



CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

UMA ESTRATÉGIA PARA A PREDIÇÃO DE COAUTORIAS EM DADOS DE PUBLICAÇÕES CIENTÍFICAS

FELIPE AFFONSO

Orientador: Thiago Magela Rodrigues Dias
Centro Federal de Educação Tecnológica de Minas Gerais

Coorientador: Alisson Marques da Silva
Centro Federal de Educação Tecnológica de Minas Gerais

BELO HORIZONTE
NOVEMBRO DE 2020

FELIPE AFFONSO

UMA ESTRATÉGIA PARA A PREDIÇÃO DE COAUTORIAS EM DADOS DE PUBLICAÇÕES CIENTÍFICAS

Dissertação apresentada ao Programa de Pós-graduação em Modelagem Matemática e Computacional do Centro Federal de Educação Tecnológica de Minas Gerais, como requisito parcial para a obtenção do título de Mestre em Modelagem Matemática e Computacional.

Área de concentração: Modelagem Matemática e Computacional

Linha de pesquisa: Sistemas Inteligentes

Orientador: Thiago Magela Rodrigues Dias
Centro Federal de Educação Tecnológica de Minas Gerais

Coorientador: Alisson Marques da Silva
Centro Federal de Educação Tecnológica de Minas Gerais

CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL
BELO HORIZONTE
NOVEMBRO DE 2020

Affonso, Felipe
A257e Uma estratégia para a predição de coautorias em dados de publicações científicas / Felipe Affonso. – 2020.
63 f.

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Modelagem Matemática e Computacional.

Orientador: Thiago Magela Rodrigues Dias.

Coorientador: Alisson Marques da Silva.

Dissertação (mestrado) – Centro Federal de Educação Tecnológica de Minas Gerais.

1. Redes de informação – Modelos matemáticos – Teses. 2. Comunicação na ciência – Teses. 3. Currículo – Teses. I. Dias, Thiago Magela Rodrigues. II. Silva, Alisson Marques da. III. Centro Federal de Educação Tecnológica de Minas Gerais. IV. Título.

CDD 519.6



SERVIÇO PÚBLICO FEDERAL
MINISTÉRIO DA EDUCAÇÃO
CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
COORDENAÇÃO DO CURSO DE MESTRADO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

***“UMA ESTRATÉGIA PARA PREDIÇÃO DE COAUTORIAS EM DADOS DE
PUBLICAÇÕES CIENTÍFICAS”***

Dissertação de Mestrado apresentada por **Felipe Affonso**, em 27 de novembro de 2020, ao Programa de Pós-Graduação em Modelagem Matemática e Computacional do CEFET-MG, e aprovada pela banca examinadora constituída pelos professores:

Prof. Dr. Thiago Magela Rodrigues Dias (Orientador)
Centro Federal de Educação Tecnológica de Minas Gerais

Prof. Dr. Alisson Marques da Silva (Co-orientador)
Centro Federal de Educação Tecnológica de Minas Gerais

Prof. Dr. Gray Farias Moita
Centro Federal de Educação Tecnológica de Minas Gerais

Prof. Dr. Jether Oliveira Gomes
Vale S/A

Visto e permitida a impressão,

Prof. Dr. Thiago de Souza Rodrigues
Coordenador do Programa de Pós-Graduação em
Modelagem Matemática e Computacional

Dedico a todos que me ajudaram ao longo destes anos.

Agradecimentos

Aos meus pais, Daniela e Márcio, por todo o apoio em todos os momentos, e por sempre me aconselharem em todas as situações.

Ao meu irmão, Moises, por me ajudar a desenvolver a minha paciência, tão importante neste momento.

Ao meu orientador, Thiago Magela Rodrigues Dias, que esteve presente em toda a minha trajetória acadêmica, e hoje posso chamar de amigo.

À todos os professores que um dia passaram pela minha vida, e se tornaram fonte de inspiração para a conquista deste título. Em especial ao meu co-orientador, Alisson Marques da Silva.

Aos meus amigos, que de diferentes formas contribuíram com conversas e tornaram este momento mais leve.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 88882.442595/2019-01.

Resumo

Ao se publicar um artigo em conjunto com outros autores, um vínculo é formado pela colaboração, o que pode ser caracterizado como uma rede de colaboração científica. Neste contexto, os trabalhos representam as arestas, e os autores representam os nós, formando uma rede. É possível compreender melhor o relacionamento e as características existentes entre os autores ao se utilizar os conceitos da análise de redes sociais. Desta forma, surge a seguinte pergunta: “É possível estimar futuras colaborações científicas a fim de compreender a evolução desta rede?”. Para ser possível responder a esta pergunta, é necessário analisar como dois nós interagem entre si, ou seja, quais fatores são essenciais para que uma nova ligação seja realizada. O trabalho em questão tem como objetivo realizar a predição de ligações em redes de colaboração científica formadas pelos doutores brasileiros, em 8 diferentes áreas do conhecimento. Como fonte de dados é adotada a Plataforma Lattes, atualmente com 6,9 milhões de currículos cadastrados, que representa um dos repositórios científicos mais relevantes e reconhecidos mundialmente. Deste conjunto, foram utilizados dados de 71.010 doutores, que realizaram, em conjunto, a autoria de 5,8 milhões de colaborações científicas no período de 1960 a 2019, tornando este trabalho inédito devido à sua representatividade. De forma a tornar possível a análise e execução do trabalho, técnicas de aprendizado de máquinas foram utilizadas. Estas possuem a capacidade de identificar padrões de comportamento em grandes conjuntos de dados. Para tanto, quatro algoritmos são utilizados: Regressão Logística, Florestas Aleatórias, K Vizinhos Mais Próximos e *Naive Bayes*. A partir da abordagem utilizada, é possível identificar diferentes características das áreas do conhecimento analisando como estas influenciam o comportamento das técnicas escolhidas. Os resultados obtidos apresentam uma média de 88% de acertos, considerando todas as áreas e todos os algoritmos, onde a técnica Florestas Aleatórias apresentou o melhor resultado, com 99% de acertos em todas as redes analisadas. Avaliando os fatores que influenciam o surgimento de colaborações científicas, o atributo Menor Caminho se destaca, sendo considerado o mais influente.

Palavras-chave: Redes. Colaboração Científica. Plataforma Lattes.

Abstract

When publishing an article together with other authors, a link is formed by collaboration, which can be characterized as a network of scientific collaboration. In this context, the works represent the edges, and the authors represent the nodes of the network. By using the concepts of social network analysis, it is possible to better understand the relationship and characteristics that exist between the authors. Therefore, the following question arises: "Is it possible to estimate future scientific collaborations in order to understand the evolution of this network?". To answer this question, it is necessary to analyze how two nodes interact with each other, that is, which factors are essential for a new connection to be made. The work in question aims to predict the connections in scientific collaboration networks formed by Brazilian Phd's, in 8 different areas of knowledge. The Lattes Platform is adopted as the data source, currently with 6.9 million registered resumes, representing one of the most relevant and recognized scientific repositories worldwide. From this data set, data from 71,010 doctors were used, who jointly authored 5.8 million scientific collaborations between 1960 and 2019, making this work unprecedented due to its representativeness. In order to enable the analysis and execution of this work, machine learning techniques were used to identify behavior patterns in large data sets. For this, four algorithms are used, namely: Logistic Regression, Random Forests, K Nearest Neighbors, and, Naive Bayes. From the approach used, it is possible to identify different characteristics of the knowledge areas analyzed, and how they influence the behavior of the chosen techniques. The results obtained show an average of 88% of correct answers, considering all areas and all algorithms, where the Random Forests technique presented the best result, with 99 % of correct answers in all analyzed networks. The factors that influence the appearance of scientific collaborations, the attribute Shortest Path stands out, being considered the most influential.

