de diversos trabalhos para ser consolidada (Mullainathan & Spiess, 2017). Também pode-se detectar os atributos mais relevantes para que um *stakeholder* seja enquadrado como "Pessimista" ou "Otimista", analisando-se assim estes fatores encontrados.

Após esta seção introdutória, o restante do artigo está dividido da seguinte forma: a Seção 2 aborda a fundamentação teórica, introduzindo as técnicas utilizadas nesta pesquisa, bem como as métricas de avaliação e validação dos resultados; a Seção 3 discorre sobre o conjunto de dados e a sequência da pesquisa; já as seções 4 e 5 mostram os resultados e comparam as técnicas utilizadas, elencando também os atributos mais relevantes encontrados pela técnica IG. Por fim, a Seção 6 descreve brevemente sobre o que foi realizado nesta pesquisa para avaliar a percepção dos *stakeholders* sobre o futuro brasileiro e os resultados obtidos a partir das técnicas de classificação aplicadas, incluindo uma discussão sobre os atributos mais relevantes, limitações e pesquisas futuras.

FUNDAMENTAÇÃO TEÓRICA

O aprendizado de máquina, ou o termo em inglês, "*machine learning*", corresponde a uma área específica da inteligência artificial em que se aplicam algoritmos para aprender com as experiências e realizar a predição de eventos futuros (Dogan & Birant, 2021). Existem várias técnicas para realizar o aprendizado, como regressão linear, regressão logística, Redes Neurais Artificias, Árvores de decisão, entre outras. Porém, o algoritmo a ser escolhido depende do tipo de aprendizado a ser realizado, e também da tarefa de mineração de dados. Para esta pesquisa, foi realizada a tarefa de classificação.

A classificação é uma das tarefas mais comuns da área de mineração de dados e se encaixa no contexto do aprendizado supervisionado. Esta tarefa consiste em predizer o rótulo (classe) de instâncias desconhecidas, utilizando amostras de treinamento em que estes rótulos são conhecidos e que orientam o aprendizado (Tan, Steinbach, Karpatne, & Kumar, 2019), como classificar e-mails em "*spam*" ou "não *spam*". Alguns dos algoritmos considerados mais importantes podem ser a regressão logística, árvores de decisão, SVM, KNN, RF e ANN (Géron, 2019). Neste artigo, foram utilizadas SVM, KNN, RF, ANN, e também NB.

Redes Neurais Artificiais

As ANNs foram desenvolvidas para simular o sistema nervoso humano, tratando as unidades computacionais de maneira similar aos neurônios humanos (Aggarwal et al., 2018). Elas também são consideradas modelos estatísticos não lineares em que as camadas que formam uma rede neural são independentes em si. Desta forma,

cada camada pode possuir um certo número de nós (neurônios), em que os neurônios da camada escondida são parâmetros selecionados pelo usuário, enquanto os neurônios de entrada e de saída correspondem aos recursos de entrada e às classes de saída (Abiodun et al., 2018). Uma das vantagens de se usar esta técnica é que esta é orientada pelos próprios dados, não necessitando de pressuposições para a sua aplicação (Fergus, Idowu, Hussain, & Dobbins, 2016).

K-Vizinhos Mais Próximos

O KNN pertence a uma classe chamada *lazy learner*, pois o processo de aprendizagem ocorre diretamente na fase de classificação (M.-L. Zhang & Zhou, 2007). O "k" representa o número de vizinhos a serem analisados, e a instância testada será classificada na categoria mais comum entre os k-vizinhos mais próximos (Pan, Wang, & Pan, 2020). Para a maioria dos casos, modelos clássicos de KNN adotam a distância euclidiana (X. Zhang & Gou, 2022), a qual pode ser calculada para analisar a distância entre dois pontos a partir de dois atributos (cálculo da hipotenusa do Teorema de Pitágoras) ou para qualquer número de atributos (Bramer, 2016). Desta forma, todo o vetor de teste é comparado com os k vetores de treino que estão mais próximos dele em um espaço de n-dimensões, onde n é a quantidade de atributos destes vetores (Han, Kamber, & Pei, 2012).

O KNN é um algoritmo bastante popular para a tarefa de classificação por ser bastante intuitivo (Oliveira, Faria, Gaio, & Reis, 2017) e conceitualmente simples, mas é considerado ineficiente em termos computacionais (Myslín, Zhu, Chapman, Conway, et al., 2013), já que em bases de dados muito grandes a necessidade de analisar todas as instâncias na base para encontrar aquelas que são as k mais próximas o torna menos competitivo que outros métodos (Kafaf, Kim, & Lu, 2017).

Máquinas de Vetores de Suporte

O SVM é uma técnica oriunda da teoria do aprendizado estatístico e tem recebido bastante atenção nas últimas décadas, tanto em estratégias de implementação como também em algoritmos (Cortes e Vapnik, 1995, Raghavendra e Deka, 2014). Essa técnica tem sido aplicada com sucesso em diversas áreas, como o reconhecimento de imagens, detecção, mineração de textos e problemas de regressão (Zendehboudi, Baseer, e Saidur, 2018, Silva, Welfer, Gioda, e Dornelles, 2017). Comumente, usam-se diferentes funções *kernel* que funcionam com um artifício para desempenhar uma classificação bidimensional de um conjunto originalmente unidimensional, realizando uma projeção dos dados de um baixo espaço dimensional para uma dimensão maior (Patle & Chouhan, 2013). Desta forma, além do *kernel* linear, escolhas comuns são os kernels de base radial, polinomial e sigmoide, sendo que muitas destas funções possuem parâmetros associados que são ajustados a partir do conjunto de treino (Aggarwal, 2015).

Floresta Aleatória

O RF foi publicado por Breiman (2001), sendo definido como um classificador do tipo *ensemble* que se utiliza de uma coleção de árvores de decisão estruturadas $\{h(x, \Theta_k), k = 1, \ldots\}$ onde são vetores aleatórios distribuídos e cada árvore elenca uma unidade de voto para a classe mais frequente da entrada x. Esta técnica é popular por superar algumas limitações das árvores de decisão, sendo robusta para conjuntos de grandes dimensões (Zhao, Henriksson, Asker, & Boström, 2015). O algoritmo RF seleciona subconjuntos de dados para cada árvore, o que permite uma predição agregada de cada árvore, oferecendo frequentemente melhores resultados do que se teria com apenas uma árvore (Speiser, Miller, Tooze, & Ip, 2019). Os métodos do tipo *ensemble* são caracterizados pela diversidade dos classificadores e o ganho que se obtém desses métodos deriva dessa diferença dos padrões de erro de seus classificadores-base (Genuer & Poggi, 2020).

