UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

DOUGLAS MARTINS DE SOUZA ROSA

PREDIÇÃO DA SATISFAÇÃO EM RELAÇÃO À DEMOCRACIA NO BRASIL CONSIDERANDO DADOS DE PESQUISA DE OPINIÃO

LONDRINA

2022

DOUGLAS MARTINS DE SOUZA ROSA

PREDIÇÃO DA SATISFAÇÃO EM RELAÇÃO À DEMOCRACIA NO BRASIL CONSIDERANDO DADOS DE PESQUISA DE OPINIÃO

Prediction Of Satisfaction Regarding Democracy In Brazil Considering Opinion Survey Data

Trabalho de conclusão de curso de graduação apresentada como requisito para obtenção do título de Bacharel em Nome do Curso de Engenharia de Produção da Universidade Tecnológica Federal do Paraná (UTFPR).
Orientador(a): Dr. Bruno Samways dos Santos

LONDRINA

2022

DOUGLAS MARTINS DE SOUZA ROSA

PREDIÇÃO DA SATISFAÇÃO EM RELAÇÃO À DEMOCRACIA NO BRASIL CONSIDERANDO DADOS DE PESQUISA DE OPINIÃO

Trabalho de conclusão de curso de graduação apresentada como requisito para obtenção do título de Bacharel em Nome do Curso de Engenharia de Produção da Universidade Tecnológica Federal do Paraná (UTFPR).

07/Junho/2022

_____

Rafael Henrique Palma Lima
Doutor
Universidade Tecnológica Federal do Paraná (UTFPR)

_____

Pedro Rochavetz De Lara
Doutor
Universidade Tecnológica Federal do Paraná (UTFPR)

_____

Bruno Samways dos Santos
Doutor
Universidade Tecnológica Federal do Paraná (UTFPR)

LONDRINA

2022

AGRADECIMENTOS

Agradeço a minha família, por proporcionar o privilégio de seguir esse caminho, pelos incentivos e apoio.

Agradeço aos meus professores que contribuíram de diversas formas para minha caminha como profissional, em especial ao meu orientador que me auxiliou de forma contínua nesses últimos passos de minha formação.

Agradeço aos meus amigos que me auxiliaram nos estudos, dúvidas e que estiveram presentes durante essa jornada.

Agradeço a minha noiva por estar ao meu lado em todos os momentos, sempre me auxiliando.

RESUMO

Pesquisas de opinião, em questionários, são comumente aplicadas em grandes grupos e a área da democracia pode ser medida com esse tipo de pesquisa, gerando possibilidades de aplicação de técnicas da área de mineração de dados para se obter conhecimento. Nesse trabalho foi utilizada uma base de dados em que o questionário buscou analisar posições sobre satisfação com a democracia, sobre economia, problemas sociais, confiança em instituições e governos. A partir desse questionário, este trabalho fez o uso de métodos de aprendizagem de máquina (AM) supervisionada buscando classificar a satisfação com a democracia no Brasil. Os métodos de classificação utilizados foram *Support Vector Classification* (SVC), *Random Forest* e Redes Neurais Artificiais (RNA), a base de dados utilizada foi de uma pesquisa de opinião do Latinobarómetro, organização privada sem fins lucrativos, ano de 2020, com 20.204 instâncias e 408 atributos. Para realizar a limpeza da base de dados e reduzir o questionário para dados mais relevantes ao trabalho foi utilizado o Índice de Democracia Local (IDL), que busca através de cinco divisões e subdivisões avaliar a qualidade da democracia. Após realizado as iterações o melhor classificador foi o Random Forest com resultados superiores aos outros métodos aplicados e acurácia de 81%. Os dois principais atributos que contribuíram para compreender melhor a escolha por estar “satisfeito” ou “insatisfeito” com a democracia foram “Satisfação com a situação econômica (em geral)” e “confiança no governo”.

Palavras-chave: Aprendizado de máquina, Pesquisa de Opinião, Classificação, Democracia.

ABSTRACT

Opinion surveys, in questionnaires, are commonly applied in large groups and the area of democracy can be measured with this type of research, generating possibilities for the application of techniques in the area of data mining to obtain knowledge. In this work, a database was used in which the questionnaire sought to analyze positions on satisfaction with democracy, on the economy, social problems, trust in institutions and governments. From this questionnaire, this work made use of supervised machine learning (MA) methods seeking to classify satisfaction with democracy in Brazil. The classification methods used were Support Vector Classification (SVC), Random Forest and Artificial Neural Networks (ANN). and 408 attributes. In order to clean the database and reduce the questionnaire to more relevant data for the work, the Local Democracy Index (IDL) was used, which seeks through five divisions and subdivisions to assess the quality of democracy. After conducting the iterations, the best classifier was Random Forest, with results superior to the other applied methods and an accuracy of 81%. The two main attributes that contributed to a better understanding of the choice for being “satisfied” or “dissatisfied” with democracy were “Satisfaction with the economic situation (in general)” and “trust in the government”.

Keywords: *Machine learning, Survey Research, Classification, Democracy.*

SUMÁRIO

1.INTRODUÇÃO....8
1.1 OBJETIVO GERAL....9
1.2 Objetivos específicos....9
1.3 Justificativa....9
2.REFERENCIAL TEÓRICO....11
2.1 Mineração de dados e Knowledge Discovery in Databases (KDD)....11
2.2 Tipos de aprendizado de máquina....12
2.3 Técnicas de classificação....15
2.3.1 Redes Neurais....15
2.3.3 Máquina de Vetores de Suporte....18
2.4 Métodos de validação....21
2.4.1 Validação cruzada (*Holdout cross-validation*)....21
2.4.2 Validação cruzada em k-*fold*....21
2.5 Métricas de avaliação....22
2.6 Opinião pública e satisfação com a democracia....23
3.METODOLOGIA....25
4.RESULTADOS E DISCUSSÃO....29
4.1 Resultados da análise descritiva....29
4.2 Discussão da análise....37
5.CONCLUSÃO....39
REFERÊNCIAS....40

1. INTRODUÇÃO

Recentemente, muito se tem falado sobre o termo *Knowledge Discovery in Databases* (Descoberta de Conhecimento em Bases de Dados), também conhecido como KDD, e *data mining* (mineração de dados), pois a partir deles, partem muitas informações e ferramentas importantes. Segundo Fayyad *et al*. (1996), os dados gerados e as análises realizadas por diferentes setores como o *marketing*, financeiro e saúde, são geradores de dados. Os setores analisam e interpretam estes dados de forma ineficiente, muitas vezes manualmente, o que gera maior lentidão no processo de exploração e interpretação, retardando o tempo para conseguir os resultados. Estes processos se tornam custosos e acabam gerando conclusões subjetivas, além do aumento do volume de dados constantemente, o que tornaria essa análise manual impraticável, tendo em vista que as variáveis podem chegar na casa dos *Zettabytes*, 10^{21} bytes de dados (NAVARRO *et al*., 2021).

Devido ao crescimento da quantidade de dados gerados, o uso da tecnologia para se armazená-los e interpretá-los se torna imprescindível, necessitando-se a utilização de metodologias e algoritmos para o processo de coleta, transformação, exploração, mineração de dados e interpretação. Uma das fases importantes do processo KDD, a mineração de dados, pode ser definida como a aplicação de algoritmos para conseguir extrair padrões de bancos de dados, padrões que podem ser desconhecidos ou ainda não identificados, e que poderiam ser encontrados através de muito tempo e esforço manual (BHOJANI & BHATT, 2016). Muitas vezes não se tem tempo e recursos disponíveis, ainda mais com a velocidade das mudanças e o crescimento, com taxas sem precedentes, de base de dados comerciais (VERMA; NASHINE, 2012).

Um dos campos de análises em que a mineração de dados pode ser aplicada é na área de pesquisa de sentimento ou também chamada de mineração de opinião. Segundo Bing Liu (2012), após o ano 2000, se tornou um campo muito ativo de pesquisa, sendo utilizada amplamente em quase todos os setores e ao grande volume de dados de opinião gerados para processamento.

O foco deste trabalho foi utilizar as técnicas de mineração de dados para a mineração de dados em um banco de dados disponibilizado pela Latinobarómetro, organização privada sem fins lucrativos, localizada no Chile, que realiza pesquisas de

opinião pública em 18 países da América Latina com mais de 20 mil pessoas entrevistadas, que busca representar a população de 600 milhões de pessoas na América Latina. Para selecionar e reduzir o questionário foi utilizado o Índice de Democracia Local (IDL). Assim, buscou-se classificar a satisfação com a democracia no Brasil e entender sobre os fatores que mais influenciam neste processo.

1.1 Objetivo geral

Aplicar técnicas de mineração de dados para a tarefa de classificação quanto a opinião pública sobre a satisfação com a democracia, a partir da base de dados da pesquisa realizada pelo Latinobarómetro Corporation.

1.2 Objetivos específicos

- Explorar uma base de dados de opinião pública;
- Realizar o pré-processamento dos dados;
- Aplicar as técnicas de mineração de dados para classificação;
- Extrair informações sobre variáveis que mais influenciam a decisão dos entrevistados referente a satisfação com a democracia;

1.3 Justificativa

De acordo com Bing Liu (2012), quando uma empresa ou organização precisa da opinião de consumidores, elas buscam realizar pesquisas com grande público, grupos de foco ou enquetes. Assim, pode-se realizar a tomada de decisões em determinados assuntos buscando a opinião de um terceiro, uma avaliação de produto ou a validação com embasamento empírico.

A pesquisa realizada pelo Latinobarómetro busca analisar o desenvolvimento econômico e democrático desses países, com uma base de dados sólida e confiável, financiada por grandes parceiros como união Europeia, *United States of Research e Inter-American Development Bank* (IADB).

Trabalhos foram realizados na área de desenvolvimento econômico e satisfação com a democracia, mas voltados para área de ciências sociais e dados do Latinóbarometro, mas com foco estatísticos como média do PIB e porcentagem da satisfação com a democracia na América Latina (RESENDE, EPITÁCIO, 2014). Gomes e Aquino (2018) buscaram verificar uma os efeitos da violência sobre a satisfação com a democracia, utilizando a hipótese de duplo vetor, sendo direto

reduzindo apoio instrumental ao regime democrático ou indireto reduzindo a confiança interpessoal, utilizando regressão linear, equações simultâneas e modelos não paramétricos na base de dados do Barômetro das américas, de 2014.

Neste trabalho o tipo de abordagem é inédita, não encontrada na literatura, em que busca aplicar métodos de classificação em pesquisa de opinião voltada para a satisfação com a democracia. Assim o objetivo foi encontrar quais os modelos que melhor classificam a variável “satisfação com a democracia” de acordo com as métricas de acurácia, *recall*, F1 Score e precisão, também será verificado quais variáveis mais influenciam para a classificar a variável de saída, “satisfeito” ou “insatisfeito”.