Keywords: Networks. Scientific collaboration. Lattes Plataform.

Lista de Figuras

Figura 1 – Exemplo de um grafo	17
Figura 2 – Graus de Separação	18
Figura 3 – Problema da predição de links	20
Figura 4 – Desbalanceamento entre as classes	22
Figura 5 – Representação de um classificador de vizinhos mais próximos.	25
Figura 6 – Função Sigmóide	27
Figura 7 – Método Proposto	30
Figura 8 – LattesDataXplorer	31
Figura 9 – Evolução das Redes de Colaboração científica	33
Figura 10 – Divisão do Conjunto de Dados	42
Figura 11 – Proporção entre dados de treino e dados de teste.	46
Figura 12 – Quantidade de observações para treino e teste	48
Figura 13 – Média da métrica área sob a curva (AUC) para todos os algoritmos	52
Figura 14 – Importância de cada atributo	53

Lista de Tabelas

Tabela 1 – Caracterização das Redes de Colaboração científica	34
Tabela 2 – Notação utilizada	39
Tabela 3 – Matriz de Confusão	43
Tabela 4 – Fatias do conjunto de dados	47
Tabela 5 – Resultado Global	49
Tabela 6 – Resultados obtidos nas áreas selecionadas	50

Sumário

1 – Introdução	1
1.1 Motivação	3
1.2 Objetivos	4
1.3 Organização do trabalho	4
2 – Trabalhos Relacionados	6
2.1 Análise de redes	6
2.2 Predição de ligações	8
2.3 Colaboração científica e a Plataforma Lattes como fonte de dados	12
2.4 Resumo do Capítulo	14
3 – Fundamentação Teórica	16
3.1 Análise de redes Sociais	16
3.2 Predição de ligações	19
3.2.1 Balancamento de Classes	22
3.3 Aprendizado de Máquinas	23
3.3.1 Algoritmos de Classificação	24
3.3.1.1 Método Geral	25
3.3.1.2 K Vizinhos Mais Próximos:	25
3.3.1.3 Regressão Logística:	26
3.3.1.4 Florestas Aleatórias:	26
3.3.1.5 <i>Naive Bayes</i>	27
3.4 Resumo do Capítulo	28
4 – Metodologia	29
4.1 Método Proposto	29
4.2 Extração dos Dados	31
4.3 Caracterização das Redes de Colaboração Científica	32
4.4 Processamento do conjunto de dados	36
4.5 Cálculo das Métricas	39
4.6 Predição de futuras colaborações	41
4.7 Avaliação	43
4.8 Resumo do capítulo	44
5 – Análise e Discussão dos Resultados	45
5.1 Caracterização dos dados	45

5.2 Resultados obtidos	47
5.3 Resumo do Capítulo	54
6 – Conclusão	55
6.1 Trabalhos Futuros	56
Referências	58

1 Introdução

Através do estudo das colaborações científicas, é possível ter uma visão geral da comunidade científica, identificar grupos quem possuem uma maior intensidade em suas colaborações e até mesmo auxiliar na aplicação de recursos financeiros de forma a impulsionar os resultados das pesquisas científicas (NEWMAN, 2004; DIAS et al., 2014). A análise de redes permite estudar como ocorrem as colaborações científicas. Neste tipo específico de rede, os autores representam os nós, e as arestas representam trabalhos publicados em conjunto. Este tipo de análise também pode ser utilizado em diversos outros tipos de rede, como, por exemplo, as pertencentes à biologia, Internet, roteadores, entre outros (BARABÁSI; ALBERT, 1999; NEWMAN, 2001; NEWMAN; PARK, 2003). Um tipo específico de rede é chamado de rede social, responsável por motivar alguns estudos específicos, que objetivaram compreender como ocorre o surgimento de novas conexões entre os objetivos em diferentes áreas do conhecimento (NEWMAN, 2003). A partir dos trabalhos citados anteriormente, foi possível representar as redes matematicamente, através da utilização de grafos, o que impulsionou ainda mais as pesquisas e estudos nesta área. Utilizando esta representação matemática, foram adotadas métricas, além de serem estabelecidos índices e propriedades para melhor explicar o comportamento dos objetos pertencentes às redes.

A partir da análise de redes, é possível explicar diversos fenômenos, como, por exemplo, o surgimento das ligações entre diferentes nós que compõem uma rede. Ao se estudar o surgimento de novas arestas, a seguinte pergunta pode ser feita: "como ocorre a evolução da rede ao longo do tempo?". Nesse contexto, Hasan e Zaki (2011) explicam que compreender a evolução de uma rede como um todo é uma tarefa complexa, devido à grande quantidade de informações envolvidas no processo, e, consequentemente, ao alto custo computacional para tanto. Partindo disto, Liben-Nowell e Kleinberg (2003) propuseram o problema da predição de ligações, que consiste na tentativa de prever futuras ligações entre os nós de uma determinada rede. Inicialmente, foram utilizados métodos que calculavam a similaridade entre dois nós, e, quanto mais similares eles fossem, maior a chance de possuírem uma ligação entre si. Posteriormente, novas abordagens e métodos foram utilizados, trazendo ao problema da predição de ligações novas possibilidades e aplicações em diversas áreas do conhecimento (ACAR; DUNLAVY; KOLDA, 2009; ZHOU; LÜ; ZHANG, 2009; LIU et al., 2011). Dentre as técnicas possíveis estão os modelos probabilísticos, os baseados em álgebra linear, e por fim, os relacionados a classificação binária, que foram utilizados neste trabalho. Dessa forma, é possível utilizar algoritmos de aprendizado de máquinas para realizar a predição de futuras ligações.

A abordagem utilizada por Liben-Nowell e Kleinberg (2003), que leva em conside-

ração a similaridade dos nós, possibilita a sua aplicação em diferentes cenários, afinal, diversos elementos do cotidiano podem ser representados a partir dos princípios relacionados às redes. Dessa forma, a predição de ligações pode ser aplicada às redes sociais na tentativa de encontrar amigos que ainda não fizeram algum tipo de contato em um ambiente *on-line*, na propagação de um vírus através da Internet e até mesmo futuras pessoas a serem infectadas por uma doença (ADAMIC; ADAR, 2003; HASAN et al., 2006).

Aplicando tais conceitos a um domínio mais específico, pode-se dirigir as atenções às redes pertencentes à comunidade científica. Ao se publicar um trabalho em conjunto com outro cientista, uma ligação é formada pela colaboração realizada. Nestas redes, os autores representam os nós, e os trabalhos escritos em conjunto representam as arestas (MARUYAMA; DIGIAMPIETRI, 2019a). Tais redes são chamadas de redes de colaboração científica, ou redes de co-autoria, e foram o objeto de estudo deste trabalho.