Naïve Bayes

NB foi criado a partir do Teorema de Bayes, simplificando a sua utilização por adotar independência entre todos os atributos e todos têm a mesma contribuição para a classe (Modu et al., 2017). Esta técnica traz uma combinação de probabilidade *a priori* e probabilidade condicional em apenas um método, calculando-se a probabilidade de cada uma das possíveis classes, escolhendo-se a classe com o maior valor retornado (Bramer, 2016).

Seleção de Atributos por Ganho De Informação

O IG é baseado na teoria da informação e entropia (Li et al., 2018), verificando-se a relevância de cada um dos atributos para a classe em avaliação, podendo ser ordenada a contribuição de cada um destes atributos (Kubat, 2017). Pode ser dito que a entropia dentro do subconjunto de treinamento é medida em "bits" de informação (Li et al., 2018). Esta técnica é considerada como sendo simples e computacionalmente eficiente, pois independe de um classificador (método de filtro).

Algoritmo de imputação *missForest*

O *Missforest* é um algoritmo de ML para imputação de dados faltantes e que utiliza o princípio da Floresta Aleatória. Ele inicia preenchendo os valores faltantes com a média, moda ou mediana e rotula essa entrada como "Predito", e as demais entradas como "Treinamento". Esses valores são colocados num modelo de RF, o qual é treinado iterativamente. A cada iteração o modelo decide se substitui o valor que está como "Predito" ou se o mantém. O desempenho do modelo é melhorado a cada iteração até que se atinja um critério de parada (Stekhoven & Bühlmann, 2011).

MATERIAIS E MÉTODOS

Nesta seção, estão descritos os passos da pesquisa, incluindo o conjunto de dados, a etapa de pré-processamento, o fluxograma de trabalho e as métricas utilizadas para a comparação das técnicas de ML.

Descrição Prévia do Conjunto de Dados

De abril a maio de 2019, 933 *stakeholders* pertencentes ao WBG no Brasil foram convidados a opinar sobre o trabalho do WBG participando da Pesquisa de Opinião. Os entrevistados receberam os questionários por correspondência ou pela plataforma de pesquisa *online*. Com uma taxa de respostas de 32%, obteve-se o conjunto de dados com 300 respondentes.

Cada instância compreende as informações preenchidas por um respondente, e as perguntas (atributos) eram em formato de múltipla escolha ou em escalas de 1 a 10. As perguntas de múltipla escolha podiam ser de resposta única ou, quando pertinente, receber até três respostas diferentes de cada respondente, e algumas questões eram compostas de outras perguntas, resultando em 344 atributos. Essa base foi obtida no *website* do WBG (https://microdata.worldbank.org/index.php/catalog/3511) e, primeiramente, foi excluída uma instância inválida, ou seja, o *stakeholder* respondeu às duas primeiras questões, interrompendo o seu preenchimento a partir da terceira pergunta, restando assim 299 instâncias válidas para a etapa do pré-processamento dos dados. As instâncias restantes, mesmo com alguns poucos dados faltantes, foram incluídas para análise, necessitando da imputação de dados para completar o conjunto. A próxima etapa incluiu a seleção inicial de atributos (manual), análise da classe (recodificação de classes e exclusão de indecisos), verificação de dados faltantes, imputação de dados, binarização e normalização.

Etapa de Pré-Processamento

A Figura 1 sintetiza os procedimentos realizados durante a etapa de pré-processamento, que teve início com n=299 instâncias e m=344 atributos.

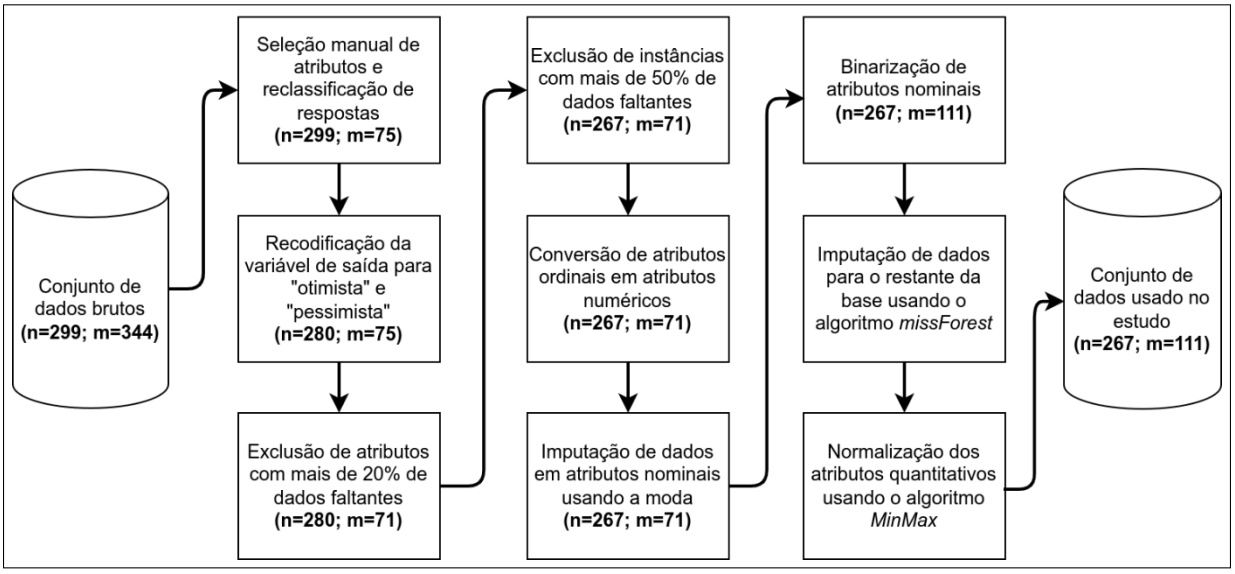

Figura 1. Etapa de pré-processamento do conjunto de dados.

Primeiramente foi realizada a seleção manual de atributos, sendo excluídos 218 atributos que não têm relação com o objetivo desta pesquisa. Os atributos restantes foram reclassificados em oito dimensões distintas: (i) Infraestrutura; (ii) Recursos Básicos; (iii) Recursos Econômicos; (iv) Bem-Estar Social; (v) Governamental; (vi) Ambiental; (vii) Mídia Local; e (viii) Mídia Internacional, o que possibilitou a exclusão de atributos que se sobrepunham. Isso reduziu a quantidade de atributos para m=75.