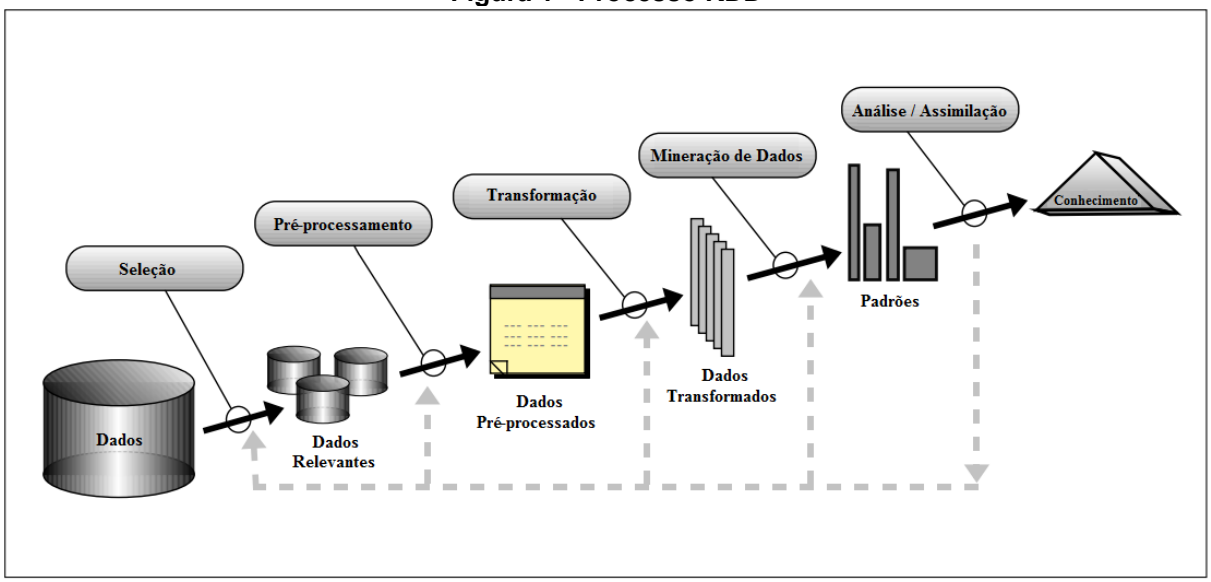
2. REFERENCIAL TEÓRICO

O capítulo atual descreve alguns tópicos muito importantes sobre mineração de dados, técnicas, métodos e referencial sobre opinião pública.

2.1 Mineração de dados e Knowledge Discovery in Databases (KDD)

A origem mineração de dados ocorreu ao final da década de 80 (COENEN, 2014). Com o passar dos anos o conceito foi se desenvolvendo, e no começo da década de 90, ficou comumente reconhecido como um subprocesso do KDD, Figura 1, que por sua vez pode ser definido como um processo não trivial de identificação válida, singular, potencialmente utilizável e gerando uma compreensão dos padrões encontrados nos dados (FAYYAD *et al*., 1996).

Figura 1 - Processo KDD



Fonte: FAYYAD *et al.* (1996, p.41)

Outro subprocesso que compõe o KDD é a preparação dos dados que envolve a forma de armazenamento e acesso dos dados, de forma a criar um processo orgânico e fluído. Inclui também a utilização sobre os dados de diferentes fontes e formas de armazenamento. A limpeza dos dados, visa retirar dados incompletos, erros e informações que possam gerar ruídos ou perturbações, que geralmente são ocasionados por conta de sua geração, captação ou entrada (FACELI *et al*., 2021).

Além da limpeza dos dados, que pode ser feita manualmente, também se utiliza algoritmos e técnicas de pré-processamento para agilizar o processo e minimizar os erros, verificando-se a sequência das informações e realizando-se mudanças caso necessário. Por exemplo, bases de dados podem ser transformadas de nominais ou texto para numéricas, podendo assim serem trabalhadas em algoritmos conhecidos que atuam somente com valores numéricos (FACELI *et al*., 2021).

Após a limpeza, organização e transformação dos dados, pode ser feita uma análise exploratória, gerando-se algumas informações prévias sobre os dados de forma descritiva e visual. Esta análise exploratória também pode vir a ser realizada de forma concomitante à limpeza e transformação.

Seguindo com o KDD, a próxima etapa é a mineração de dados, que difere dos algoritmos de aprendizado de máquina (AM). De acordo com Coenen (2014), mineração e dados é focada nos dados, em todos os formatos, visto como um campo de aplicação. Já AM é voltado ao algoritmo, com o qual o computador consegue aprender e replicar esse aprendizado, podendo prever se irá, por exemplo “chover”, de acordo com as entradas fornecidas e os dados que geraram o aprendizado, ou até “jogar xadrez”, onde os trabalhos iniciais em AM começaram. Assim, o AM pode ser visto como uma tecnologia e mineração e dados uma aplicação.

Para analisar os dados na mineração um dos primeiros passos seria a escolha do algoritmo, realização de testes e acurácia, validação, avaliação para verificar se o resultado está congruente. Uma definição para o resultado da mineração de dados seria obter conhecimento na forma de regras que oriente uma decisão, se feito de modo correto pode-se fazer previsões ou descobrir novas associações (GONZALES, ZAMPIROLLI, 2014)

A Figura 1 representa de forma visual aquilo que foi falado anteriormente, demonstrando os subprocessos do KDD, de forma a desenvolver uma metodologia consistente, coerente e padronizada. Assim como Fayyad *et al.*, (1996) afirma sobre os princípios de um bom algoritmo representação do modelo, validação do modelo e a busca.

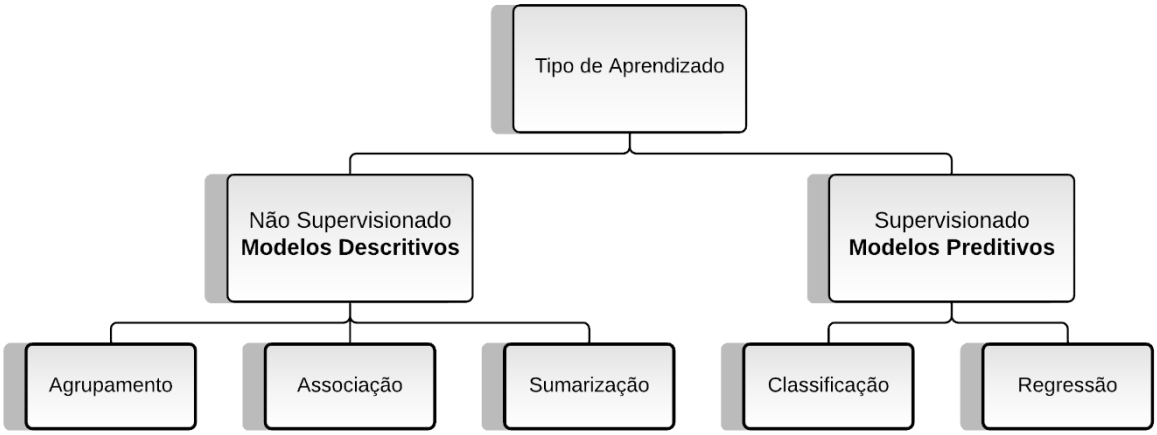
2.2 Tipos de aprendizado de máquina

Algoritmos de AM geralmente são divididos em duas categorias: preditivos e descritivos. Os modelos descritivos e não supervisionados buscam o agrupamento,

associação ou a sumarização dos dados, assim segundo Faceli *et al*., (2021), essas tarefas buscam extrair padrões de um conjunto de dados apresentados.

Em contrapartida os modelos preditivos utilizam um banco de dados como treinamento buscando predizer um valor alvo final, de acordo com determinada entrada e os parâmetros da base de dados. Nesse modelo, é seguido o aprendizado supervisionado, pois são necessários dados passados ou já existentes para treinar o algoritmo. Comumente é dividido em classificação e regressão, Figura 2.

Figura 2 - Tipos de aprendizagem de máquina



Fonte: Faceli *et al.* (2021, p. 3)

Entre os três modelos descritivos, agrupamento, associação e sumarização, o agrupamento, também conhecido como clusterização, possui o objetivo de encontrar e estruturar em grupos a base de dados trabalhada (FACELI *et al.,* 2021). Por exemplo no agrupamento, uma base de dados pode ocorrer agrupamento por idade, sexo, altura, peso ou de uma relação entre essas informações.

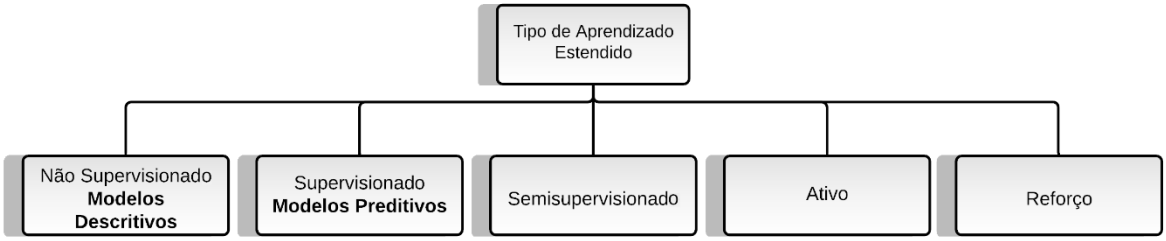
A associação se refere à busca de padrões recorrentes que se associam de alguma maneira entre as características (variáveis) de um conjunto de dados (FACELI *et al.*, 2021). Geralmente, esse modelo é utilizado no setor financeiro, para associar e oferecer serviços que possam ser vendidos em conjunto ou para *marketing* selecionado.

Sumarização tem o objetivo mais simples, pois busca descrever de forma simples e compacta os dados (MIRKIN, 2011). Assim, pode ser utilizado estatística simples e inferencial, com o intuito de sumarizar as informações.

Em relação aos modelos preditivos supervisionados, tem-se a classificação e regressão. Esses modelos buscam antever o que uma instância se tornará (predição), de acordo com os dados de entrada, sendo um estimador. Na classificação, o algoritmo é treinado na base de dados conhecida com valores nominais, no caso da regressão se essa base de dados é finita e ordenada, possuindo alguma tendência de crescimento ou decrescimento, pode ser considerado um problema de regressão. No caso da classificação e da regressão dado uma nova entrada, exemplo, que não está na base de dados eles atribuem a uma das possíveis classes, assim podem ser considerados também uma função (DIETTERICH, 1998).

Apesar dos modelos descritivos e preditivos serem os mais utilizados, também existem outros três, que seriam o semissupervisionado, aprendizado ativo e aprendizado por reforço, Figura 3.

Figura 3 - Tipos de aprendizado de máquina estendido



Fonte: Faceli *et al.* (2021, p. 4)

O modelo semissupervisionado pode ser uma tarefa tanto de agrupamento, classificação ou regressão. Ele é utilizado onde somente parte dos dados estão rotulados, assim ele procura aumentar o número de objetos rotulados. Essa ação de rotular pode ser difícil e dispendiosa, por isso de sua utilização, no entanto pode causar redundância que não contribuem para um bom modelo (FACELI, *et al*., 2021).

O problema de redundância é suprido pelo modelo ativo, que seleciona iterativamente os dados que serão rotulados e seu rótulo, assim somente o que não possuir atributo similar aos existentes serão escolhidos, do contrário, não serão utilizados. O modelo por reforço pune em ações que são consideradas negativas e recompensa em atividades consideradas positivas, de acordo com a métrica utilizada (FACELI, *et al*., 2021). Para este trabalho, foram utilizados os modelos para classificação, descrito com maior profundidade na sequência.

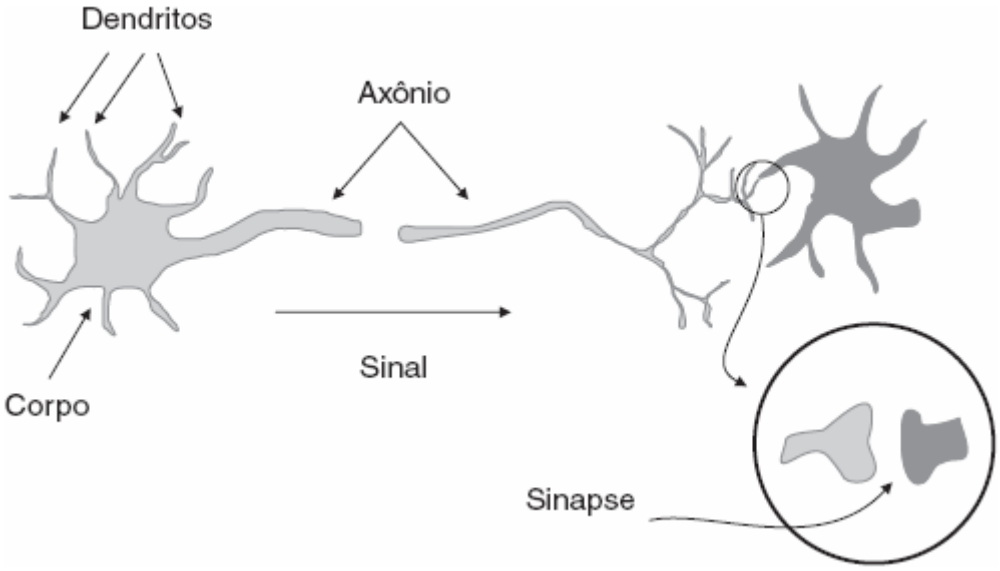
2.3 Técnicas de classificação

As técnicas de classificação utilizadas nesta pesquisa foram: Redes Neurais, *Random Forest* (Floresta Aleatória) e *Support Vector Machines* (Máquina de Vetores de Suporte).

