PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
Programa de Pós-graduação em Informática

Andressa Paula Castro de Oliveira

SELEÇÃO DE INSTÂNCIAS REPRESENTATIVAS PARA A CARACTERIZAÇÃO DOS PERFIS DE USUÁRIOS DE REDES SOCIAIS

Uma análise dos perfis de interação social do Facebook

Belo Horizonte
2021

Andressa Paula Castro de Oliveira

SELEÇÃO DE INSTÂNCIAS REPRESENTATIVAS PARA A CARACTERIZAÇÃO DOS PERFIS DE USUÁRIOS DE REDES SOCIAIS

Uma análise dos perfis de interação social do Facebook

Dissertação apresentada ao Programa de Pós-Graduação em Informática da Pontifícia Católica de Minas Gerais, como requisito parcial para obtenção do título de Mestre em Informática.

Orientadora: Prof[a]. Dr[a]. Cristiane Neri Nobre

Área de concentração: Inteligência Artificial e Mineração de Dados

Belo Horizonte
2021

FICHA CATALOGRÁFICA
Elaborada pela Biblioteca da Pontifícia Universidade Católica de Minas Gerais

Oliveira, Andressa Paula Castro de
O48s Seleção de instâncias representativas para a caracterização dos perfis de usuários de redes sociais: uma análise dos perfis de interação social do facebook / Andressa Paula Castro de Oliveira. Belo Horizonte, 2021.
74 f. : il.

Orientadora: Cristiane Neri Nobre
Dissertação (Mestrado) – Pontifícia Universidade Católica de Minas Gerais. Programa de Pós-Graduação em Informática

1. Facebook (Rede social on-line) - Pesquisa. 2. Redes sociais on-line. 3. Algorítmos computacionais. 4. Percepção de padrões. 5. Usuários da Internet. 6. Banco de dados da Web. I. Nobre, Cristiane Neri. II. Pontifícia Universidade Católica de Minas Gerais. Programa de Pós-Graduação em Informática. III. Título.

CDU: 681.3.011
Ficha catalográfica elaborada por Elizângela Ribeiro de Azevedo - CRB 6/3393

Andressa Paula Castro de Oliveira

SELEÇÃO DE INSTÂNCIAS REPRESENTATIVAS PARA A CARACTERIZAÇÃO DOS PERFIS DE USUÁRIOS DE REDES SOCIAIS

Uma análise dos perfis de interação social do Facebook

Dissertação apresentada ao Programa de Pós-graduação em Informática da Pontifícia Universidade Católica de Minas Gerais, como requisito parcial para obtenção do título de Mestre em Informática.

Área de concentração: Inteligência Artificial e Mineração de Dados

---

Profª. Drª. Cristiane Neri Nobre (Orientador) – PUC Minas

Prof. Dr Luis Enrique Zarate – PUC Minas

Profª. Drª. Jonice de Oliveira Sampaio – UFRJ

Belo Horizonte, 01 de Setembro de 2021

RESUMO

A grande maioria dos usuários da internet acessam as redes sociais, Seu elevado número de acessos leva pesquisadores a identificarem o que faz uma rede social ter sucesso e se manter latente na *web* por um longo período, principalmente para encontrar padrões que possibilitem o desenvolvimento de novas redes sociais de sucesso. Considerada a maior rede social da atualidade, o Facebook vem tentando influenciar as pessoas a adotarem determinados comportamentos para que continuem utilizando suas funcionalidades. Um dos fatores que levam uma rede social a proporcionar melhores funcionalidades é conhecer profundamente seus usuários e especificar quais funções ele executa na rede. Estudos anteriores identificaram a presença de três perfis de usuários no Facebook quanto aos aspectos de interação social, são eles: Espectador, Participante e Produtor de Conteúdo, e observaram que esses usuários são influenciados por estratégias persuasivas utilizadas pela rede social no intuito de fazer com que as pessoas interajam mais na rede e consequentemente passe mais tempo navegando. Sabendo disso, é possível considerar as características específicas de cada perfil de usuário, a fim de desenvolver e direcionar estratégias específicas para cada um deles, e assim atingir o número máximo possível de pessoas. Sabendo disso, o objetivo deste trabalho é caracterizar, por meio de regras de classificação e técnicas de visualização multivariada, estes três perfis de interação, descrevendo como eles interagem na rede social. Os resultados mostram que existem diferenças significativas entre os perfis de Participante e Espectador. Enquanto os Participantes frequentemente comentam e compartilham publicações com conteúdo com o qual concordam; os Espectadores se comportam de maneira oposta. Para a amostra considerada, observou-se que os Produtores de conteúdo possuem características dos dois perfis anteriores e que esses perfis podem não ser mutuamente exclusivos. Apesar disso, a caracterização realizada, considerando as instâncias mais representativas de cada perfil, mostra as características mais importantes de cada um deles. Esses resultados contribuem principalmente para a área de *design* de redes sociais persuasivas, colaborando para o desenvolvimento de aplicações com maior probabilidade de sucesso.

Palavras-chave: Facebook. Perfil de interação. Redes sociais. Seleção de Instância. Visualização multivariada.

ABSTRACT

The vast majority of internet users access social networks. Its high number of hits leads researchers to identify what makes a social network successful and remain latent on the web for an extended period, mainly to find patterns that enable the development of new successful social networks. Considered the largest social network today, Facebook has been trying to influence people to adopt certain behaviors to continue to use its features. One factor that leads a social network to provide better functionality is to know its users deeply and specify what functions it performs on the web. Previous studies identified the presence of three user profiles on Facebook regarding aspects of social interaction, namely: Spectator, Participant, and Content Producer, and observed that these users are influenced by persuasive strategies used by the social network to make the people interact more on the web and consequently spend more time browsing. Knowing this, it is possible to consider the specific characteristics of each user profile to develop and direct particular strategies for each one of them and thus reach the maximum number of people possible. Knowing this, the objective of this work is to characterize, through classification rules and multivariate visualization techniques, these three interaction profiles, describing how they interact in the social network. The results show that there are significant differences between the Participant and Spectator profiles. While Participants often comment and share posts with content they agree with; Spectators behave oppositely. For the sample considered, it was observed that Content Producers have characteristics of the two previous profiles and that these profiles may not be mutually exclusive. Despite this, the characterization performed, considering the most representative instances of each profile, shows the most important characteristics of each one of them. These results contribute mainly to the design area of persuasive social networks, contributing to applications with a greater probability of success.

Keywords: Facebook. Instance Selection. Interaction profile. Multivariate visualization. Social networks.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A sociedade está cada vez mais orientada para a informação devido às novas tecnologias que fornecem dados em uma velocidade e escala muito altas e a análise desses dados é uma parte importante para responder a diversos questionamentos sobre seu funcionamento. Essas base de dados com registros do mundo real, por sua vez, estão muito suscetíveis à possuir informações confusas, mal estruturadas e com ruídos, o que influencia diretamente a análise desses dados. Assim, uma maneira de melhorar essa análise é encontrar instâncias que são mais representativas dentro da base de dados para obter as informações desejadas.

Um ambiente em que a produção de dados é muito rápida, extensa e precisa ser avaliada são as redes sociais. Redes sociais (em inglês, *Social Networks* - SNSs) são espaços onde as pessoas podem interagir através do sistema e criar vínculos: se conectar, compartilhar informação, conversar, conhecer pessoas, visualizar tópicos de interesse, compartilhar fotos e vídeos, dentre outros. Elas oferecem oportunidades para o desenvolvimento de redes de contato, fomentando a criação de comunidades e mecanismos de compartilhamento de informações; desempenhando um papel ativo e essencial no engajamento do usuário (SILVA; PRADO, 2015). Estima-se que aproximadamente 90% dos usuários da internet acessam redes sociais (BANKS, 2015).

Consideradas representações *online* do que as pessoas são ou pretendem ser na sociedade, as redes sociais podem influenciar de maneiras distintas seus usuários, pois a forma com que as pessoas utilizam as ferramentas é um reflexo de suas características pessoais. Assim, ao se adaptar aos desejos e necessidade de seus usuários, estimulando-os a se envolverem com suas ferramentas e recursos, a rede social, para essas pessoas, permanecerá relevante e terá um poder persuasivo maior.

A maioria das redes sociais possuem três principais objetivos: engajamento, para aumentar o uso e manter os usuários navegando; crescimento, para que os usuários convidem outras pessoas para fazerem parte; e publicidade, para gerar renda para a empresa (HARRIS, 2020). Por isso, a rede social precisa se manter sempre atualizada, oferecendo funcionalidades que agradem seus usuários para garantir que seu ciclo de vida seja longo.

Pensando nisso, utilizar-se de estratégias persuasivas pode ser visto como uma maneira de manter seus clientes mais ativos. Assim, cada rede social oferece serviços e recursos específicos que têm como alvo uma comunidade bem definida do mundo real, mas para que as estratégias adotadas sejam efetivas e as redes sociais evoluam e proporcionem melhores funcionalidades é fundamental conhecer muito bem o público-alvo (FERNANDES; SIQUEIRA, 2013).

Portanto, para que haja um bom funcionamento do sistema, é importante identificar o usuário e caracterizá-lo, especificando quais funções ele executa na rede social

para ajustar o comportamento do sistema à realidade e expectativas do usuário. Sistemas que conhecem seus usuários podem personalizar suas funcionalidades de acordo com o modelo do usuário, pois as diferenças em algumas características das pessoas afetam a utilidade dos serviços ou informações fornecidas a elas (SOSNOVSKY; DICHEVA, 2010). De acordo com Cao et al. (2014), conhecendo o contexto do usuário, o sistema pode se adaptar a ele, e o usuário também pode adaptar suas interações com o sistema para ajustar seu comportamento.

Visando estimular seus usuários a se manterem conectados utilizando seus recursos com frequência, o Facebook, considerado a maior e mais difundida rede social, adota diversas estratégias persuasivas nas funcionalidades que disponibiliza para seus usuários no intuito de fazer com que o usuário retorne à rede social várias vezes ao dia (FOGG; IIZAWA, 2008; RUAS DA SILVEIRA; NOBRE; CARDOSO, 2014; OLIVEIRA et al., 2017; NASCIMENTO et al., 2018). Com isso, somente no primeiro trimestre de 2021, o Facebook atingiu cerca de 2,85 bilhões de usuários ativos* por mês (STATISTA, 2021c), destacando-se muito em relação a outros *sites* do mesmo seguimento. O Facebook é considerado um fenômeno mundial que mudou a forma como as pessoas se comunicam (ALOUDAT; AL-SHAMAILEH; MICHAEL, 2019).

De acordo com o *site* Alexa (2021)†, o Facebook também está entre os principais *sites* do mundo. Dentre os países com o maior número de usuários do Facebook, o Brasil ocupa a quarta posição, com 130 milhões de usuários. O país com mais usuários é a Índia (330 milhões), seguido de Estados Unidos (190 milhões) e Indonésia (140 milhões) (STATISTA, 2021b).

Em relação aos aspectos de interação, RUAS DA SILVEIRA et al. (2019) afirmam que o Facebook possui três perfis de usuário: 1) *Espectador*, usuário cujo comportamento predominante é apenas visualizar a rede social; 2) *Participante*, usuário cujo comportamento típico é interagir com os conteúdos publicados por seus contatos na rede; e 3) *Produtor de conteúdo*, usuário que, predominantemente, produz conteúdo na rede social. Os autores sugerem que a rede social influencia estes usuários de maneira diferente.

No trabalho de RUAS DA SILVEIRA et al. (2019) esses perfis de usuário foram identificados por meio de uma coleta de interações que os usuários realizaram na rede social e os resultados mostram que sua distribuição no Facebook é desbalanceada, sendo que 48% são Participantes, 34% são Espectadores e apenas 18% são Produtores de conteúdo. Porém, ainda não se sabe quais as características específicas de cada perfil, portanto também não é possível afirmar quais estratégias persuasivas, visando aumentar as interações na rede, podem ser aplicadas para cada tipo de usuário.

Pensando nisso, foi desenvolvido e disponibilizado um questionário com o objetivo de adquirir informações sobre o uso do Facebook. Neste questionário uma das perguntas

* Usuários ativos são aqueles que logaram no Facebook nos últimos 30 dias. † Um site que mede os níveis de interação em sites de todo o mundo.

era com qual perfil (Participante, Espectador ou Produtor de conteúdo) o usuário se identificava e essa resposta foi utilizada para a rotulação das instâncias da base de dados.

Neste questionário foram obtidas 1100 respostas, sendo 426 Participantes, 623 Espectadores e 51 Produtores de conteúdo. Seguindo o que foi exposto por RUAS DA SILVEIRA et al. (2019), a quantidade de usuários que se classificaram como Produtores de conteúdo foi a mais baixa, dificultando a definição de regras para esse perfil e trazendo a necessidade de realização da seleção de instâncias para identificação das instâncias mais representativas do conjunto de dados.

Realizar uma análise desses dados de comportamento dos usuários é muito relevante, pois podem apresentar relações importantes entre as características de cada um desses perfis. E saber quais são essas características pode auxiliar no desenvolvimento de ferramentas que tenham mais chance de sucesso, pois, com essa informação em mãos, os idealizadores de redes sociais podem focar em desenvolver funcionalidades direcionadas para cada perfil e até mesmo disponibilizar gatilhos para que os usuários utilizem a rede social com frequência maior e assim mantenham a rede social viva.

Uma maneira de extrair informações de um conjunto de dados é por meio da visualização, pois, ao exibir esses dados em gráficos de forma otimizada pode ser possível encontrar padrões que auxiliem na descrição de suas características. Para isso é importante que sejam desenvolvidos gráficos que transmitam muito mais do que apenas imagens estatísticas; que também comuniquem ideias, que sejam capazes de mostrar uma verdade que até então não podia ser vista e que possam induzir as pessoas à acreditarem na verdade que a visualização mostra (BERINATO, 2016).

Uma outra forma de extrair informações de um conjunto de dados é através da criação de modelos por meio de técnicas de *machine learning*. Como nesse trabalho queremos obter as características de cada perfil e como estas características estão relacionadas, utilizamos árvores de decisão, algoritmos conhecidos pela sua alta capacidade de interpretabilidade, já que o conhecimento adquirido pelo modelo de aprendizado, representado por regras de classificação, pode ser facilmente obtido a partir da árvore. Vale ressaltar que, além de árvores de decisão, também foram feitos experimentos com métodos mais robustos do tipo caixa preta (Rede Neural, por exemplo) e os resultados encontrados foram semelhantes aos da árvore de decisão.

Assim, utilizamos os algoritmos de árvore de decisão C4.5 (QUINLAN, 1993) e CART (BREIMAN et al., 1984) para geração das regras de classificação dos três perfis: Espectador, Participante e Produtor de conteúdo. Optamos por trabalhar com dois algoritmos de árvore de decisão para avaliando se as regras geradas por ambos são semelhantes e assim validar e definir com mais segurança as características de cada perfil.

Como a quantidade de instâncias da classe Produtor de conteúdo obtidas pelo questionário foi muito baixa, os algoritmos de árvore de decisão podem apresentar dificuldade em classificá-la, e devido a isso utilizamos também a regressão logística para avaliar

essa classe.

1.1 Problema abordado

A subjetividade do usuário na rotulação e a baixa quantidade de usuários do perfil Produtor de conteúdo fazem com que a criação de regras para esse conjunto de dados seja uma tarefa complexa, pois criar modelos e selecionar instâncias em classes minoritárias não é uma tarefa trivial.

No caso da base de dados utilizada nesse trabalho, a hipótese é que ela contenha ruídos e *outliers* que precisam ser analisados para que apenas instâncias representativas sejam utilizadas na etapa de mineração de dados. Para isso, a proposta é utilizar métodos de seleção de instância visando criar modelos de aprendizado de máquina que avaliem os três perfis conjuntamente.

Diante do exposto, o problema dessa dissertação é sintetizado nas seguintes perguntas:

- Quais *insigths* podemos inferir sobre os perfis de usuário do Facebook a partir de uma visualização multivariada do conjunto de dados? E como validar se esses *insigths* são verdadeiros?
- Como determinar as principais características dos perfis de usuário do Facebook utilizando as instâncias mais representativas da base de dados por meio de *machine learning*?

1.2 Justificativa

Compreender como os usuários se comportam quando se conectam a redes sociais permite avaliar o desempenho de sistemas existentes e criar oportunidades para um melhor *design* de interface (BENEVENUTO et al., 2012). Saber quais são as características de cada grupo de usuário pode auxiliar na previsão do ciclo de vida da rede social (RUAS DA SILVEIRA et al., 2019) e também no desenvolvimento de ferramentas que tenham mais chance de sucesso, proporcionando tecnologias persuasivas mais eficientes.

Empresas que utilizam as redes sociais também podem se beneficiar dessa informação como mecanismo de *marketing*, podendo assim direcionar suas propagandas para os grupos de usuários por meio de sistemas de recomendação que são grandes aliados da personalização de sistemas computacionais. Esses sistemas coletam informações sobre

preferências de usuários com objetivo de ajudá-los a encontrar conteúdos, produtos ou serviços relevantes para si (PARK et al., 2012).

Para Tuna et al. (2016), a compreensão das características dos padrões de comportamento do usuário em redes sociais também pode fornecer informações sobre o comportamento das pessoas e, de posse dessas informações, desenvolvedores de redes sociais podem focar em funcionalidades que se adaptem ao seu comportamento.

Assim, ao compreender o comportamento das pessoas, a rede social pode oferecer melhores serviços, aumentando cada vez mais a qualidade da experiência do usuário, podendo ser reconhecida como um instrumento de muito valor para ele.

Portanto, as informações encontradas neste trabalho podem servir de inspiração para a construção de diversas novas aplicações no contexto de redes sociais, visando contribuir no processo de desenvolvimento de tecnologias persuasivas tornando-as mais eficazes. Pois, ao adaptar o sistema às características definidas pelos grupos de usuário, para essas pessoas, o sistema possuirá um valor maior (FERNANDES; SIQUEIRA, 2013).

1.3 Objetivos

O objetivo deste trabalho é caracterizar os três perfis de interação do Facebook utilizando visualização multivariada e algoritmos de aprendizado de máquina. Para atingir o objetivo proposto, foram desenvolvidos os seguintes objetivos específicos:

a) Desenvolvimento e aplicação de questionário para análise dos perfis de interação dos usuários da rede social;

b) Seleção de instâncias para identificação das instâncias mais representativas do conjunto de dados;

c) Visualização de dados para levantamento de *insights* sobre o comportamento de cada perfil de usuário;

d) Identificação das características de cada perfil de interação por meio de algoritmos de classificação, mais especificamente árvore de decisão e regressão logística;

1.4 Visão geral do trabalho

Esta dissertação está organizada da seguinte forma: o Capítulo 2 apresenta a revisão bibliográfica no qual essa pesquisa se baseia. No Capítulo 3 são apresentados os

trabalhos relacionados. O Capítulo 4 mostra a metodologia utilizada para o seu desenvolvimento. No Capítulo 5 são apresentados os resultados da seleção de instâncias, da visualização multivariava e da caracterização dos perfis de usuário. O Capítulo 6 apresenta as considerações finais, limitações da pesquisa e propostas de trabalhos futuros.

2 REVISÃO BIBLIOGRÁFICA

Este capítulo apresenta a revisão bibliográfica que aborda tópicos importantes para a compreensão deste trabalho, tais como: comportamento e perfis de usuário, seleção de instância, visualização multivariada, descoberta de conhecimento em base de dados e algoritmos para classificação.

2.1 Comportamento e perfil de usuário

Originalmente projetados para armazenar, processar e manipular dados, os computadores passaram a ser usados também como ferramenta persuasiva nos anos 70 (FOGG, 2003). O sistema desenvolvido na época tinha como objetivo fazer com que seus usuários adquirissem hábitos saudáveis. A partir daí, os computadores passaram a ser cada vez mais utilizados para influenciar o comportamento dos usuários.

Com o rápido desenvolvimento da vida social e digital, especialmente nas redes sociais, a mineração e a análise de dados estão cada vez mais sendo usadas para resolução de problemas e se tornando cada vez mais úteis (CAO et al., 2014). Devido à grande quantidade de dados sendo gerados na internet, a análise comportamental e a customização do serviço ganharam importância significativa.

Os dados comportamentais gerados fornecem uma nova perspectiva que permite a observação e análise dos comportamentos individuais e coletivos dos usuários. Os sistemas personalizados, por sua vez, procuram cada vez mais fornecer acesso eficiente e informações relevantes que correspondam aos interesses de cada usuário. Sabendo disso, a criação de perfis de usuário é crucial para a personalização por ser uma informação estruturada que contém preferências, comportamento e contexto do usuário (KADIMA; MALEK, 2010).

Além disso, um dos fatores que contribuem para o desempenho dos negócios é a personalização do sistema, pois eles afetam a forma como os produtos e serviços são vendidos e como as empresas atendem às necessidades dos seus clientes. A personalização do sistema também pode ser usada por anunciantes para marketing direcionado e a maioria das redes sociais usa dados personalizados para este fim (YASLAN; GÜLAÇAR; KOÇ, 2017).