Após isso, a variável de saída foi reclassificada, mapeando os cinco valores existentes em apenas dois ("otimista" ou "pessimista"), excluindo-se as instâncias em que os *stakeholders* responderam "*not sure*". Isso eliminou 19 instâncias da base, passando a contar com n=280 linhas. Também foram eliminados atributos com mais de 20% de dados faltantes e instâncias com mais de 50% de valores em branco, reduzindo a base para 267 instâncias e 71 atributos.

Alguns ajustes precisaram ser feitos nos valores da base de dados. No caso de atributos ordinais, os valores foram convertidos em escalas numéricas. Valores nominais faltantes foram primeiramente preenchidos pela moda do atributo correspondente e, posteriormente, binarizados, o que elevou a quantidade de atributos para 111. A binarização consiste na criação de um novo atributo binário para cada valor distinto que o atributo nominal possui, o que é necessário para que os algoritmos de ML possam processá-lo. Por fim, os dados faltantes foram imputados usando o algoritmo *missForest* (Stekhoven & Bühlmann, 2011), e os atributos quantitativos foram normalizados usando o algoritmo *MinMax*, o qual converte os valores originais para o intervalo contínuo entre 0 e 1.

Para a separação dos conjuntos de "treino" e "teste" foi utilizado o método de validação cruzada *k-fold*, o qual divide o conjunto de dados em k subconjuntos. Com esta técnica, k treinamentos do algoritmo são realizados e, em cada um desses treinamentos, um dos k subconjuntos é utilizado como a base de "teste", e os demais formam a base de "treino". Assim, garante-se que todas as instâncias da base farão parte da base de "teste" exatamente uma vez. Para esta pesquisa, foi escolhido um k igual a 10. Para cada modelo gerado pela validação cruzada foram calculadas as métricas de acurácia, precisão, *recall* e *f1-score*. Considerando-se VP (verdadeiro positivo), VN (verdadeiro negativo), FP (falso positivo) e FN (falso negativo), as Equações 1 a 4 mostram como essas métricas são calculadas.

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{1}$$

$$Precisão = \frac{VP}{VP + FP} \tag{2}$$

$$Recall = \frac{VP}{VP + FN} \tag{3}$$

$$F1 - score = 2 \times \frac{Precisão \times Recall}{Precisão + Recall} \tag{4}$$

Em relação às técnicas utilizadas, foram feitos testes com variações dos parâmetros da ANN, KNN, RF e SVM, descrito na Tabela 1. Os parâmetros com melhor desempenho estão destacados na tabela e foram utilizados nos experimentos subsequentes.

Técnica	Parâmetro	Variações
NB	*Default* (Bernoulli)	—
ANN	Neurônios na camada escondida	[2, **10**, 30]
KNN	Número de *k* vizinhos	[2, 5, 10, 20, **50**]
RF	Árvores estimadoras	[50, **100**, 500]
SVM	*Kernel*	[linear, polinomial, **função de base radial**, sigmoide]

Tabela 1. Parâmetros testados para cada uma das técnicas.

Desta forma, todos os testes realizados podem ser resumidos de acordo com a Figura 2, incluindo os dois conjuntos de dados avaliados.

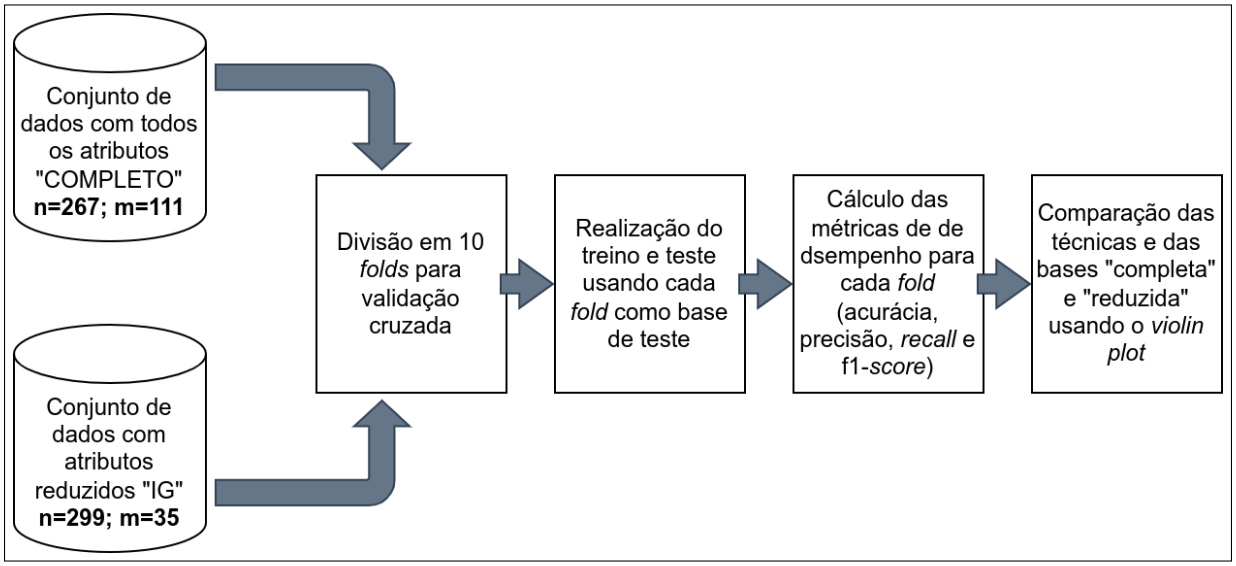

Figura 2. Separação e comparação dos testes utilizados.

As ferramentas utilizadas nesta pesquisa foram a linguagem de programação Python 3 (versão 3.7, disponível em https://www.python.org/) e também o *software* WEKA para implementação do IG, versão 3.8.4. (*Waikato Environment for Knowledge Analysis*), elaborado pela Universidade de Waikato, Nova Zelândia (WEKA, 2021).

RESULTADOS

Esta seção apresenta os resultados obtidos, discutindo-se tanto as métricas de ML como os atributos mais bem ranqueados pela técnica IG. Conforme apresentado na Figura 1, a base de dados completa possui 111 atributos no total. Após a aplicação da técnica IG, apenas 35 atributos foram considerados relevantes para o atributo-classe, ou seja, a percepção "pessimista" ou "otimista" quanto ao Brasil.

Para a avaliação dos resultados, foram gerados gráficos conhecidos como *violin plots*, os quais possuem informações sobre o *kernel density estimation* (KDE), que estima a função densidade de probabilidade (em azul), juntamente com as informações de um *boxplot*. O *boxplot* apresentado no *violin plot* divide os quartis usando linhas tracejadas: o primeiro quartil é representado pelo tracejado inferior, enquanto a mediana ou segundo quartil é ilustrado pela linha tracejada central, e a linha superior diz respeito ao terceiro quartil.