Neste contexto, a Plataforma Lattes, mantida pelo CNPq¹, tem sido fonte de dados de diversos trabalhos que visam analisar redes de colaboração científicas, principalmente por englobar dados de grande parte da produção científica nacional (PEREZ-CERVANTES et al., 2013; DIAS et al., 2013; DIAS et al., 2014; MARUYAMA; DIGIAMPIETRI, 2019a). Atualmente, a Plataforma Lattes conta com 6.9 milhões de currículos de indivíduos e representa uma das fontes de dados científicos mais relevantes e reconhecidas mundialmente (LANE, 2010). O conjunto de dados registrado nos currículos cadastrados na Plataforma Lattes possui atributos como: nome, formação acadêmica, experiência profissional, projetos, publicações científicas, entre outros. O grande volume de dados presente nos currículos pode fornecer informações valiosas e até então desconhecidas (DIAS et al., 2013).

As informações, descritas anteriormente, contidas nos currículos dos indivíduos cadastrados na Plataforma Lattes, podem ser extraídas dessa base, de forma a realizar diversos estudos e análises. Em especial, por meio destas informações, é possível que redes de colaboração científica sejam geradas e analisadas, a fim de compreender como ocorre o surgimento de novas colaborações científicas. Entretanto, para que seja possível compreender como ocorre a evolução da rede como um todo, é necessário compreender, inicialmente, como dois nós interagem entre si.

A rede é formada pelo relacionamento entre os nós, e, portanto, busca-se uma forma de estimar quais pesquisadores produzirão um trabalho científico em conjunto no futuro. Tal comportamento está presente em diversas redes sociais através das "ligações sugeridas". Dessa forma, é possível utilizar as mesmas técnicas aplicadas em nosso dia a dia para sugestão de amigos, páginas e perfis para as redes de colaboração científica estudadas neste trabalho.

Dentre as possíveis técnicas a serem utilizadas para realizar a sugestão de novas

¹ Conselho Nacional de Desenvolvimento Científico e Tecnológico

conexões em redes sociais, o aprendizado de máquinas, que representa uma abordagem específica dos conceitos de inteligência artificial, também pode ser utilizado. Através do emprego dessa abordagem, é possível compreender quais atributos dos nós os fazem mais similares uns aos outros, e, portanto, possuem uma maior chance de criarem uma ligação entre si no futuro. A partir disso, é possível utilizar diferentes tipos de atributos, uma vez que o próprio algoritmo atribuirá pesos maiores àqueles que mais influenciam o resultado final. Sendo assim, os atributos são divididos em duas categorias: topológicos e específicos do domínio. Para os atributos topológicos, são considerados aqueles referentes à própria rede, como, por exemplo, o caminho mínimo entre dois nós, a quantidade de conexões em comum, entre outros. Já os atributos específicos do domínio são aqueles que estão diretamente vinculados aos nós da rede, como, por exemplo, a cidade, instituição e estado de atuação profissional do indivíduo.

Isto posto, pretende-se realizar a predição de ligações em redes de colaboração científica formada pelos dados de indivíduos com o doutorado concluído, de 8 diferentes áreas do conhecimento, que possuem seus currículos cadastrados na Plataforma Lattes. Com isso, será possível compreender o comportamento dessas redes e acompanhar a sua evolução ao longo do tempo, identificando também os pesquisadores que poderão colaborar no futuro.

1.1 Motivação

A predição de ligações pode ser utilizada para diversos fins. De acordo com [Newman \(2018\)](#), essa técnica pode ser utilizada para identificação de arestas que deveriam estar na rede, porém não estão presentes no momento da observação. Em muitas redes biológicas, tais como teias alimentares, redes de interação proteína-proteína e redes metabólicas, para comprovar a existência de uma aresta entre dois nós, experiências em campo e / ou laboratoriais, normalmente muito dispendiosas, devem ser realizadas ([LÜ; ZHOU, 2011](#)). Como resultado, diversas redes estudadas estão incompletas. Uma alternativa atraente à verificação exaustiva de uma conexão entre cada par de vértices em uma rede é tentar prever, com antecedência e com base nas conexões já observadas, quais nós, provavelmente, estão conectados ([CLAUSET; MOORE; NEWMAN, 2008](#)). O estudo da predição de ligações também impacta diretamente na compreensão dos relacionamentos entre indivíduos no mundo real, uma vez que, cada vez mais, as pessoas participam de comunidades virtuais *on-line* ([TANG, 2017](#)).

Além de auxiliar a identificar arestas faltantes nas redes, algoritmos para predição de ligações também podem ser utilizados para prever quais conexões serão feitas no futuro ([LÜ; ZHOU, 2011](#)). Esta informação pode ser útil para futuros projetos de pesquisa e estratégias de fomento à pesquisa. Este tipo de abordagem pode, assim, otimizar a produção dos

pesquisadores por meio da sugestão de pesquisadores com os quais as parcerias sejam promissoras (DIGIAMPIETRI et al., 2015).

Este trabalho se mostra extremamente relevante por utilizar um conjunto de dados até então pouco utilizado em sua completude. Através dos resultados aqui apresentados, será possível estabelecer a evolução da rede de colaborações científicas dos pesquisadores em nível nacional, auxiliando as agências de fomento na escolha de futuros pesquisadores que podem trabalhar em colaboração.

1.2 Objetivos

Com o objetivo de propor uma estratégia para predição de futuras colaborações científicas, foram utilizados os currículos de indivíduos com doutorado concluído, de 8 diferentes áreas do conhecimento, cadastrados na Plataforma Lattes.

Para alcançar tal objetivo, são realizadas as seguintes tarefas:

- Obtenção dos currículos da Plataforma Lattes;
- Seleção do conjunto de interesse para desenvolvimento do trabalho;
- Geração das redes de colaboração científica dos doutores pertencentes à 8 áreas específicas do conhecimento;
- Análise baseada em redes para identificação de atributos relevantes à colaboração científica;
- Análise de métricas estruturais de análise de redes sociais para obtenção de dados topológicos das mesmas;
- Filtragem e organização dos dados referentes a pesquisadores com publicações no período determinado, e pertencentes às áreas previamente estabelecidas em diferentes períodos de tempo;
- Aplicação de diferentes algoritmos de predição de futuras colaborações;
- Análise e comparação das redes dos representantes de cada uma das 8 áreas do conhecimento selecionadas, no que diz respeito à predição e às futuras contribuições científicas previstas para determinada área.

1.3 Organização do trabalho

O texto está organizado da seguinte forma: no [Capítulo 2](#) são apresentados os estudos previamente realizados, técnicas utilizadas e resultados obtidos. Já o [Capítulo 3](#) é dedicado à definição de alguns conceitos que serão importantes para compreensão do estudo aqui realizado. O [Capítulo 4](#) apresenta a metodologia aplicada, onde são explicadas todas as técnicas e decisões tomadas para que os objetivos propostos sejam atingidos. Os

resultados obtidos a partir dessa metodologia são apresentados no [Capítulo 5](#). Por fim, uma conclusão e alguns trabalhos futuros são apresentados no [Capítulo 6](#).

2 Trabalhos Relacionados

Este capítulo tem como objetivo apresentar os trabalhos relacionados a este estudo. Inicialmente, a [Seção 2.1](#), apresenta estudos relacionados à análise de redes sociais, bem como alguns exemplos de redes que já foram objetos de pesquisa anteriormente. Posteriormente, na [Seção 2.2](#), são apresentados os estudos na área de predição de ligações. A partir disso, é possível conhecer as principais referências no tema, assim como alguns estudos que possuem objetivos similares aos aqui apresentados. A [Seção 2.3](#) apresenta trabalhos relacionados às redes de colaboração científica e estudos realizados com foco na Plataforma Lattes. Este entendimento se faz necessário uma vez que serão utilizados os conceitos de análise de redes sociais, juntamente com as técnicas de predição de ligações para prever futuras colaborações científicas no conjunto de indivíduos com o doutorado concluído, de 8 diferentes áreas, com currículos cadastrados na Plataforma Lattes.