Existem diversas pesquisas sobre perfis de usuários na literatura nas quais a maioria das abordagens são desenvolvidas para sistemas de recomendação. Nesses casos, os perfis de usuário encontrados são declarados pelos usuários ou inferidos de hábitos de navegação, como páginas visitadas e fluxos de cliques. No caso deste trabalho será feita

a identificação de perfis de usuários em redes sociais, mais especificamente no Facebook, buscando identificar regras que classificam os usuários nos perfis de interação anteriormente encontrados (Participante, Espectador e Produtor de conteúdo) baseado na visão que eles possuem de seu próprio uso da ferramenta.

2.2 Visualização multivariada

A visualização de dados consiste basicamente na representação gráfica de um conjunto de dados a fim de compreender as informações apresentadas e ser capaz de deduzir novos conhecimentos (FREITAS et al., 2001), auxiliando a análise e a compreensão de um conjunto de dados. O termo multivariada diz respeito ao estudo do comportamento de três ou mais variáveis simultaneamente. Portanto, a visualização multivariada trabalha com elementos gráficos para analisar o comportamento de diversas variáveis ao mesmo tempo.

Uma boa visualização pode ajudar muito na etapa de análise de dados, pois ela pode fornecer uma visão global do problema, revelar relações interessantes entre classes e atributos e fornecer informações sobre semelhanças e diferenças entre as classes. Essas visualizações podem não ser triviais de se compreender, pois podem conter muitos atributos, por isso é importante criar e interpretar corretamente o resultado.

Uma visualização multivariada para análise exploratória pode ser feita utilizando o FreeViz (DEMSAR; LEBAN; ZUPAN, 2008), um método para visualização de conjuntos de dados multidimensionais rotulados por classe disponível no Orange*. Ele faz uma otimização que encontra a projeção linear e o gráfico de dispersão associado que melhor separa instâncias de classes diferentes.

O FreeViz é considerado um algoritmo simples e relativamente rápido, entretanto para um conjunto de dados muito grandes ele pode ser executado de forma mais lenta, pois sua complexidade de tempo é $O(N^2 + NA)$, onde N é o número de instâncias e A o número de atributos.

Ao utilizar o Freeviz, é preciso ter em mente que sua finalidade é auxiliar na exploração dos dados, sugerindo padrões e relações potenciais para auxiliar na formulação de hipóteses, cabendo à estatística confirmá-las ou rejeitá-las, uma vez que a projeção é feita em baixa dimensão, o que limita seus resultados.

Em um gráfico resultante do FreeViz as linhas originadas do centro do gráfico são projeções dos vetores de base dos atributos, ou seja, cada vetor representa um atributo presente no conjunto de dados que é rotulado no ponto final da sua projeção. Além disso,

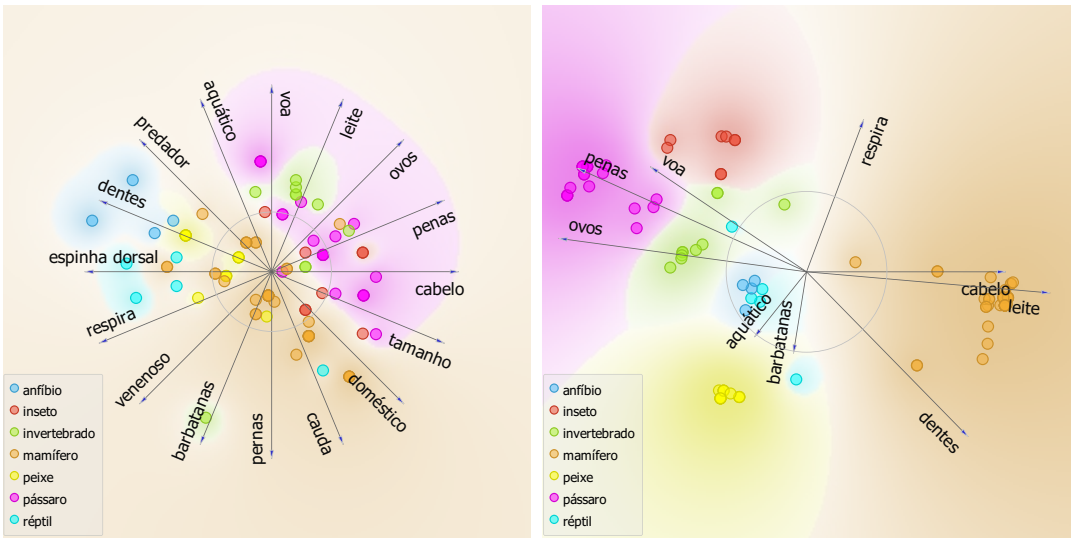
* O Orange é um kit de ferramentas de visualização de dados, aprendizado de máquina e mineração de dados de código aberto disponível em: https://orangedatamining.com.

como o FreeViz otimiza a projeção em relação à classificação, os atributos que são mais importantes para a classificação geralmente terão projeções mais longas. A direção da projeção também deve ser observada, pois quanto mais perpendicular à linha que separa as classes, mais útil ela é para distingui-las. Além disso, caso haja uma quantidade muito grande de atributos no conjunto de dados, é possível expor apenas os mais importantes, ou seja, visualizar somente os vetores de base com pontos finais que excedem uma certa distância do centro.

Cada instância do conjunto de dados é representada por um ponto e sua cor indica a qual classe pertence. Portanto, instâncias de classes que se assemelham são colocadas mais próximas. Além disso, caso haja uma região em que a predominância de uma classe é maior, essa região pode ser colorida com sua respectiva cor. Neste caso, quando uma região do gráfico é preenchida majoritariamente por instâncias de uma determinada classe, os atributos naquela direção podem ser bons indicadores da caracterização dessa classe.

A Figura 1 exibe um exemplo simples do uso do FreeViz em uma base de dados de zoologia do repositório *UCI Machine Learning* (DUA; GRAFF, 2017) com 101 instâncias e 17 atributos, dentre eles o tipo, usado como atributo de classe. Os gráficos mostram a posição inicial dos vetores e das instâncias e o resultado final, com a projeção otimizada.

Figura 1 – Visualização da base de dados de zoologia utilizando o FreeViz.



(a) Visualização inicial (b) Visualização otimizada

A visualização gerada pelo FreeViz e exibida na Figura 1(b) revela várias descobertas importantes sobre as classes (DEMSAR; LEBAN; ZUPAN, 2008):

- Ter cabelo está fortemente relacionado com dar leite e um pouco menos com ter dentes;
- Ter penas, voar e botar ovos parecem estar correlacionados e são todos propriedade dos pássaros;

- Os dois grupos de características acima estão em lados opostos, o que, na interpretação, significa que os animais que dão leite não põem ovos nem voam;

- Os mamíferos normalmente dão leite e possuem cabelo e dentes;

- Da mesma forma, existem os peixes que possuem barbatanas, não respiram e não possuem pernas;

- Parece haver dois grupos de animais que voam, pássaros e insetos, os primeiros são os que botam ovos e possuem penas;

- A domesticidade, o veneno e o tamanho não são muito informativos, pois não estão sendo exibidos na visualização.

No FreeViz todo gráfico é criado baseado nas informações de cada instância. A posição de uma instância e na projeção gerada é calculada pela Equação 2.1:

$$e = (e_x, e_y) \qquad e_x = \sum_i e^i A_x^i \qquad e_y = \sum_i e^i A_y^i \tag{2.1}$$

onde $e =$ corresponde à uma instância de dados descrita por valores de n características $= [e^1, e^2, ..., e^n]$; e $A =$ corresponde a matriz de projeção em que a linha $A^i = [A_x^i, A_y^i]$ representa a projeção do i-ésimo vetor base, isto é, os componentes x e y do vetor correspondente ao i-ésimo atributo.

Para o processo de otimização do algoritmo é utilizada a metáfora física que afirma que: "assim como as partículas físicas, as instâncias também exercerão forças umas sobre as outras dependendo de sua distância", e a direção dessa força depende dos tipos de cargas, ou seja, as classes das instâncias. Portanto, instâncias da mesma classe se atrairão e instâncias de classes diferentes se repelirão, e o objetivo da otimização é encontrar a projeção com o mínimo de energia potencial.

O algoritmo, por sua vez, calcula a força entre cada par de instâncias apenas uma vez e a adiciona à soma das forças para ambas as instâncias, porém com direções diferentes ($F_{f \to e} = -F_{f \to e}$). A força ($F\_ef$) é separada nos componentes x e y ($Fefx$ e $Fefy$) multiplicando-a por projeções para o eixo x e o eixo y, dx/r e dy/r, respectivamente.

Para evitar que a projeção exploda ou imploda, ela é re-centralizada e re-normalizada após cada etapa, de modo que a soma de todas as projeções do vetor base seja zero e o vetor base mais longo tenha comprimento unitário. O procedimento é repetido até que não haja diminuição considerável da energia potencial por algumas etapas consecutivas.

2.2.1 Técnicas de visualização multivariada de dados

O grande volume de dados que vem sendo produzido atualmente traz a necessidade de utilizar boas ferramentas para visualização, com isso é perceptível os grandes avanços que essas área tem alcançado. As ferramentas de visualização precisam trazer informações de forma clara e objetiva, gerando gráficos de alta qualidade que possibilitem a identificação de *insights* iniciais para direcionar pesquisas futuras.

O Radviz (visualização radial) (HOFFMAN et al., 1997) é uma técnica de visualização multidimensional não linear que pode exibir dados definidos em diversas variáveis em uma projeção bidimensional. Os atributos são apresentadas como pontos de ancoragem igualmente espaçados em torno do perímetro de um círculo unitário e são chamados de âncoras dimensionais (DAs). As DAs podem ser movidas interativa ou algoritmicamente para revelar diferentes padrões significativos no conjunto de dados.

As instâncias são mostradas como pontos dentro do círculo, com suas posições determinadas por uma metáfora da física: cada ponto é mantido no lugar com molas que são conectadas na outra extremidade às âncoras variáveis. A rigidez de cada mola é proporcional ao valor da variável correspondente e o ponto termina na posição onde as forças da mola estão em equilíbrio, ou seja, iguais à zero. A localização no espaço visual depende muito da organização dos atributos ao redor do círculo unitário.

A *Star Coordinates - SC* (KANDOGAN, 2000) é uma técnica de projeção de redução de dimensionalidade que organiza e representa variáveis numéricas em um *layout* radial. A SC gera mapeamentos lineares computando combinações lineares de um conjunto de vetores de baixa dimensão que representam eixos radiais, a orientação determina a direção em que uma variável aumenta, e o comprimento especifica a quantidade de contribuição de uma determinada variável na visualização resultante, dado que todas as variáveis têm uma escala semelhante.

O GridViz (DZEMYDA; KURASOVA; ZILINSKAS, 2013) é uma extensão do Radviz que coloca as âncoras dimensionais em uma grade retangular em vez de um círculo. O sistema de molas é o mesmo do Radviz em que as instâncias são posicionados onde a força das molas atinge um equilíbrio.

Long (2018) desenvolveu o ArcViz, um aprimoramento do RadViz oferecendo mais espaço para exibir conjuntos de dados de alta dimensão. Nesse método a definição das âncoras no círculo unitário e a posição das instâncias são feitas como no RadViz tradicional. A diferença é que as posições das âncoras não são mais representadas apenas por um ponto, mas por um arco ao redor do círculo.

2.3 Eliminação de ruídos

Algoritmos de aprendizagem tem como objetivo fazer uma generalização a partir de um conjunto de instâncias de treinamento de modo que sua precisão de classificação em instâncias não observados anteriormente seja maximizada (ZHU; WU; CHEN, 2003).

Um fator muito importante para garantir que a precisão seja máxima é a qualidade dos dados de treinamento. Em base de dados com informações do mundo real é comum a presença de ruídos, por isso fazer uma limpeza na base tende a melhorar significativamente a precisão da classificação. Além disso, as informações em base de dados podem não ser todas igualmente informativas e além de ruídos, algumas instâncias também podem ser consideradas *outliers*, que provavelmente influenciarão de forma negativa o desempenho da classificação final. Em árvores de decisão, por exemplo, o ruído de rótulo afeta muito o desempenho do classificador, podendo gerar árvores muito grande e tornando-as excessivamente complicadas (QUINLAN, 1986).

Assim, para que o treinamento de um modelo seja mais robusto pode ser necessário identificar as instâncias mais representativas eliminando os possíveis ruídos e *outliers* da base de dados. Para isso as técnicas de seleção de instância podem ser utilizadas, pois um dos seus objetivos é reduzir ou eliminar ruídos e valores discrepantes (CAVALCANTI; SOARES, 2020). Os ruídos podem ser classificados em dois tipos (WU, 1995):

- Ruído de atributo: São erros presentes nos valores dos atributos da base de dados;
- Ruído de classe: Podem ser encontrados em exemplos contraditórios, ou seja, instâncias iguais com rótulos de classe diferente ou em instâncias rotuladas com classes incorretas.

Avaliar a presença de ruído de classe na classificação é uma etapa muito importante durante o pré-processamento da base de dados, pois a presença de ruídos pode trazer muitas consequências para a construção do modelo, como diminuir sua precisão geral e aumentar sua complexidade (FRENAY; VERLEYSEN, 2014). Portanto, é necessário implementar técnicas que eliminem o ruído ou reduzam suas consequências.

A fonte do ruído de classe pode vir de diversas maneiras (FRENAY; VERLEYSEN, 2014). As informações fornecidas podem ser insuficientes, ruins ou possuir qualidade variável para realizar uma rotulagem confiável. Quando a tarefa de rotulagem é subjetiva, por meio de um questionário por exemplo, ela pode ser entendida de maneiras diferentes, causando variabilidade na rotulagem para instâncias semelhantes.

As consequências negativas que a presença do ruído podem trazer para o aprendizado são muitas, dentre elas, a diminuição no desempenho na classificação, a necessidade de aumentar o número de amostras necessárias para o aprendizado e o aumento na complexidade do modelo (FRENAY; VERLEYSEN, 2014).

Para tratar o ruído de classe podem ser adotadas três estratégias diferentes (FRENAY; VERLEYSEN, 2014):

1. Utilizar algoritmos que não sejam muito sensíveis à presença de ruídos, como por exemplo métodos *ensemble*: LogitBoost (FRIEDMAN et al., 2000) e BrownBoost (FREUND, 1999);

2. Utilizar algoritmos que consideram o ruído de uma forma incorporada, como por exemplo CN2 (CLARK; NIBLETT, 1989) e PAM (KHARDON; WACHMAN, 2007);

3. Melhorar a qualidade dos dados de treinamento identificando e tratando os ruídos.

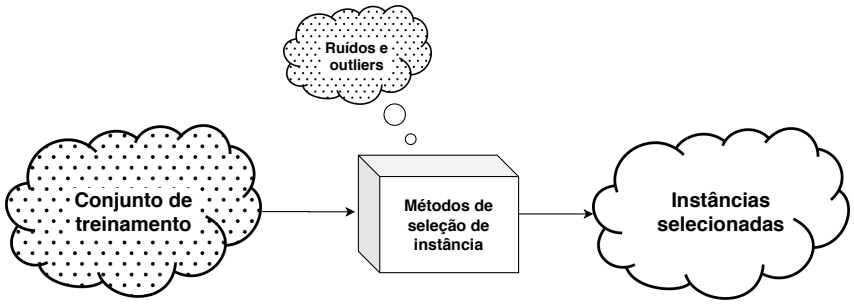
Para esta pesquisa adotaremos a terceira estratégia, ou seja, por meio de seleção de instâncias iremos eliminar os ruídos e manter as instâncias mais representativas visando melhorar a qualidade dos dados disponíveis.

2.4 Seleção de instâncias

Também conhecida como técnica de redução ou método de seleção de protótipo, a seleção de instância visa obter um conjunto de treinamento representativo que seja menor em comparação com o original mas que possua uma precisão de classificação semelhante ou ainda maior para novos dados de entrada. A principal vantagem da seleção de instâncias é a capacidade de escolher amostras relevantes sem gerar novos dados artificiais (GARCIA; LUENGO; HERRERA, 2015).

Dado um conjunto de treinamento, o objetivo da seleção de instâncias é obter um subconjunto que não contenha instâncias supérfluas e que a precisão da classificação seja aproximadamente a mesma da base de dados completa (OLVERA-LOPEZ et al., 2010). A Figura 2 ilustra esse processo de seleção de instância.

Figura 2 – Processo de seleção de instância.



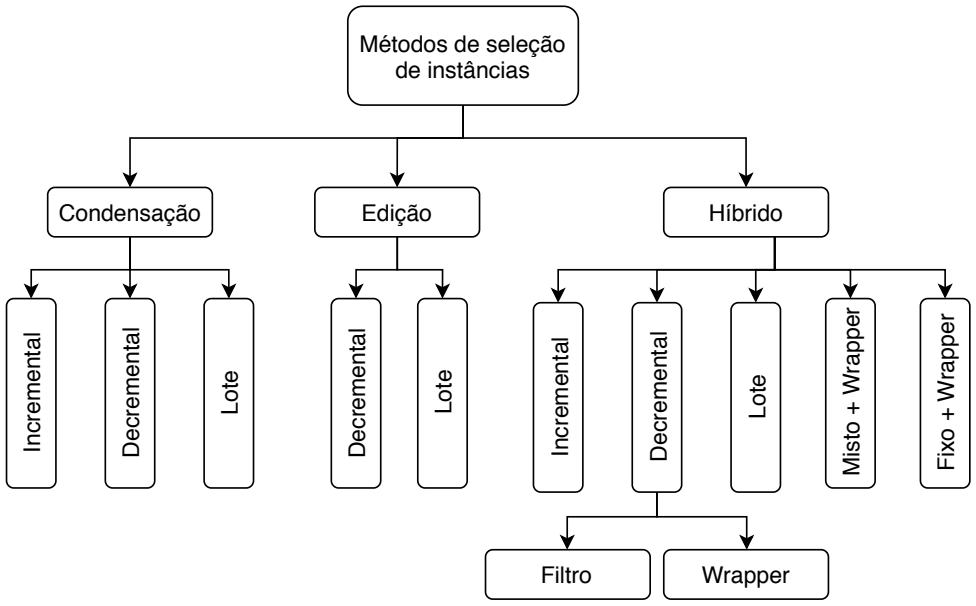
Fonte: Adaptado de Olvera-Lopez et al. (2010)

Para Cavalcanti e Soares (2020), os algoritmos de seleção de instância têm dois objetivos principais: (1) reduzir ou eliminar ruídos e valores discrepantes (*outliers*) e (2) reduzir a carga computacional de algoritmos de aprendizagem baseados em instância. Portanto, a redução de instâncias visa remover automaticamente dados não representativos do conjunto de treinamento mantendo a integridade do conjunto de dados original (YOU et al., 2020).

2.4.1 Taxonomia da seleção de instâncias

Garcia, Luengo e Herrera (2015) afirmam que os métodos de seleção de instância podem ser categorizados de acordo com a direção de pesquisa (incremental, decremental, lote, misturado e fixo), com o tipo de seleção (condensação, edição e híbrido) e com a avaliação da pesquisa (filtro e *wrapper*). A Figura 3 mostra uma taxonomia para categorizar os métodos de seleção de instância.

Figura 3 – Taxonomia de seleção de instâncias.



Fonte: Adaptado de Garcia, Luengo e Herrera (2015)

- **Direção de pesquisa:** A seleção de instâncias pode prosseguir por várias direções durante a busca do subconjunto, podendo iniciar o subconjunto contendo dados do conjunto original ou vazio. Garcia, Luengo e Herrera (2015) afirmam que há cinco direções:
  - **Incremental:** Os métodos incrementais se iniciam com o subconjunto vazio e adiciona as instâncias durante o processo de seleção caso ela cumpra alguns

critérios. Um fator que pode ser muito importante neste caso é a ordem de apresentação das instâncias, pois os primeiros casos analisados podem ter uma probabilidade diferente de serem incluídos no subconjunto do que se fossem visitados posteriormente. Portanto a ordem de apresentação das instâncias deve ser aleatória. Uma vantagem dessa abordagem é que instâncias que forem disponibilizadas posteriormente podem continuar a ser inseridas no subconjunto de acordo com os mesmos critérios. A principal desvantagem é que eles precisam tomar decisões com base em poucas informações e, portanto, estão sujeitos a erros até que mais informações estejam disponíveis. Alguns algoritmos examinam todas as instâncias disponíveis para ajudar a selecionar qual instância adicionar a seguir, isso pode melhorar muito o desempenho do algoritmo, mas torna-o não verdadeiramente incremental.