2.3.1 Redes Neurais

Na busca por melhoramentos no desenvolvimento de máquinas inteligentes, um modelo proposto foi baseado no cérebro humano, tido como uma estrutura complexa que torna tarefas consideradas difíceis, fáceis, como Facelli *et al*., (2021), evidencia na dificuldade de ensinar robôs a andar, correr e pegar objetos. Dessa forma, o estudo foi direcionado para compreender o funcionamento do cérebro humano e aquilo que é a base para seu funcionamento, os neurônios, esse também é a base para a técnica de Redes Neurais Artificiais (RNAs). Um neurônio biológico é representado pela Figura 4.

Figura 4 - Neurônio biológico simplificado

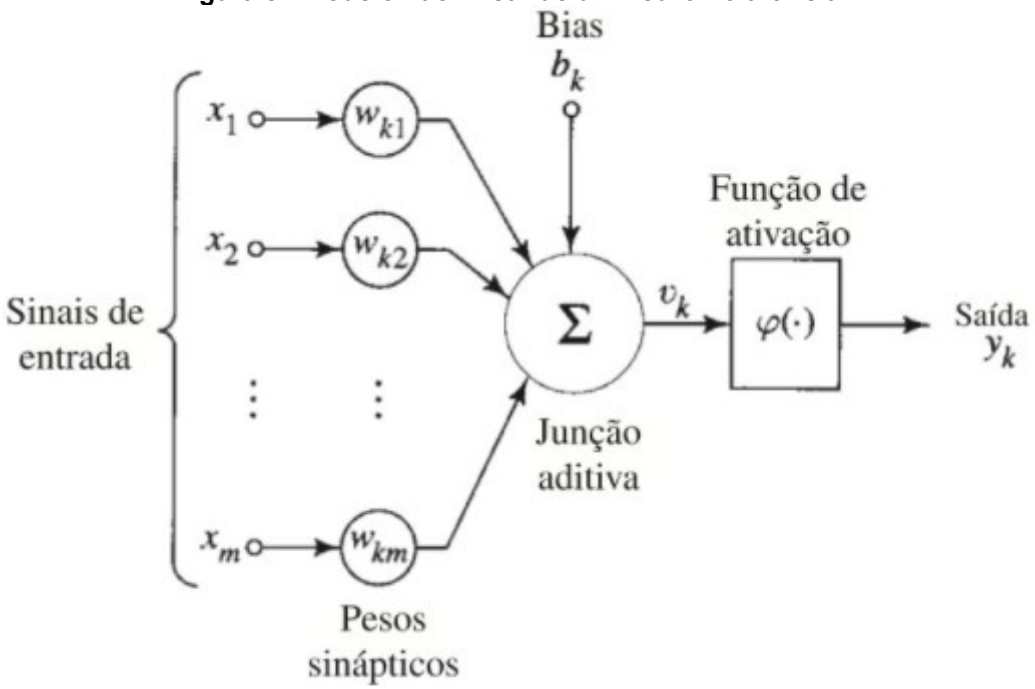


Fonte: Faceli *et al.* (2021, p. 102)

Com base nos neurônios biológicos, as RNAs foram criadas pensando em um sistema interligado de neurônios artificiais de processamento simples, organizados em uma ou mais camadas, interligados por um alto número de conexões, acionando funções matemáticas e funcionando paralelamente (FACELI *et al*., 2021).

Para compor uma RNA são necessários vários neurônios e unidade de processamento da RNA, fundamental na RNA. Na Figura 5 são mostrados os três elementos básicos: o conjunto de sinapses ou elos de conexão, o combinador linear e a função de ativação (HAYKIN, 2011).

Figura 5 - Modelo não-linear de um neurônio artificial



Fonte: Haykin (2011, p.36)

O conjunto de sinapses é caracterizado por um peso. Segundo Haykin (2011) Um sinal xi no início da sinapse j atrelada ao neurônio k é multiplicada pelo peso sináptico Wkj, no caso do peso sináptico o índice i refere-se ao neurônio e o índice $j$ao terminal de entrada da sinapse a que o peso está atrelado, podendo o peso sináptico estar em um intervalo positivo ou negativo.

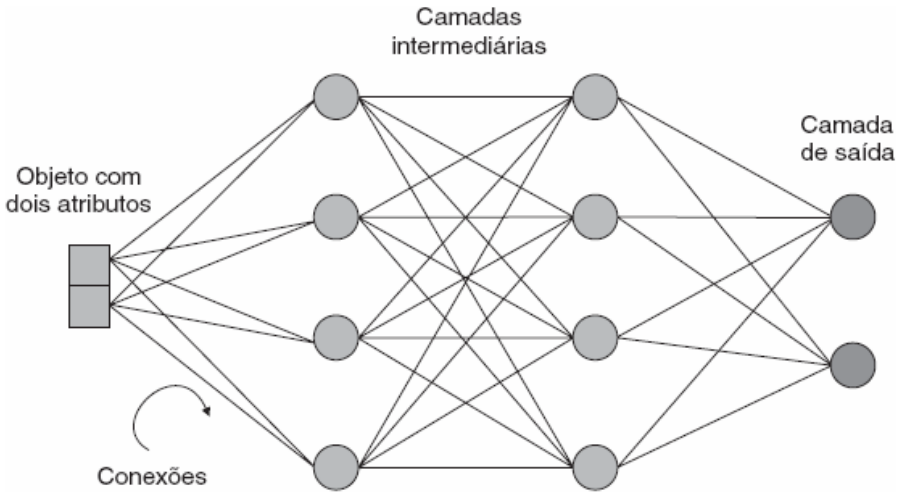
O somador representado por ∑, realiza a somatória dos sinais de entrada de forma ponderada pelas sinapses dos neurônios, o que forma uma combinação linear. De acordo com Haykin, (2011), a função e ativação ou função restritiva, chamada assim, pois limita o intervalo permitido de amplitude da saída do neurônio. O modelo também adiciona um *bias* (ou viés, em português), bk, aplicado de forma externa, possuindo o efeito de aumentar caso positivo ou diminuir, caso negativo, a entrada a função de ativação (HAYKIN, 2011).

A função de ativação define qual será a saída de um neurônio (HAYKIN, 2011). Temos as mais conhecidas: linear identidade, a limiar de McCuloch e Pitts que

desenvolveram em 1943, a função Sigmoidal, a função tangente hiperbólica uma variação da Sigmoidal, a gaussiana e a função linear retificada (ReLu, do inglês, *Rectified Linear Unit*) (FACELI et al., 2021).

Na Figura 6, tem-se a representação de como um neurônio disposto em três camadas, nesse caso a RNA recebe de dois atributos de entrada e gera dois de saída.

Figura 6 - RNA multicamadas



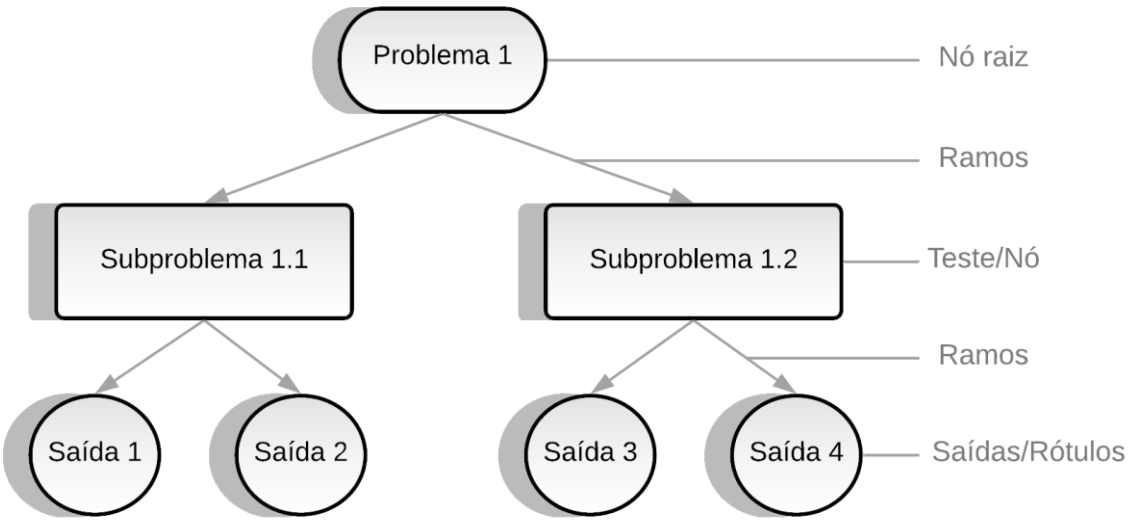
Fonte: Faceli *et al*. (2021, p.105)

Os neurônios podem possuir uma ou mais camadas dependendo da necessidade, quanto mais camadas, mais profunda se torna e maior a capacidade de aumentar a aprendizagem para situações complexas, conhecida também por redes neurais profundas (ou do inglês, *deep neural networks*). Quando um neurônio possui mais de uma camada, ele pode receber no terminal de entrada a saída de um neurônio de camada anterior ou enviar sua saída para a próxima camada (FACELI *et al*., 2021).

2.3.2 Floresta Aleatória

Árvore de decisão é a base para a floresta aleatória, e na árvore de decisão, o intuito é dividir problemas complexos em problemas mais simples e, se necessário, reaplicar essa estratégia até atingir um objetivo satisfatório (FACELI *et al*., 2021). Ela é composta por um fluxograma montado similar à estrutura de uma árvore, Figura 7, cada nó interno, é referente ao teste de um atributo, o ramo seria a saída do teste e cada nó folha ou terminal seria o rótulo da classe, o nó superior representaria o nó raiz (HAN et al., 2011).

Figura 7 - Exemplo árvore de decisão



Fonte: Adaptado Faceli *et al*. (2021, p.79)

O algoritmo das florestas aleatórias foi introduzido por Ho (1995), em que aplicava um novo método nas generalizações, aumentando a acurácia no AM (BREIMAN, 2001). Esse método foi conhecido como Florestas aleatórias que utilizava diversas árvores de decisões, onde as combinações geravam um resultado único de acordo com a decisão do mais popular.

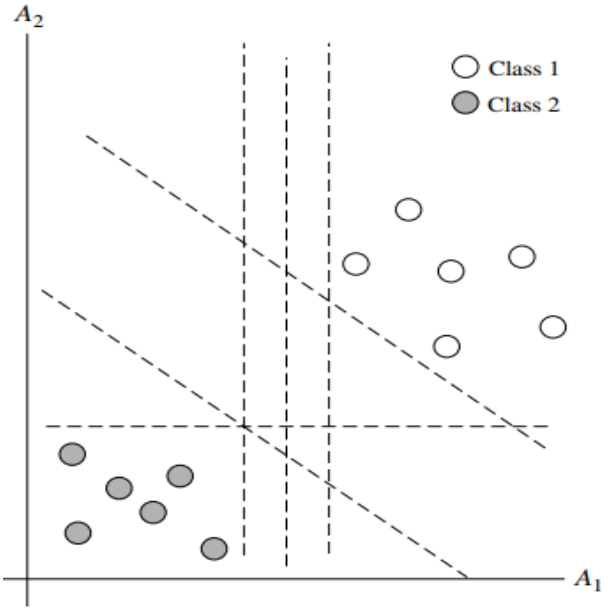
Esse algoritmo é baseado em *bagging,* utilizando como exemplo a árvore de decisão, a qual possui uma resposta A, B, C ou D, esse método iria gerar n árvores de decisão e as respostas seriam dadas de acordo com a maior ocorrência de resposta, de forma majoritária. *Bagging* é um tipo de *ensemble methods,* onde ocorre uma composição de modelos e uma combinação de classificadores, buscando aumentar a acurácia na classificação (HAN *et al.*, 2011). Segundo Faceli *et al.,* (2021) esse é considerado um dos algoritmos mais competitivos.