Como primeira métrica de análise, a acurácia verifica o poder de classificação do algoritmo como um todo, ou seja, analisa o total de acertos entre todas as instâncias testadas. A Figura 3 ilustra os resultados para a acurácia em cada uma das técnicas, estratificada por conjuntos.

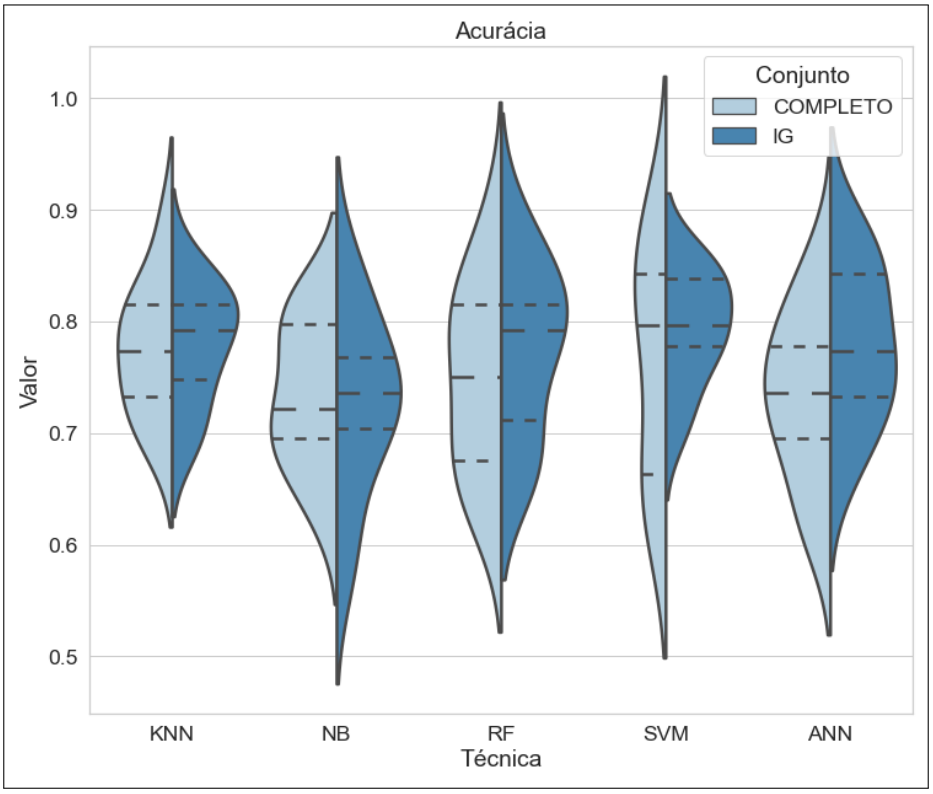

Figura 3. Resultados de acurácia para cada técnica.

A Figura 3 mostra não apenas as medianas dos resultados, mas também a dispersão dos testes realizados. A mediana da acurácia foi próxima de 80% para o KNN, RF e SVM, todas no conjunto IG, mostrando que estas técnicas foram superiores para esta métrica utilizando um conjunto reduzido de atributos. Os resultados também mostram uma diferença entre as bases de dados, verificando-se uma mediana superior para todas as técnicas, exceto para o SVM, ao usar a base com atributos reduzidos (IG). Para a acurácia, percebe-se que NB obteve os menores resultados, obtendo uma mediana não maior do que 74% para qualquer conjunto. Em termos de dispersão, observa-se um valor mais acentuado para o caso do SVM quando aplicado com o conjunto completo, com uma menor dispersão para o mesmo SVM quando aplicado em conjunto reduzido de atributos. O KNN também mostrou ter uma menor dispersão dos resultados quando comparado às outras técnicas, porém ainda podem ser considerados altos para um modelo de ML.

Para a métrica do *recall*, que mede o poder de classificação de uma determinada classe específica de interesse, neste caso a classe "pessimista", a Figura 4 mostra os resultados obtidos.

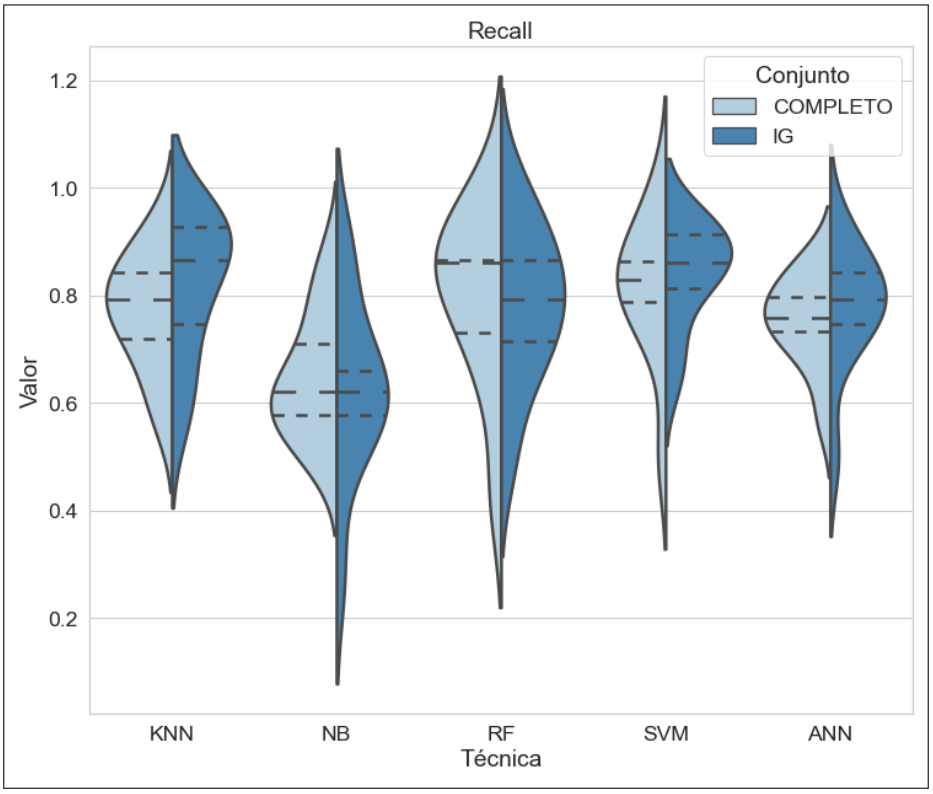

Figura 4. Resultados de *recall* para cada técnica.