2.1 Análise de redes

Iniciando os estudos a respeito de redes complexas, [Barabási e Albert \(1999\)](#), apresentam conceitos-chave, onde são discutidos principalmente os fenômenos que levam à existência de um alto grau de organização entre os nós da rede, caracterizando, assim, a escalabilidade de redes complexas. Como será visto posteriormente, os graus de uma rede representam uma característica extremamente importante na análise destes.

Tentando entender o efeito "mundo pequeno" — um antigo ditado popular conhecido por atestar que todos estamos à 6 pessoas de distância uns dos outros — [Newman e Watts \(1999\)](#) realizam uma análise das redes sociais para tentar identificar a distância entre os nós da rede. Em uma afirmação importante para a época, os autores destacam que redes sociais não são aleatórias, muito pelo contrário, existem localidades na rede onde a probabilidade de existir uma conexão entre dois nós pode ser maior ou menor.

Ainda tentando compreender características fundamentais das redes, [Newman e Park \(2003\)](#) salientam as principais diferenças entre redes sociais e não sociais. Especificamente, é destacado que a relação entre os graus dos nós adjacentes das redes são positivamente correlacionados nas redes sociais, enquanto nos outros tipos de rede essa correlação é negativa. Em segundo lugar, redes sociais mostram alto nível de agrupamento. Como conclusão, é possível afirmar que as redes sociais são divididas em comunidades, enquanto as não sociais, não apresentam esta divisão.

Já em [Newman \(2003\)](#) são estudados os padrões das redes e as suas tendências em

possuir conexões com nós que estejam mais próximos. Ao longo do trabalho, o conceito de assortatividade é demonstrado ao longo de diversas redes no mundo real. Tal conhecimento pode ser aplicado, por exemplo, no combate à epidemias. Ao compreender o comportamento da rede, é possível entender como a doença se espalha e retirar os nós importantes, impedindo, conseqüentemente, a proliferação da mesma. É criada uma métrica, chamada de coeficiente de assortatividade, que objetiva mensurar e avaliar a tendência de ligações com nós similares ocorrerem.

Com o objetivo de compreender a importância da heterogeneidade em redes, [Han \(2009\)](#) apresenta um trabalho a respeito da influência entre os nós. O autor propõe quatro métodos para resolução de problemas em redes: distinção de objetos semanticamente iguais, veracidade de informações em uma rede, análise *on-line* de informações, e, por último, uma metodologia para ranqueamento e agrupamento de objetos similares. Como conclusão, é destacado que existe muito conhecimento a ser descoberto na análise de redes, e que há necessidade de realizar mais estudos sistemáticos na área.

Já com o objetivo de compreender a influência entre os nós de uma rede, [Wang e Groth \(2010\)](#) criaram um *framework* que tem como objetivo medir o impacto de um conjunto de variáveis no desenvolvimento da rede. Os modelos propostos resultam em uma rede de influência, onde são apresentados os impactos das variáveis ao longo do tempo. Em um dos estudos de caso, os autores analisam uma rede de co-autoria, com o objetivo de compreender a influência entre os tópicos de pesquisa e os autores de projetos. Através da análise, foi constatado que um tópico popular em um ano, tende a ser menos popular no ano seguinte. Ou seja, os autores dos trabalhos da conferência estudada tendem a ser influenciados pelos tópicos de pesquisa que representam uma lacuna no conhecimento, e, quando conseguem auxiliar no preenchimento deste espaço, passam para outro desafio que ainda não foi completamente resolvido. Esse tipo de análise propicia uma nova fonte para compreensão do relacionamento entre os nós de uma rede.

Para entender como redes se desenvolvem e relacionamentos surgem, [Qiao et al. \(2014\)](#) criaram um modelo para simular a evolução de uma rede social. Esse tipo de técnica auxilia no entendimento da interação entre as estruturas da rede no processo. Os resultados da simulação mostraram que é possível reproduzir o comportamento das redes reais tanto em um nível macro, quanto em um nível micro. A partir do momento em que o modelo de simulação consegue se comparar com um modelo real, é possível entender os mecanismos de evoluções das redes sociais e analisar as diferenças e similaridades entre diferentes tipo de redes. Ainda tratando das redes sociais, [Stattner e Collard \(2015\)](#), utilizaram as estruturas topológicas das redes como fonte de conhecimento. Dentre os objetivos abordados, os autores buscam criar um modelo para a extração de comunidades nas redes sociais, que podem ser entendidos como grupos homogêneos de nós e arestas.

[Parand, Rahimi e Gorzin \(2016\)](#) destacam a importância da análise de redes sociais

e as possíveis estratégias para investigar estruturas sociais através de redes e teoria dos grafos. Os autores combinam duas técnicas utilizadas na literatura, inferência *fuzzy* e centralidade *eigenvector*, para obter a importância de nós específicos na rede. Ainda baseado na centralidade, porém dessa vez na centralidade das arestas, [Li et al. \(2019\)](#), utilizam essa métrica para fazer a detecção de comunidades em redes sociais. De acordo com os autores, encontrar, em uma rede, grupos de nós que possuem interesses em comum é um tópico relevante para os pesquisadores devido a sua aplicação em *e-commerces* e redes sociais.

2.2 Predição de ligações

[Liben-Nowell e Kleinberg \(2003\)](#) são pioneiros na área de predição de ligações em redes. Seus estudos são até hoje considerados o ponto de partida para este campo. O tema é introduzido com foco nas redes sociais e no seu dinamismo. Ao longo do tempo, novas arestas são adicionadas às redes, o que representa o surgimento de novas interações na estrutura social. Os autores definem o problema da predição de *links* como: dada uma rede social em um momento t , o objetivo é prever, com precisão, as arestas que serão adicionadas à rede durante o intervalo t e um tempo futuro t' . Além disso, é realizada a seguinte pergunta: até onde a evolução de uma rede social pode ser modelada usando características intrínsecas da própria rede? A predição de *links*, ou ligações, nesse contexto, permite descobrir indivíduos que já estão trabalhando juntos, porém a sua interação ainda não foi diretamente observada ([KREBS, 2002](#)). Ao longo do trabalho, são utilizados diversos métodos para tentar prever as arestas faltantes em diferentes redes. A performance dos preditores escolhidos é bem baixa, e, portanto, os autores destacam que ainda existe um longo trabalho a ser realizado nesta área.

Ainda em 2003, [Adamic e Adar \(2003\)](#) realizam um estudo a fim de investigar a partir de qual fonte de informação seria possível de indicar relacionamentos entre usuários. Os autores consideraram os seguintes pontos: textos similares nas páginas pessoais; *links* referenciando a página de outros usuários; *links* recebidos, ou seja, quem escreveu sobre determinada pessoa em sua página pessoal; e, por último, listas de *e-mail*. Cabe ressaltar que a Internet no ano de 2003 não apresentava as mesmas características de hoje. Portanto, essas foram as fontes de informação disponíveis para compreender os relacionamentos em uma rede. Ao longo do trabalho em questão são realizadas diversas etapas para entender a ligação de um usuário com outro. Para isso, o autor se refere ao problema como "predição de relacionamentos", e, utiliza um ranqueamento de pessoas similares para prever as ligações faltantes. Ao final do estudo, uma parcela dos indivíduos participantes recebeu a listagem das pessoas mais similares a eles, e, frequentemente, os reconheciam. O autor salienta que o grande desafio de tais análises é possuir apenas um pequeno conjunto de dados, que representa uma ínfima porção dos dados reais.