2.3.3 Máquina de Vetores de Suporte

Support Vector Machines (SVMs) ou Máquina de Vetores de Suporte, é um método de classificação ou regressão para dados lineares ou não lineares. Esse algoritmo é baseado em maximização de margens conhecidas como “hiperplanos” (FACELI *et al*, 2021). O tempo de treinamento desses algoritmos são lentos, mas possuem alta acurácia, devido a sua habilidade de modelar complexos limites de fronteira (HAN *et al.*, 2011).

As SVMs se originaram pela utilização da teoria do aprendizado estatístico (FACELI *et al.*, 2021). SVMs linear com margens rígidas possui o intuito de definir uma fronteira linear a partir dos dados linearmente separáveis, separando por um hiperplano, e o conjunto de dados possuindo somente duas classes. A Figura 8 ilustra as possibilidades de separação dos planos, entre as classes que supostamente poderiam ser separadas, verticais e transversais.

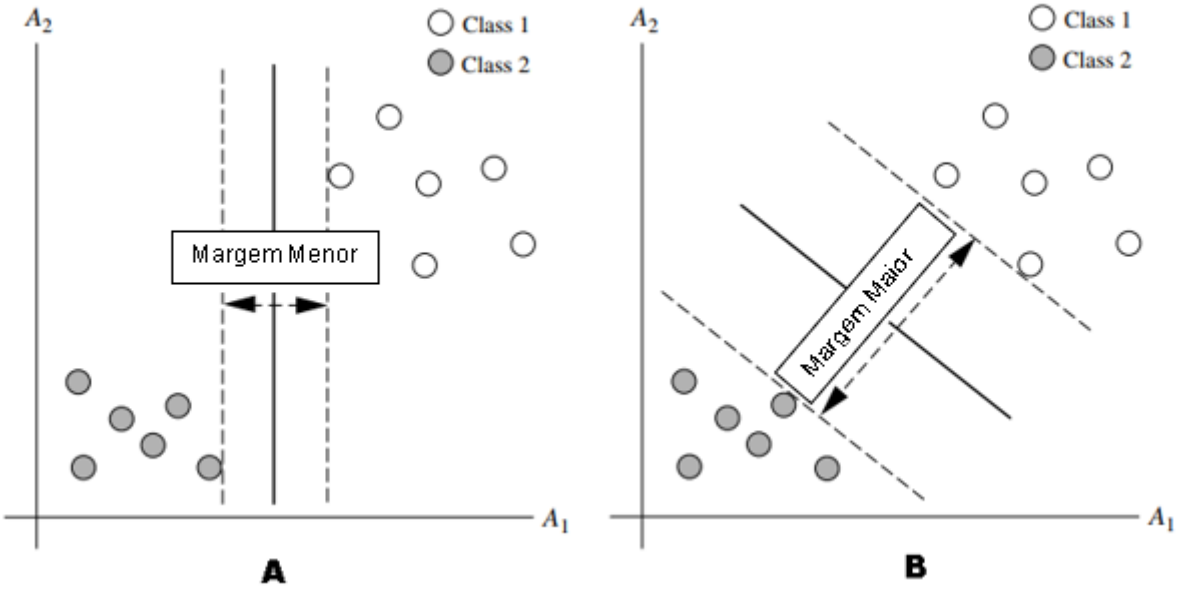
Figura 8 - Conjunto de dados SVM Linear



Fonte: Adaptado Han et al. (2011, p.409)

Na Figura 9, ocorre a separação pelos hiperplanos e as margens associadas. Na margem (A) tem-se um método de separação e na (B) foi realizado outro tipo, buscando a maximização das margens. Neste contexto, o método com melhor resultado encontrado é o (B) (HAN *et al.*, 2011).

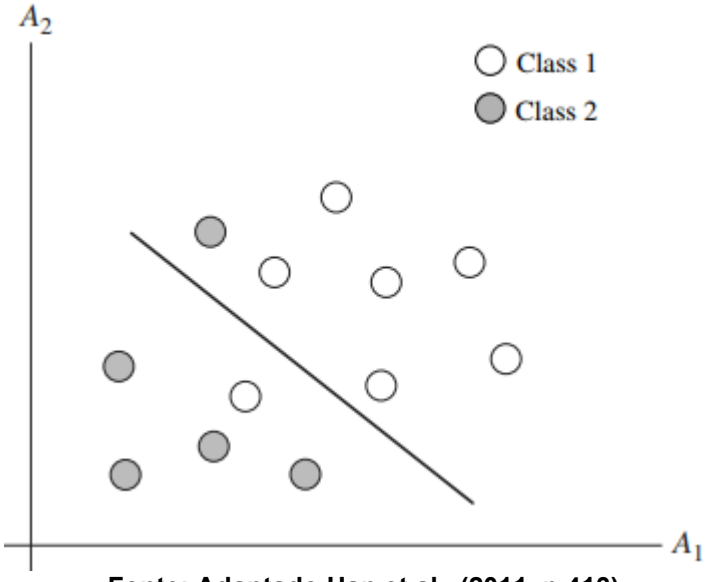
Figura 9 - Possibilidades de separação de hiperplanos SVM Linear



Fonte: Adaptado Han et al., (2011, p.410)

Quando os dados não estão linearmente separáveis (como no exemplo da Figura 10) é necessária uma nova forma de abordagem.

Figura 10 - Dados não linearmente separáveis



Fonte: Adaptado Han et al., (2011, p.413)

Nos casos não linearmente separáveis é necessário realizar uma transformação do espaço para um de maior dimensão, dessa forma ele será transformado em de 3D para 6D, por exemplo, resultando em um maior plano. Matematicamente, são utilizadas funções chamadas kernels para otimizar os cálculos

nesses casos. Ressalta-se que as SVM também podem ser combinados para casos multiclasses (HAN et al., 2011).

2.4 Métodos de validação

A validação cruzada pode ser utilizada como técnica de cessar o treinamento de modelos preditivos, mas a aplicação mais comum é como um método de validação de desempenho (FERRARI; DE CASTRO, 2016).

2.4.1 Validação cruzada (*Holdout cross-validation*)

Segundo Kovac (1995), o método *Holdout* separa os dados em dois grupos mutuamente exclusivos, um chamado de grupo de treinamento, comumente utilizando 2/3 dos dados, e o outro chamado de grupo de teste com 1/3 dos dados. A estimativa do método *holdout* pode ser encontrada após a repetição do modelo, k vezes, e tirada a média dessas medidas.

2.4.2 Validação cruzada em k-*fold*

A validação cruzada em k-*fold* (ou do inglês, k-*fold cross validation*) consiste em dividir a base de dados em k subconjuntos, uma *fold* (pasta) sendo para teste e as outras $k - 1$ *folds* sendo para treinamento. Esse procedimento de treino é repetido com todos os k subconjuntos. A média dos desempenhos é tomada como um indicador da qualidade do modelo (FERRARI; DE CASTRO, 2016).

Segundo Castro (2016), na divisão da base um aspecto importante é a validação cruzada estratificada, onde as classes são distribuídas uniformemente entre as *folds*, ou seja, sempre é necessário manter a proporção dos conjuntos de objetos ao se classificar as bases de treinamento e teste.

Existem dúvidas na quantidade adequada de *folds* para bons resultados, o usual seriam 10-*folds*, exemplificado na Figura 11. Também é utilizada a validação em 10-*folds* 10 vezes, gerando assim um melhor resultado (FERRARI; DE CASTRO, 2016). A estratificação, segundo Maimom (2010), gera uma grande melhora nos resultados.

Figura 11 - Validação cruzada

Fonte: Adaptado Ferrari; de Castro (2016, p158)

2.5 Métricas de avaliação

O desempenho de um bom método de classificação está diretamente ligado à flexibilidade, viés e a qualidade do treinamento, e à sua variância (validação). As medidas de avaliação de desempenho baseiam-se em uma taxa de erro ou acerto do método de classificação (FERRARI; DE CASTRO, 2016).

Uma forma usual de construção das métricas de avaliação é por meio da matriz de confusão, também chamada de matriz de contingência ou conhecida ainda como matriz erro, mostrado na Tabela 1.

Tabela 1 - Exemplo matriz de confusão binária

Matriz de Confusão		
	Classe Predita	
Classe Original	Positiva	Negativo
Positiva	VP	FN
Negativo	FP	VN

Fonte: Adaptado Ferrari; de Castro (2016, p.160)

De acordo com Ferrari e de Castro (2016):

- VP (Verdadeiro Positivo): Objeto de classe positiva classificado como positivo;
- VN (Verdadeiro Falso): Objeto de classe negativa classificado como negativo;
- FP (Falso Positivo): Objeto de classe negativa classificado como positivo, que seria um alarme falso ou erro tipo 1

- FN (Falso Negativo): objeto de classe positiva classificado como negativo ou erro tipo 2.

A taxa global de sucesso do algoritmo, também conhecido como acurácia (ACC), é dada pela fórmula 1:

$$ACC = \frac{VP+VN}{VP+VN+FP+FN} \quad (1)$$

A precisão (Pr) mede a qualidade ou a exatidão do algoritmo, podendo ser entendido como a possibilidade de um item recuperado ser relevante, está ligada dessa forma a relevância do algoritmo e é definida como:

$$Pr = \frac{VP}{VP + FP} \quad (2)$$

Recall, sensibilidade ou revocação (Re) também está ligada a relevância do algoritmo, a probabilidade de recuperação de um item relevante. Seu objetivo é mensurar a completude do algoritmo e é dada como:

$$R = \frac{VP}{VP + FN} \quad (3)$$

F1 Score é definido como a média harmônica da precisão e do recall, desenvolvido para trabalhar bem com dados desbalanceados:

$$F1\ Score = 2 * \frac{Pr * R}{Pr + R} \quad (4)$$

2.6 Opinião pública e satisfação com a democracia

O vocábulo “público” se origina do latim *publicus,* que pode ser traduzido como relativo ao povo. A opinião pública pode ser definida de formas diferentes de acordo com cada autor, mas o que se sabe é que ela é varia de acordo com diversos fatores e sobre isso Lazarsfeld (1972), diz que a opinião pública não pode estar presa a uma fórmula.

A opinião pública ainda pode ser muitas vezes manipulada ou em outros casos espontânea, mas ainda sim é um complexo de pronunciamentos (LAZARSFELD, 1972). A pesquisa de opinião pode ser entendida como algo a promover a autoconsciência, assim como Martin (1985) utilizou como base para validar junto a população o senso comum.

Atualmente de acordo com Ferreira (2015) a opinião pública surge com a ascensão da classe média, com o desenvolvimento dos órgãos democráticos,

aumento da taxa de alfabetização e dos meios de comunicação. Essas opiniões são captadas de duas formas *Surveys*, pesquisas com questionários bem estruturados, ou através da coleta de dados de forma online, através de opinião em redes sociais (ROMANINI, 2021).

Outra pesquisa sobre a democracia no Brasil além do realizado pelo Latinobarómetro, é feito pelo Instituto da Democracia e da Democratização da Comunicação (INCT), que no ano de 2018 já demostrou que a população estava diminuindo a satisfação com a democracia, levada pelo baixo grau de confiança das instituições, partidos e políticos (Meneguello et al., 2018).

Estudos mais focalizados também estão sendo feitos em grandes cidades como São Paulo e Curitiba. O instituto Atuação em parceria com diversos pesquisadores nacionais e internacionais criaram o Índice de Democracia Local (IDL), que é composto de diversas métricas e indicadores, ao final retorna uma análise geral da democracia, dando uma nota de 0 a 10 (SILVA *et al*.,2021).