A partir da Figura 4, nota-se que há um benefício ao se usar o conjunto reduzido por IG, pois verifica-se um ganho no *recall* para os casos do KNN, SVM e ANN. O *recall* ultrapassou os 86% para o KNN e SVM com o IG, mas a técnica RF também atingiu uma mediana que supera os 86%, mas com o conjunto completo. Verifica-se, assim como na métrica de acurácia, que houve alta dispersão nos resultados encontrados.

Não é recomendado analisar o *recall* sem considerar também a métrica de *precisão*, pois esta última avalia os modelos a partir das instâncias que foram classificadas na predição como "pessimista", observando quantas, de fato, eram da classe "pessimista". Neste contexto, foram gerados os gráficos da Figura 5 sobre a métrica da precisão para todas as técnicas.

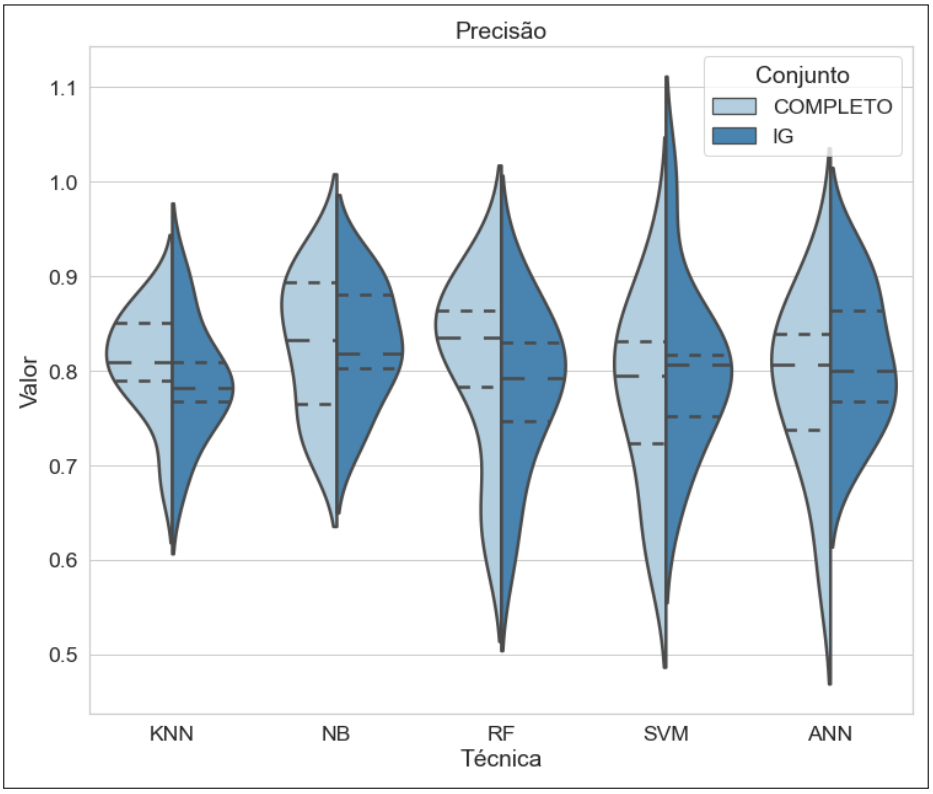

Figura 5. Resultados de precisão para cada técnica.

A Figura 5 ilustra a diminuição da métrica de precisão com a utilização do conjunto reduzido de atributos. Esta situação ocorreu para todas elas, exceto para o SVM, que obteve um ganho sutil, subindo de 79,51% para 80,63%. A mediana das técnicas que tiveram perda diminuiu, para 78,17%, 81,82%, 79,29% e 80% para KNN, NB, RF e ANN, respectivamente. Vale notar que todas as técnicas aqui possuem grande dispersão em seus resultados, com amplitudes que ultrapassam 30% para os casos do SVM (nos dois conjuntos) e ANN, no conjunto completo.

Também é possível notar que os modelos da Figura 5 obtiveram muitos falsos positivos para o conjunto IG. Isso pode ser visto com o ganho do *recall* e a perda da precisão em parte dos modelos, ao mesmo tempo em que estes melhoraram a classificação de *stakeholders* pessimistas como um todo.

Por fim, a métrica *f1-score* busca balancear as métricas de *recall* e precisão calculando a média harmônica entre essas duas métricas (vide Equação 4). Os resultados obtidos estão ilustrados na Figura 6.

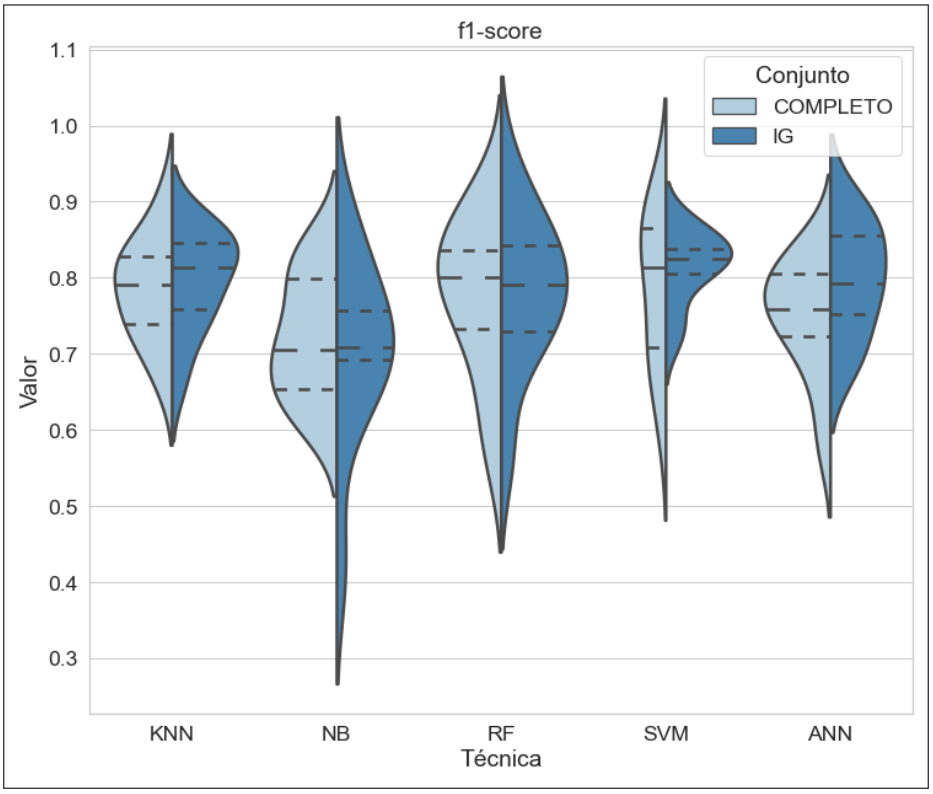

Figura 6. Resultados de *f1-score* para cada técnica.