É importante compreender que ambos os trabalhos de [Liben-Nowell e Kleinberg \(2003\)](#) e [Adamic e Adar \(2003\)](#) foram publicados no mesmo ano, e sugerem o mesmo tópico de pesquisa em aberto.

Posteriormente, [Hasan et al. \(2006\)](#), através de técnicas de aprendizado de máquinas, desenvolvem um modelo de classificação responsável por identificar o conjunto de autores que possuirão um trabalho em conjunto no futuro. O autor destaca que esse tipo de trabalho pode auxiliar em diversas áreas do conhecimento, como, por exemplo: encontrar nós escondidos em uma rede social formada por terroristas, e analisar como uma doença, notícia ou vírus se espalharão pela rede. São utilizadas duas bases de dados, uma contendo 156.561 autores, outra contendo 1.564.617. Como resultados, os autores atingiram uma precisão de 90,56% para a primeira base de dados, e, 83,18% para a segunda. Essa diferença é explicada pelo fato da primeira base ser mais balanceada do que a segunda. Outro fator importante é que a primeira rede utilizava um período de 5 anos, já a segunda possuía 15 anos de publicações. Dessa maneira, é possível que o interesse dos pesquisadores tenha mudado ao longo deste período, influenciando no comportamento do algoritmo.

Já [Zhou, Lü e Zhang \(2009\)](#) tem como objetivo responder a seguinte questão: é possível encontrar um meio-termo que garanta uma alta precisão e utilize um baixo poder computacional no problema da predição de ligações? Para isso, é utilizado um *framework* simples, que tem como base a similaridade dos nós. O *framework* é aplicado em seis redes diferentes e nove métricas locais são testadas para chegar a uma conclusão. Como resultado, a característica mais simples, chamada de vizinhos em comum, possui a melhor performance. Além disso, é proposta uma nova métrica chamada *Resources Allocation(RA)*, motivada pelo processo de alocação de recursos. O autor destaca que outras medidas podem oferecer predições melhores, porém o custo computacional seria mais alto.

[Lü e Zhou \(2009\)](#) demonstram que a área de predição de ligações tem sido bastante estudada ao longo dos anos, mas a maioria dos estudos não leva em consideração o peso das arestas. Neste estudo, são utilizadas três redes ponderadas diferentes, duas de co-autoria e uma do sistema de transporte aéreo dos Estados Unidos. Nas duas primeiras redes, o peso das arestas é referente a quantidade de artigos publicados em conjunto. Já na rede de transportes aéreos, o peso é a frequência com que voos ocorrem de um aeroporto ao outro. Para a surpresa dos autores, ao se utilizar uma rede ponderada, a performance é pior do que utilizando grafos não ponderados. Tal descoberta vai de acordo com a teoria dos laços fracos, proposta por [Granovetter \(1977\)](#); os "laços fracos", em algumas redes, não são tão fracos quanto os pesos sugerem.

Inspirados no conceito da homofilia — que diz que pessoas com interesses similares tendem a se conectar uns com os outros, e também, pessoas com interesses similares possuem uma maior probabilidade de se tornarem amigos — [Yang et al. \(2011\)](#) apresentam

um modelo que leva em conta a propagação das amizades e interesses na rede, integrando, assim, o aprendizado por interesses e a predição de ligações em um único processo. Esse modelo realiza a junção de uma rede de amizades e uma de interesses, de forma a conectar um usuário aos itens de potencial interesse, e, posteriormente, realizar a ligação entre usuários com comportamentos similares. Ainda utilizando os mesmos conceitos, [Gong et al. \(2011\)](#) trabalham de forma a utilizar simultaneamente a estrutura da rede e as informações dos atributos dos nós para melhorar a performance dos resultados da predição de ligações. O trabalho é motivado pela interação observada e a homofilia encontrada entre a estrutura da rede e os atributos dos nós. Os autores conseguem demonstrar que é possível inferir atributos via predição de ligações. Para isso, um modelo de análise de redes sociais é utilizado, e, através dessa metodologia, a performance do estado da arte é alcançada. Ao combinar a inferência de atributos com a predição de ligações, uma precisão ainda mais alta é atingida.

Por ter ganhado a atenção dos pesquisadores, trabalhos que objetivam analisar o problema da predição de ligações mais profundamente também foram realizados, como é o caso de [Lichtnwalter e Chawla \(2012\)](#), que pesquisam os maiores desafios ao se trabalhar nessa área. Entre eles podemos destacar: (i) extração dos dados para teste e treino, uma vez que é necessário manter a mesma estrutura da rede; (ii) direcionalidade; caso seja utilizado um grafo direcionado, os resultados podem sofrer um grande impacto; (iii) desbalanceamento de classes, ou seja, a proporção de conexões existentes é muito pequena se comparada com todas as possibilidades possíveis. Tais fatores impossibilitam a comparação entre os estudos existentes. Portanto, o objetivo deste trabalho é proporcionar referencial relacionado à esses desafios, bem como uma série de padrões e recomendações para aqueles que estejam realizando uma pesquisa relacionada à área.

Analisando a importância dos atributos topológicos das redes ao se realizar um trabalho de predição de ligações, [Hu et al. \(2012\)](#) destacam que esses métodos podem propiciar o descobrimento de vestígios ou evidências importantes a respeito dos mecanismos que controlam a evolução das redes. Em virtude da dificuldade de aquisição de informações relacionadas ao domínio, normalmente informações relacionadas a topologia da rede em si são utilizadas ao se trabalhar com o problema da predição de ligações. Os autores salientam que a performance da metodologia proposta depende fortemente da heterogeneidade dos graus da rede. Se todos os nós possuírem basicamente o mesmo grau, conseqüentemente serão geradas predições ruins. Isso indica a importância de se considerar a topologia como fator fundamental para a predição de ligações.

Em um trabalho similar ao desta dissertação, [Digiampietri et al. \(2015\)](#) tem como objetivo realizar a predição de ligações na rede de colaboração científica formada pelos pesquisadores dos programas de pós graduação em Ciência da Computação. Foram utilizados os currículos de 657 pesquisadores, que atuaram nos triênios de 2004-2006 e

2007-2009. São combinados atributos estruturais da rede e também atributos específicos do domínio. Após utilizar diversos módulos para resolução dos problemas relacionados a predição de ligações citados acima, o autor alcança uma precisão média acima de 99%. Como é destacado no trabalho, além das considerações citadas anteriormente, três outros fatores podem ter impactado positivamente os resultados: (i) foram utilizados atributos estruturais diferentes dos encontrados na literatura; (ii) foram utilizados atributos específicos do domínio; (iii) os dados utilizados nos testes representam um conjunto homogêneo, e todos os professores trabalham nos programas de pós graduação na mesma área há dois triênios.