  - **Decremental:** Os métodos decrementais iniciam-se com o conjunto contendo todos as instâncias e vai removendo-as ao longo da seleção. A ordem de apresentação continua sendo importante, mas ao contrário do processo incremental, todos os exemplos de treinamento estão disponíveis para auxiliar durante o processo de análise. Uma desvantagem do método decremental é que ele apresenta um custo computacional maior do que algoritmos incrementais. No entanto, se a aplicação de um algoritmo decremental resultar em uma redução maior, então o cálculo extra durante o aprendizado pode valer a pena se puder alcançar uma maior precisão de generalização.
  - **Lote:** Nos métodos em lote todas as instâncias são analisadas previamente em relação aos critérios de remoção. Logo em seguida todas aquelas que atendam aos critérios são removidas de uma única vez. Assim como nos métodos decrementais, o processamento em lote sofre com o aumento da complexidade do tempo em relação aos métodos incrementais.
  - **Misto:** Os métodos mistos começam com um subconjunto pré-selecionado (aleatoriamente ou selecionado por um processo incremental ou decremental) e pode adicionar ou remover iterativamente qualquer instância que atenda ao critério específico. A principal vantagem desse métodos é facilitar a obtenção de subconjuntos de instâncias adequados à boa precisão. Geralmente sofre das mesmas desvantagens relatadas em algoritmos decrementais.
  - **Fixo:** Os métodos fixos funcionam como o misto, a diferença é que o número de adições e remoções permanece o mesmo. Assim, o número de instâncias finais é determinado no início da fase de aprendizagem e nunca é alterado.

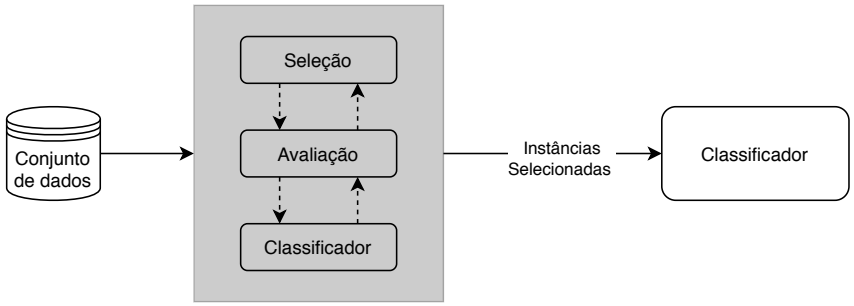
- **Tipo de seleção:** A seleção de instâncias também pode ser de tipos diferentes, buscando reter pontos de fronteira, pontos centrais ou algum outro conjunto de pontos. Garcia, Luengo e Herrera (2015) afirmam que há três tipos:

  - **Condensação:** Os métodos de condensação visam obter uma alta taxa de redução do conjunto de dados. Para isso retem os pontos mais próximos dos limites de decisão, também chamados de pontos de fronteira, e elimina os pontos internos por não afetar os limites de decisão tanto quanto os pontos de fronteira, podendo ser removidos com relativamente pouco efeito na classificação. A capacidade de redução desses métodos geralmente é alta devido ao fato de que há menos pontos de fronteira do que pontos internos na maioria dos conjuntos de dados e a precisão do conjunto de treinamento é preservada, porém a precisão da generalização no conjunto de teste pode ser afetada negativamente.
  - **Edição:** Os métodos de edição visam melhorar o desempenho de classificação do algoritmo de aprendizagem e, ao contrário dos métodos de condensação, eliminam as instâncias nos limites de decisão, enquanto as instâncias nos pontos internos são mantidas no conjunto de dados. Métodos de edição podem ser utilizados para limpar as instâncias ruidosas e *outliers* sem qualquer compromisso para obter uma alta taxa de redução do conjunto de dados, sendo assim seus resultados aumentam a precisão da generalização nos dados de teste, mas a taxa de redução de dados permanece baixa.
  - **Híbrido:** Os métodos híbridos combinam o melhor dos dois métodos anteriores, ou seja, visam aumentar tanto a precisão da classificação quanto a taxa de redução de dados. Para isso, eles selecionam as instâncias mais representativas, independentemente do fato de que as instâncias estão na fronteira ou posicionadas em regiões mais internas do espaço de recurso.

- **Avaliação da pesquisa:** Os algoritmos de seleção de instâncias também podem ser classificados em relação à avaliação da pesquisa (GARCIA; LUENGO; HERRERA, 2015):

  - ***Wrapper*:** Nos métodos *wrapper*, as instâncias são selecionadas com base no desempenho de um algoritmo de aprendizagem (como uma caixa-preta) em subconjuntos de instâncias candidatas. Ou seja, nestes métodos um classificador é utilizado para avaliar a qualidade dos subconjuntos durante o processo de seleção e sua acurácia é usada como medida para avaliar esses subconjuntos. Sendo assim, as melhores instâncias para uma rede neural, por exemplo, podem não ser as melhores para uma árvore de decisão. Geralmente esses métodos tendem a obter uma ótima estimativa da precisão da generalização, no entanto seus custos computacionais podem ser altos. A Figura 4 ilustra como é feito esse processo.
  - **Filtro:** Em métodos de filtro o processo de escolha do subconjunto acontece antes do processo de aprendizado. A ideia é filtrar instâncias irrelevantes, se-

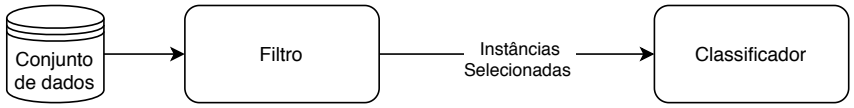
Figura 4 – Método Wrapper.



Fonte: Elaborado pelo autor

gundo algum critério, antes de iniciar a indução. Assim, esse método considera características gerais do conjunto de exemplos para selecionar algumas instâncias e excluir outras. Com isso, ao contrário dos métodos *wrapper*, não se beneficiam dos resultados de nenhum algoritmo de aprendizagem durante esse processo e sua precisão pode não ser aprimorada. A Figura 5 ilustra como é feito esse processo.

Figura 5 – Método Filtro.



Fonte: Elaborado pelo autor

2.5 Algoritmo genético

Desenvolvido por Holland (1992), um algoritmo genético (AG) é uma técnica inspirada no princípio darwiniano de seleção natural e reprodução genética (hereditariedade, mutação, seleção natural e *crossing over*), um método usado para encontrar soluções aproximadas em problemas de busca e otimização.

De acordo com a teoria da evolução de Charles Darwin (DARWIN, 1859), os seres sofrem mutações, e os indivíduos que apresentam características mais aptas ao meio ambiente têm maior probabilidade de sobreviver e deixar seus herdeiros. Assim, ao longo das gerações, a população de uma espécie evolui para se adaptar melhor ao meio ambiente onde reside.

Portanto, investigar como essa seleção natural ocorre e realizar simulações permite gerar dados com as características desejadas. Esses princípios são imitados na construção de algoritmos computacionais que buscam uma melhor solução para um determinado

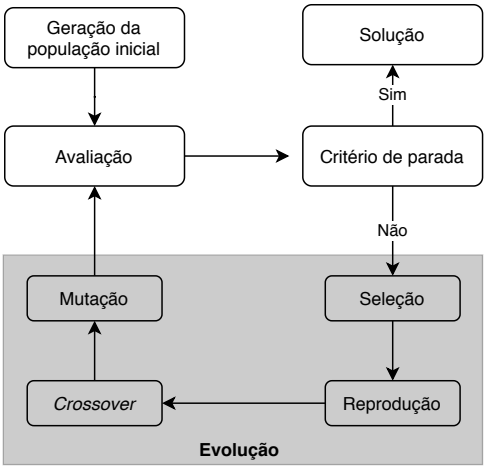
problema por meio da evolução de populações de soluções codificadas por meio de cromossomos artificiais.

O algoritmo genético funciona iterativamente em um conjunto de soluções possíveis para um determinado problema. Esse conjunto constitui uma população, e cada solução possível corresponde a um indivíduo dessa população, também conhecido como cromossomo.

A função de aptidão direciona o processo de evolução atribuindo o valor de aptidão a cada indivíduo na população. Esta função varia de acordo com o problema a ser resolvido.

A Figura 6 mostra um diagrama de atividades usado para representar um algoritmo genético básico. O processo iterativo de um algoritmo genético começa com a criação da população inicial de cromossomos. Esses cromossomos são avaliados quanto à adequação. Se um critério de parada não for atendido, a evolução da população começa.

Figura 6 – Diagrama de um algoritimo genético.



A primeira fase é a seleção, em que são escolhidos os indivíduos que formarão a próxima geração da população. Os indivíduos selecionados iniciam a fase reprodutiva em pares, na qual são realizadas operações de cruzamento genético para gerar seus descendentes.

Depois que todos os indivíduos são gerados, uma política de substituição é usada para determinar como a próxima geração será formada. Formada a nova população, cada indivíduo pode passar pelo processo de mutação que causa mudanças aleatórias nos genes do indivíduo. A função de aptidão, por sua vez, reavalia essa nova população até que o critério de parada seja satisfeito.

2.6 Classificação por meio de árvore de decisão

Uma técnica de mineração de dados bem conhecida é a árvore de decisão. Faceli et al. (2011) afirmam que "uma árvore de decisão é um grafo direcionado a-cíclico no qual cada nó é um nó de partição, com dois ou mais nós sucessores, que podem ser outros nós de partição ou nós folha".

Cada nó de partição indica um teste de divisão usando o valor de um dos recursos de entrada (RUSSELL; NORVIG, 2013). A escolha do recurso de divisão geralmente é feita com base em uma medida de ganho de informação. Uma vez que a árvore é produzida, pode-se seguir seus ramos de acordo com os valores dos atributos para classificar as instâncias não rotuladas, testando apenas as características do modelo de árvore. Em nosso estudo, as árvores de decisão foram construídas usando os algoritmos C4.5 e CART.

Proposto por Quinlan (1993), o C4.5 é considerado uma extensão do algoritmo ID3 (QUINLAN, 1986) que gera um modelo de classificação de árvore de decisão que avalia a entropia dos subconjuntos gerados em cada partição. As partições são escolhidas selecionando a combinação de recurso e valor de teste que incorre na melhor taxa de ganho de informação, considerando os subconjuntos gerados pelos nós filhos. O método C4.5 também utiliza o conceito de entropia, medindo a pureza de um conjunto. A estratégia do algoritmo é reduzir a dificuldade de predição da variável alvo (rótulo da classe) após cada partição (FACELI et al., 2011). Considerando uma variável aleatória A, sua entropia é dada pela Equação 2.2.

$$H(A) = -\sum_{i=1}^{n} P_i * log_2 p_i \tag{2.2}$$

onde p_i é a probabilidade de cada valor em A.

O algoritmo CART (*Classification and Regression Trees*) foi desenvolvido por Breiman et al. (1984) e pode pesquisar relações nos dados, prevendo valores de variáveis alvo discretas (classificação) e variáveis alvo contínuas (regressão). Ao construir uma árvore de decisão usando este algoritmo, todos os nós pais são particionados em dois nós filhos com menos variabilidade em seus subconjuntos.

Ao contrário do C4.5, o algoritmo CART utiliza o índice de Gini como um critério para escolher o recurso de divisão para fazer a partição em cada nó não folha. Este índice mede a heterogeneidade nos dados; assim, ele também pode medir o grau de impureza de um nó. Se um conjunto de dados B tem amostras de classes N, o índice Gini é dado pela Equação 2.3:

$$gini(B) = 1 - \sum_{j=1}^{n} p_j^2 \tag{2.3}$$

onde p_j é a frequência relativa da classe j em B.

2.7 Regressão logística para identificação de atributos mais relevantes

O principal objetivo da regressão logística é explorar a relação entre uma ou mais variáveis explicativas e uma variável resposta. É uma ferramenta estatística multivariada muito utilizada quando se quer explicar ou predizer a ocorrência de determinado evento buscando modelar a relação "logística" entre uma variável resposta e uma serie de variáveis explicativas (CORRAR; PAULO; FILHO, 2007). Em outras palavras, essa técnica tenta predizer valores de uma variável em função de valores conhecidos de outras variáveis.

Para estimar os parâmetros da equação logística essa técnica utiliza um processo estatístico que estima a probabilidade máxima de ocorrência de determinado evento, o método da verossimilhança. E, ao final da execução do modelo logístico, uma equação contendo atributos do conjunto de dados com seus respectivos coeficientes ajustados é apresentada como resultado.

A regressão logística se divide em duas: Regressão Logística Binária e Regressão Logística Multinomial (RLM). A diferença entre as duas é que na RLM podem haver mais de duas possíveis classes para a variável dependente, enquanto na RLB a variável dependente precisa ser dicotômica, ou seja, apresentar exatamente duas classes, tais como: sim ou não, aceitar ou rejeitar e positivo ou negativo.

Além disso, a regressão logística também estima a porcentagem de ocorrência de determinado evento, agindo como um facilitador em problemas que envolvem não apenas discriminação de classes mas também estimativa de probabilidade (CORRAR; PAULO; FILHO, 2007).

Essa técnica também pode ser utilizada para identificar variáveis que contribuem para ocorrência de determinado evento, conforme já utilizado por Horta et al. (2010), o que auxiliará no desenvolvimento deste trabalho. Para isso aplicamos o teste *Wald* que mede o grau de significância de cada coeficiente da equação logística (CORRAR; PAULO; FILHO, 2007). Ele verifica se cada parâmetro estimado é significativamente diferente de zero, ou seja, testa a hipótese de que um determinado coeficiente é nulo. Por fim, o grau de importância de um atributo é medido pelo *p-value* de Wald. O *p-value* é uma probabilidade que mede a evidência contra a hipótese nula, portanto, as probabilidades inferiores fornecem evidências mais fortes contra a hipótese nula. Assim, quanto menor for este valor mais importante será considerado o atributo.

A regressão logística pode ser aplicada em diversas áreas atuando como uma forte ferramenta de análise de dados de resposta categórica, auxiliando a estimava da probabilidade de ocorrência de eventos e a avaliação de fatores que contribuem para o mesmo.

3 TRABALHOS RELACIONADOS

Este capítulo apresenta as pesquisas que já foram desenvolvidos sobre os principais pontos desse trabalho: (1) comportamento de usuário em redes sociais, (2) visualização multivariada e (3) algoritmos de seleção de instâncias.

3.1 Comportamento do usuário em redes sociais

Benevenuto et al. (2010) realizaram uma análise do serviço de vídeo do UOL (Universo Online), e apresentaram uma caracterização das sessões de seus usuários, solicitações de servidor e perfis de navegação. Um dos resultados do estudo foi a descoberta de diferentes perfis de usuários que acessam o sistema. Segundo os autores, essas informações poderiam ser utilizadas pelos administradores do sistema para personalizar serviços. Eles também propuseram um método para categorizar os usuários de acordo com seu perfil de navegação: um desses perfis, denominado *Viewers*, visualiza predominantemente vídeos, enquanto outro exemplo de perfil, *Searchers*, são usuários que geralmente iniciam suas sessões em busca de conteúdo específico.

Li, Bernoff e Groot (2011) usam dados demográficos para agrupar usuários em *personas* por suas atividades na rede social. Para eles, saber como acessar os diferentes usuários é extremamente benéfico para a estratégia social de uma empresa. Os autores afirmam que existem sete perfis de usuário: 1) *Creators*: pelo menos uma vez por mês publica um blog ou artigo *online*, mantém uma página da *web* ou faz *upload* de vídeo ou áudio; 2) *Conversationalists*: expressam suas opiniões a outros usuários e empresas usando as redes sociais; 3) *Critics*: posta comentários em blogs, fóruns *online*, escreve resenhas e edita *wikis*; 4) *Coletores*: armazena URLs, usa RSS* ou vota em *sites* favoritos; 5) *Joiners*: usuários de redes sociais como Facebook e MySpace; 6)*Spectators*: usuários que consomem o que os demais produzem e são o maior grupo; e 7) *Inactives*: usuários que não participam mesmo estando *online*.

Em Vasconcelos et al. (2012), os autores analisaram como os usuários do Foursquare† exploram três recursos disponíveis na plataforma: 1) *Tips*: dicas, avaliações ou recomendações que são postadas por outros usuários, refletindo suas impressões positivas ou negativas sobre um determinado local; 2) *Dones*: marcações feitas pelos usuários concordando com dicas, como um *feedback*; e 3) *ToDos*: lista onde os usuários adicionam locais para visitar. Os usuários foram agrupados em quatro perfis de comportamento dis-

* RSS: *Rich Site Summary* é um formato de distribuição de informações em tempo real pela internet usado principalmente em *sites* de notícias e blogs. † Uma rede social baseada em localização, que combina SNSs com compartilhamento de informações geográficas. Disponível em: https://pt.foursquare.com/.

tintos nessa rede social por meio da análise desses recursos. Dois desses perfis correspondem a usuários regulares, diferenciando-se em termos de engajamento, como frequência de postagem de dicas. Um outro grupo é composto por usuários influentes, que postam várias dicas na rede (empresas ou marcas administram algumas dessas contas de usuário). O último perfil de usuário é o de *spammers*, caracterizado por postagens não relacionadas à localização das dicas, normalmente com *links* em suas postagens (VASCONCELOS et al., 2012).

Çelik e Karaaslan (2014) usaram as redes sociais (Facebook, Twitter, Foursquare e Instagram) para determinar os preditores de sua frequência de uso. Para isso, foi aplicado um questionário a 822 alunos de universidades turcas. Seus resultados revelaram que a frequência de *login*, o tempo gasto em redes sociais e os eventos foram os preditores que mais impactaram significativamente o uso das redes sociais.

RUAS DA SILVEIRA, Nobre e Cardoso (2014) analisaram as estratégias de persuasão empregadas no Facebook, seguindo as definições propostas por Fogg (2003). Os autores também avaliaram por meio de um questionário respondido por 296 usuários do Facebook as influências dessas estratégias no comportamento do usuário. Além disso, propuseram a existência de três tipos de interação de perfis para usuários do Facebook: *Espectador*, *Participante* e *Produtor de conteúdo*. De acordo com os resultados do questionário, 48% dos usuários do Facebook se consideram Participantes, 34% Espectadores e apenas 18% Produtores de conteúdo.

Em outro trabalho, RUAS DA SILVEIRA et al. (2019) utilizaram métricas de redes complexas como uma estratégia de extração de recursos para enriquecer um conjunto de dados coletados do Facebook, agrupar seus usuários e identificar seus perfis a partir dos resultados do agrupamento. Eles identificaram três grupos nos dados coletados que confirmaram ainda mais a definição dos perfis no trabalho anterior.

Há estudos publicados que caracterizam perfis de usuários de redes sociais e estudos que mencionam estratégias de persuasão usadas no Facebook para diferentes perfis de usuários. No entanto, uma caracterização baseada nas regras desses perfis com base na experiência do usuário e como eles se identificam na rede social ainda é um assunto que não foi discutido. Diferentemente dos estudos citados acima, fizemos um trabalho sobre perfis de usuários do Facebook quanto aos seus aspectos de interação social (como curtidas, comentários e compartilhamentos feitos na rede social), para caracterizá-los com base em regras de classificação.

3.2 Visualização multivariada

As técnicas de visualização já foram aplicadas em diversos campos como bioinformática, biologia, saúde, engenharia e finanças e são muito úteis para várias tarefas exploratórias de análise de dados, incluindo descoberta de estrutura de *cluster*, detecção de *outliers* e tendência, extração de recursos ou tarefas de suporte à decisão (RUBIO-SANCHEZ et al., 2016).

Campbell et al. (2010) dissertam sobre o desenvolvimento de um sistema para gerar novas hipóteses terapêuticas para doenças humanas e utilizam a visualização para melhor transmitir o significado geral de um conjunto integrado de dados, incluindo associação de doenças, drogabilidade, inteligência do concorrente, genômica e mineração de texto. Os autores afirmam que, embora já existam várias abordagens de visualização nesta área, a maioria considera apenas uma gama limitada de dados e reforçam a necessidade de mais ferramentas neste contexto.

Meydan e Sezerman (2011) utilizam a visualização para identificar a importância de cada atributo e sua relação com as instâncias em um conjunto de dados de pacientes infectados com o vírus HIV (vírus da imunodeficiência humana). Além disso, eles fazem uma seleção dos cinco principais atributos a partir de um resultado otimizado que é apresentado na visualização para depois submeter o conjunto de dados a um algoritmo de classificação e levantam alguns *insights* iniciais sobre a correlação que existe entre alguns atributos e também sobre outros que teoricamente seriam importantes para a classificação, porém, de acordo com a visualização não teriam influencia no resultado.

Noor et al. (2011) desenvolveram uma aplicação visando aumentar a eficiência na investigação de crimes e a acessibilidade às informações usando a visualização. Por apresentar informações muito complexas, a genética forense requer uma boa forma de representação para analisar os perfis de DNA na investigação de um crime, por exemplo. Os autores utilizam então a visualização para analisar dados de perfis de DNA suspeitos e desconhecidos para tornar a interpretação e a busca de detalhes mais fácil e rápida.