A Figura 6 mostra que as duas melhores medianas obtidas foram com o do KNN (81,33%) e SVM (82,56%), ambas obtidas no conjunto reduzido por IG. Considerando o *dataset* reduzido pelo IG em relação ao conjunto completo, houve pouca diferença com a mudança de conjuntos de atributos. Apenas as técnicas KNN e ANN obtiveram melhoria mais visível no gráfico. Aqui permaneceu a alta dispersão dos modelos de aprendizado de máquina, com destaque para a alta variação de NB para o conjunto IG.

DISCUSSÃO

Realizando-se a comparação do desempenho das técnicas para cada um dos dois conjuntos testados (completo e reduzido por IG), em alguns momentos houve grandes diferenças nas métricas avaliadas. Os ganhos e perdas obtidos pelo uso do conjunto reduzido por IG em relação ao conjunto completo são apresentados no *heatmap* da Figura 7.

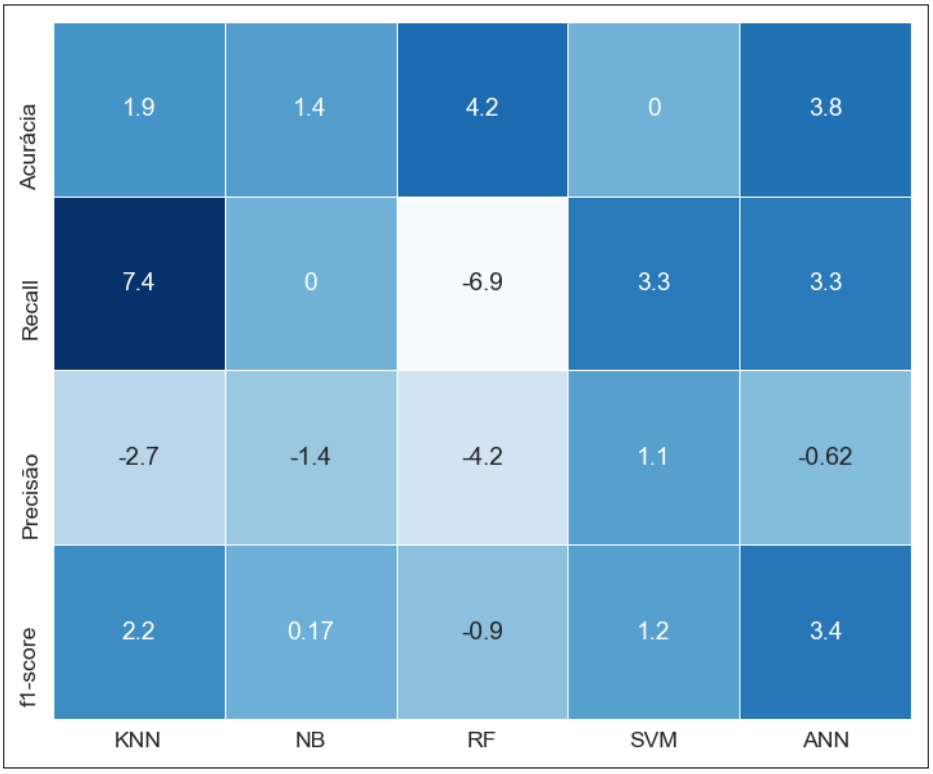

Figura 7. Diferença (em %) do uso do conjunto reduzido por IG em relação ao conjunto completo de atributos.

Verifica-se na Figura 7 que a única técnica que obteve resultados iguais ou melhores a zero foi SVM, enquanto RF teve uma queda expressiva para *recall* e precisão. Ressalta-se que a técnica RF possui uma técnica de filtro embutida para se escolher os melhores atributos para a árvore, enquanto as outras não possuem tais métodos. Isso acaba mostrando que o método de seleção de atributos da RF fez a diferença para esta técnica ter melhores resultados com o conjunto completo. Isto também pode significar que houve um aumento na especificidade do modelo, ou seja, melhorou o seu poder de classificação de *stakeholders* otimistas, uma vez que a acurácia aumentou enquanto as outras métricas diminuíram. A diferença positiva de 7,4 pontos percentuais para o *recall* do KNN mostra a melhora do modelo em classificar *stakeholders* pessimistas, mas a acurácia teve um pequeno ganho. No caso do NB, os ganhos e perdas foram baixos, porém vale destacar que esta técnica mostrou grande variabilidade nos resultados em comparação com as demais técnicas. Deste modo, destacam-se os seguintes resultados preliminares:

- O conjunto reduzido de atributos (identificados por IG) é mais interessante para a classificação da percepção dos *stakeholders*, pois retornou um melhor resultado para a maioria dos testes. Por ter uma menor dimensão do conjunto de dados, reduz também a complexidade dos modelos criados, evitando em muitos casos o *overfitting* do modelo;
- Em termos de técnicas, NB mostrou-se mais limitada em relação às outras técnicas.
- Houve uma alta variação nos resultados dos 10-*folds* testados, havendo técnicas que atingiram, para um mesmo conjunto, uma amplitude de mais de 70%, com o restante dos valores distribuídos não próximos às medianas, criando intervalos largos dos quartis Q1 e Q3 nos *violin plots*.

Importância dos atributos

Como já citado anteriormente, o IG é um método de filtro para a verificação da contribuição de cada atributo para a classe de interesse. Dos 35 atributos considerados relevantes, os 10 que mais contribuíram para os modelos se encontram na Tabela 2.

Ranking	Atributo	Nome	Valor IG	Dimensão	IG da Dimensão
1	a2	Percepção sobre a perspectiva de oportunidades econômicas	0,3404	Oportunidades econômicas	0,3404
2	a7_1	Desenvolvimento humano (saúde, educação e proteção social)	0,2371	Escala de avaliação da direção do Brasil	1,2280
3	a7_3	Crescimento econômico que favorece a criação de empregos	0,2053		
4	a7_2	Oportunidades iguais para homens e mulheres	0,1975		
5	a7_8	Gestão de recursos naturais	0,16		
6	a7_4	Lacuna entre ricos e pobres	0,1464		
7	a7_7	Atração de novas fontes de investimento estrangeiro	0,1411		
8	a7_9	Mudança climática	0,1406		
9	a6_9	Confiança no setor privado internacional	0,1251	Escala de confiança em instituições	0,2404
10	a6_1	Confiança no governo federal	0,1153		

Tabela 2. Atributos mais relevantes e suas dimensões.