Utilizando os mesmos conceitos relacionados à predição de ligações, é possível analisar os nós mais influentes de uma rede. Levando isso em consideração, [Daud et al. \(2015\)](#) tentam encontrar indivíduos em ascensão em uma rede através da utilização de uma abordagem de aprendizado de máquinas. Mais especificamente, o objetivo é prever futuros nós em ascensão no domínio das redes de co-autoria. O resultado do classificador deve ser: "O acadêmico em questão tem potencial para se tornar uma estrela em ascensão ou não". São utilizados os dados DBLP¹, com a inclusão da base de dados Arnetminer². Foram selecionados 44.167 pesquisadores, onde 1.000 instâncias foram criadas. O autor também realiza o balanceamento das classes escolhendo nós aleatórios que deixam as instâncias utilizadas balanceadas. Como conclusão, a quantidade de dados influenciou positivamente os resultados, indicando que quanto maior o conjunto de dados, maior a precisão encontrada. A lista dos autores ranqueados é fiel ao mundo real. A aplicação deste conceito, utilizando de conhecimentos do domínio da base de dados, foi considerada funcional.

Ainda utilizando o conceito de influência, [Wu et al. \(2019\)](#) propõem uma nova métrica, tendo como base características locais e globais. Para tanto, um *framework* baseado na similaridade dos nós, que leva em consideração a influência de cada um sobre a rede, é criado. Dessa forma, sabendo quais nós são mais importantes para a rede como um todo, pode ser realizado um aperfeiçoamento das métricas locais. A proposta é testada em redes ponderadas e não ponderadas, e, por fim, é validada.

Realizando uma revisão sistemática na área de predição de ligações, ([MARUYAMA; DIGIAMPIETRI, 2019b](#)) destacam que entre os anos de 2010-2011 houve um aumento de 71% nas publicações a respeito do assunto. Neste contexto, [Lü e Zhou \(2011\)](#) tem como objetivo apresentar uma discussão detalhada dos novos desenvolvimentos na área da predição de ligações. De acordo com o autor, muitos trabalhos não levaram em conta o estudo de redes complexas, desconsiderando inclusive, as características estruturais das redes, que podem gerar informações úteis para a descoberta de ligações faltantes. Em

¹ *Digital Bibliography & Library Project*

² <https://aminer.org/>

contraponto, alguns trabalhos desenvolvidos por físicos que utilizam estudos probabilísticos apresentam resultados promissores. Ainda buscando compreender as melhores técnicas para o problema da predição de ligações, [Hasan e Zaki \(2011\)](#) destacam que compreender toda a evolução de uma rede é uma tarefa complexa, mas, entender a associação entre dois nós específicos é mais simples. Para tanto, algumas perguntas podem ser feitas: Como o padrão de ligações muda com o tempo? Quais fatores orientam essas ligações? Como a ligações entre dois nós é afetada por outros nós? Para responder as perguntas, o autor utiliza a formulação do problema dada por [Liben-Nowell e Kleinberg \(2003\)](#) e realiza uma pesquisa das abordagens existentes com foco, principalmente, em grafos formados por redes sociais.

2.3 Colaboração científica e a Plataforma Lattes como fonte de dados

[Newman \(2001\)](#) apresenta um estudo de uma rede de colaborações científicas composta por um milhão de pesquisadores e seus trabalhos publicados. São utilizados quatro conjuntos de dados: pesquisadores na área bio-médica, pesquisadores de física teórica, pesquisadores da área de física de altas energias, e, por último, uma rede de pesquisadores da área de ciência da computação. O autor analisa diversas propriedades das redes sociais, que serão explicadas no [Capítulo 3](#), como, por exemplo: os graus de separação entre os nós da rede, o agrupamento dos nós, entre outros. De acordo com o autor, as comunidades científicas constituem um "pequeno mundo". Também foi observado que algumas as redes formadas por algumas áreas são altamente agrupadas, enquanto outras não. Nesse caso, as propriedades das comunidades científicas podem influenciar este comportamento. Outro ponto a ser destacado é que o trabalho apresentado consiste apenas uma primeira análise das redes de colaboração citadas.

Ainda utilizando as redes de colaboração científica, [Newman \(2004\)](#) apresenta um estudo a respeito de três redes em particular: biomedicina, física, e, matemática. O autor apresenta diversas características das redes de colaboração científica, e realiza diversas análises para compreender o comportamento dos nós nessa rede. Existem similaridades e diferenças entre elas, como, por exemplo, todas parecem possuir uma distribuição similar com relação aos coautores. A maioria possui um pequeno grupo de pesquisadores, que trabalham juntos há algum tempo, enquanto um pequeno grupo possui centenas, ou as vezes, milhares de coautores. É destacada a importância das redes de colaboração científica, e como elas apresentam informações meticulosas, bem documentadas e até mesmo com eventos temporais do relacionamento social e profissional dos cientistas.

Buscando destacar a importância da colaboração científica, [Abbasi, Altmann e Hwang \(2010\)](#) realçam que a colaboração leva a trabalhos melhores, aumentando o impacto

dos resultados, diminuindo os recursos utilizados, entre outros. Portanto, avaliar os pesquisadores, com base em sua produtividade se tornou um fator importante. Com esse objetivo, o autor propõe a criação de um índice para quantificar as atividades de colaboração dos pesquisadores e outro para as atividades de colaboração na comunidade. Tais índices levam em conta a quantidade de resultados gerados pela pesquisa, e, também, a qualidade destes resultados, para indicar o quão importante é a colaboração de determinado pesquisador. Através da aplicação dos índices criados em uma rede contendo 1.815 autores e 2.139 publicações, foi possível identificar a importância dos pesquisadores, bem como o quanto de colaboração existe para os pesquisadores fora da comunidade. Os índices propostos pelos autores do trabalho também são mais precisos do que os utilizados na literatura apresentada na época. Apesar disso, possuem duas limitações: exigem uma quantidade maior de informações na rede e são mais caros computacionalmente. Apesar disso, os índices apresentados no artigo em questão oferecem um primeiro auxílio ao identificar pesquisadores que são adequados a liderar um projeto de pesquisa, por exemplo.

Já utilizando a Plataforma Lattes, [Digiampietri et al. \(2012\)](#) abordam a importância deste repositório no compartilhamento do conhecimento científico. O objetivo do trabalho é realizar a extração da maior quantidade possível de currículos. Para isso, os autores dividem o trabalho em duas etapas, na primeira delas, são realizadas 80 buscas na própria Plataforma Lattes por diferentes áreas do conhecimento. Nesse momento, são obtidos os identificadores de cada um dos currículos retornados nessa busca, que posteriormente, durante a segunda etapa, serão utilizados para extração dos dados correspondentes. Através dessa estratégia, foram extraídos 1.236.548 currículos, que foram analisados, estudados e utilizados como conjunto de dados para caracterização da rede de colaboração científica obtida. Como conclusão, os autores destacam a grande quantidade de informações que podem ser utilizadas tendo como base a Plataforma Lattes, afirmando que esta é uma das bases mais completas do mundo.

[Perez-Cervantes et al. \(2013\)](#) apresentam uma abordagem que tem como objetivo obter a influência de um indivíduo em uma rede de colaboração. Para tanto, é utilizado um preditor de ligações baseado em métricas locais da estrutura da rede. A influência individual colaborativa é obtida levando em consideração a influência de um determinado pesquisador na predição de ligações da rede como um todo.