Zhang et al. (2014) utilizam visualização para discriminar entre dois tipos de carcinomas do pulmão. As análises realizadas mostram que, ao ser combinada com um método de classificação, a visualização desempenha um papel importante na seleção dos atributos relevantes e melhora a parcimônia, enquanto o modelo final sofre nenhuma ou menos perda de precisão. Os autores também sugerem que a representação gráfica é mais facilmente compreensível para um clínico do que métodos estatísticos ou equações matemáticas, por exemplo, o que pode auxiliar os diagnósticos clínicos.

Bellocchi et al. (2019) investigam o microbioma intestinal e o metaboloma do plasma em pacientes com doenças autoimunes sistêmicas e utilizam a visualização para representar os agrupamentos construídos a partir das informações selecionadas do con-

junto de dados.

Ferraz et al. (2020) avaliam as propriedades químicas, físicas, anatômicas e mecânicas de painéis de cimento reforçados com materiais lignocelulósicos (eucalipto, bagaço de cana-de-açúcar, casca de coco, casca de café e pseudocaule de banana). Eles utilizam a visualização para observar a composição química dos resíduos de material, analisando quais componentes têm maior ou menor efeito na composição do material lignocelulósico e, com base nos componentes químicos médios, eles determinam a aptidão do material.

3.3 Algoritmos de seleção de instâncias

Visando encontrar técnicas de redução de instância que forneçam tolerância a ruído, alta precisão de generalização, insensibilidade à ordem de apresentação das instâncias e redução significativa de armazenamento, diversos algoritmos de seleção de instância foram propostos.

Wilson e Martinez (2000) propuseram seis algoritmos de redução de instância chamados DROP1-DROP5 (procedimento de otimização de redução decremental, do inglês, *decremental reduction optimization procedure*) e DEL (comprimento de codificação decremental, do inglês, *decremental encoding length*) que são baseados na distância entre os objetos. Os resultados do trabalho dos autores mostram que, quando comparados com outros algoritmos que fornecem redução substancial de armazenamento, os algoritmos DROP têm precisão de generalização média mais alta, especialmente na presença de ruído de classe uniforme.

Zhu e Wu (2006) propuseram o 3STAR (seleção e classificação de instâncias representativas escaláveis, do inglês, *scalable representative instance selection and ranking*) para seleção e classificação de instâncias representativas escalonáveis que fornece uma lista de classificação de instâncias representativas, de modo que os usuários podem sempre selecionar instâncias de cima para baixo, com base no número de exemplos de sua preferência. O algoritmo agrupa as instâncias em pequenas células de dados com comportamentos semelhantes e avalia progressivamente as células de dados e suas combinações, organizando-as em uma lista de forma que os modelos construídos a partir das células superiores sejam mais precisos.

Tsai et al. (2019) apresentam uma nova abordagem para seleção de instâncias chamada CBIS (seleção de instância baseada em cluster, do inglês, *cluster-based instance selection*) que é composto por dois componentes: análise de agrupamento e seleção de instâncias. A análise de agrupamento é usada para agrupar amostras de dados semelhantes no conjunto de dados da classe majoritária em uma série de grupos que podem ser considerados como "subclasses" da classe majoritária. A seleção de instância é usada para

filtrar amostras de dados não representativas de cada uma das "subclasses". O objetivo dos autores é realizar uma subamostragem eficaz do conjunto de dados da classe majoritária e os resultados encontrados demonstram a eficácia da abordagem proposta.

Garcia-Pedrajas, del Castillo e Cerruela-Garcia (2020) propõem um novo algoritmo de seleção de instância e atributos, simultaneamente. Segundo os autores, a seleção de instância é um dos métodos mais amplamente usado para redução de dados. Porém os problemas atuais envolvem conjuntos de dados altamente complexos com um grande número de atributos que dificulta significativamente as tarefas de classificação e reconhecimento, pois a velocidade de muitos algoritmos de classificação é muito melhorada quando a complexidade dos dados é reduzida. Porém, de acordo com os autores, poucos algoritmos foram propostos para abordar simultaneamente as tarefas de seleção de instância e atributos. Além disso, a maioria desses métodos é baseada em heurísticas complexas. Por isso, o método proposto pelos autores usa uma pesquisa heurística simples e vários mecanismos de escalonamento que podem ser aplicados a conjuntos de dados com milhões de atributos e instâncias. Esse método atinge melhor redução de armazenamento e erro do que as abordagens anteriores.

Também pensando no tratamento conjunto de seleção de instâncias e recursos, Du et al. (2020) propõem um novo *framework* para difundir efetivamente o processo de seleção de atributos e instâncias. O método trabalha de forma simultânea e iterativa nas duas seleções. Para fazer a seleção de instâncias utiliza-se resultados intermediários da seleção de atributos, e vice-versa. Ao potencializar as interações entre essas duas tarefas de seleção, acredita-se que o método de aprendizagem dual pode alcançar melhores resultados em ambos os lados. Além disso, os autores afirmam que o método escolhe instâncias para reconstruir todos os dados no espaço de recursos reduzido e preserva a estrutura global que é caracterizada pelo coeficiente de reconstrução esparso entre as instâncias selecionadas.

Orlinski e Jankowski (2020) propõem uma nova versão mais rápida dos algoritmos DROP com complexidade reduzida a $O(m \log m)$, enquanto a complexidade original é $O(m^3)$. Os novos algoritmos desenvolvidos usam florestas de *hashing* e outras estruturas de dados para manter a complexidade computacional o mais baixa possível. Esses algoritmos podem ser usados para grandes conjuntos de dados, com a classificação permanecendo inalterada, conforme comprovado por uma análise estatística em vários conjuntos de dados.

A maior parte das aplicações disponíveis atualmente costumam envolver um grande volume de dados e alta dimensionalidade, trazendo grandes desafio de armazenamento e custo computacional. Pensando nisso, a seleção de instâncias torna-se uma grande aliada para o tratamento desses dados, uma vez que reduzindo a quantidade de informação e encontrando instâncias que são mais representativas os resultados podem ser encontrados de forma mais rápida sem perder qualidade.

4 METODOLOGIA

Com o propósito de auxiliar na elaboração de projetos de pesquisa, Gil (2017) apresenta formas de classificar a pesquisa. De acordo com a autor, elas podem ser classificadas segundo a área de conhecimento, a finalidade, o nível de explicação e os métodos adotados. A classificação da metodologia adotada nessa pesquisa, seguindo as definições de Gil (2017), pode ser visualizada no Quadro 1.

Quadro 1: Quadro metodológico

Quanto à área de conhecimento	**Quanto à finalidade**	**Quanto aos propósitos**	**Quanto aos métodos empregados**
Ciências Exatas e da Terra	Aplicada	Descritiva	Levantamento

Fonte: Elaborado pelo autor

Quanto à área de conhecimento, essa pesquisa se enquadra na área das **Ciências Exatas e da Terra**, pois utiliza ferramentas de tecnologia da informação para seu desenvolvimento.

Quanto à finalidade da pesquisa ela é considerada **aplicada**, pois o conhecimento adquirido por meio dos seus resultados oferecem a oportunidade de desenvolver ferramentas mais eficientes e eficazes.

Quanto aos propósitos, essa pesquisa possui caráter **descritivo**, pois nesse tipo de pesquisa o principal objetivo é descrever as características de determinadas populações ou fenômenos (GIL, 2017). Nossa pesquisa busca identificar os fatores que contribuem para caracterização de perfis de usuário por meio de uma análise de dados obtidos de um questionário.

Quanto aos procedimentos técnicos utilizados, essa pesquisa se caracteriza como um **levantamento**, pois esse tipo de pesquisa tem como característica a interrogação de pessoas cujo comportamento será avaliado (GIL, 2017). Nossa pesquisa faz uma análise de dados de um questionário aplicado em uma comunidade específica, os usuários da rede social Facebook, buscando explorar a persuasão nesse meio.

Visando alcançar os objetivos inicialmente propostos, os procedimentos metodológicos adotados foram desenvolvidos observando as etapas do KDD (descoberta de conhecimento em base de dados, do inglês, *knowledge-discovery in databases*). Esse processo é definido por Fayyad et al. (1996) como "o processo, não trivial, de extração de informações implícitas, previamente desconhecidas e potencialmente úteis, a partir de dados armazenados em um banco de dados". Ou seja, é a extração de informações por meio de alguma técnica de busca ou inferência que traga consigo algum benefício para o sistema e também para o usuário.

O processo de KDD é dividido em etapas que são executadas de forma 1) interativa,

possibilitando que o usuário interfira e controle o curso das atividades e 2) iterativa, por ser uma sequência finita de operações onde o resultado de cada uma é dependente dos resultados das que a precedem. As etapas do KDD são as seguintes:

1. **Seleção**: Na primeira fase do processo são escolhidos os dados contendo todas as possíveis variáveis e registros que farão parte da análise, ou seja, os mais importantes para se alcançar o objetivo;

2. **Pré-processamento e limpeza**: Os dados são filtrados e os que estiverem errôneos, redundantes, ausentes ou inconsistentes são retirados ou transformados;

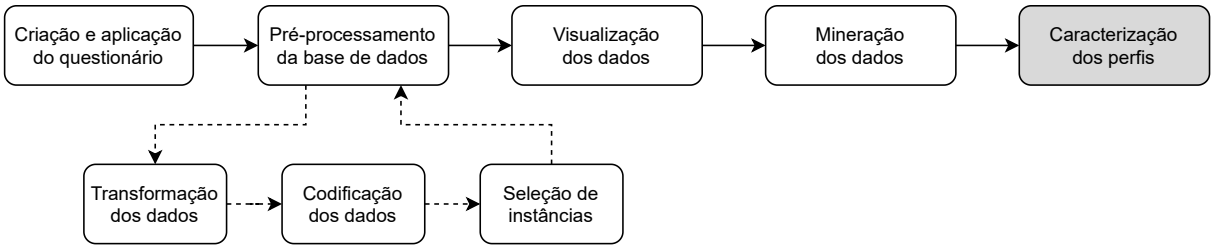
3. **Transformação**: Os dados são formatados e armazenados para que o *software* de mineração de dados consiga processá-los;

4. **Mineração de dados (*data mining*)**: É a etapa mais importante do KDD, nela são aplicados algoritmos de descoberta de padrões. As técnicas de *data mining* podem ser aplicadas a tarefas como associação, classificação, predição, segmentação e sumarização (FAYYAD et al., 1996);

5. **Interpretação e avaliação**: Nessa fase as informações são direcionadas para a realização de análise dos dados com o objetivo de verificar se possuem validade para o problema proposto.

Assim, seguindo o processo de KDD, a metodologia adotada para essa pesquisa é apresentada na Figura 7. As subseções seguintes descrevem o que será realizado em cada uma dessas etapas.

Figura 7 – Metodologia adotada



Fonte: Elaborado pelo autor

4.1 Métodos que não resolveram o problema

Os métodos utilizados para o desenvolvimento dessa pesquisa e que serão descritos nas seções seguintes foram aplicados após uma quantidade grande de experimentos, pois

ao nos depararmos com a dificuldade de classificação dos três perfis simultaneamente, fizemos várias tentativas para resolver o problema: utilizamos métodos de balanceamento, métodos caixa preta e também diferentes métodos de seleção de instância.

Como a quantidade de instâncias dos perfis não era balanceada, principalmente quando olhamos para os Produtores de conteúdo, inicialmente utilizamos estratégias de balanceamento para aproximar a quantidade de perfis e assim fizemos testes tanto com balanceamento *undersample* quanto *oversample* e ambos não trouxeram melhorias significativas nos resultados da classificação

Também fizemos alguns testes utilizamos métodos caixa preta, como rede neural, *machine learning* automatizado e florestas aleatórias, além de fazer várias combinações dos resultados desses algoritmos. Porém, a dificuldade em separar a classe Produtor de conteúdo das demais permanecia.

Submetemos a base de dados à alguns métodos de seleção de instâncias como os métodos RIS (CAVALCANTI; SOARES, 2020), que também não trouxeram melhorias quanto a classificação dos Produtores de conteúdo.

Assim, a cada teste sem sucesso percebíamos a complexidade da base de dados que tínhamos em mãos e que alcançar os objetivos propostos nesse trabalho não seria uma tarefa trivial.

4.2 Descrição do questionário

Uma das primeiras etapas desta pesquisa foi a aquisição do conjunto de dados. O questionário aplicado continha 16 questões para investigar como os usuários do Facebook respondiam a diferentes funções da rede social. Este questionário* esteve disponível *online* entre 22 de outubro de 2018 e 2 de julho de 2020 obtendo 1100 respostas.

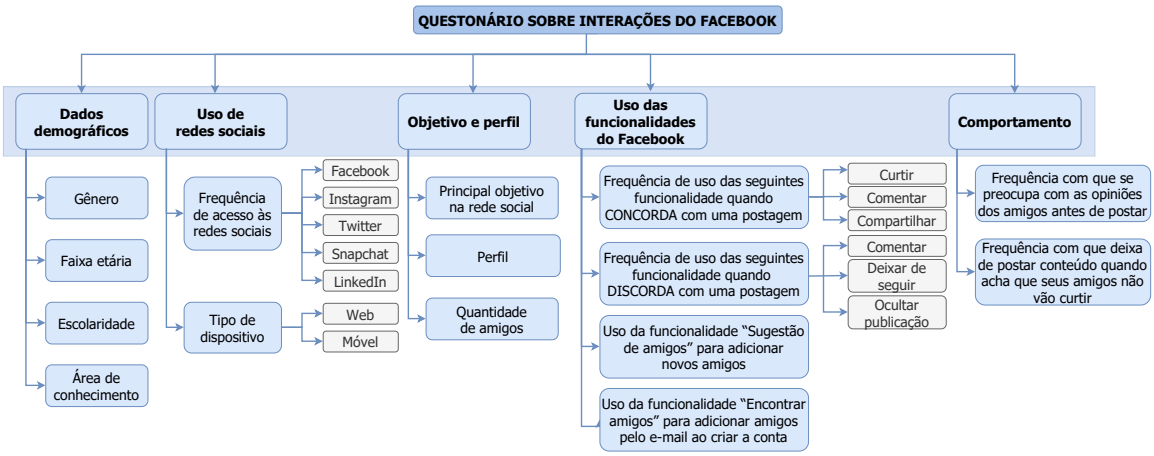
Para alcançar essa amostra, o questionário foi compartilhado via Facebook e e-mail. Foi feita uma rede de compartilhamento iniciando pelos contatos próximos dos autores, que também compartilharam com seus contatos mais próximos e assim por diante. Além disso, também divulgamos em comunidades do Reddit†.

Um diagrama com todas as perguntas feitas no questionário pode ser visto na Figura 8.

As primeiras perguntas do questionário referem-se às informações demográficas dos usuários. Essas informações podem ser usadas para determinar os perfis dos usuários e também para mostrar a diversidade dos entrevistados. Essas questões são descritas a

* O questionário pode ser acessado em: http://bit.ly/pesquisaface. † O Reddit é uma rede social com notícias do mundo inteiro em que usuários podem postar conteúdo na plataforma para que as pessoas votem positiva ou negativamente fazendo com que a informação seja mais ou menos relevante.

Figura 8 – Pesquisa sobre interações do Facebook.



seguir.

1. Sexo: 'Masculino', 'Feminino' ou 'Outro';

2. Faixa etária: 'menos que 18 anos', '18 a 23', '24 a 29', '30 a 35', '36 a 41', '42 a 47', '48 a 53', '54 a 59' ou '60 ou mais';

3. Educação: 'Ensino Fundamental Incompleto', 'Ensino Fundamental Completo', 'Ensino Médio Incompleto', 'Ensino Médio Completo', 'Ensino Superior Incompleto', 'Ensino Superior Completo', 'Pós Graduação Incompleta' ou 'Pós Graduação Completa';

4. Área de estudo: 'Ciências Exatas e da Terra', 'Engenharias', 'Ciências Biológicas', 'Ciências da Saúde', 'Ciências Sociais Aplicadas', 'Ciências Humanas', 'Linguística, Artes e Letras' ou 'Outros';

As perguntas a seguir referem-se à frequência com que os usuários usam as principais redes sociais:

5. Frequência com que os usuários acessam as redes sociais Facebook, Instagram, Twitter, Snapchat e Linkedin.

   Para essas perguntas, os usuários podem escolher entre as seguintes opções de resposta para cada rede social: 'Não inscrito', 'Todos os dias', 'Até 1 hora por dia', 'Entre 1 e 3 horas por dia', '3 a 6 horas por dia' e 'Mais de 6 horas por dia'.

A próxima pergunta era sobre a forma preferida do usuário de acessar o Facebook.

6. Qual é a sua forma preferida de acessar o Facebook: dispositivos móveis (celular, tablet, etc.) ou web (Chrome, Firefox, Internet Explorer, etc.).

A pergunta a seguir era sobre os três principais objetivo do usuário ao acessar a rede social. Para essa pergunta ele deveria escolher as três melhores opções a partir de uma lista predefinida.

7. Qual é o seu objetivo no Facebook?

   A lista de objetivos fornecida foi: 1) Acompanhar o dia-a-dia de meus amigos e familiares; 2) Ficar por dentro das novidades; 3) Compartilhar minhas fotos e vídeos; 4) Comunicar-me com amigos e familiares; 5) Conhecer diferentes opiniões sobre diversos assuntos; 6) Conhecer novas pessoas; 7) Encontrar e participar de promoções; 8) Compartilhar as informações que considero relevantes; 9) Ler sobre assuntos de meu interesse; 10) Acessar os jogos disponíveis na rede social; 11) Compartilhar minha opinião sobre diversos temas, e 12) Produzir e compartilhar conteúdo próprio na rede social. Os usuários também podem marcar a resposta 13) Outros e especificá-los*.

A pergunta seguinte foi feita para determinar com qual dos três perfis (*Espectador*, *Participante* e *Produtor de conteúdo*) o usuário se identificou. Foi fornecida uma breve descrição dos três perfis de acordo com o que foi definido por RUAS DA SILVEIRA et al. (2019), podendo o usário escolher apenas um deles.

8. Com qual perfil você se classifica no Facebook? As opções apresentadas foram: 1) *Espectador* (tem como comportamento principalmente ver o que está publicado na rede social); 2) *Participante* (curtir, compartilhar e comentar o conteúdo da rede social); e 3) *Produtor de conteúdo* (posta conteúdo próprio na rede social). Esta questão foi usada como atributo de classe.

   A próxima pergunta é sobre o número de conexões que o usuário tem no Facebook.

9. Quantos amigos você tem no Facebook? As opções de resposta foram as seguintes: '0 a 99', '100 a 199', '200 a 299', '300 a 499', '500 a 699', '700 a 999', '1.000 a 1499' ou '1.500 ou mais'.

As perguntas a seguir procuraram determinar com que frequência o usuário utiliza alguns recursos ou executou ações específicas no Facebook. Para todos eles, o usuário tinha os seguintes graus de concordância: 'Nunca', 'Raramente', 'Às vezes', 'Quase sempre' e 'Sempre'.

10. Quando você encontra conteúdo no Facebook com o qual CONCORDA, com que frequência você usa os seguintes recursos: 'curtir', 'compartilhar' e 'comentar'.

* Alguns objetivos mencionados como "Outros" incluíram: ser informado sobre eventos, oportunidades de emprego e concursos; compartilhar meu trabalho e entretenimento.

11. Quando você encontra conteúdo no Facebook com o qual você DISCORDA, com que frequência você usa os seguintes recursos: ‘comentar’, ‘ocultar postagem’ e ‘deixar de seguir’.

A próxima pergunta é sobre o uso do recurso “Sugestão de amigos”.

12. Com que frequência você adiciona novos amigos sugeridos pelo Facebook?

As perguntas a seguir perguntam com que frequência o usuário avalia se seus amigos vão gostar de sua postagem antes de fazê-lo e com que frequência eles param de postar conteúdo porque acreditam que seus amigos não vão gostar.

13. Antes de postar algo em seu mural, você se preocupa com o que seus amigos vão pensar?

14. Você deixa de postar um conteúdo quando acha que seus amigos não vão curtir?

Por fim, a última pergunta feita aos usuários é sobre o uso da funcionalidade “Encontrar amigos”.

15. Quando você criou seu perfil do Facebook, você usou o recurso “Encontrar amigos” para encontrar amigos em seus contatos de e-mail? As opções de resposta foram as seguintes: ‘Sim’, ‘Não’ e ‘Não me lembro’.

Desta forma, podemos descrever formalmente o conjunto de atributos considerados para prever o comportamento do usuário, conforme apresentado na Equação 4.1*.

$$Atributos = \{D_4, U_6, O_3, F_8, B_2\} \tag{4.1}$$

onde D = corresponde às informações demográficas dos usuários (questões 1 a 4); U = corresponde às informações sobre o uso das redes sociais pelos usuários (questões 5† e 6); O = corresponde aos objetivos e perfil dos usuários (questões 7 a 9); F = corresponde ao uso de recursos do Facebook (questões 10‡, 11§, 12 e 15), e B = corresponde ao comportamento do usuário do Facebook diante da opinião de amigos (questões 13 e 14).