O atributo "a2" mede a percepção dos *stakeholders* quanto às oportunidades econômicas no Brasil e foi considerado pelo IG como o fator mais relevante na determinação do sentimento "otimista" ou "pessimista" dos entrevistados. Isso é um indício de que a existência de oportunidades econômicas torna empresas e investidores mais otimistas e mais propensos a realizar investimentos no Brasil.

Também vale destacar que a dimensão que avalia a Direção do Brasil teve sete atributos entre os 10 mais importantes segundo a técnica IG, demonstrando que o sentimento dos *stakeholders* é influenciado por sua percepção de que fatores estratégicos também estão melhorando no país. Por exemplo, a opinião dos entrevistados quanto à melhora ou piora na educação, saúde, distribuição de renda e emprego tem alta importância na determinação de seu posicionamento "pessimista" ou "otimista".

Vale destacar que o atributo "a6_1" foi o único representante da dimensão que mede a confiança dos *stakeholders* nas instituições a figurar entre os 10 mais importantes. Isso é um indicativo de que a confiança dos *stakeholders* no governo federal tem grande importância em seu otimismo com relação ao futuro do país, sendo mais importante que a confiança em outras instituições como o FMI, ONU ou o próprio Banco Mundial.

CONCLUSÃO

Este trabalho comparou técnicas de aprendizado de máquina para classificar a percepção de *stakeholders* sobre o futuro do Brasil, classificando-os em "otimistas" ou "pessimistas". Os dados utilizados foram coletados a partir de uma pesquisa realizada pelo WBG, no ano de 2019. Inicialmente os dados foram pré-processados e técnicas de classificação de ML foram aplicadas a duas versões do conjunto de dados: a primeira consistindo em todos os atributos selecionados da base de dados original, e a segunda consistindo em 35 atributos considerados relevantes pela técnica de ganho de informação (IG).

As técnicas ANN, SVM, RF e KNN foram aplicadas e avaliadas, as quais obtiveram um desempenho satisfatório para a classificação dos entrevistados. Não foi possível afirmar que há diferenças significativas entre as técnicas estudadas. Entretanto, as medianas das métricas indicão que o uso da base de dados com atributos selecionados pelo IG é melhor para a classificação dos *stakeholders* quanto ao futuro do país.

Com relação à importância dos atributos, foi possível observar que a existência de oportunidades econômicas é o principal determinante da opinião dos entrevistados com relação ao seu posicionamento otimista ou pessimista. Também merecem destaque os fatores ligados ao direcionamento país, indicando que a visão positiva ou negativa quanto a educação, saúde, e emprego são fatores relevantes para a classificação da percepção dos *stakeholders*.

Pesquisa futuras podem trabalhar com a aplicação do questionário do WBG para atualizar a percepção dos *stakeholders* e aumentar o tamanho da base de dados, o que pode propiciar melhor treinamento das técnicas de ML. Isso pode melhorar os resultados das métricas e reduzir a alta dispersão observada em todos os experimentos realizados.

REFERÊNCIAS

Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A., & Arshad, H. (2018). State-of-the-art in artificial neural network applications: A survey. *Heliyon*, *4*(11). doi: https://doi.org/https://doi.org/10.1016/j.heliyon.2018.e00938

Aggarwal, C. C. (2015). *Data mining: the textbook* (v. 1). Springer International Publishing. doi: https://doi.org/10.1007/978-3-319-14142-8

Aggarwal, C. C., et al. (2018). Neural networks and deep learning. In *Neural networks and deep learning*. Springer International Publishing. doi: https://doi.org/10.1007/978-3-319-94463-0

Banco Central do Brasil. (2020). *Relatório de mercado focus*. Recuperado de https://www.bcb.gov.br/publicacoes/focus

Bramer, M. (2016). *Principles of data mining* (3a. ed.). Springer London. doi: https://doi.org/10.1007/978-1-4471-7307-6

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32. doi: https://doi.org/10.1007/9781441993267_5

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Chemical Biology & Drug Design*, *20*, 273–297. doi: https://doi.org/10.1111/j.1747-0285.2009.00840.x

Dogan, A., & Birant, D. (2021). Machine learning and data mining in manufacturing. *Expert Systems with Applications*, *166*, 114060. doi: https://doi.org/10.1016/j.eswa.2020.114060

Featherstone, J. D., Ruiz, J. B., Barnett, G. A., & Millam, B. J. (2020). Exploring childhood vaccination themes and public opinions on twitter: A semantic network analysis. *Telematics and Informatics*, *54*, 101474. doi: https://doi.org/10.1016/j.tele.2020.101474

Fergus, P., Idowu, I., Hussain, A., & Dobbins, C. (2016). Advanced artificial neural network classification for detecting preterm births using ehg records. *Neurocomputing*, *188*, 42–49. doi: https://doi.org/10.1016/j.neucom.2015.01.107

Genuer, R., & Poggi, J.-M. (2020). *Random forests with r*. Springer Cham. doi: https://doi.org/10.1007/978-3-030-56485-8

Géron, A. (2019). *Mãos à obra: Aprendizado de máquina com scikit-learn & tensorflow* (1a. ed.). Alta Books.

Haihong, E., Yingxi, H., Haipeng, P., Wen, Z., Siqi, X., & Peiqing, N. (2019). Theme and sentiment analysis model of public opinion dissemination based on generative adversarial network. *Chaos, Solitons & Fractals*, *121*, 160–167. doi: https://doi.org/10.1016/j.chaos.2018.11.036

Han, J., Kamber, M., & Pei, J. (2012). *Data mining concepts and techniques third edition* (Third ed.). Morgan Kaufmann.