Em [Perez-Cervantes \(2015\)](#) a descrição dos métodos utilizados no trabalho citado anteriormente é detalhada. São utilizados dados de 47.555 pesquisadores da Plataforma Lattes, que foram obtidos através da utilização do *ScriptLattes* ([MENA-CHALCO; JUNIOR, 2009](#)). Como resultado, as medidas da influência colaborativa apresentam correlação inversa significativa quando comparadas com as medidas de centralidade mais conhecidas. Tal fato demonstra a efetividade das métricas propostas. Outro fator importante é que a metodologia descrita pode ser calculada independentemente para cada vértice, sem a necessidade de

um cálculo global, reduzindo, assim, o custo computacional.

Dias et al. (2013) descrevem uma abordagem para extração dos currículos da Plataforma Lattes e construção de uma rede de colaboração científica. A relação entre os colaboradores é realizada através da presença de um ou mais trabalhos em conjunto. Através do *framework* construído, são apresentadas redes que possuem palavras-chave em comum, participaram de um mesmo congresso ou até mesmo de uma mesma área. Já em Dias et al. (2014), os autores apresentam um método detalhado para formação das redes e análise das propriedades presentes nelas. Além das redes abordadas em Dias et al. (2013), são geradas as redes de um programa de pós graduação específico, redes de autores que utilizaram as mesmas palavras chaves, e, também, uma rede formada apenas por palavras-chave. De acordo com os autores, esta última rede possui uma grande relevância, pois representa o quão importante uma palavra-chave é no conjunto de dados analisados e quais seus principais vínculos. Também é importante destacar que o algoritmo utilizado para criação das redes de colaboração possui um custo computacional mais baixo do que os outros métodos propostos anteriormente na literatura. Como resultado, é possível se construir grandes redes, contendo milhões de registros.

Com o objetivo de identificar como ocorre a colaboração científica entre os principais pesquisadores do Brasil, Dias, Moreira e Dias (2018) realizam um estudo utilizando os currículos dos bolsistas de produtividade. Tal bolsa é paga pelo CNPq pela excelência nas pesquisas e trabalhos desenvolvidos. Dessa forma, os bolsistas pertencem a um seleto grupo de pesquisadores que se destacam na realização de pesquisas científicas. Através da análise conduzida, é possível perceber, por exemplo, que nas modalidades mais altas de bolsas, a colaboração entre os cientistas ocorre de forma mais intensa. Também é possível analisar os tipos de publicações (anais de congresso ou periódicos) dadas as modalidades. Diante disto, é destacada a importância da colaboração científica para o avanço do conhecimento.

Tendo em vista os estudos aqui apresentados, este trabalho se destaca na utilização dos conceitos de análise de redes sociais para predição de ligações, utilizando um conjunto de dados, até então, pouco explorado pelos pesquisadores.

2.4 Resumo do Capítulo

Neste capítulo foram apresentadas as principais referências utilizadas para o desenvolvimento do trabalho aqui realizado. Inicialmente, foram apresentados alguns trabalhos sobre análise de redes sociais, com o objetivo de fornecer uma base a respeito do domínio do problema a ser resolvido. Posteriormente, foram abordados trabalhos que tiveram como objetivo a predição futuras ligações em redes. A partir da apresentação de tais conceitos, o objeto de estudo deste trabalho, redes de colaboração científica, é abordado. Alguns

trabalhos que analisaram tais redes são apresentados, bem como a Plataforma Lattes, fonte de dados a ser utilizada.

3 Fundamentação Teórica

Neste capítulo, será apresentada a fundamentação teórica referente aos temas abordados ao longo da dissertação. Inicialmente, a [Seção 3.1](#) diz respeito a conceitos relacionados a redes, explicitando como estas podem ser representadas, bem como suas diferentes características e propriedades. Tais conceitos se fazem importantes, uma vez que algumas propriedades das redes serão utilizadas como atributos para a predição de ligações. Posteriormente, na [Seção 3.2](#), o problema da predição de ligações será explicado formalmente, onde serão elencadas as diferentes abordagens para resolução deste e vantagens e desvantagens de cada uma delas. Por fim, a [Seção 3.3](#) apresenta as técnicas de aprendizado de máquinas utilizadas no trabalho.

3.1 Análise de redes Sociais

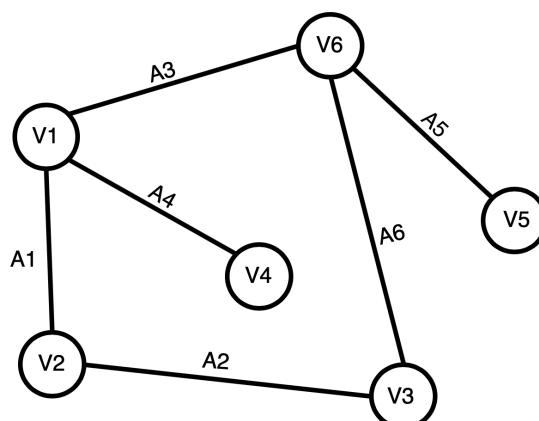
A análise de redes sociais (ARS) tem como objetivo entender a complexidade social por meio da representação das redes através da utilização de grafos ([POTGIETER et al., 2009](#)). Do ponto de vista da ARS, o ambiente social pode ser representado pelos padrões ou regularidades nos relacionamentos entre unidades que interagem entre si ([WASSERMAN; FAUST, 1994](#)). A predição de ligações influencia crucialmente no tráfego e nos rendimentos de uma rede social, e normalmente é reconhecida como uma das tarefas mais importantes na análise de redes sociais ([YANG et al., 2011](#)). Neste contexto, grafos podem ser utilizados como ferramenta de representação visual das redes.

Uma rede pode ser representada como um grafo $G = (V, A)$, que consiste em um conjunto finito, não vazio de vértices $V = V(G)$ e um conjunto $A = A(G)$, formado por pares de elementos de V , chamados de arestas de G ([BONDY; MURTY et al., 1976](#)). A [Figura 1](#) apresenta um grafo, onde os pontos $V1, V2, \dots, Vn$ representam os nós; no cenário específico deste trabalho, estes representariam os pesquisadores que possuem seu currículo cadastrado na Plataforma Lates, e as arestas $A1, A2, \dots, An$ representariam a ligação entre os nós. Nesse caso, as arestas podem ser entendidas como uma colaboração científica entre dois pesquisadores.

Como pode ser visualizado na [Figura 1](#), os vértices $V1$ e $V2$ estão ligados pela aresta $A1$. Neste contexto, os dois vértices representariam dois autores, que juntos publicaram o artigo $A1$. Da mesma forma, as demais arestas e vértices representam autores e seus artigos publicados, formando, a partir disso, uma rede de colaboração científica.