O IDL é dividido em cinco áreas: processo eleitoral, funcionamento do governo local, participação política, cultura democrática e direitos e liberdade civis. Dentro dessas áreas ocorre uma subdivisão, com perguntas orientadas a elas, buscando gerar informações para orientar políticas e verificar a qualidade da democracia local (GAZETA DO POVO, 2018).

3. METODOLOGIA

O conjunto de dados foi obtido por meio do *site* Latinobarómetro https://www.latinobarometro.org, especificamente dados de opinião pública que abrange aproximadamente 20 mil entrevistados em 16 países no ano de 2020, a mais recente até o início deste trabalho. A partir do conjunto dados do Latinobarómetro, foi verificado o questionário aplicado e cada pergunta foi encaixada de acordo com sua similaridade com as áreas do Índice de Democracia Local (IDL), desenvolvido pelo instituto Atuação, que trata das seguintes dimensões e atributos detalhados:

- Processo eleitoral
    - Escolha Democrática
    - Integridade
    - Inclusão
- Funcionamento do Governo Local
    - Freios e contrapesos
    - Transparência
    - Controle
    - Responsividade
    - Segurança Pública
- Participação política
    - Sentido amplo
    - Sentido Estrito
- Cultura Democrática
    - Dimensão cognitiva
    - Vida Comunitária
    - Normas e Valores
- Direitos e Liberdades Civis
    - Liberdades Civis
    - Liberdade de Expressão
    - Liberdade Econômica
    - Acesso à Justiça
    - Tratamento Justo

A primeira etapa do estudo envolveu uma seleção de perguntas presentes na base de dados do Latinobarómetro que se encaixavam com a métrica do IDL, assim, de um total de 408 perguntas, foi possível reduzir para 53 perguntas (atributos), retirando perguntas como: “ Quanta confiança você tem de que essa empresa opera para melhorar nossa qualidade de vida: Organizações multilaterais/internacionais”; “Quanta confiança você tem de que essa empresa opera para melhorar nossa qualidade de vida: Empresas como Facebook”; “Recentemente, fingiu estar doente para não ir trabalhar”; “Integração do seu país com outros países da América Latina”; “Parecer sobre a União Europeia”; “Os produtos nacionais têm, em geral, melhor qualidade do que os produtos importados” e “Inteligência artificial e robôs farão desaparecer mais locais de trabalho do que criarão”. O Latinobarómetro, responsável pela pesquisa, realiza uma limpeza prévia dos dados para disponibilizar junto com as informações da pesquisa e do instrumento de coleta, conhecido por *codebook*.

Mesmo com essa limpeza, foi necessária uma segunda etapa que consistiu em retirar dados referentes às respostas não efetivas, como: “não sabe”, “não respondeu”, “não aplicável”, “não perguntou” e “não sei”, remanescentes do conjunto de dados referente à variável “Satisfação com a democracia” variável de saída. Nessa etapa, também foram considerados somente os dados para o contexto do Brasil, inicialmente com 20.204 linhas (instâncias), para 1.165, buscando delimitar o estudo apenas para o país Brasil, onde reside o autor do trabalho.

Na terceira etapa, fase de verificação da variável de saída, código “P11STGBS_A” - Satisfação com a Democracia, era respondida como “Muito satisfeito” e “Bastante satisfeito” se tornou “Satisfeito”, já o “Não muito satisfeito” e “Nada satisfeito” se tornou “Insatisfeito”. Para melhorar os resultados, ela foi dicotomizada, tornando-se em uma variável binária, com 257 “Satisfeitos” e 908 “Insatisfeitos”.

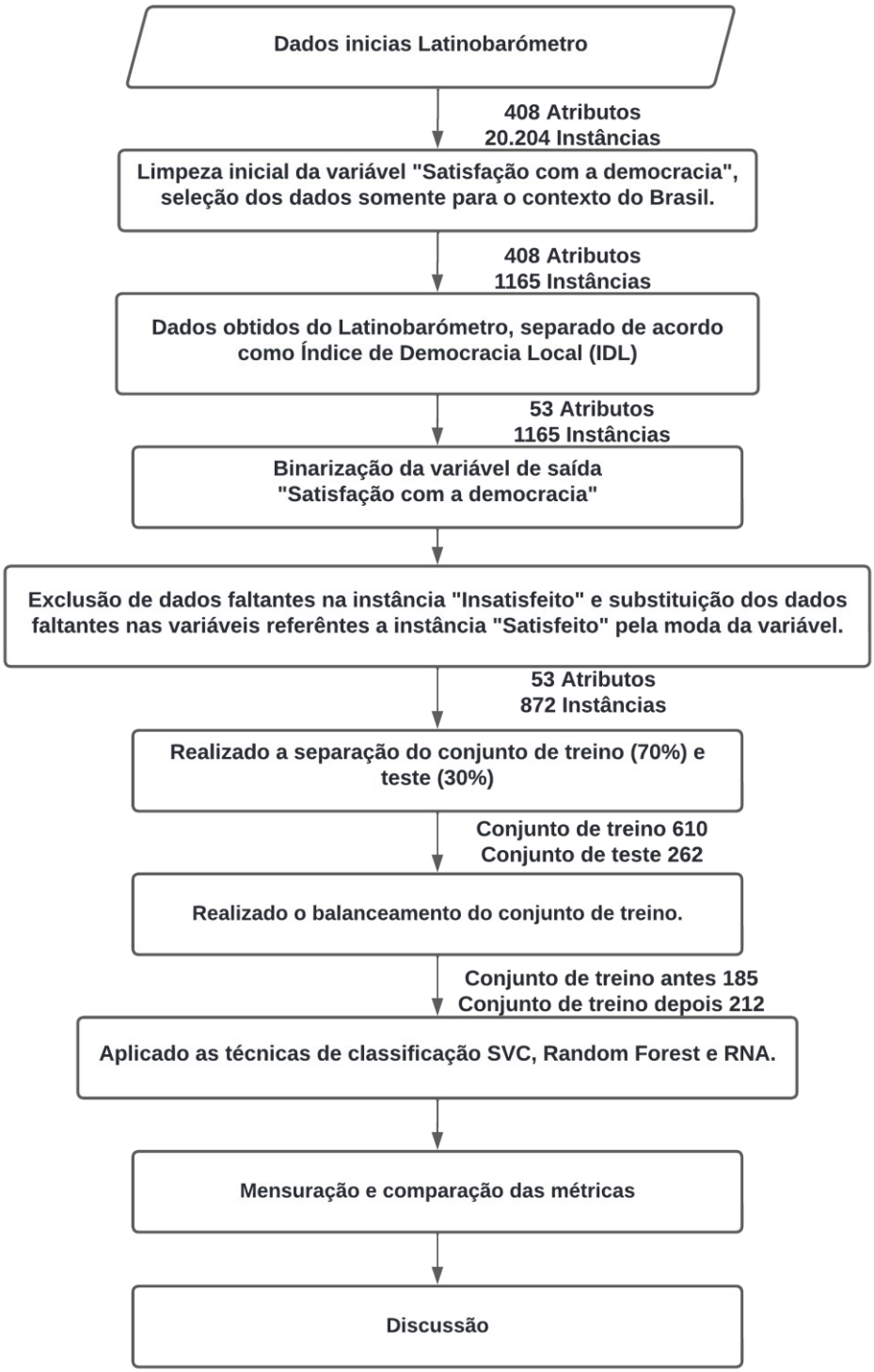
Na quarta etapa foi realizada a exclusão de todas as instâncias referentes à variável de saída “Insatisfeitos” nas quais as variáveis, atributos, que possuíam como resposta não efetiva: “não sabe”, “não respondeu”, “não aplicável”, “não perguntou” e “não sei”, inicialmente com 908 para 615 instâncias. No caso dos “Satisfeitos” com a democracia, buscando manter o conjunto balanceado, essas respostas não efetivas foram substituídas pela moda dos dados da respectiva variável, mantendo as 257 respostas, visando diminuir o desbalanceamento do conjunto. Por exemplo no caso da variável “Confiança no Judiciário” se estava na iteração da classe “Satisfeitos” e

respondida como “não sei” essa resposta foi substituída pela moda da “Confiança no Judiciário”, caso fosse da classe “Insatisfeito” toda a instância seria excluída.

Para a quinta etapa foi utilizada a linguagem *Python* 3 no *Jupyter Notebook,* uma plataforma de computação interativa baseada em *web browser*. Para executar as técnicas de AM foram utilizadas as bibliotecas *Pandas, Numpy Scikit-learn* e *Imblearn*, para a visualização de alguns dados foram utilizadas as bibliotecas *Matplotlib, Seaborn* e o *software Microsoft Excel.* Ao importar o arquivo inicial em csv os dados foram 872 linhas e 53 colunas, ou seja, 872 instâncias e 53 classes ou variáveis.

Esses dados foram então divididos em treino e teste, 70% e 30% respectivamente. Após essa etapa foi realizado o balanceamento do conjunto de teste devido à grande disparidade da quantidade de instâncias da variável de saída, com a técnica *Smote (Synthetic Minority Oversampling Technique).* Após o balanceamento os dados foram modelados e classificados por três métodos *Support Vector Classification* (*SVC*), *Random Forest* e Redes Neurais Artificiais (RNA) e mensurados e avaliados por precisão, acurácia, recall e f1-score, Figura 12.

Figura 12 - Passos da Metodologia da pesquisa



Fonte: Autoria própria (2022)

4. RESULTADOS E DISCUSSÃO

4.1 Resultados da análise descritiva

Após a realização da separação do conjunto de teste 30% e treino 70%, validação *holdout*, foi realizado o balanceamento através da técnica *Smote (Synthetic Minority Oversampling Technique),* importada da biblioteca *Imblearn*, gerando dados sintéticos e diferente das amostras existentes, mas partindo da base real de treinamento para aumentar o número de instâncias da classe “Satisfeito”, menor classe. Aplicando o parâmetro “*sampling_strategy*” (estratégia de amostragem) a 0.5, resultou em um aumento das instâncias relacionadas a “Satisfeitos”. Assim, foram implementadas as técnicas de *SVC*, *Random Forest* e RNA, e como métricas de análise dos algoritmos foram usadas a Precisão, *Recall*, F1 Score e acurácia, como mostrado na Tabela 2.

Tabela 2 - Resultados *Sampling_strategy* igual 0.5

SVC			
	Precisão	**Recall**	**F1 Score**
Insatisfeito (0)	80%	89%	84%
Satisfeito (1)	62%	45%	52%
Acurácia		**76%**	
Random Forest			
	Precisão	**Recall**	**F1 Score**
Insatisfeito (0)	79%	92%	85%
Satisfeito (1)	67%	38%	49%
Acurácia		**77%**	
RNA			
	Precisão	**Recall**	**F1 Score**
Insatisfeito (0)	84%	82%	83%
Satisfeito (1)	57%	61%	59%
Acurácia		**76%**	

Fonte: Autoria própria (2022)

Buscando melhorar os resultados foi realizado uma investigação de hiperparâmetros para árvore de decisão, com o “*GridSearchCV*”, utilizando o *k-fold*, entretanto, não houve grande melhora nos resultados, mas foi possível analisar a variação dos resultados de acordo com a mudança da divisão do conjunto de treino,

em que a primeira coluna seria o ranking dos 36 modelos testados. Assim, foi verificado que, dependendo do conjunto separado como teste e treino, o resultado das métricas serão consideravelmente afetados, variando a média dos testes de aproximadamente 68% a 76% e com desvio padrão de 1% a 4.8%, aproximadamente, como mostrado na Tabela 3.