Haupt, M. R., Jinich-Diamant, A., Li, J., Nali, M., & Mackey, T. K. (2021). Characterizing twitter user topics and communication network dynamics of the "liberate" movement during covid-19 using unsupervised machine learning and social network analysis. *Online Social Networks and Media*, *21*, 100114. doi: https://doi.org/10.1016/j.osnem.2020.100114

Instituto Brasileiro de Geografia e Estatística. (2020). *Produto interno bruto – pib*. Recuperado de https://www.ibge.gov.br/explica/pib.php

Kafaf, D. A., Kim, D.-K., & Lu, L. (2017). B-knn to improve the efficiency of knn. *Proceedings of the 6th international conference on data science, technology and applications*, 126–132. doi: https://doi.org/10.5220/0006393301260132

Kang, Y., Wang, Y., Zhang, D., & Zhou, L. (2017). The public's opinions on a new school meals policy for childhood obesity prevention in the us: A social media analytics approach. *International Journal of Medical Informatics*, *103*, 83–88. doi: https://doi.org/https://doi.org/10.1016/j.ijmedinf.2017.04.013

Kubat, M. (2017). *An introduction to machine learning*. Springer International Publishing. doi: https://doi.org/10.1007/978-3-319-63913-0

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2018). Feature selection: A data perspective. *ACM computing surveys (CSUR)*, *50*(6), 1–45. doi: https://doi.org/10.1145/3136625

Liu, K., Ergu, D., Cai, Y., Gong, B., & Sheng, J. (2019). A new approach to process the unknown words in financial public opinion. *Procedia Computer Science*, *162*, 523–531. doi: https://doi.org/10.1016/j.procs.2019.12.019

Modu, B., Polovina, N., Lan, Y., Konur, S., Asyhari, A. T., & Peng, Y. (2017). Towards a predictive analytics-based intelligent malaria outbreak warning system. *Applied Sciences*, *7*(8). doi: https://doi.org/10.3390/app7080836

Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106. doi: https://doi.org/10.1257/jep.31.2.87

Myslín, M., Zhu, S.-H., Chapman, W., Conway, M., et al. (2013). Using twitter to examine smoking behavior and perceptions of emerging tobacco products. *Journal of medical Internet research*, *15*(8), e2534. doi: https://doi.org/10.2196/jmir.2534

Oliveira, A., Faria, B. M., Gaio, A. R., & Reis, L. P. (2017). Data mining in hiv-aids surveillance system: application to portuguese data. *Journal of medical systems*, *41*(4). doi: https://doi.org/10.1007/s10916-017-0697-4

Pan, Z., Wang, Y., & Pan, Y. (2020). A new locally adaptive k-nearest neighbor algorithm based on discrimination class. *Knowledge-Based Systems*, *204*, 106185. doi: https://doi.org/10.1016/j.knosys.2020.106185

Patle, A., & Chouhan, D. S. (2013). Svm kernel functions for classification. In *2013 international conference on advances in technology and engineering (icate)* (p. 1–9). doi: 10.1109/ICAdTE.2013.6524743

Puri, M., & Robinson, D. T. (2007). Optimism and economic choice. *Journal of financial economics*, *86*(1), 71–99. doi: https://doi.org/10.1016/j.jfineco.2006.09.003

Raghavendra, N. S., & Deka, P. C. (2014). Support vector machine applications in the field of hydrology: a review. *Applied soft computing*, *19*, 372–386. doi: https://doi.org/10.1016/j.asoc.2014.02.002

Silva, C., Welfer, D., Gioda, F. P., & Dornelles, C. (2017). Cattle brand recognition using convolutional neural network and support vector machines. *IEEE Latin America Transactions*, *15*(2), 310–316. doi: https://doi.org/10.1109/TLA.2017.7854627

Souza, J. G., & Spinola, N. D. (2017). Medidas do desenvolvimento econômico. *RDE-Revista de Desenvolvimento Econômico*, *1*(39), 78. doi:

https://doi.org/10.21452/rde.v1i36.4697

Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. *Expert systems with applications*, *134*, 93–101. doi: https://doi.org/10.1016/j.eswa.2019.05.028

Stekhoven, D. J., & Bühlmann, P. (2011). Missforest—nonparametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112–118. doi: https://doi.org/10.1093/bioinformatics/btr597

Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2019). *Introduction to data mining* (2a. ed.). Pearson Prentice Hall.

Viana, G., & Lima, J. F. (2010). Capital humano e crescimento econômico. *Interações*, *11*(2), 137–148. doi: https://doi.org/10.1590/S1518-70122010000200003

Wang, G., Chi, Y., Liu, Y., & Wang, Y. (2019). Studies on a multidimensional public opinion network model and its topic detection algorithm. *Information Processing & Management*, *56*(3), 584–608. doi: https://doi.org/10.1016/j.ipm.2018.11.010

WBG. (2020). *World bank group country survey 2019.* Recuperado de https://microdata.worldbank.org/index.php/catalog/3511/get-microdata

WEKA. (2021). *Waikato environment for knowledge analysis.* University of Waikato. Recuperado de https://www.cs.waikato.ac.nz/ml/weka/

Zendehboudi, A., Baseer, M. A., & Saidur, R. (2018). Application of support vector machine models for forecasting solar and wind energy resources: A review. *Journal of cleaner production*, *199*, 272–285. doi: https://doi.org/10.1016/j.jclepro.2018.07.164

Zhang, M.-L., & Zhou, Z.-H. (2007). Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, *40*(7), 2038–2048. doi: https://doi.org/10.1016/j.patcog.2006.12.019

Zhang, X., & Gou, H. (2022). Statistical-mean double-quantitative k-nearest neighbor classification learning based on neighborhood distance measurement. *Knowledge-Based Systems*, *250*, 109018. doi: https://doi.org/10.1016/j.knosys.2022.109018

Zhao, J., Henriksson, A., Asker, L., & Boström, H. (2015). Predictive modeling of structured electronic health records for adverse drug event detection. *BMC medical informatics and decision making*, *15*(4). doi: https://doi.org/10.1186/1472-6947-15-S4-S1

Como citar este artigo (APA):
Silva, A. O., Raminelli, D. G. T. L., Santos, B. S.,& Lima, R. H. P. (2023). Classificação das percepções de stakeholders sobre o futuro do Brasil utilizando aprendizado de máquina. *AtoZ: novas práticas em informação e conhecimento, 12*, 1 – 14. Recuperado de: http://dx.doi.org/10.5380/atoz.v12.84075

NOTAS DA OBRA E CONFORMIDADE COM A CIÊNCIA ABERTA

CONTRIBUIÇÃO DE AUTORIA

Papéis e contribuições	Amauri Ornellas da Silva	Daniele Gonçalves de Toledo Luchetta Raminelli	Bruno Samways dos Santos	Rafael Henrique Palma Lima
Concepção do manuscrito	X	X		
Escrita do manuscrito		X	X	X
Metodologia	X	X		
Curadoria dos dados	X		X	
Discussão dos resultados		X		X
Análise dos dados	X		X	X

EQUIPE EDITORIAL