---

* Os números subscritos correspondem ao número de atributos em cada categoria. † A questão 5 foi feita para cada uma das redes sociais presentes no questionário (Facebook, Instagram, Twitter, Snapchat e Linkedin), portanto, foi transformada em cinco atributos no banco de dados. ‡ A questão 10 foi feita para cada uma das três principais funcionalidades que podem ser usadas ao concordar com o conteúdo (‘curtir’, ‘compartilhar’ e ‘comentar’), por isso é transformada em três atributos no banco de dados. § A questão 11 foi feita para cada uma das três principais funcionalidades que podem ser utilizadas ao discordar do conteúdo (‘comentar’, ‘ocultar postagem’ e ‘deixar de seguir’), então ela foi transformada em três atributos no banco de dados.

4.3 Pré-processamento da base de dados

O conjunto de dados contendo as respostas dos usuários ao questionário foi pré-processado antes da aplicação das técnicas de mineração de dados. Dentre as etapas do pré-processamento se aplicaram a este trabalho a transformação e codificação dos dados e a seleção de instâncias.

4.3.1 Transformação e codificação dos dados

Devido a baixa quantidade de respostas obtidas em algumas opções foi necessário a junção de determinadas categorias para melhorar a qualidade dos dados. Com isso a faixa etária que antes possuía nove categorias passou a ter três: 23 ou menos, 24 a 35 anos e 36 ou mais. O mesmo foi feito com a escolaridade que antes possuía seis categorias e passou a ter três: Fundamental a Médio, Superior e Pós Graduação.

As nove áreas de conhecimento* (incluindo a opção 'Outros') foram unidas da seguinte forma:

- **Grupo A:** Ciências Exatas e da Terra e Engenharias;
- **Grupo B:** Ciências Biológicas e Ciências da Saúde;
- **Grupo C:** Ciências Sociais Aplicadas e Ciências Humanas;
- **Grupo D:** Outros.

A quantidade de amigos que era divida em oito faixas passou a definida em apenas três, são elas: '0 a 299', '300 a 999' e '1000 ou mais'.

O questionário perguntava ao usuário, em ordem de relevância, quais seriam seus três principais objetivos ao acessar a rede social, para esta pesquisa iremos utilizar apenas o primeiro objetivo, o principal deles.

Alguns algoritmos trabalham apenas com dados numéricos, como é o caso dos métodos utilizados para selecionar as instâncias, portanto foi necessário adaptar todos os dados nominais e para isso foi aplicada uma conversão simbólico-numérico. A aplicação dessa conversão depende se o atributo nominal é ordinal ou não. No caso de atributos nominais ordinais a codificação deve preservar essa relação. Para isso, basta ordenar os valores, de acordo com sua relevância, e codificar cada valor de acordo com sua posição. Assim, a distância entre os valores varia de acordo com a proximidade entre eles.

* É importante ressaltar que a área Ciências Agrárias não foi mencionada em nenhuma das respostas, portanto não é considerada neste estudo.

Neste trabalho, atributos como idade, escolaridade e frequência de uso foram convertidos seguindo essa técnica.

Caso não haja uma relação de ordem entre os valores do atributo, a inexistência dessa relação deve ser mantida para os valores numéricos gerados, ou seja, a diferença entre quaisquer valores deve ser a mesma. Neste caso, uma maneira de fazer a conversão é codificar cada valor nominal por uma sequência de n bits, onde n é a quantidade de categorias possíveis para o atributo. Neste trabalho, atributos como objetivo, gênero e área de conhecimento foram convertidos seguindo essa técnica.

4.3.2 Algoritmo genético para seleção de instâncias

Para identificar os perfis dos usuários, inicialmente consideramos todas as instâncias do conjunto de dados ($N = 1100$, com 426 Participantes, 623 Espectadores e 51 Produtores de conteúdo). Porém identificamos que os resultados obtidos não foram satisfatórios, principalmente para o perfil Produtor de conteúdo. Na maioria dos casos, o algoritmo de árvore de decisão classificou todas as instâncias desse perfil incorretamente, o que nos levou a fazer uma análise mais detalhada do conjunto de dados e aplicar métodos de seleção de instância para identificação das instâncias mais representativas de cada classe.

Portanto, neste trabalho, a seleção de instâncias tem como foco melhorar a qualidade dos dados de treinamento, identificando e eliminando instâncias com rótulos incorretos, ruídos e *outliers*, antes de aplicar o algoritmo de aprendizagem, aumentando assim a precisão da classificação. De acordo com Brodley e Friedl (1999), o erro de rotulagem pode ocorrer por diversos motivos, inclusive subjetividade. Acreditamos que tenha sido esse o caso para este trabalho.

Adotamos uma estratégia de seleção de instâncias usando um algoritmo genético multiobjetivo no qual a implementação foi feita em Python da seguinte forma, utilizando a biblioteca DEAP (FORTIN et al., 2012):

1. Cada indivíduo é representado como uma *string* de bits que indica qual instância está selecionada ou não: um valor igual a 0 indica que a instância correspondente no conjunto de dados não foi selecionada. Um valor igual a 1 indica uma instância selecionada. Por exemplo, uma sequência igual a 001101001 indica que as instâncias 3, 4, 6 e 9 foram selecionadas, e as instâncias 1, 2, 5, 7 e 8 não foram selecionadas em uma base fictícia de apenas 9 instâncias. Inicialmente, é gerada uma população de indivíduos na qual os bits são atribuídos aleatoriamente.

2. As duas funções de *fitness* (“objetivos”) são implementadas da mesma maneira:

para cada indivíduo, dois conjuntos de dados são gerados: um com as instâncias selecionadas (bits iguais a 1, a "base selecionada") e o outro com os não selecionados (bits iguais a 0, a "base complementar").

3. O primeiro objetivo é maximizar a medida F (F1) do classificador do Vizinho Mais Próximo (k-NN, *k-Nearest Neighbor*), com k variando de 1 a 15 por meio de validação cruzada de 10 vezes com a base selecionada, e o segundo objetivo é calculado da mesma forma, mas em uma base complementar.

4. O cruzamento e a mutação dos indivíduos são feitos normalmente nos algoritmos genéticos. A evolução da população é feita até que a aptidão do melhor indivíduo pare de melhorar ou os dez melhores indivíduos atinjam a aptidão igual a 1.

O método de seleção de instâncias adotado neste trabalho pode ser classificado como: (1) *Misto*, uma vez que começa com um subconjunto pré-selecionado aleatoriamente e, durante sua execução, adiciona e remove instâncias de acordo com critérios pré-definidos; (2) *Híbrido*, pois visa aumentar a precisão e a taxa de redução; e (3) *Wrapper*, pois utiliza o classificador do Vizinho Mais Próximo (k-NN) para avaliar a qualidade dos subconjuntos gerados durante o processo de seleção.

4.4 Algoritmo de classificação

Na próxima fase do processo de KDD, submetemos o conjunto de dados pré-processado ao algoritmo de aprendizagem. Como o objetivo é identificar as características dos perfil optamos por utilizar algoritmos de *machine learning* caixa branca que apresenta modelos em que é possível interpretar e explicar os resultados obtidos. Caso fosse utilizado um algoritmo caixa preta, como uma rede neural, por exemplo, talvez tivéssemos um modelo com melhor precisão, porém a explicabilidade desse modelo não seria trivial, uma vez que não é possível compreender a função modelada pelo treinamento desse algoritmo.

Assim, optamos por utilizar árvores de decisão, pois são algoritmos que possuem como resultado um modelo em que podemos seguir a lógica aplicada e extrair os atributos que levaram uma determinada instância ser atribuída a uma classe ou a outra, revelando assim as características de cada perfil. Para gerar as árvores de decisão neste trabalho são utilizados o C4.5 (QUINLAN, 1993) e o CART (BREIMAN et al., 1984).

Como nosso objetivo é caracterizar os perfis de usuário do Facebook, optamos por utilizar dois algoritmos de árvore de decisão para comparar seus resultados, avaliando se as regras geradas são semelhantes e assim ter mais confiança na definição das características de cada perfil.

A versão do C4.5 utilizada é o VTJ48*, uma implementação de código aberto Java disponível na ferramenta Weka† que ajusta automaticamente os parâmetros do C4.5 para construir uma árvore de decisão menor e fácil de visualizar. Os parâmetros ajustados pelo VTJ48 são os seguintes: C - Limite de confiança para poda da árvore e M - Número mínimo de instâncias por folha da árvore. Neste trabalho os parâmetros foram ajustados para: $M = 2$ e $C = 0{,}046875$. Para executar o algoritmo CART (BREIMAN et al., 1984), utilizamos a implementação em R‡, denominado RPART§, com os parâmetros *default*.

4.4.1 Métricas de avaliação do classificador

As métricas utilizadas para avaliar a qualidade dos modelos de classificação gerados foram as seguintes (FACELI et al., 2011):
Acurácia: porcentagem de instâncias preditas corretamente pelo classificador.

$$A = \frac{VP + VN}{VP + VN + FP + FN} \tag{4.2}$$

Precisão: porcentagem de instâncias previstas como uma classe, que pertencem a essa classe.

$$P = \frac{VP}{VP + FP} \tag{4.3}$$

Sensibilidade: porcentagem de instâncias de uma classe que foram preditas corretamente.

$$S = \frac{VP}{VP + FN} \tag{4.4}$$

F-Measure: Média harmônica entre precisão e sensibilidade.

$$FM = \frac{2 \times P \times S}{P + S} \tag{4.5}$$

onde, para a classe considerada, Verdadeiro Positivo (VP) é o número de instâncias corretamente rotuladas como parte da classe; Verdadeiro Negativo (VN) é o número de instâncias corretamente rotuladas como não pertencentes à classe; Falso Positivo (FP) é o número de instâncias rotuladas como da classe, mas que não pertencem a ela; e Falso Negativo (FN) é o número de instâncias que pertenciam à classe, mas foram rotuladas

* *Visually Tuned* J48 é um pacote de código aberto para descoberta rápida de conhecimento usando árvores de decisão em ambiente WEKA. Disponível para download em http://www.ri.fzv.um.si/vtj48.
† *Waikato Environment for Knowledge Analysis* é um programa gratuito e de código aberto dedicado ao estudo de inteligência artificial e aprendizado de máquina. Disponível para download em http://www.cs.waikato.ac.nz/ml/weka. ‡ O R é uma linguagem de programação e ambiente de *software* livre para computação estatística e gráfica. Informações sobre a linguagem R estão disponíveis em https://www.r-project.org. § Particionamento Recursivo e Árvores de Regressão, do inglês, *Recursive Partitioning And Regression Trees*

incorretamente como outra classe. Ambos os modelos foram gerados usando um processo de validação cruzada k-fold, com $k = 10$ (KOHAVI, 1995).

5 RESULTADOS E DISCUSSÕES

Este capítulo apresenta os resultados deste trabalho, sendo: 1) uma análise por meio da visualização da base de dados; 2) a caracterização dos perfis de usuário do Facebook utilizando algoritmo de árvore de decisão; 3) uma descrição qualitativa dos Produtores de conteúdo; e 4) uma análise das instâncias desconsideradas pelo algoritmo genético.

5.1 Visualização da base de dados e seleção de instâncias

Esta seção apresenta os resultados inferidos a partir da análise do gráfico gerado com as informações dos usuários do Facebook. A utilização de uma ferramenta de visualização multivariada foi fundamental na busca de padrões qualitativos e tendências no conjunto de dados, pois ela fornece uma visão geral do problema de classificação que está sendo estudado, separando instâncias de diferentes classes da melhor maneira possível.

A Figura 9 mostra a distribuição das classes de forma otimizada utilizando o FreeViz, expondo os atributos que mais influenciam cada classe. Por meio dessa visualização não foi possível separar a classe Produtor de conteúdo das demais, mesmo após a seleção de instâncias, pois suas instâncias estão amplamente difundidas entre as demais classes.

A Figura 9(a) foi desenvolvida usando o conjunto de dados original completo com 1100 instâncias. Nela é possível observar apenas duas regiões, o que indica que o conjunto de dados possui duas classes bem definidas (Participante e Espectador), embora ainda vejamos algumas inconsistências nessas classes. O Produtor de conteúdo não está bem caracterizado, pois não possui uma área específica para suas instâncias, indicando que não há diferença significativa dele para as demais classes. Essa visualização revela ainda descobertas importantes em relação aos perfis de usuário do Facebook:

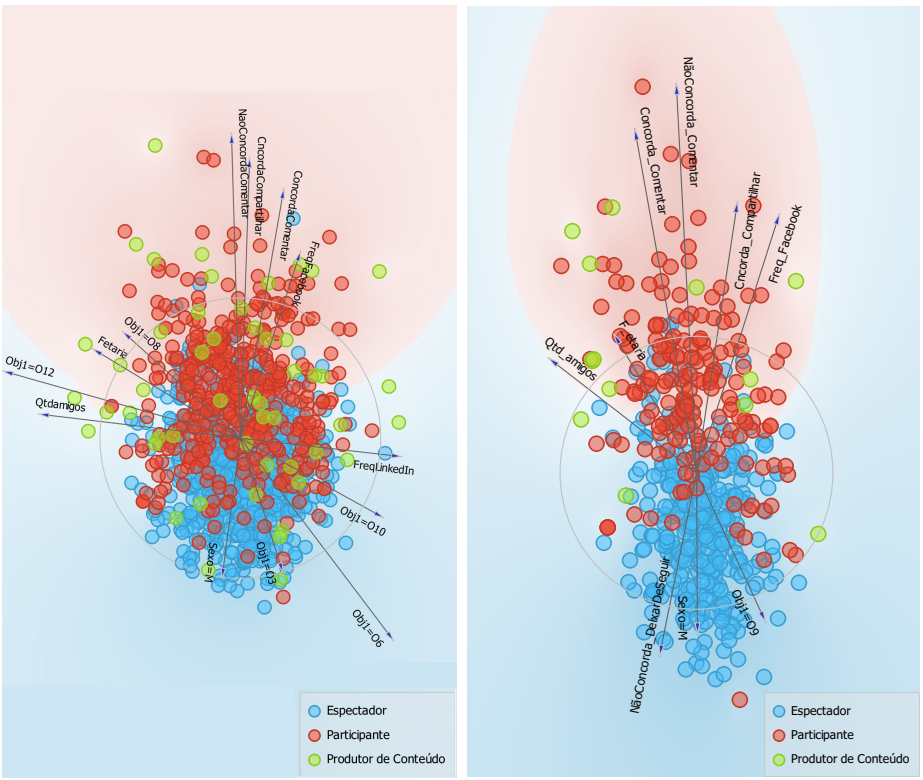
- Utilizar a funcionalidade de comentar quando o usuário concorda ou discorda está fortemente relacionado à frequência de uso do Facebook e também ao uso da funcionalidade compartilhar, e todas essas características pertencem aos Participantes;
- Ser do sexo masculino e ter o Objetivo 3 (Compartilhar minhas fotos e vídeos) ou Objetivo 6 (Conhecer pessoas novas) ao usar o Facebook parecem correlacionados e são características dos Espectadores;
- O filtro aplicado pelo raio ocultou a exibição de algumas propriedades, indicando que não são muito informativas nessa base, dentre elas: escolaridade, área de conhecimento, frequência de uso de outras redes sociais (Instagram, Twitter e Snapchat), tipo de acesso ao utilizar a rede social, dentre outros.

Apesar de apresentarem duas regiões diferentes, essas regiões não estão bem definidas e, portanto, não foi possível organizar as classes de uma maneira que deixe os perfis mais segregados na visualização. Na região central as classes são muito mistas e nas regiões periféricas ainda há muitos indivíduos de ambas as classes. Portanto, analisando essas instâncias, podemos concluir que há ruídos nessa base que podem dificultar a caracterização dos perfis.

Em relação aos Produtores de conteúdo, essa figura mostra que eles possuem características dos perfis Participante e Espectador, que também foi confirmado pelos algoritmos de árvore de decisão que apresentaram um desempenho de 0% para essa classe quando treinada com as três classes (dados não mostrados), das 51 instâncias dos Produtores de conteúdo, 24 foram classificadas como Participantes e 27 como Espectadores.

A partir dessa análise, percebemos que talvez as instâncias desse conjunto de dados tivessem uma classificação multi-rótulo, ou seja, o usuário pode ter características de mais de um perfil e, portanto, ter dificuldades em se classificar em apenas um. Essa categoria de dificuldade traz subjetividade à resposta do usuário, demandando a seleção das instâncias representativos de cada classe.

Figura 9 – Visualização das classes: Em a), temos o conjunto de dados inicial com 1100 instâncias. Em b), temos o conjunto de dados após selecionar as instâncias utilizando o AG.



(a) Original: 1100 instâncias (b) Selecionada: 526 instâncias

Foi feita então a seleção de instâncias utilizando o algoritmo genético descrito na

Seção 4.3.2, e a quantidade de instâncias resultantes na base de dados é de 526 ($N = 526$, com 329 Espectadores, 184 Participantes e 13 Produtores de conteúdo). Assim, 574 instâncias foram consideradas ruídos, eliminadas da base de dados e a Seção 5.5 fornece uma análise dessas instâncias ($N = 574$, 294 Espectadores, 242 Participantes e 38 Produtores de conteúdo).

A Figura 9(b) foi desenvolvida utilizando o conjunto de dados após a seleção de instâncias e mostra as classes Participante e Espectador mais segregadas em comparação com a Figura 9(a), não há mais uma grande concentração de instâncias na região central e o ruído nas áreas periféricas diminuiu consideravelmente. Os Produtores de conteúdo ainda estão distribuídos entre as duas classes, indicando que sua caracterização realmente é mais complexa. Portanto, por apresentar resultados ruins durante os testes iniciais e por não possuir uma região bem definida, esta classe foi retirada do conjunto de dados para aplicação dos algoritmos de aprendizado de máquina.

Pela Figura 9(b) também é possível perceber que há uma proximidade maior entre os *Participantes* e os *Produtores de conteúdo*. Enquanto os *Espectadores* estão organizados em uma região mais isolada da imagem, *Produtores de conteúdo* e *Participantes* dividem a região oposta. Em alguns momentos é possível perceber que se misturam e as características de ambos são muito próximas, indicando que esses dois perfis se comportam de forma similar e fazer uma distinção entre eles pode se tornar uma tarefa muito complexa.

Nessa visualização temos uma quantidade reduzida de dados, mas ela também revela descobertas importantes e podemos extrair algumas informações em relação aos perfis de usuário:

- As funcionalidades de comentar, compartilhar e a frequência de uso do Facebook continuam possuindo uma forte relação e são características dos Participantes;
- Ser do sexo masculino continua sendo ligado aos Espectadores, porém agora notamos que o objetivo dessa classe ao utilizar a rede social seria o Objetivo 9 (Ler sobre assuntos do meu interesse), que também esta fortemente relacionado com a funcionalidade de deixar de seguir seu amigo ao não concordar com sua publicação;
- O filtro aplicado pelo raio também ocultou a exibição de algumas propriedades, indicando que não são muito informativas nessa base, dentre elas: escolaridade, frequência de uso de outras redes sociais (Instagram, Twitter, Linkedin e Snapchat), tipo de acesso ao utilizar a rede social, dentre outros.

Em ambos os gráficos podemos notar que os vetores relacionados à funcionalidade 'Comentar' se sobressaem em relação aos demais, indicando que essa funcionalidade possui grande influência na classificação dessa base de dados. Avaliando o tamanho e a posição desses vetores podemos notar que o uso das funcionalidades 'Comentar' e 'Compartilhar'

são características muito influentes na classe Participante, pois são os maiores vetores e também seu posicionamento é perpendicular à classe oposta. Já para os Espectadores, os principais vetores apresentam diferenças entre as duas imagens. Na Figura 9(a) o Objetivo 6 (Conhecer pessoas novas) é uma característica muito importante para sua classificação, já na Figura 9(b) o Objetivo 9 (Ler sobre assuntos do meu interesse) juntamente com a funcionalidade 'Deixar de seguir' possuem uma relevância maior.

Vale lembrar que nem todas as hipóteses levantadas decorrentes da interpretação da visualização do FreeViz são necessariamente exatas, pois há a limitação do uso de uma projeção de baixa dimensão, porém é possível sugerir alguns padrões de comportamento e relações potenciais entre as propriedades que ajudam a formular hipóteses para a caracterização desses perfis.

5.2 Avaliação dos Produtores de conteúdo

Ao executar alguns testes iniciais na base de dados notamos a dificuldade que os algoritmos de aprendizado* apresentavam ao classificar os Produtores de conteúdo. Na maioria dos casos todas as instâncias dessa classe eram classificadas de forma incorreta, reduzindo consideravelmente a precisão do modelo. Foram feitos então vários ajustes na intenção de melhorar as métricas para os Produtores de conteúdo, porém ao fazer isso notamos que o aumento da precisão para essa classe ocasionava em uma diminuição para as demais. Portanto, optamos por fazer uma análise separadamente e mais especifica dos Produtores de conteúdo, uma vez que a visualização já indicava que a separação dessa classe das demais seria uma tarefa complexa.