Tabela 3 - *GridSearchCV* Resultados

	Parâmetros							
Rank	**Critério**	**Mínimo de amostras para folha**	**Mínimo de amostras para divisão**	**Divisor**	**Mín**	**Máx**	**Méd**	**Desv**
1	*gini*	10	5	*best*	0,745902	0,770492	0,757377	0,011119
2	*entropy*	5	2	*best*	0,721311	0,770492	0,755738	0,018255
4	*entropy*	5	10	*best*	0,721311	0,770492	0,754098	0,017958
23	*gini*	5	5	*random*	0,663934	0,770492	0,713115	0,037741
27	*entropy*	5	10	*random*	0,647541	0,737705	0,706557	0,031703
32	*gini*	1	2	*random*	0,672131	0,713115	0,693443	0,015203
36	*entropy*	1	2	*random*	0,606557	0,737705	0,683607	0,04852

Legenda: Mín (valor mínimo); Máx (valor máximo); Méd (média); Desv (desvio-padrão).
Observação: todos os valores são em relação aos cinco *folds* da validação cruzada.
Fonte: Autoria própria (2022)

Verificou-se por meio do recurso *feature importances* (importância dos atributos) do método *Random Forest* que os atributos que mais influenciaram no resultado, importância acima de 0,04, estão de acordo com o Quadro 1.

Quadro 1 - Atributos com importâncias acima de 0,04

Classificação	**Valor**	**Atributos**
1	0,1123	Satisfação com a situação econômica (em geral)
2	0,0582	Confiança no Governo Nacional
3	0,0484	Idade
4	0,0440	Próprio posicionamento na escala Esquerda-Direita
5	0,0434	Situação econômica futura do país

Fonte: Autoria própria (2022)

Após verificar os atributos mais importantes foi gerado uma nova iteração considerando no *dataset* somente os atributos do Quadro 1 e com o parâmetro "*sampling_strategy*" a 0.5. Os resultados encontrados estão na Tabela 4.

Tabela 4 - Resultados Atributos Mais Importantes com *Sampling_strategy* igual a 0.5

SVC			
	Precisão	**Recall**	**F1 Score**
Insatisfeito (0)	**79%**	**90%**	**84%**
Satisfeito (1)	**63%**	**42%**	**50%**
Acurácia		**76%**	
Random Forest			
	Precisão	**Recall**	**F1 Score**
Insatisfeito (0)	**82%**	**85%**	**83%**
Satisfeito (1)	**59%**	**54%**	**57%**
Acurácia		**76%**	
RNA			
	Precisão	**Recall**	**F1 Score**
Insatisfeito (0)	**80%**	**79%**	**80%**
Satisfeito (1)	**51%**	**53%**	**52%**
Acurácia		**71%**	

Fonte: Autoria própria (2022)

A diferença entre o resultado das métricas com todos os atributos (Tabela 2) e com somente os atributos mais importantes (Tabela 4), estão na Tabela 5. Foi verificado uma leve melhora para as instâncias “Satisfeito”, no método *Random Forest*, mas uma piora nos casos do SVC e RNA. O *Recall* ainda se encontra muito abaixo quando comparado com os resultados dos “Insatisfeitos”.

Tabela 5 – Diferença todos os atributos contra atributos de maior importância

SVC			
	Precisão	**Recall**	**F1 Score**
Insatisfeito (0)	**-1%**	**1%**	**0%**
Satisfeito (1)	**1%**	**-3%**	**-2%**
Acurácia		**0%**	
Random Forest			
	Precisão	**Recall**	**F1 Score**
Insatisfeito (0)	**3%**	**-7%**	**-2%**
Satisfeito (1)	**-8%**	**16%**	**8%**
Acurácia		**-1%**	
RNA			
	Precisão	**Recall**	**F1 Score**
Insatisfeito (0)	**-4%**	**-3%**	**-3%**
Satisfeito (1)	**-6%**	**-8%**	**-7%**
Acurácia		**-5%**	

Fonte: Autoria própria (2022)

Buscando um melhor resultado para as métricas e para o aprendizado do algoritmo foram realizadas diferentes variações no "*sampling_strategy*", sem *Smote* (0) e depois variando de 0.5 a 1, com passo de 0.1. A Tabela 6 representa a métrica Precisão, que busca mensurar o quanto o modelo do algoritmo realmente acertou, nas variações do *Smote*.

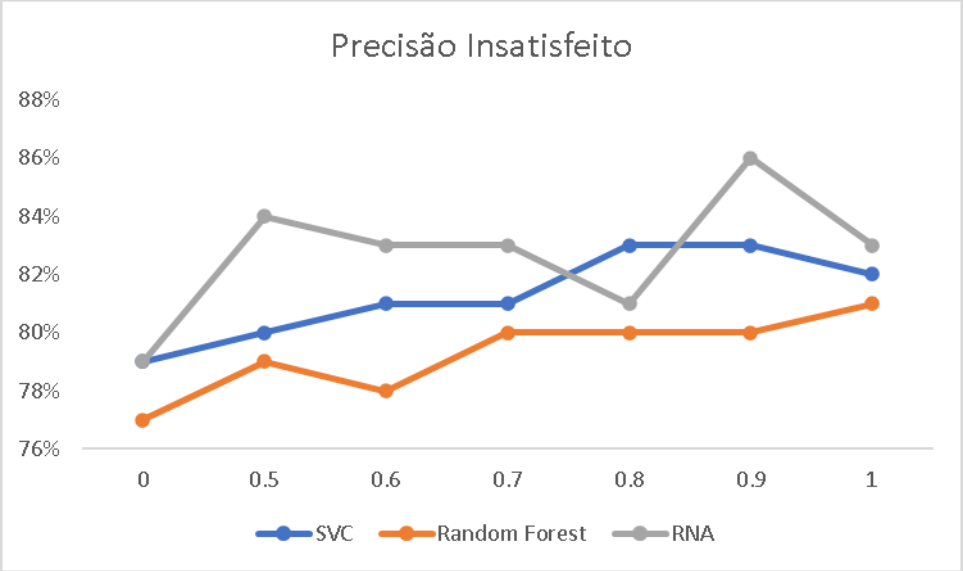
Tabela 6 - Variação da precisão para o *Sampling_strategy*

Smote	**Precisão Insatisfeito**			**Precisão Satisfeito**		
0 - 1	**SVC**	**Random Forest**	**RNA**	**SVC**	**Random Forest**	**RNA**
0	79%	77%	79%	63%	68%	67%
0.5	80%	79%	84%	62%	67%	57%
0.6	81%	78%	83%	63%	61%	58%
0.7	81%	80%	83%	61%	67%	57%
0.8	83%	80%	81%	65%	61%	54%
0.9	83%	80%	86%	64%	65%	55%
1	82%	81%	83%	60%	68%	56%

Fonte: Autoria própria (2022)

Para a classe "Insatisfeito" houve uma melhora de modo geral sobre a precisão, com melhores resultados para o método RNA e SVC próximo, Figura 13.

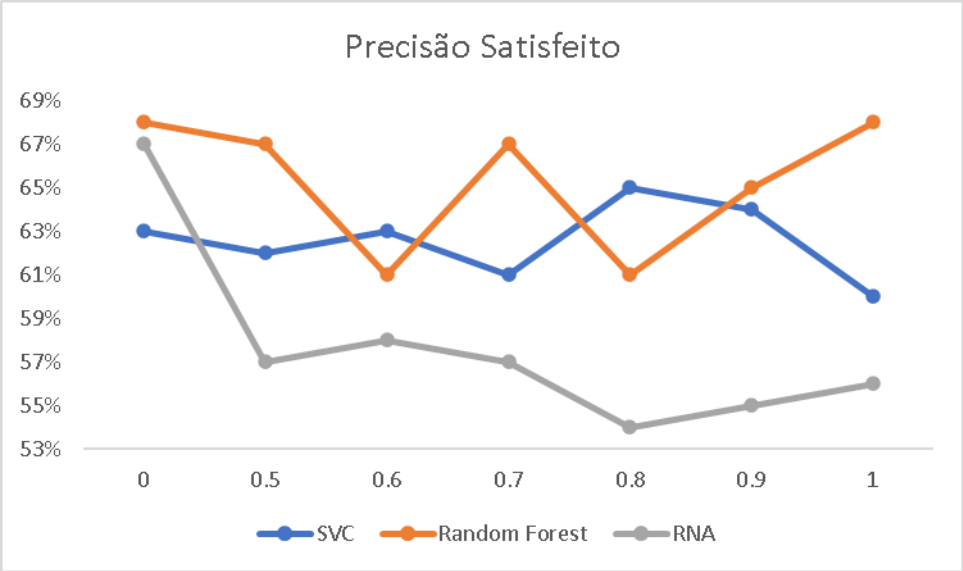
Figura 13 - Precisão classe "Insatisfeito" Variando o *Sampling_strategy*



Fonte: Autoria própria (2022)

Em relação à classe "Satisfeito", a classe que se busca melhorar os resultados devido à baixa quantidade de instâncias, gerou novos resultados próximos, com exceção do método RNA em que ocorreu uma queda significativa, Figura 14.

Figura 14 – Precisão classe "Satisfeito" Variando o *Sampling_strategy*



Fonte: Autoria própria (2022)

A Tabela 7 representa a variação do *Recall,* que busca mensurar o quão bom o algoritmo é para acertar a classe de saída, analisando a proporção de amostras da classe de interesse que foram classificadas corretamente, para os grupos "Satisfeito" e "Insatisfeito".

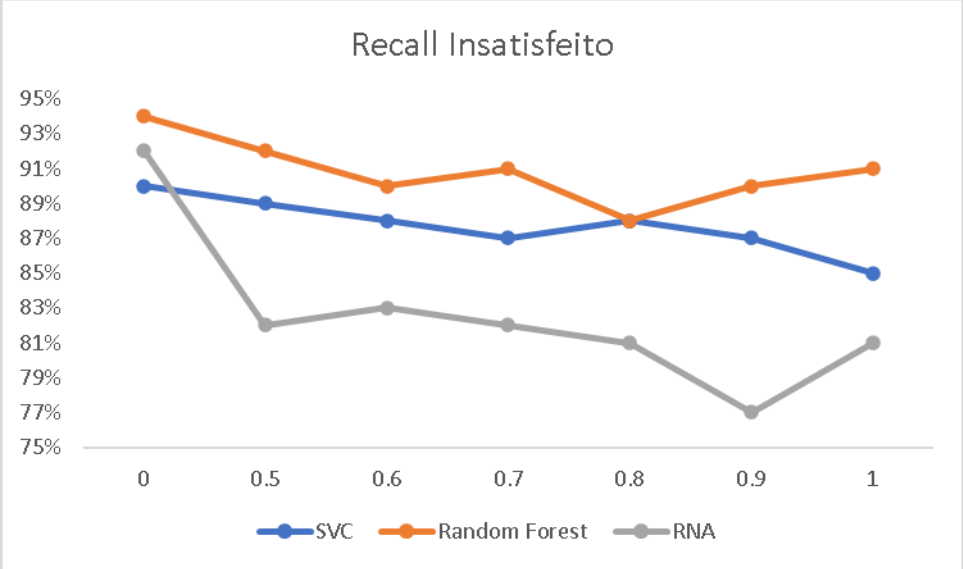
Tabela 7 - Recall Variação *Sampling_strategy*

Smote	Recall Insatisfeito			Recall Satisfeito		
0 - 1	**SVC**	**Random Forest**	**RNA**	**SVC**	**Random Forest**	**RNA**
0	90%	94%	92%	42%	30%	41%
0.5	89%	92%	82%	45%	38%	61%
0.6	88%	90%	83%	49%	39%	58%
0.7	87%	91%	82%	50%	45%	58%
0.8	88%	88%	81%	57%	47%	54%
0.9	87%	90%	77%	55%	46%	68%
1	85%	91%	81%	53%	47%	59%

Fonte: Autoria própria (2022)

O *recall* para a classe “Insatisfeito” ocorre uma queda para todos os métodos de classificação, Figura 15, contrastando com a classe “Satisfeito” que houve uma melhora.