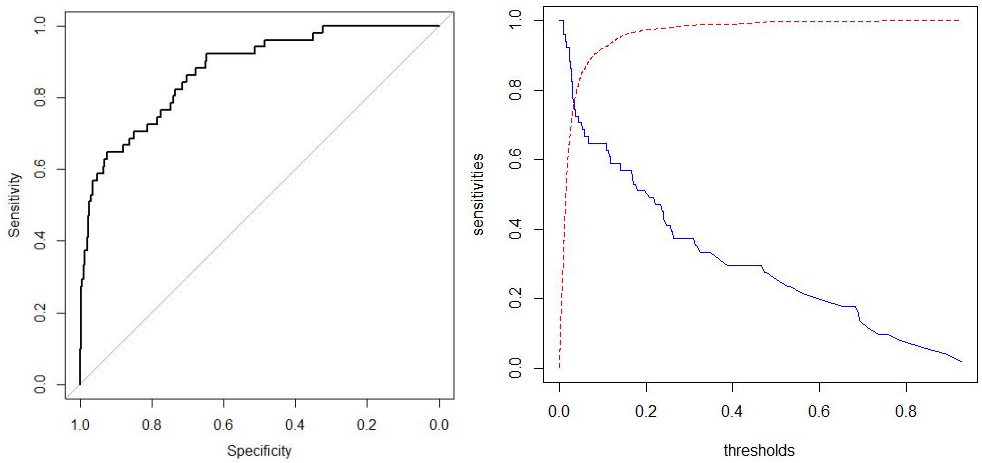
Foi feita uma análise a partir de um modelo logístico binário entre a classe Produtor de conteúdo e uma segunda classe em que unimos as demais categorias majoritárias e chamamos de Não-produtores (Participante e Espectador). Para avaliar os resultados do modelo logístico utilizamos a Curva ROC (*"Receiver Operating Characteristic")* e a AUC (*"Area Under the ROC Curve"*). A Curva ROC é um gráfico simples que permite estudar a variação da sensibilidade e especificidade para diferentes limiares de classificação na probabilidade estimada (*thresholds*), ou seja, ela nos mostra o quão bom o modelo criado pode ser para distinguir entre duas classes. A AUC simplifica ainda mais a análise da Curva ROC e é um dos índices de precisão mais utilizados para avaliar a qualidade da curva. Seu resultado é um valor único calculado a partir da "área sob a curva" que agrega todos os limiares da ROC e, quanto maior o valor da AUC, melhor o modelo.

O resultado do modelo logístico pode ser avaliado por meio da Figura 10. A Figura

* Em nossos testes iniciais utilizamos árvore de decisão e rede neural para avaliar a qualidade do conjunto de dados. Os resultados obtidos para ambos algoritmos não foram bons para o perfil Produtor de conteúdo, indicando a necessidade do pré-processamento dessa base.

10(a) mostra a Curva ROC e sua AUC ficou em 87,58%, indicando que o modelo gerado é considerável, o que não implica que ele consegue discriminar exatamente a classe, pois ainda há categorias específicas que desfavorecem a ocorrência dessa classe pequena. O gráfico da Figura 10(b) nos mostra que, apesar de ter uma sobreposição dos dados, é possível identificar as instâncias da classe minoritária, porém classificá-las corretamente implica em classificar incorretamente as instâncias das demais classes.

Figura 10 – Avaliação do modelo logístico



(a) (b)

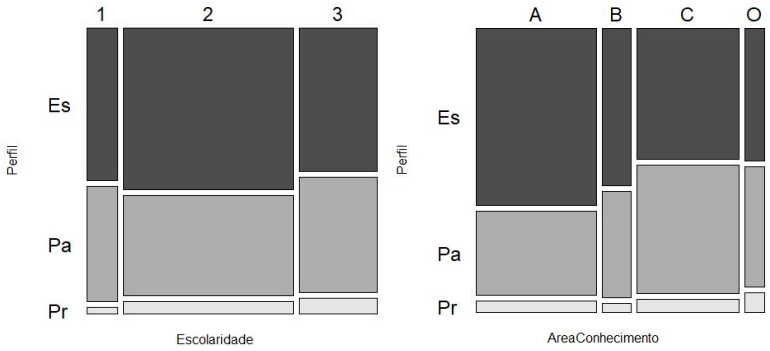
Aplicando o modelo logístico obtemos uma taxa de acerto de 76,47% para os Produtores de conteúdos, mas errando cerca de 23% das demais classes, um valor considerado alto já que essas classes possuem um número maior de instâncias. Assim podemos concluir que há algumas categorias que favorecem a ocorrência da classe minoritária, mas não é possível separar as três classes.

A partir do modelo logístico gerado podemos destacar os principais atributos dessa base de dados, aqueles que possuem significância de 95%. Para isso avaliamos o *p-value* de cada atributo e utilizamos apenas os que possuem valor menor ou igual à $0,05$* e os demais não serão considerados. As Figuras 11, 12, 13 e 14 apresentam gráficos de mosaico para cada um desses atributos e mostram a proporção de cada perfil de acordo com o atributo escolhido. Nas figuras, no eixo y estão representados os três perfis de usuário (Es = Espectador; Pa = Participante; e Pr = Produtor de conteúdo), que também estão separados em três tons de cinza para melhor visualização. No eixo x temos as opções de resposta para o atributo em questão. No mosaico, cada campo retangular indica a proporção de instâncias do cada perfil em relação à uma opção de resposta. A partir desses gráficos também podemos inferir algumas informações sobre o perfil Produtor de conteúdo.

Pela Figura 11(a) podemos notar que, em proporção, usuários que são Produtores

* Um nível de significância de $0,05$ indica um risco de 5% de se concluir que existe uma associação quando não existe uma associação real.

Figura 11 – Proporção dos atributos demográficos de cada Perfil



(a) Escolaridade (b) Área de conhecimento

de conteúdo apresentam uma escolaridade mais elevada, pois grande parte das instâncias desse perfil estão nas categorias 2 e 3, ou seja, possuem ensino superior ou pós-graduação, respectivamente.

Em relação à área de conhecimento mostrada na Figura 11(b), os Produtores de conteúdo possuem uma proporção maior para categoria 'Outros'. Ao avaliar o que foi informado nessa categoria temos alguns usuários com nível de escolaridade inferior à graduação que não responderam essa pergunta e outros usuários que transitam em diferentes áreas e, portanto, não se classificaram em apenas uma.

A Figura 12 mostra a proporção de uso de outras redes sociais. A partir dessa figura notamos que a proporção de Produtores de conteúdo que utilizam as demais redes sociais com muita frequência é elevado quando comparado com os demais perfis.

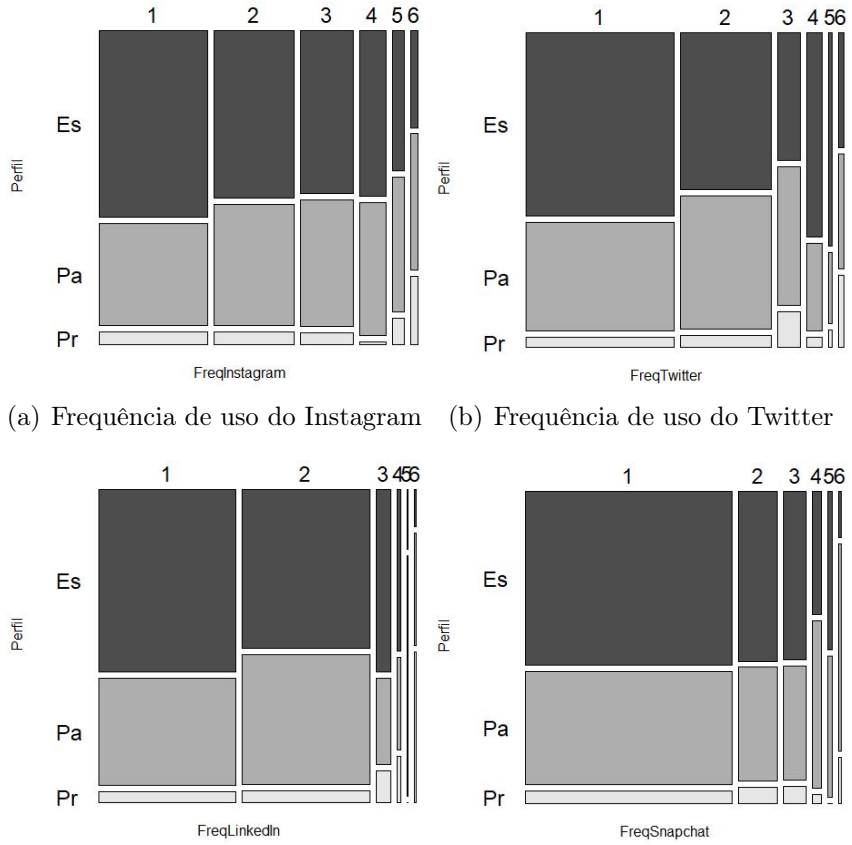
Quanto aos objetivos ao utilizar a rede social, a Figura 13 mostra que para Produtores de conteúdo, o objetivo que apresenta maior proporção é o Objetivo 6 (Conhecer pessoas novas).

Sobre a utilização de ferramentas que demonstram seu descontentamento com o algum conteúdo publicado, a Figura 14 mostra que os Produtores de conteúdo apresentam grandes proporções para o uso das funcionalidades Comentar e Deixar de seguir.

Essa análise nos mostra que, apesar da dificuldade em separar esse perfil dos demais, eles apresentam algumas características que são relevantes, ou seja, nesse conjunto de dados existem atributos que possuem uma proporção maior ou menor para o perfil Produtor de conteúdo, porém fazer essa diferenciação entre as três classes não é uma tarefa trivial, por isso optamos por remover a classe de produtores da base de dados.

A Figura 15 mostra a retirada dos Produtores de conteúdo da base de dados, resultando em um conjunto com 513 instâncias ($N = 513$, com 329 Espectadores e 184 Participante). Nessa figura as classes são apresentadas com melhor distribuição, exibindo duas regiões mais definidas. As características principais de ambas as classes permanecem semelhantes a da Figura 9(b); para os Participantes, o uso dos recursos de comentário e compartilhamento ainda é proeminente e para os Espectadores, o uso do recurso de parar

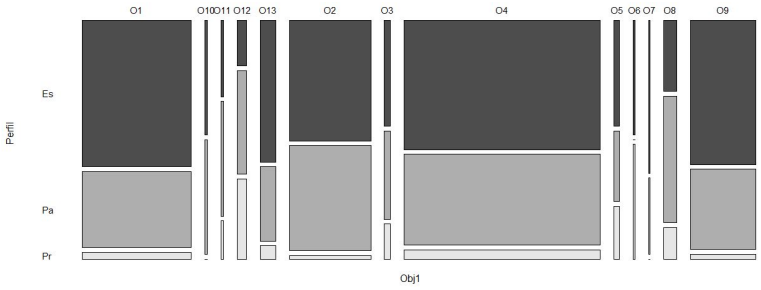
Figura 12 – Proporção dos atributos de frequência de uso de redes sociais de cada Perfil



(a) Frequência de uso do Instagram (b) Frequência de uso do Twitter

(c) Frequência de uso do Linkedin (d) Frequência de uso do Snapchat

Figura 13 – Proporção dos objetivos de cada Perfil

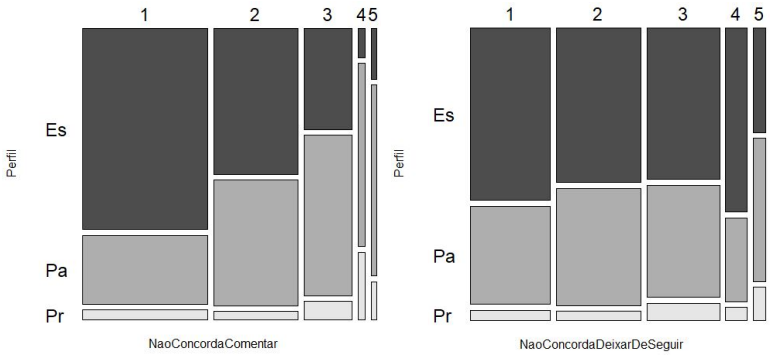


de seguir continua prevalecendo.

5.3 Descrição da base de dados após a seleção de instâncias

Para a caracterização dos perfis de interação dos usuários Participantes e Espectadores, temos um conjunto de dados com 513 usuários ($N = 513$, com 329 Espectadores e 184 Participantes). Dividimos o conjunto de dados da seguinte forma: reservamos 10%

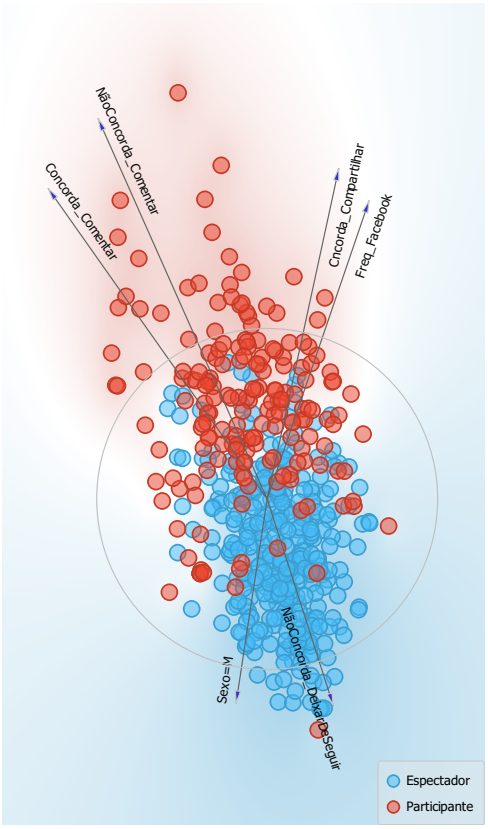
Figura 14 – Proporção de uso de funcionalidades ao discordar de um conteúdo de cada Perfil



(a) Comentar quando não concorda

(b) Deixa de seguir quando não concorda

Figura 15 – Base de dados final.



de cada perfil para teste (Teste 1 da Tabela 1) e o restante para criação dos modelos de aprendizagem. Além disso, testamos os modelos com as instâncias consideradas ruído pelo AG (Teste 2 da Tabela 1).

A partir da base de dados selecionada (com 513 instâncias), temos uma descrição das informações presentes em cada atributo. A Tabela 2 mostra os dados demográficos para entender a composição do conjunto de dados usado nesta pesquisa.

A Tabela 3 mostra a frequência de acesso a algumas das principais redes sociais:

Tabela 1 – Separação do conjunto de dados em treinamento e teste.

Classes	**Treinamento**	**Teste 1**	**Teste 2**	**Total**
Participante	166	18	242	426
Espectator	297	32	294	623
Total	463	50	536	1049

Tabela 2 – Questões Demográficas.

QN*	**Questão**	**N**	**%**
1	**Gênero**		
	Masculino	280	54,6
	Feminino	232	45,2
	Outro	1	0,2
2	**Faixa etária**		
	29 ou menos	344	67,1
	30 a 47 anos	129	25,1
	48 ou mais	40	7,8
3	**Escolaridade**		
	Fundamental ou médio	53	10,3
	Graduação	340	66,3
	Pós-graduação	120	23,4
4	**Área de conhecimento**		
	Ciências Exatas e da Terra - Engenharias	242	47,2
	Ciências Biológicas - Ciências da Saúde	58	11,3
	Ciências Sociais Aplicadas - Ciências Humanas - Linguística, Letras e Artes	185	36,1
	Outra	28	5,4

*QN: Número da pergunta no questionário

Facebook, *Instagram*, *Twitter*, *Linkedin* e *Snapchat*. Pelos dados, é possível perceber que o Facebook é o que apresenta maior frequência de acesso quando comparado às demais redes. Geralmente as principais redes sociais possuem um grande número de contas ativas ou um forte envolvimento do usuário, estão disponíveis em vários idiomas e permitem que os usuários se conectem com amigos ou pessoas além de fronteiras geográficas, políticas ou econômicas. O uso de redes sociais é muito diversificado, há plataformas focadas na comunicação entre amigos e familiares que estimulam a interação por meio de compartilhamento de fotos ou *status* e jogos sociais, como o Facebook, por exemplo. Há outras que tratam de comunicação rápida, como o *Twitter*, e tem também aquelas que destacam e exibem conteúdo gerado pelo usuário.

Tabela 3 – Frequência de acesso às principais redes sociais.

		Facebook		Instagram		Twitter		Linkedin		Snapchat	
QN	**Options**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
5	A	0	0.0	200	39.0	262	51.1	245	47.76	368	71.7
	B	120	23.4	139	27.1	175	34.1	233	45.4	74	14.4
	C	160	31.2	90	17.5	31	6.0	21	4.1	36	7.0
	D	124	24.2	48	9.4	27	5.3	8	1.6	18	3.5
	E	56	10.9	24	4.7	9	1.5	2	0.4	10	1.9
	F	53	10.3	12	2.3	9	1.7	4	0.8	7	1.4

As frequências utilizadas para esta análise foram:

A: usuário não cadastrado na rede social; **B:** usuário que não usa a rede social todos os dias; **C:** usuário que usa a rede social por até 1 hora por dia; **D:** usuário que usa a rede social entre 1 e 3 horas por dia; **E:** usuário que usa a rede social entre 3 e 6 horas por dia; **F:** usuário que usa a rede social mais de 6 horas por dia.

O acesso do Facebook através de dispositivos móveis desempenha um grande papel em seu sucesso. Em abril de 2021, mais de 98% das contas de usuários ativos em todo o

mundo acessaram a rede social através de qualquer tipo de dispositivo móvel (STATISTA, 2021a). A Tabela 4 mostra o tipo de dispositivo que os usuários usam ao acessar o Facebook, seus objetivos principais na rede, o número de amigos que possuem e também o perfil de interação com o qual se que identificam: Espectador ou Participante.

Tabela 4 – Dispositivo de acesso, objetivos, perfil de interação e quantidade de amigos.

QN	**Questão**	**N**	**%**
6	**Principal forma de acesso ao Facebook**		
	Mobile (cellphone, tablet, etc.)	349	68.0
	Web (Chrome, Firefox, Internet Explorer, etc.)	164	31.9
7	**Principal objetivo ao utilizar o Facebook**		
	1: Acompanhar o dia a dia dos meus amigos/familiares	124	24.2
	2: Atualizar-me com notícias e acontecimentos	89	17.3
	3: Compartilhar minhas fotos e vídeos	6	1.2
	4: Comunicar com amigos/familiares	185	36.1
	5: Conhecer a opinião dos outros sobre vários assuntos	4	0.8
	6: Conhecer pessoas novas	1	0.2
	7: Descobrir e participar de promoções	3	0.6
	8: Disseminar informações que considero relevantes	14	2.7
	9: Ler sobre assuntos do meu interesse	65	12.7
	10: Para ter acesso aos jogos disponíveis na rede social	2	0.4
	11: Possibilidade de opinar sobre diversos assuntos	1	0.2
	12: Produzir e comunicar conteúdos próprios na rede	3	0.6
	13: Outro	16	3.1
8	**Perfil de interação**		
	Participante	184	35.9
	Espectador	329	64.1
9	**Quantidade de amigos no Facebook**		
	0 a 299	199	38.8
	300 a 999	250	48.7
	1000 ou mais	64	12.5

Durante o uso da rede social, ao *concordar* com um conteúdo exibido no Facebook, o usuário pode usar três dos principais recursos da rede social para expressar sua opinião: 'Curtir', 'Comentar' e 'Compartilhar'. Da mesma forma, ao *discordar* de um conteúdo exibido no Facebook, o usuário pode usar alguns recursos da rede social para expressar sua opinião: 'Comentar', 'Deixar de seguir' e 'Ocultar Conteúdo'. A frequência de uso desses recursos quando *concorda* ou *discorda* do conteúdo pode ser vista na Tabela 5.

Tabela 5 – Frequência de uso das funcionalidades quando concorda/discorda com o conteúdo na rede social.

		Concorda com o conteúdo					
		Curtir		Comentar		Compartilhar	
QN	**Frequência**	**N**	**%**	**N**	**%**	**N**	**%**
10	Nunca	18	3.5	58	11.3	108	21.1
	Raramente	59	11.5	208	40.5	187	36.5
	Ás vezes	180	35.1	194	37.8	151	29.4
	Quase sempre	138	26.9	35	6.8	51	9.9
	Sempre	118	23.0	18	3.5	16	3.1
		Não concorda com o conteúdo					
		Comentar		Deixar de seguir		Ocultar publicação	
QN	**Frequência**	**N**	**%**	**N**	**%**	**N**	**%**
11	Nunca	256	49.9	158	30.8	188	36.6
	Raramente	149	29.0	159	31.0	135	26.3
	Ás vezes	86	16.8	132	25.7	105	20.5
	Quase sempre	14	2.7	47	9.2	49	9.6
	Sempre	8	1.6	17	3.3	36	7.0

A Tabela 6 mostra com que frequência o usuário adiciona novos amigos através do recurso 'Sugestão de amigos'. Também mostra com que frequência o usuário avalia se seus amigos vão gostar de sua postagem antes de fazê-la e com que frequência ele deixa de postar conteúdo porque acredita que seus amigos não vão curtir. Por fim, esta tabela também mostra se o usuário usou a funcionalidade 'Encontrar amigos' ao criar sua conta do Facebook.

Tabela 6 – Questões apresentadas aos usuários sobre interações no Facebook.

QN	**Questão**	**N**	**%**
12	**Frequência que adiciona novos contatos a partir da funcionalidade "Sugestão de Amigos"**		
	Eu nunca adiciono	125	24.3
	Eu raramente adiciono	236	46.0
	Às vezes eu adiciono	128	25.0
	Quase sempre eu adiciono	16	3.1
	Sempre adiciono	8	1.6
13	**Frequência com que avalia se seus contatos irão curtir sua publicação antes de fazê-la**		
	Eu nunca avalio	126	24.6
	Eu raramente avalio	99	19.3
	Às vezes eu avalio	138	26.9
	Quase sempre eu avalio	76	14.8
	Sempre avalio	74	14.4
14	**Frequência com que deixa de publicar algo por acreditar que seus contatos não irão curtir certo conteúdo**		
	Eu nunca deixo de postar	163	31.8
	Eu raramente deixo de postar	118	23.0
	Às vezes deixo de postar	157	30.6
	Quase sempre deixo de postar 42	8.2	
	Sempre deixo de postar	33	6.4
15	**Quando criou a conta no Facebook utilizou a função para encontrar os amigos a partir dos contatos do e-mail**		
	Sim	216	42.1
	Não	228	44.4
	Não lembro	69	13.5

A frequência de uso das principais funcionalidades disponibilizadas pela rede social refletem um pouco da sua aceitação por parte dos usuários, podendo indicar quais são suas preferências e por qual caminho a rede social deve seguir para aumentar seu engajamento.

5.4 Caracterização dos perfis de usuário

A caracterização dos perfis Participante e Espectador foi realizada utilizando os algoritmos de árvores de decisão (C4.5 e CART). As métricas de avaliação com os resultados da classificação de ambos são apresentadas na Tabela 7. Percebe-se que os modelos gerados obtiveram bons resultados tanto para C4.5 quanto para CART, indicando que para o conjunto de dados utilizado, as árvores construídas classificam corretamente a maioria das instâncias. Porém, vemos que os dois algoritmos se comportam melhor para a classe do Espectador, com uma diferença de sensibilidade de 18,8% em relação ao perfil do Participante.

Tabela 7 – Métricas de avaliação obtidas pelos algoritmos C4.5 e CART, em porcentagem.

	Precisão	**Sensibilidade**	***F-Measure***
C4.5			
Espectador	82.5	87.5	84.9
Participante	75.0	66.9	70.7
Média	78.8	77.2	77.8
CART			
Espectador	86.2	92.3	89.1
Participante	84.1	73.5	78.5
Média	85.1	82.9	83.8

A partir das regras obtidas pela árvore de decisão, é possível caracterizar os perfis de interação dos usuários *Espectador* e *Participante*. As Tabelas 8 e 9 mostram as principais regras geradas pelos modelos de árvore de decisão C4.5 e CART, respectivamente.

Tabela 8 – Principais regras geradas para cada classe pelo C4.5.

Regras	**Número e porcentagem de instâncias cobertas**
Regra 1: SE ConcordaComentar = (Nunca ou Raramente) ENTÂO ***Espectador***	243 (81.8%)
Regra 2: SE ConcordaComentar = (Ás vezes) E FrequenciaFacebook ≠ (Entre 1 e 6 horas por dia) E FaixaEtária = (47 ou menos) E NãoConcordaComentar ≠ (Nunca) E AvaliarAntesPublicar ≠ (Sempre) E ***Participante***	42 (25.3%)
Regra 3: SE ConcordaComentar = (Ás vezes, Quase sempre ou Sempre) E FrequenciaFacebook = (Mais que 6 horas por dia) ENTÃO ***Participante***	41 (24.7%)

Número de Espectadores = 297
Número de Participantes = 166

Observando a Regra 1, concluímos que os *Espectadores* utilizam o recurso de comentário 'nunca' ou 'raramente' quando concordam com o conteúdo com 81,8% de cobertura. Isso significa que somente esta regra gerada pelo algoritmo C.45 classifica a grande maioria das instâncias dessa classe. É uma regra simples e possui uma representatividade significativa, pois a maioria se enquadra nessa definição. Avaliando a regra do CART, temos uma cobertura de 72,7%, com igual importância. Isso mostra que esses usuários não interagem com frequência no Facebook, uma vez que geralmente não comentam as publicações que possuem conteúdo com o qual concordam na rede social.

Por outro lado, para classificar os *Participantes*, as regras são mais complexas. É necessário utilizar um número maior de atributos e a abrangência das regras é mais modesta. A Regra 2 e a Regra 3 (juntas têm cobertura de 50,0% e 59,7% para C4.5 e CART, respectivamente) mostram que os *Participantes* frequentemente interagem com a

rede social, comentando as publicações com frequências: 'sempre', 'quase sempre' e 'às vezes'. Portanto, os *Participantes* são considerados o perfil de usuário mais ativo na rede social.

Tabela 9 – Regras principais geradas para cada classe pelo CART

Regras	**Número e porcentagem de instâncias cobertas**
Regra 1: SE ConcordaComentar = (Nunca ou Raramente) ENTÃO ***Espectador***	216 (72.7%)
Regra 2: SE ConcordaComentar = (Ás vezes, Quase sempre ou Sempre) E ConcordaCompartilhar = (Ás vezes, Quase sempre ou Sempre) E FrequenciaTwitter ≠ (Não cadastrado) ENTÃO ***Participante***	65 (39.2%)
Regra 3: SE ConcordaComentar = (Ás vezes, Quase sempre ou Sempre) E ConcordaCompartilhar = (Ás vezes, Quase sempre ou Sempre) E FrequenciaTwitter = (Não cadastrado) E Gênero = (Masculino) ENTÃO ***Participante***	34 (20.5%)

Número de Espectadores = 297
Número de Participantes = 166

Essas conclusões foram inferidas principalmente da questão 10 do questionário (Tabela 5), que pede ao usuário que informe a frequência com que usa as funções de comentário para o conteúdo com o qual concorda.

Ao comparar as árvores geradas pelos dois modelos, a regra principal extraída para os *Espectadores* é semelhante. Ambas as árvores de decisão usaram o mesmo recurso para particionar a maioria das instâncias (ConcordaComentar), mas elas diferem nos nós mais profundos da árvore. No entanto, as medições da árvore CART produzem resultados melhores do que a árvore C4.5.

Durante o desenvolvimento da pesquisa, esperava-se que os dados demográficos (sexo, faixa etária, escolaridade e área de estudo) influenciassem significativamente na definição do perfil do usuário. No entanto, apenas informações de gênero e idade apareceram nos nós mais profundos da árvore, não expressando a relevância que acreditávamos que teriam.

A Regra 2 de C4.5 usa o campo Faixa etária como condição para que o usuário seja considerado um *Participante* após analisar alguns outros atributos, indicando que há um grupo de *Participantes* que possui 47 anos ou menos e usa o Facebook por pelo menos 1 hora por dia.

No CART, a frequência de uso do Twitter é uma condição nas Regras 2 e 3 que classificam *Participantes*. Nesse caso, as duas regras apresentam condições semelhantes, pois afirmam que o usuário comenta e compartilha o conteúdo quando concorda com alta

frequência (às vezes, quase sempre ou sempre). No entanto, observe que essas regras diferem apenas sobre o uso do Twitter: se o usuário não possuir cadastro nesta rede social, é considerado *Participante* com cobertura de aproximadamente 39%; se o usuário possui conta no Twitter, é necessário analisar seu gênero: se for Masculino, também é considerado *Participante*, mas esta regra tem menos cobertura (aproximadamente 20%).

A função 'curtir', uma das principais características do Facebook, não é mencionada nas regras geradas nesta pesquisa. Por se tratar de uma funcionalidade de apenas um clique, acredita-se que não haja diferenças significativas no seu uso para os perfis de interação utilizados. Ou seja, diferentes usuários usam a funcionalidade 'comentar' com frequência semelhante.

5.4.1 Caracterização dos Produtores de conteúdo

Após o processo de seleção de instâncias, apenas 13 instâncias da classe *Produtor de conteúdo* permaneceram na base de dados. Devido à discrepância desta quantidade para as demais classes (*Participantes* e *Espectadores*), a caracterização dessa classe foi feita de forma diferente.

Assim, ao contrário dos perfis *Participantes* e *Espectadores*, em que a caracterização foi realizada utilizando um algoritmo de árvore de decisão, para o perfil *Produtores de conteúdo*, a caracterização foi feita qualitativamente, analisando cada uma das instâncias pertencentes para esta classe.

Analisando os 13 *Produtores de conteúdo* presentes na base de dados selecionada, encontramos as seguintes características:

- São usuários com até 47 anos de idade e com graduação, mestrado ou doutorado;
- Provêm de áreas do conhecimento não diretamente relacionadas com a matemática (Ciências Biológicas, Ciências da Saúde, Ciências Sociais Aplicadas, Humanidades e Linguística, Letras e Artes);
- Têm alta frequência de uso do Facebook e baixa frequência de uso outras redes sociais;
- Quanto ao número de amigos no Facebook, eles possuem no máximo 1000 amigos;
- Afirmam que utilizaram a funcionalidade disponibilizada para encontrar amigos de seus contatos de e-mail ao criar sua conta.

Pela análise descritiva, concluímos que os Produtores de conteúdo são: usuários jovens e adultos, possuem alta escolaridade, estão em áreas de conhecimento não relacionadas à matemática, apresentam alta frequência de uso do Facebook e baixa das demais

redes sociais, possuem no máximo 1000 amigos e utilizaram a funcionalidade para encontrar amigos por e-mail.

As características levantadas pela analise descritiva trazem algumas informações já encontradas pela análise visual dos gráficos de mosaico construídos a partir das variáveis selecionadas na regressão logística e acrescentam outras diferentes.

As conclusões inferidas a partir da regressão foram que Produtores de conteúdo seria usuários que: apresentam escolaridade alta, muitos classificam sua área de conhecimento como "outra", a frequência de uso de algumas redes sociais (Instagram, Twitter, LinkedIn e Sanpchat) é alta, o objetivo 6 (Conhecer pessoas novas) apresenta bastante relevância para essa classe e a frequência de uso das funcionalidades Comentar e Deixar de seguir ao discordar de um conteúdo publicado na rede social também é relevante.

Esses resultados nos trazem informações muito importantes a respeito do perfil Produtor de conteúdo, possibilitando o entendimento do comportamento desses usuários e a definição de algumas de suas características. Por mais que não tenha sido possível classificá-los utilizando as árvores de decisão, como as demais classes, essas informações trazem *insigths* que podem ser aprofundados em estudos futuros.

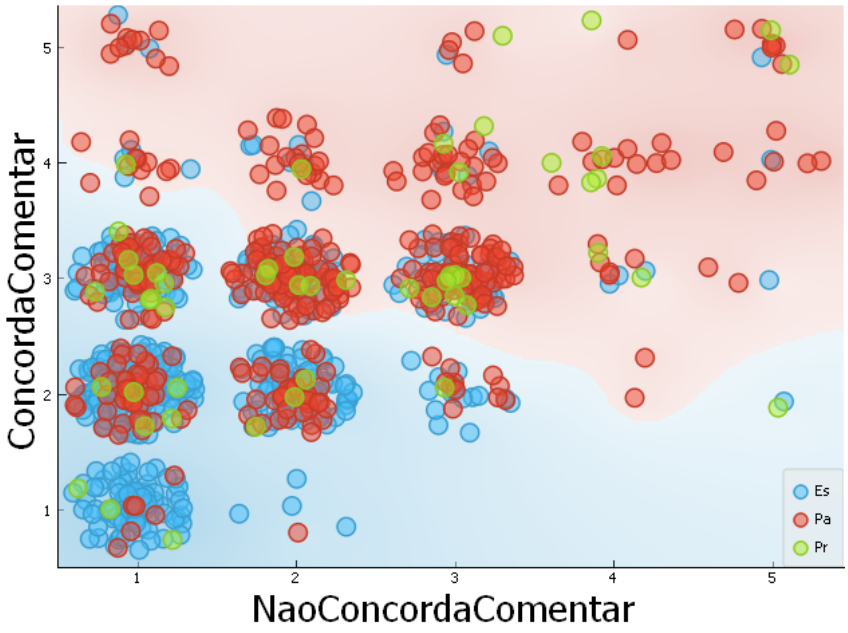
5.5 Avaliação das instâncias desconsideradas

Analisando o conjunto de dados completo com 1100 instâncias notamos que, embora haja uma maioria de usuários que correspondem à caracterização feita por RUAS DA SILVEIRA et al. (2019), também há instâncias de usuários que se comportam de forma contrária, conforme mostrado na Figura 16, que exibe as instância de acordo com a frequência de uso da funcionalidade 'comentar' ao concordar ou não com um conteúdo na rede social.

Pela Figura 16 podemos notar áreas mais bem definidas para duas classes: os *Participantes*, representados pela cor vermelha estão na região superior direita; e os *Espectadores* representados pela cor azul estão na região inferior esquerda. Isso indica que, ao comparar esses dois perfis, os *Espectadores* utilizam essa funcionalidade com uma frequência baixa enquanto os *Participantes* a utilizam com uma frequência mais alta. *Os Produtores de conteúdo*, por sua vez, não possuem uma área definida, estão espalhados entre essas duas classes, indicando que para esse perfil não há um padrão de utilização dessa funcionalidade.

A Figura 9 também mostra uma distribuição mais heterogênea nas áreas periféricas, com algumas instâncias da classe oposta à esperada. Com isso, utilizamos o AG para selecionar os exemplos mais representativos da base de dados, pois nos interessa conhecer o perfil dos usuários que são predominantemente *Participantes*, *Espectadores*, e *Produtores*

Figura 16 – Distribuição das instâncias. A coordenada x corresponde à frequência de uso da função 'comentar' ao CONCORDAR com o conteúdo postado, e a coordenada y corresponde à frequência de uso da função 'comentar' ao DISCORDAR do conteúdo postado. No gráfico, um ponto mais próximo da origem tem menor frequência de utilização da função. Os valores variam de 1 a 5, de acordo com as opções de respostas do questionário: 'Nunca', 'Raramente', 'Às vezes', 'Quase sempre' ou 'Sempre'.



de conteúdo. Assim, mantivemos apenas as instâncias mais representativas para encontrar regras capazes de classificar os exemplos que representassem a essência de cada perfil.

Ao observar as 574 instâncias desconsideradas ($N = 574$, 294 *Espectadores*, 242 *Participantes* e 38 *Produtores de conteúdo*), notamos que esses usuários seguem um determinado padrão, mas se identificam como pertencentes a uma classe diferente. Isso significa que esses usuários se comportam como esperado de outras classes, de acordo com as regras nas Tabelas 8 e 9, e com as definições feitas por RUAS DA SILVEIRA et al. (2019).

Ou seja, alguns usuários se declararam *Participantes*, mas responderam que comentam 'raramente' ou 'nunca' uma postagem na rede social quando concordam ou discordam dela. Por outro lado, outros usuários se autodeclararam como *Espectadores*, mas responderam que comentam 'quase sempre' ou 'sempre' uma postagem na rede social quando concordam ou discordam dela.

Em relação aos *Produtores de conteúdo*, a Figura 9 mostra que eles possuem características dos demais perfis, mas tendem a se assemelhar aos *Participantes*, que também são usuários mais ativos na rede; no entanto, ainda não há como separá-los.

5.6 Avaliação da qualidade dos testes

Para analisar a qualidade dos modelos gerados, avaliamos o modelo utilizando o conjunto de dados de teste com 10% do conjunto de dados original (Teste 1), e as instâncias não selecionadas pelo AG durante a seleção de instâncias (Teste 2). Os resultados são mostrados na Tabela 10.

Tabela 10 – Avaliação do conjunto de dados de teste e instâncias não selecionadas, em porcentagem

		Precisão	**Sensibilidade**	***F-Measure***
Teste 1	**C4.5**			
	Espectador	80.0	87.5	83.6
	Participante	73.3	61.1	66.7
	Média	73.3	61.1	66.7
	CART			
	Espectador	78.9	93.8	85.7
	Participante	83.3	55.6	66.7
	Média	81.1	74.7	76.2
		Precisão	**Sensibilidade**	***F-Measure***
Teste 2	**C4.5**			
	Espectador	63.7	80.4	71.1
	Participante	64.2	43.4	51.8
	Média	63.9	61.9	61.5
	CART			
	Espectador	62.8	75.1	68.4
	Participante	59.5	45.1	51.3
	Média	61.1	60.1	59.8

Teste 1: Base de teste
Teste 2: Instâncias não selecionados pelo AG

Usando o conjunto de dados de teste, 66,7% e 76,2% (média *F-measure*) deles foram classificados corretamente pelo C4.5 e CART, respectivamente. Esse fato mostra que os modelos obtidos classificam razoavelmente bem as instâncias desconhecidas.

Além disso, também observamos que o modelo aprendeu, de forma mais eficiente, o perfil *Espectador* nos dois algoritmos avaliados. Observe, por exemplo, que das 32 instâncias desta classe, CART atingiu 93,8% delas, enquanto o desempenho da classe *Participante* se mostra pior. Esses resultados são semelhantes aos obtidos com a construção dos modelos (ver Tabela 7).

Esses resultados vão até mesmo contra os resultados apresentados nas Tabelas 8 e 9, nas quais se observa que a cobertura da regra obtida para o perfil *Espectador* pelos dois algoritmos, CART e C4.5, é sempre muito superior às regras obtidas para o perfil *Participante*. Isso reforça o fato de que esses algoritmos de aprendizado identificam mais facilmente a classe *Espectador*.

Utilizando as instâncias desconsideradas pelo AG, 61,5% e 59,8% foram corretamente classificados pelo C4.5 e CART, respectivamente. Nesse caso, o modelo também respondeu melhor ao perfil do *Espectador*.

6 CONSIDERAÇÕES FINAIS

Neste trabalho, para termos sistemas mais personalizados, buscamos caracterizar os perfis dos usuários do Facebook (Espectador, Participante e Produtor de conteúdo), uma das principais redes sociais da atualidade. Para isso, utilizamos dados de um questionário disponibilizado *online* em um processo de KDD utilizando algoritmos de árvore de decisão. Nos baseamos em duas perguntas para auxiliar o desenvolvimento da pesquisa, são elas:

1. Quais *insigths* podemos inferir sobre os perfis de usuário do Facebook a partir de uma visualização multivariada do conjunto de dados? E como validar se esses *insigths* são verdadeiros?

2. Como determinar as principais características dos perfis de usuário do Facebook utilizando as instâncias mais representativas da base de dados por meio de *machine learning*?

Em relação à primeira questão, ao utilizar o FreeViz (DEMSAR; LEBAN; ZUPAN, 2008) levantamos alguns *insigths* sobre a distribuição para os perfis Espectadores e Participantes que foram validados com o resultado obtido pelos modelos criados por *machine learning*, árvore de decisão e modelo logístico. O resultado das árvores de decisão e do modelo logístico, por sua vez, nos permitiram extrair as principais características, respondendo assim oque foi definido na segunda questão.

Esses resultados mostram que, para a seleção das instâncias mais representativas dos perfis analisados, o algoritmo genético foi eficiente, trazendo melhores resultados na construção do modelo preditivo. Assim, foi possível identificar que os perfis possuem características de interação muito diferentes e utilizam as ferramentas fornecidas pelo Facebook com frequências significativamente distintas: enquanto os usuários *Participante* frequentemente comentam e compartilham postagens feitas na rede social, *o*s Espectadores raramente o fazem.

De posse dessas informações os gestores e desenvolvedores do Facebook vão conhecendo cada vez mais o usuário e podem deixar seu serviço cada vez mais personalizado, trabalhando de forma alinhada com os interesses do seu público-alvo e atendendo suas expectativas, fazendo com que ele continue utilizando a rede social.

Realizar uma análise desses dados de comportamento dos usuários é muito relevante, pois podem apresentar relações importantes entre as características de cada um desses perfis. E saber quais são essas características pode auxiliar no desenvolvimento de ferramentas que tenham mais chance de sucesso, pois, com essa informação em mãos, os idealizadores de redes sociais podem focar em desenvolver funcionalidades direcionadas para cada perfil e até mesmo disponibilizar gatilhos para que os usuários utilizem a rede social com frequência maior e assim mantenham a rede social viva.