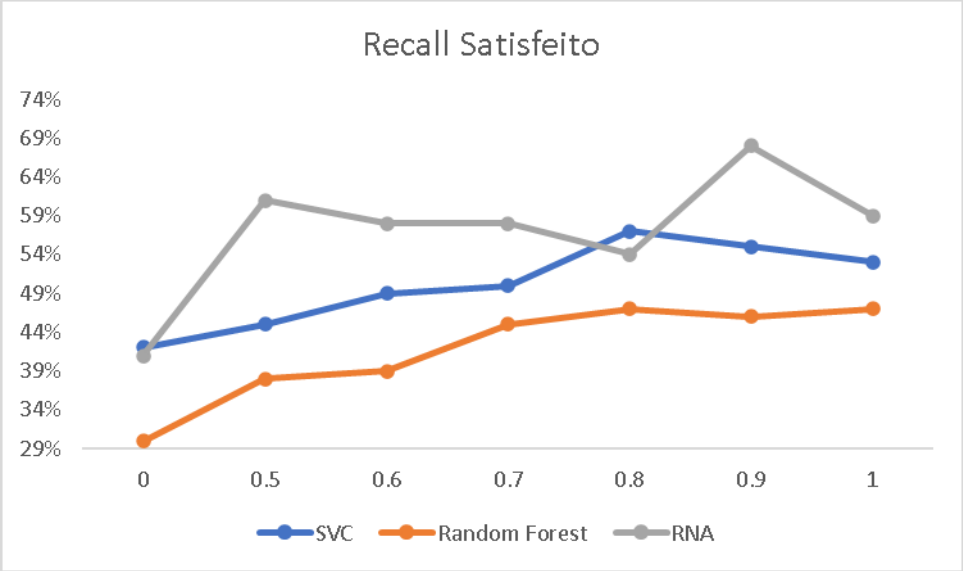
Figura 15 - Recall classe “Insatisfeito” Variando o *Sampling_strategy*



Fonte: Autoria própria (2022)

Em relação aos “Satisfeitos” houve um aumento significativo no *Recall* em relação ao aumento das instâncias com o *Smote,* Figura 16, saindo no método RNA de 41% para 59% com *Smote* igual a 1, aumento de 18%, considerável para uma mudança somente na quantidade do número de instâncias.

Figura 16 - Recall classe "Satisfeito" Variando o *Sampling_strategy*



Fonte: Autoria própria (2022)

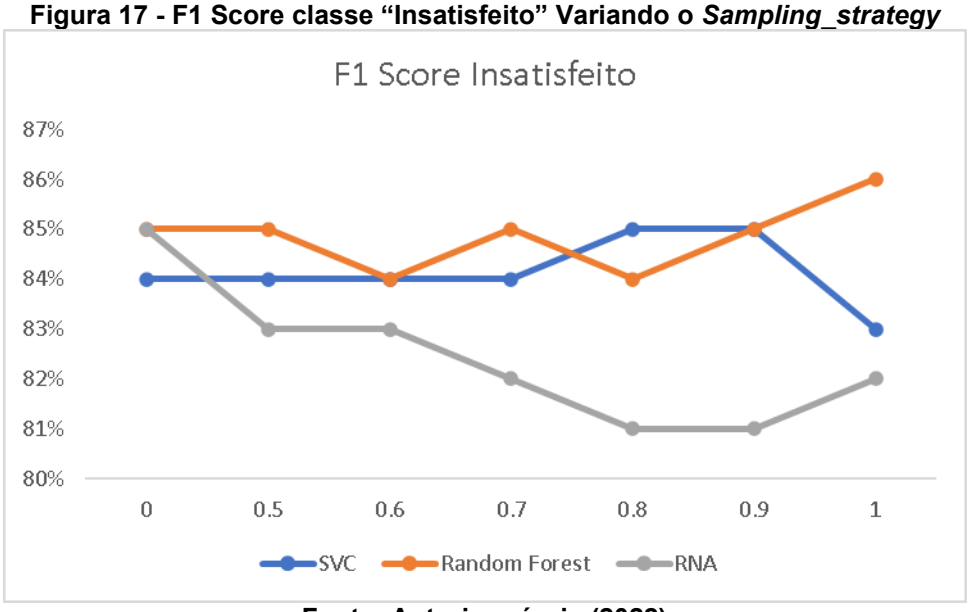
A Tabela 8 e as Figura 17 e Figura 18 representam a variação do F1 *Score,* definido como a média harmônica da precisão e *recall,* para os grupos "Satisfeito" e "Insatisfeito".

Tabela 8 - F1 Score Variação *Sampling_strategy*

Smote	F1 Score Insatisfeito			F1 Score Satisfeito		
0 - 1	**SVC**	**Random Forest**	**RNA**	**SVC**	**Random Forest**	**RNA**
0	84%	85%	85%	50%	42%	51%
0.5	84%	85%	83%	52%	49%	59%
0.6	84%	84%	83%	55%	48%	58%
0.7	84%	85%	82%	55%	54%	58%
0.8	85%	84%	81%	61%	53%	54%
0.9	85%	85%	81%	59%	54%	61%
1	83%	86%	82%	56%	56%	58%

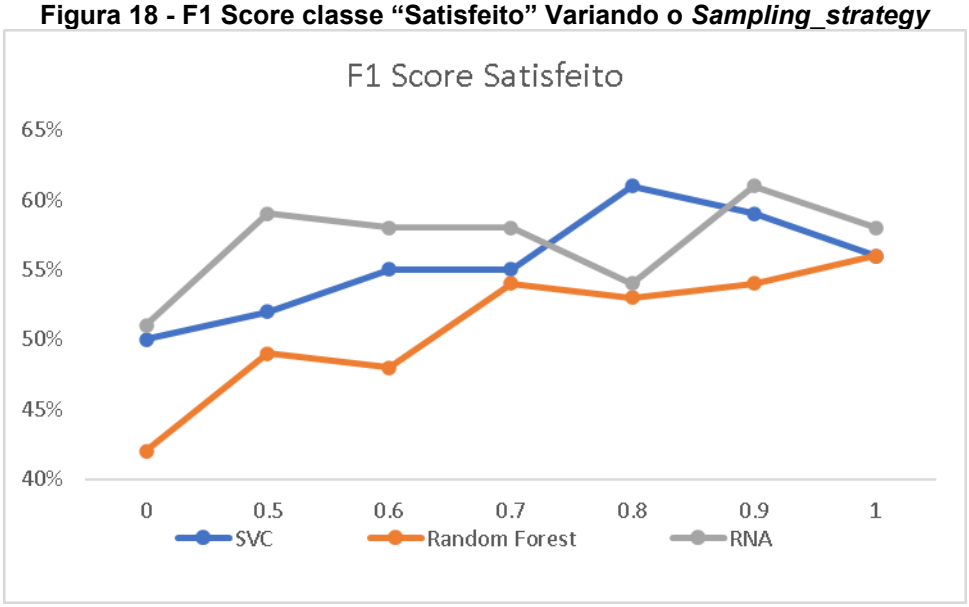
Fonte: Autoria própria (2022)

Figura 17 - F1 Score classe “Insatisfeito” Variando o *Sampling_strategy*



Fonte: Autoria própria (2022)

Figura 18 - F1 Score classe “Satisfeito” Variando o *Sampling_strategy*



Fonte: Autoria própria (2022)

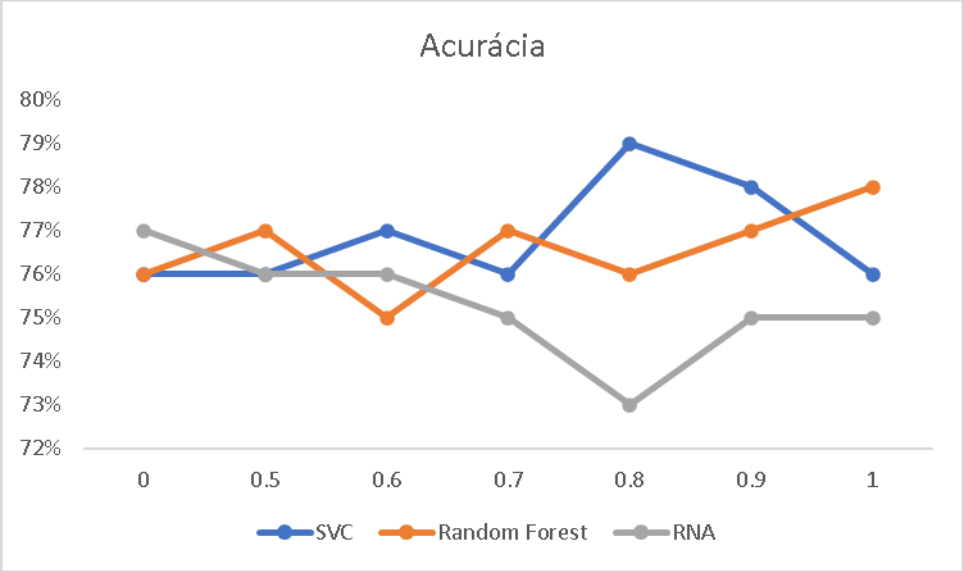
A Tabela 9 e a Figura 19 representam a variação da Acurácia para os grupos “Satisfeito” e “Insatisfeito”.

Tabela 9 - Acurácia variação *Sampling_strategy*

Smote	Acurácia		
0 - 1	**SVC**	**Random Forest**	**RNA**
0	76%	76%	77%
0.5	76%	77%	76%
0.6	77%	75%	76%
0.7	76%	77%	75%
0.8	79%	76%	73%
0.9	78%	77%	75%
1	76%	78%	75%

Fonte: Autoria própria (2022)

Figura 19 - Acurácia variando o Smote



Fonte: Autoria própria (2022)

4.2 Discussão da análise

Partindo dos resultados, é possível verificar no Quadro 1 alguns dos fatores que mais influenciaram no resultado do algoritmo *Random Forest*, buscando saber se um indivíduo está satisfeito ou não com a democracia. Neste caso, é possível observar que o fator econômico possui um grande peso na decisão, seguido da confiança no governo, idade e posicionamento político. Assim, com essas informações poderia ser realizado uma verificação na campanha política de candidatos ao governo (Federal, estadual ou municipal), buscando pontos para explorar melhorar o desempenho nas campanhas.

Uma aplicação seria buscar eleitores insatisfeitos, analisar os atributos de maior importância, que fizeram eles estarem insatisfeitos com a democracia e seguir uma estratégia similar à que Jonathan Albright (2016) descreve sobre as eleições

presidenciais dos Estado Unidos da América para o primeiro mandato de Donald Trump, em que foram utilizados dados para verificar os eleitores que poderiam ser convencidos a alterar o voto, focalizando a estratégia para convencer essas pessoas a saírem de suas casas e votarem a favor. Assim esses eleitores insatisfeitos seriam o foco de campanhas de marketing, otimizando seu alcance, utilizando os atributos identificados com maior importância na satisfação com a democracia.

A tática utilizada por eles analisando informações de diversas fontes, questionários de opinião, comportamental e de redes sociais, foram eficazes nas eleições de 2016, mudando o panorama que chegou a 15% de chance de vitória para Donald Trump Katz NY Times (2016), mas culminando em uma vitória ao final das apurações dos votos.

Além de poderem ser utilizados em campanhas, esta pesquisa mostra algumas das variáveis mais importantes no contexto da democracia, podendo ser utilizados para melhorar políticas e planos de desenvolvimento nos âmbitos da liberdade civil, mais relacionada a liberdade econômica. Participação política em amplo sentido abrindo espaço para participação em partidos e conselhos municipais. No funcionamento do Governo local aumentando a transparência e responsabilidade dos três poderes do governo aumentando assim a confiança da população.

Buscando através da saída “Insatisfeitos” que, de acordo com os resultados encontrados nesta pesquisa, foram muito superiores aos “Satisfeitos”, seria possível a implementação dessa categoria, visando saber mais o porquê está Insatisfeito e, com base nos dados, verificar no questionário e nas importâncias onde seria melhor agir para elevar a satisfação, melhorando assim os indicadores da satisfação com a democracia.

Os dados gerados pelo *Smote,* não resultou em melhoras nas métricas para a classe de saída “Satisfeitos” se comparado ao dos “Insatisfeitos”, mesmo sabendo que essa técnica gera novas instâncias sintéticas a partir da base de dados e ainda em outra iteração utilizando apenas os atributos de maior importância nos métodos de classificação (Tabela 4), era esperado um melhor resultado.

5. CONCLUSÃO

A utilização de técnicas de AM em conjuntos de dados referentes a questionários de opinião pode trazer a concretização de suposições, gerar um modelo possível ser implementado para predizer novas opiniões e revelar os fatores que mais influenciam a satisfação dos entrevistados com a democracia, variável de saída.

Dessa forma, neste trabalho foi possível verificar que as variáveis sobre satisfação com a economia, confiança no governo e idade acabam influenciando mais a satisfação com a Democracia, do que o posicionamento político, direita ou esquerda, de acordo com o valor dos atributos gerados pelo algoritmo *Random Forest.*

Os modelos de classificação geraram resultados aceitáveis se considerar somente os relacionados a categoria de saída “Insatisfeitos”. Em contrapartida a categoria de saída “Satisfeitos” gerou resultados abaixo do satisfatório, comparando aos “Insatisfeitos”, em uma das melhores iterações do método RNA a precisão ficou em 55%, o recall em 68% e o F1 Score em 61% com acurácia para o modelo de 75%, considerando um *Smote* de 0.9, demonstrando o quanto o desbalanceamento do conjunto influência na saída.

Pensando em uma maneira de contornar essa deficiência no conjunto, uma das possíveis opções seria usar conjuntos passados, por exemplo 2019, somente para treino e utilizar o conjunto de 2020 somente para teste, podendo também realizar a junção dos dois conjuntos e dividir aleatoriamente, buscando balancear as classes com dados reais para poder implementar em futuro buscando direcionamento de campanhas políticas, pontos de melhoria pelo governo de modo geral, buscando ajustar a confiança no governo, melhora da economia e assim aumentando a satisfação com a democracia.

REFERÊNCIAS

ALBRIGHT, J. ***How Trump's campaign used the new data-industrial complex to win the election.*** *LSE Phelan US Centre*, 2016. Disponível em: https://blogs.lse.ac.uk/usappblog/2016/11/26/how-trumps-campaign-used-the-new-data-industrial-complex-to-win-the-election/. Acesso em: 29 maio 2022.

BREIMAN, L. (2001). ***Random forests. Machine learning***, 45(1), p. 5-32.

BHOJANI, S.; BHATT, N. (2016). ***Data Mining Techniques and Trends – A Review.*** *GJRA, Volume-5, Issue-5, May* – 2016, ISSN No 2277 – 8160.

COENEN, F. (2011*). **Data mining:** Past, present and future. Knowledge Eng. Review*. 26. 25-29. 10.1017/S0269888910000378.

FACELI, K. et al. **Inteligência Artificial**: Uma Abordagem de Aprendizado de Máquina. Rio de Janeiro: Grupo GEN, 2021.

FAYYAD, U.; SHAPIRO, G.P.; SMYTH, P. ***From Data Mining to Knowledge Discovery in Databases****. AI Magazine*, v. 17. Estados Unidos da América. No. 3. P. 37-54, 1996. Disponível em: https://www.aaai.org/ojs/index.php/aimagazine/article/view/1230. Acesso em: 14 ago. 2021.

FERRARI, D. G.; DE CASTRO, L. N. **Introdução à Mineração de Dados**: Conceitos Básicos, Algoritmos e Aplicações. Editora Saraiva, 2016.

FERREIRA, F. V. **Raízes históricas do conceito de opinião pública em comunicação**. Universidade de Brasília (UnB). 2015.

Gazeta do Povo**. Qualidade da democracia na sua cidade (2018)**. Disponível em: https://especiais.gazetadopovo.com.br/qualidade-da-democracia-no-brasil/. Acesso em: 03 mar. 2022.

GOMES, M. L. AQUINO, J. A. **Violência e satisfação com a democracia no Brasil**. Opinião Pública. 2018, v. 24, p. 209-238. Disponível em: https://doi.org/10.1590/1807-01912018241209. Acesso em: 28 maio 2022.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining:** concepts and techniques. 3 ed. Waltham, MA, EUA: Morgan Kaufmann, 2012.

HAYKIN, S. **Redes neurais princípios e prática.** Porto Alegre: Grupo A, 2001. 9788577800865.

HO, T. K. (1995*). **Random decision forests.** In: Proceedings of 3rd international conference on document analysis and recognition*, vol. 1, p. 278–282. IEEE.

KATZ, J. ***2016 Election Forecast:*** *Who Will Be President?*. *New York Times*, 2016. Disponível em: https://www.nytimes.com/interactive/2016/upshot/presidential-polls-forecast.html?mtrref=undefined&gwh=3FAF92CCDCB82D2FFF3CAE7F5450971E&gwt=regi&assetType=REGIWALL. Acesso em: 29 maio 2022.

LAZARSFELD, P. F. **A opinião pública e a tradição clássica.** In: STEINBERG, Charles S. (org). Meios de Comunicação de Massa. São Paulo: Cultrix, 1972.

LIU, B. ***Opinion mining. In Encyclopedia of Database Systems****, pages* 1986–1990. Springer US, 2009.

MARTIN-BARÓ, I. (1985). **La encuesta de opinión pública como instrumento desideologizador**. Cuadernos de Psicologia, 1-2, vol.7: 93-108.

MIRKIN, B. (2011). ***Core Concepts in Data Analysis****: Summarization, Correlation and Visualization. Springer.*

NAVARRO, V. *et al*. ***Data Fusion and Machine Learning for Innovative GNSS Science Use Cases****. In: Proceedings of the 34th International Technical Meeting of the Satellite Division of The Institute of Navigation* (ION GNSS+ 2021). St. Louis, Missouri 2021. P. 2656-2669.

QUILICI-GONZALEZ, J. A.; ZAMPIROLLI, F. A. **Sistemas Inteligentes e Mineração de Dados.** Santo André: Triunfal Gráfica e Editora, 2014.

RESENDE, R. C. EPITÁCIO, S. S. F. **Desenvolvimento econômico e satisfação com a democracia**: uma análise da América Latina. Ciências Sociais Unisinos. 2014, 50(2), 117-126. Disponível em: https://www.redalyc.org/articulo.oa?id=93832099003. Acesso em: 29 maio 2022.

ROMANINI, V.; CALDAS, P. (2021). **Opinião pública e tecnologia**: Os impactos do Big Data nos estudos de opinião pública sob o olhar do pragmatismo. TRANS/FORM/AÇÃO: Revista De Filosofia, 44, 375–398.

SILVA, D. R. M. *et al*. **Índice De Democracia Local**: Estudos A Partir Da Experiência de São Paulo. 2021.

VERMA, D.; NASHINE, R. **"*Data Mining****: Next Generation Challenges and Future Directions". International Journal of Modeling and Optimization* (2012): 603-608.
MENEGUELLO, R.; AMARAL, O. E. et al. Satisfação com a democracia e conjuntura política no Brasil (INCT), maio, 2018.

ANEXO A - Lei n. 9.610, de 19 de fevereiro de 1998

Presidência da República
Casa Civil
Subchefia para Assuntos Jurídicos

LEI Nº 9.610, DE 19 DE FEVEREIRO DE 1998[1].

Altera, atualiza e consolida a legislação sobre direitos autorais e dá outras providências.

O PRESIDENTE DA REPÚBLICA Faço saber que o Congresso Nacional decreta e eu sanciono a seguinte Lei:

Título I - Disposições Preliminares

Art. 1º Esta Lei regula os direitos autorais, entendendo-se sob esta denominação os direitos de autor e os que lhes são conexos.

Art. 2º Os estrangeiros domiciliados no exterior gozarão da proteção assegurada nos acordos, convenções e tratados em vigor no Brasil.

Parágrafo único. Aplica-se o disposto nesta Lei aos nacionais ou pessoas domiciliadas em país que assegure aos brasileiros ou pessoas domiciliadas no Brasil a reciprocidade na proteção aos direitos autorais ou equivalentes.

Art. 3º Os direitos autorais reputam-se, para os efeitos legais, bens móveis.

Art. 4º Interpretam-se restritivamente os negócios jurídicos sobre os direitos autorais.

Art. 5º Para os efeitos desta Lei, considera-se:

I - publicação - o oferecimento de obra literária, artística ou científica ao conhecimento do público, com o consentimento do autor, ou de qualquer outro titular de direito de autor, por qualquer forma ou processo;

II - transmissão ou emissão - a difusão de sons ou de sons e imagens, por meio de ondas radioelétricas; sinais de satélite; fio, cabo ou outro condutor; meios óticos ou qualquer outro processo eletromagnético;

III - retransmissão - a emissão simultânea da transmissão de uma empresa por outra;

IV - distribuição - a colocação à disposição do público do original ou cópia de obras literárias, artísticas ou científicas, interpretações ou execuções fixadas e fonogramas, mediante a venda, locação ou qualquer outra forma de transferência de propriedade ou posse;

V - comunicação ao público - ato mediante o qual a obra é colocada ao alcance do público, por qualquer meio ou procedimento e que não consista na distribuição de exemplares;

VI - reprodução - a cópia de um ou vários exemplares de uma obra literária, artística ou científica ou de um fonograma, de qualquer forma tangível, incluindo qualquer armazenamento permanente ou temporário por meios eletrônicos ou qualquer outro meio de fixação que venha a ser desenvolvido;

VII - contrafação - a reprodução não autorizada;

VIII - obra:

a) em co-autoria - quando é criada em comum, por dois ou mais autores;

b) anônima - quando não se indica o nome do autor, por sua vontade ou por ser desconhecido;

c) pseudônima - quando o autor se oculta sob nome suposto;

d) inédita - a que não haja sido objeto de publicação;

e) póstuma - a que se publique após a morte do autor;

f) originária - a criação primígena;

g) derivada - a que, constituindo criação intelectual nova, resulta da transformação de obra originária;

h) coletiva - a criada por iniciativa, organização e responsabilidade de uma pessoa física ou jurídica, que a publica sob seu nome ou marca e que é constituída pela participação de diferentes autores, cujas contribuições se fundem numa criação autônoma;

i) audiovisual - a que resulta da fixação de imagens com ou sem som, que tenha a finalidade de criar, por meio de sua reprodução, a impressão de movimento, independentemente dos processos de sua captação, do suporte usado inicial ou posteriormente para fixá-lo, bem como dos meios utilizados para sua veiculação;

IX - fonograma - toda fixação de sons de uma execução ou interpretação ou de outros sons, ou de uma representação de sons que não seja uma fixação incluída em uma obra audiovisual;

X - editor - a pessoa física ou jurídica à qual se atribui o direito exclusivo de reprodução da obra e o dever de divulgá-la, nos limites previstos no contrato de edição;

XI - produtor - a pessoa física ou jurídica que toma a iniciativa e tem a responsabilidade econômica da primeira fixação do fonograma ou da obra audiovisual, qualquer que seja a natureza do suporte utilizado;

XII - radiodifusão - a transmissão sem fio, inclusive por satélites, de sons ou imagens e sons ou das representações desses, para recepção ao público e a transmissão de sinais codificados, quando os meios de decodificação sejam oferecidos ao público pelo organismo de radiodifusão ou com seu consentimento;

XIII - artistas intérpretes ou executantes - todos os atores, cantores, músicos, bailarinos ou outras pessoas que representem um papel, cantem, recitem, declamem, interpretem ou executem em qualquer forma obras literárias ou artísticas ou expressões do folclore.

Art. 6º Não serão de domínio da União, dos Estados, do Distrito Federal ou dos Municípios as obras por eles simplesmente subvencionadas.

---

[1] Disponível em: http://www.planalto.gov.br/ccivil_03/leis/l9610.htm.