Ao classificar os usuários do Facebook, o fator de maior importância foi: se o usuário, ao concordar com o conteúdo postado no Facebook, usará a função de comentário para interagir com aquela postagem. Essas informações se encaixam perfeitamente nas definições de perfil propostas por RUAS DA SILVEIRA et al. (2019), que definiu *Espectadores* como usuários que, predominantemente, visualizam o que está acontecendo na rede social, enquanto *Participantes* usariam as funcionalidades curtir, comentar e compartilhar com mais frequência.

Observamos ainda que a base de dados possui um percentual significativo de usuários que declararam pertencer a um determinado perfil, embora possuam todas as características de outro perfil. Isso provavelmente se deve à subjetividade intrínseca de qualquer método de pesquisa que usa questionários ou entrevistas, por exemplo. Neste trabalho, consideramos que essas instâncias eram ruídos, pois buscávamos as características mais representativas de cada perfil. No entanto, ainda acreditamos que esses casos podem ser considerados como exceções a uma regra mais geral. Assim, quando uma instância é uma exceção ao caso geral, pode parecer rotulada incorretamente.

Portanto, os resultados encontrados podem:

- Ajudar a compreender a real influência que as mídias sociais têm em nossas vidas hoje;

- Auxiliar em novas pesquisas para investigar o futuro do Facebook: se esta rede social se estabilizará em termos do número de usuários, se vai simplesmente se desfazer devido à competição ou outros fatores imprevisíveis, ou se continuará a crescer;

- Orientar o desenvolvimento de ferramentas e funções para redes sociais que contemplem as particularidades de cada perfil, aumentando o engajamento do usuário e a frequência de uso das redes sociais, apontando a importância de adaptar o comportamento do sistema ao mundo do usuário.

Esperamos que as regras obtidas ajudem no *design* de redes sociais, seja criando novas ou melhorar as existentes, a partir de informações de uma análise do Facebook, atualmente uma das maiores redes sociais.

Consequentemente, os sucessos e fracassos de uma rede social, como em qualquer sistema, não podem ser considerados aleatórios. Ambos trazem lições que podem ser aprendidas para lançamentos futuros, especialmente a reprodução de casos de sucesso. Assim, acreditamos que nossas conclusões implicam em oportunidades para projetos de redes sociais com maior probabilidade de sucesso.

6.1 Limitações

A compreensão do comportamento humano nas redes sociais está apenas começando. Este trabalho usa pequenos grupos em uma pequena escala para tratar o comportamento na rede social. No entanto, é essencial generalizar essas teorias para populações maiores e desenvolver novas técnicas de análise comportamental nas redes sociais (CAO et al., 2014), inclusive investigando outras redes sociais. Assim, embora tenhamos caracterizado os perfis dos usuários quanto aos aspectos de interação social, é imprescindível mostrar algumas limitações do nosso trabalho:

- Como essa pesquisa é uma continuação de trabalhos anteriores, estamos considerando os perfis que foram definidos por RUAS DA SILVEIRA, Nobre e Cardoso (2014). Com o passar do tempo o perfil de usuários de rede sociais pode sofrer alterações, portanto, os usuários do Facebook já podem apresentar comportamentos diferentes dos esperados nos trabalhos anteriores;

- Nossa maior limitação foi o fato do próprio respondente do questionário julgar a qual perfil pertencia com base em uma breve descrição dos três perfis de interação. Isso porque os usuários podem não conseguir se classificar com precisão, ou podem não ter parâmetros de comparação para fazê-lo, deixando a classificação subjetiva;

- Além disso, os usuários só podiam selecionar uma opção para o seu perfil de interação, o que é uma limitação para usuários que se identificam com mais de um perfil;

- Com isso, consideramos as instâncias duvidosas como ruídos e elas foram eliminadas, uma vez que estávamos tentando identificar as regras mais específicas dos perfis. Porém, algumas dessas instâncias podem ser uma exceção à regra mais representativa do perfil, e com isso podemos estar perdendo informações relevantes sobre os perfis de interação;

- Outra limitação significativa foi a quantidade de usuários do perfil Produtores de conteúdo, havia um número reduzido de ocorrências no conjunto de dados e sua qualidade não era tão boa. Com isso foi necessário aplicar a seleção de instâncias, o que diminuiu ainda mais a quantidade de instâncias dessa classe.

As limitações desse trabalho podem trazer inspirações para que novas descobertas sejam realizadas visando mitigar as dificuldades encontradas durante seu desenvolvimento.

6.2 Proposta de continuidade

Podemos elencar as seguintes propostas para trabalhos futuros:

- Melhor definição do que é um 'conteúdo' na rede social, para auxiliar o usuário em sua auto-classificação;
- Avaliar os perfis por tópico de interesse, pois um usuário pode apresentar perfis diferentes para cada conteúdo. Por exemplo, o usuário pode ser produtor para o tema 'esporte' e espectador para 'política', o que pode gerar um conflito na classificação do usuário;
- Uma melhor classificação do perfil do Produtor de conteúdo a partir do aumento do conjunto de dados para este perfil, que pode ser feita por meio de entrevistas diretas com pessoas que visivelmente se comportam como produtoras, seguindo as regras definidas por RUAS DA SILVEIRA et al. (2019);
- Adicionar ao questionário perguntas comportamentais que auxiliem na classificação dos perfis, como por exemplo, comparações entre o uso das funcionalidades: funcionalidade utiliza mais frequentemente, comentar ou curtir, dentre outras;
- Sugerimos a aplicação do agrupamento *fuzzy* no conjunto de dados, antes de sua classificação, utilizando os resultados do agrupamento para rotular os usuários ao invés de sua auto-rotulação, o que permitiria uma classificação inteiramente baseada em dados dos perfis dos usuários; ou seja, o objetivo, neste caso, seria investigar o problema como sendo de classificação multi-rótulo;
- Combinar a visualização com a técnica estatística multivariada de Hotelling (HOTELLING, 1947), que mostra se duas amostras pertencem a mesma população, para obter uma comprovação mais robusta dos *insigths* da visualização;
- Classificação de usuários de outras redes sociais e comparação de seus diferentes perfis e características com os usuários do Facebook, buscando padrões e semelhanças entre eles.

REFERÊNCIAS

ALEXA. *The TOP 500 GLOBAL SITES*. 2021. Access in 27.07.2021. Disponível em: <http://www.alexa.com/topsites>. Acesso em: 21 jul. 2021. 10

ALOUDAT, A.; AL-SHAMAILEH, O.; MICHAEL, K. Why some people do not use facebook? *Journal of Social Network Analysis and Mining*, v. 9, 12 2019. 10

BANKS, A. *Brazil Digital Future in Focus*. Evento comScore, 2015. Disponível em: <https://www.comscore.com/por/Insights/Apresentacoes-e-documentos/2015/2015-Brazil-Digital-Future-in-Focus>. 9

BELLOCCHI, C. et al. Identification of a shared microbiomic and metabolomic profile in systemic autoimmune diseases. *Journal of Clinical Medicine*, v. 8, n. 9, 2019. ISSN 2077-0383. Disponível em: <https://www.mdpi.com/2077-0383/8/9/1291>. 31

BENEVENUTO, F. et al. Characterization and analysis of user profiles in online video sharing systems. *Journal of Information and Data Management*, v. 1, n. 2, p. 261, 2010. 29

BENEVENUTO, F. et al. Characterizing user navigation and interactions in online social networks. *Information Sciences*, Elsevier, v. 195, p. 1–24, 2012. 12

BERINATO, S. *Good charts: The HBR guide to making smarter, more persuasive data visualizations*. Estados Unidos: Harvard Business Review Press, 2016. 11

BREIMAN, L. et al. Classification and regression trees (chapman y hall, eds.). *Monterey, CA, EE. UU.: Wadsworth International Group*, 1984. 11, 27, 43, 44

BRODLEY, C. E.; FRIEDL, M. A. Identifying mislabeled training data. *Journal of artificial intelligence research*, v. 11, p. 131–167, 1999. 42

CAMPBELL, S. J. et al. Visualizing the drug target landscape. *Drug Discovery Today*, v. 15, n. 1, p. 3–15, 2010. ISSN 1359-6446. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1359644609003407>. 31

CAO, L. et al. Behavior informatics: A new perspective. *IEEE Intelligent Systems*, v. 29, n. 4, p. 62–80, July 2014. ISSN 1941-1294. 10, 15, 67

CAVALCANTI, G. D.; SOARES, R. J. Ranking-based instance selection for pattern classification. *Expert Systems with Applications*, v. 150, p. 113269, 2020. ISSN 0957-4174. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417420300944>. 20, 22, 37

ÇELIK, A. K.; KARAASLAN, A. Predictors of college students' willingness to use social network services: The case of two turkish universities. *Campus-Wide Information Systems*, Emerald Group Publishing Limited, v. 31, n. 5, p. 304–318, 2014. 30

CLARK, P.; NIBLETT, T. The CN2 induction algorithm. *Machine learning*, Springer, v. 3, n. 4, p. 261–283, 1989. 21

CORRAR, L.; PAULO, E.; FILHO, J. *Análise multivariada para os cursos de administração, ciências contábeis e economia*. São Paulo: Atlas, 2007. ISBN 9788522447077. 28

DARWIN, C. *On the origin of species by means of natural selection*. London: John Murray, 1859. 564 p. 25

DEMSAR, J.; LEBAN, G.; ZUPAN, B. Freeviz—an intelligent multivariate visualization approach to explorative analysis of biomedical data. *Journal of biomedical informatics*, v. 40, p. 661–71, 01 2008. 16, 17, 65

DU, L. et al. Unsupervised dual learning for feature and instance selection. *IEEE Access*, v. 8, p. 170248–170260, 2020. 33

DUA, D.; GRAFF, C. *UCI Machine Learning Repository*. 2017. Disponível em: <http://archive.ics.uci.edu/ml>. 17

DZEMYDA, G.; KURASOVA, O.; ZILINSKAS, J. Multidimensional data visualization. *Methods and applications series: Springer optimization and its applications*, Springer, v. 75, p. 122, 2013. 19

FACELI, K. et al. *Inteligência artificial: uma abordagem de aprendizado de máquina*. São Paulo, Brasil: Grupo Gen - LTC, 2011. ISBN 978-85-216-1880-5. 27, 44

FAYYAD, U. M. et al. *Advances in knowledge discovery and data mining*. Cambridge, Massachusetts, USA, v. 21, 1996. 35, 36

FERNANDES, G.; SIQUEIRA, S. Building users profiles dynamically through analysis of messages in collaborative systems. In: *Proceedings of International Conference WWW/Internet*. Forth Worth, Texas, USA: IADES, 2013. p. 1–8. 9, 13

FERRAZ, P. F. P. et al. Agricultural residues of lignocellulosic materials in cement composites. *Applied Sciences*, v. 10, n. 22, 2020. ISSN 2076-3417. Disponível em: <https://www.mdpi.com/2076-3417/10/22/8019>. 32

FOGG, B. J. *Persuasive Technology: Using Computers to Change What We Think and Do*. San Francisco, USA: Morgan Kaufmann Publishers, 2003. 15, 30

FOGG, B. J.; IIZAWA, D. Online persuasion in facebook and mixi: A cross-cultural comparison. In: *Proceedings of the 3rd International Conference on Persuasive Technology*. Berlin, Heidelberg: Springer-Verlag, 2008. (PERSUASIVE '08), p. 35–46. ISBN 978-3-540-68500-5. Disponível em: <http://dx.doi.org/10.1007/978-3-540-68504-3_4>. 10

FORTIN, F.-A. et al. Deap: Evolutionary algorithms made easy. *J. Mach. Learn. Res.*, JMLR.org, v. 13, n. 1, p. 2171–2175, jul. 2012. ISSN 1532-4435. 42

FREITAS, C. M. D. S. et al. Introdução à visualização de informações. *Revista de informática teórica e aplicada. Porto Alegre. Vol. 8, n. 2 (out. 2001), p. 143-158*, v. 8, p. 143–158, 2001. 16

FRENAY, B.; VERLEYSEN, M. Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, v. 25, n. 5, p. 845–869, 2014. 20, 21

FREUND, Y. An adaptive version of the boost by majority algorithm. Association for Computing Machinery, New York, NY, USA, p. 102–113, 1999. Disponível em: <https://doi.org/10.1145/307400.307419>. 21

FRIEDMAN, J. et al. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, Institute of Mathematical Statistics, v. 28, n. 2, p. 337–407, 2000. 21

GARCIA-PEDRAJAS, N.; del Castillo, J. A. R.; CERRUELA-GARCIA, G. Si(fs)2: Fast simultaneous instance and feature selection for datasets with many features. *Pattern Recognition*, v. 111, p. 107723, 2020. ISSN 0031-3203. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0031320320305264>. 33

GARCIA, S.; LUENGO, J.; HERRERA, F. *Data Preprocessing in Data Mining*. Spain: Springer, 2015. v. 72. 195-243 p. ISBN 978-3-319-10246-7. 21, 22, 23, 24

GIL, A. C. *Como elaborar projetos de pesquisa*. São Paulo: Atlas, 2017. v. 6. 35

HARRIS, T. *The Social Dilemma*. Direção de Jeff Orlowski. Estados Unidos: Netflix, 2020. Documentário (89 min.). 9

HOFFMAN, P. et al. DNA visual and analytic data mining. In: *Proceedings. Visualization '97 (Cat. No. 97CB36155)*. Phoenix, AZ, USA: IEEE, 1997. p. 437–441. 19

HOLLAND, J. H. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence*. Cambridge, MA, USA: MIT Press, 1992. ISBN 0262082136. 25

HORTA, R. A. M. et al. Comparação de técnicas de seleção de atributos para previsão de insolvência de empresas brasileiras no período 2005-2007. EnANPAD, Rio de Janeiro, RJ, 2010. 28

HOTELLING, H. Multivariate quality control. *Techniques of statistical analysis*, McGraw-Hill, New York, 1947. 68

KADIMA, H.; MALEK, M. Toward ontology-based personalization of a recommender system in social network. In: *2010 International Conference of Soft Computing and Pattern Recognition*. Cergy-Pontoise, France: IEEE, 2010. p. 119–122. 15

KANDOGAN, E. Star coordinates: A multi-dimensional visualization technique with uniform treatment of dimensions. In: CITESEER. *Proceedings of the IEEE information visualization symposium*. Salt Lake City, Utah: IEEE, 2000. v. 650, p. 22. 19

KHARDON, R.; WACHMAN, G. Noise tolerant variants of the perceptron algorithm. *Journal of Machine Learning Research*, v. 8, n. Feb, p. 227–248, 2007. 21

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995. (IJCAI'95), p. 1137–1143. ISBN 1-55860-363-8. Disponível em: <http://dl.acm.org/citation.cfm?id=1643031.1643047>. 45

LI, C.; BERNOFF, J.; GROOT, M. *groundswell*. Boston, USA: Thema, 2011. 29

LONG, T. V. Arcviz: An extended radial visualization for classes separation of high dimensional data. In: *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*. Ho Chi Minh City, Vietnam: IEEE, 2018. p. 158–162. 19

MEYDAN, C.; SEZERMAN, U. Short-term hiv therapy response prediction using sequence information. *BIT*, p. 234, 2011. 31

NASCIMENTO, M. L. et al. Uma análise do fator cultural em tecnologias persuasivas: um estudo de caso da rede social facebook. *Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2018)*, Natal, Brasil, 2018. 10

NOOR, N. M. M. et al. Supporting decision-making for forensic dna analysis in crime investigation using visualization. In: *Recent Researches in Applied Mathematics and Informatics*. Montreux, Switzerland: WSEAS Press, 2011. 31

OLIVEIRA, A. P. C. d. et al. Persuasion strategies in mobile systems: A case study of facebook application. In: *Proceedings of the XVI Brazilian Symposium on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2017. (IHC 2017), p. 42:1–42:10. ISBN 978-1-4503-6377-8. Disponível em: <http://doi.acm.org/10.1145/3160504.3160546>. 10

OLVERA-LOPEZ, J. et al. A review of instance selection methods. *Artif. Intell. Rev.*, v. 34, p. 133–143, 08 2010. 21

ORLINSKI, M.; JANKOWSKI, N. O (m log m) instance selection algorithms—rr-drops. *2020 International Joint Conference on Neural Networks (IJCNN)*, p. 1–8, 2020. 33

PARK, D. H. et al. A literature review and classification of recommender systems research. *Expert Systems with Applications*, Elsevier, v. 39, n. 11, p. 10059–10072, 2012. 13

QUINLAN, J. R. Induction of decision trees. *Machine learning*, Springer, v. 1, n. 1, p. 81–106, 1986. 20, 27

QUINLAN, J. R. *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993. ISBN 1-55860-238-0. 11, 27, 43

RUAS DA SILVEIRA, P. H. B. et al. Identification and characterisation of facebook user profiles considering interaction aspects. *Behaviour and Information Technology*, v. 39, p. 15, 01 2019. 10, 11, 12, 30, 39, 61, 62, 66, 68

RUAS DA SILVEIRA, P. H. B.; NOBRE, C. N.; CARDOSO, A. M. P. A influência das estratégias persuasivas no comportamento dos usuários no facebook. *Proceedings of the 13th Brazilian Symposium on Human Factors in Computing Systems*, p. 255–264, 2014. 10, 30, 67

RUBIO-SANCHEZ, M. et al. A comparative study between radviz and star coordinates. *IEEE Transactions on Visualization and Computer Graphics*, v. 22, n. 1, p. 619–628, 2016. 31

RUSSELL, S.; NORVIG, P. *Inteligência artificial*. Rio de Janeiro: Elsevier, 2013. ISBN 9788535211771. 27

SILVA, C. M. C.; PRADO, E. P. V. Estudo sobre o engajamento de usuários de uma mídia social disponibilizada pelo governo. In: *XI Brazilian Symposium on Information System*. Goiânia, Brasil: Sociedade Brasileira de Computação, 2015. p. 26–29. 9

SOSNOVSKY, S.; DICHEVA, D. Ontological technologies for user modelling. *International Journal of Metadata, Semantics and Ontologies*, Inderscience Publishers, v. 5, n. 1, p. 32–71, 2010. 10

STATISTA. *Device usage of Facebook users worldwide*. 2021. Access in 21.07.2021. Disponível em: <https://www.statista.com/statistics/377808/distribution-of-facebook-users-by-device/>. Acesso em: 21 jul. 2021. 56

STATISTA. *Leading countries based on number of Facebook users (in millions)*. 2021. Access in 21.07.2021. Disponível em: <https://www.statista.com/statistics/268136/top-15-countries-based-on-number-of-facebook-users/>. Acesso em: 21 jul. 2021. 10

STATISTA. *Number of monthly active Facebook users worldwide as of 1st quarter 2021 (in millions)*. 2021. Access in 21.07.2021. Disponível em: <https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/>. Acesso em: 21 jul. 2021. 10

TSAI, C.-F. et al. Under-sampling class imbalanced datasets by combining clustering analysis and instance selection. *Information Sciences*, v. 477, p. 47 – 54, 2019. ISSN 0020-0255. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0020025518308478>. 32

TUNA, T. et al. User characterization for online social networks. *Social Network Analysis and Mining*, v. 6, p. 104, 11 2016. 13

VASCONCELOS, M. A. et al. Tips, dones and todos: Uncovering user profiles in foursquare. In: *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*. New York, NY, USA: ACM, 2012. (WSDM '12), p. 653–662. ISBN 978-1-4503-0747-5. Disponível em: <http://doi.acm.org/10.1145/2124295.2124372>. 29, 30

WILSON, D. R.; MARTINEZ, T. R. Reduction techniques for instance-based learning algorithms. *Machine Learning*, Kluwer Academic Publishers, Boston, v. 38, n. 3, p. 257–286, 2000. 32

WU, X. *Knowledge acquisition from databases*. New Jersey: Intellect books, 1995. 20

YASLAN, Y.; GÜLAÇAR, H.; KOÇ, M. N. User profile analysis using an online social network integrated quiz game. *Journal Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, v. 21, n. 3, p. 696–702, 2017. 15

YOU, Z.-H. et al. Integrating feature and instance selection techniques in opinion mining. *International Journal of Data Warehousing and Mining*, v. 16, p. 168–182, 07 2020. 22

ZHANG, A. et al. Visualization-aided classification ensembles discriminate lung adenocarcinoma and squamous cell carcinoma samples using their gene expression profiles. *PloS one*, Public Library of Science San Francisco, USA, v. 9, n. 10, p. e110052, 2014. 31

ZHU, X.; WU, X. Scalable representative instance selection and ranking. In: *18th International Conference on Pattern Recognition (ICPR'06)*. Hong Kong, China: IEEE, 2006. v. 3, p. 352–355. 32

ZHU, X.; WU, X.; CHEN, Q. Eliminating class noise in large datasets. In: *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*. Washington: AAAI Press, 2003. (ICML'03), p. 920–927. ISBN 1577351894. 20