Em uma das contribuições mais importantes para a área, [Wasserman e Faust \(1994\)](#) apresentam um vasto estudo a respeito de características, propriedades e conceitos

Figura 1 – Exemplo de um grafo



Fonte: Autoria Própria

utilizados para análise de redes sociais. Eles serão apresentados a seguir:

- **Ator:** Os atores e seus relacionamentos são elementos fundamentais na constituição de uma rede social (STRÖELE; ZIMBRÃO; SOUZA, 2012). Alguns exemplos são: pessoas em um grupo, departamentos de uma empresa, agências de serviço público em uma cidade, ou nações ao redor do mundo. Apesar da sugestão trazida pelo termo, os atores não possuem, necessariamente, a habilidade de agir. A maioria das redes sociais focam em conjuntos de atores que são todos do mesmo tipo;
- **Ligações Relacionais:** Atores estão ligados uns aos outros por ligações relacionais. A característica mais importante destes objetos é que eles estabelecem uma ligação entre um par de atores. Durante o restante do trabalho, as ligações relacionais serão referidas apenas como ligações;
- **Díade:** A díade é constituída por um par de atores e uma, possível, relação entre eles. Por estabelecer um vínculo entre dois atores, uma ligação é considerada uma propriedade do par, e não de cada um individualmente.
- **Triade:** Também é possível realizar uma análise levando em consideração um conjunto maior de atores. Muitos métodos de análise de redes sociais focam nas tríades, que podem ser entendidas como um subconjunto de três atores e as, possíveis, ligações entre eles;
- **Subgrupo:** Pode-se definir um subgrupo como qualquer subconjunto de atores e todas as ligações entre eles;
- **Grupo:** Ao realizar a análise de redes, não apenas os relacionamentos entre díades, tríades ou subgrupos que importam, existem outros elementos a ser considerados. Um grupo é uma coleção de todos os atores nos quais os relacionamentos devem ser medidos. Em alguns casos, um grupo pode ser constituído por um conjunto finito

de atores, que são tratados separadamente por alguma razão. No presente trabalho, os grupos são formados pelas áreas do conhecimento, onde apenas o grupo de pesquisadores pertencentes àquela determinada área é analisado;

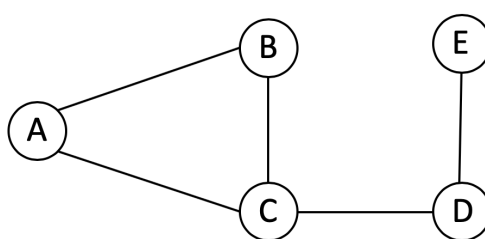
- **Relação:** A relação se refere ao conjunto de vínculos de um tipo específico medido em pares de atores de um determinado grupo;

Após entender o que é um ator, um grupo, e uma relação, pode-se compreender melhor a definição de rede social. Uma rede social consiste de um conjunto finito de atores e o seu relacionamento. A presença de informação relacional é uma característica crítica e precisa da rede social.

Devido às suas diferenças nos padrões de correlação entre os graus dos vértices adjacentes e os altos níveis de agrupamento encontrados, as redes sociais são fundamentalmente diferentes de outros tipos de sistemas em redes (NEWMAN; PARK, 2003). Portanto, serão apresentadas algumas propriedades específicas das redes sociais, que serão posteriormente utilizadas para predição de ligações.

- **Graus de Separação:** É a distancia mínima, em termos de números de nós na rede, entre todos os pares de nós na rede, pelos quais uma conexão existe (NEWMAN, 2001). Pode-se entender esse conceito como a quantidade de passos necessários para fazer com que o pesquisador "A" consiga interagir com o pesquisador "E", conforme demonstrado na Figura 2. Nesse caso, seria necessário passar pelos nós C e D para posteriormente chegar ao nó E.

Figura 2 – Graus de Separação



Fonte: Autoria Própria

- **Assortividade:** Segue o princípio de que tendemos a ser semelhantes aos nossos amigos. Dessa forma, vistos coletivamente, um grupo de amigos possui características em comum (PEREZ-CERVANTES, 2015). Também pode ser entendido como grau de correlação entre os nós e é apresentado numericamente, entre 1 e -1, onde um valor positivo indica que pesquisadores com mais colaboradores tendem a colaborar com outros que também possuem uma grande quantidade de colaboradores. Já um valor negativo indica o oposto (NEWMAN, 2004). Diversas métricas podem ser utilizadas

para realizar esse cálculo. Normalmente, na teoria dos grafos, utiliza-se o grau dos vértices. Outra forma de compreender a assortividade é pensar no trabalho em equipe, ou seja, uma equipe com muitos membros tende a colaborar com uma outra equipe que possui um número de membros similar.

- **Mundo pequeno:** [Milgram \(1967\)](#) apresenta o primeiro estudo destinado a entender o fenômeno conhecido como "Mundo Pequeno". Neste primeiro estudo, o autor chega a conclusão de que, nos Estados Unidos, as pessoas estão a 5 passos uma das outras. Buscando entender o efeito deste fenômeno em outras redes, [Kleinberg \(2000\)](#) explica que o efeito mundo pequeno possui algumas características diferentes a depender da rede sendo estudada. A partir destes estudos, pode-se concluir que, na maioria das redes, a distância típica entre os nós é surpreendente pequena ([NEWMAN, 2018](#)).
- **Homofilia:** É o princípio que diz que uma interação entre pessoas com interesses similares ocorre com uma frequência maior do que com pessoas com interesses diferentes ([MCPHERSON; SMITH-LOVIN; COOK, 2001](#)). Este é um dos mecanismos fundamentais que dirigem o dinamismo das redes sociais ([YANG et al., 2011](#)). Nas redes de colaboração científica, existem diversas formas de identificar a similaridade entre os nós, como, por exemplo, a participação nos mesmos eventos científicos, publicação de trabalhos nos mesmos periódicos, entre outros.
- **Componente Gigante:** Conforme explicado por [Newman \(2004\)](#), essa propriedade normalmente ocorre em grandes redes, onde existe um grande número de conexões, e um grupo de indivíduos que estão todos conectados entre si através de nós intermediários.
- **Agrupamento:** Uma das características que diferencia as redes sociais de outras redes é que elas possuem grupos, também chamados de comunidades locais, nas quais uma média maior de nós possuem arestas entre si (um número maior de pessoas se conhecem) ([NEWMAN, 2001](#)). O coeficiente de agrupamento mede a probabilidade de que dois nós possuam uma aresta entre si ([NEWMAN, 2004](#)).

Com base nos trabalhos citados nesta seção, foi possível aprofundar os estudos relacionados a redes, gerando interesse pelo dinamismo apresentado por elas. Dessa forma, pesquisadores buscaram entender como novos nós são adicionados à rede, e, também, como novas ligações surgem, fazendo com que os primeiros trabalhos a respeito da predição de ligações fossem iniciados.

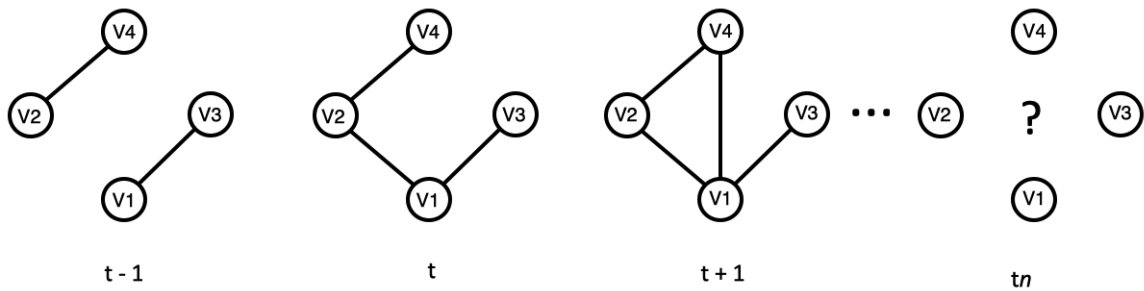
3.2 Predição de ligações

Ao se realizar a análise de redes sociais, um dos grandes desafios é compreender como se dá a evolução destas. Entretanto, entender este dinamismo é uma tarefa complexa, devido ao grande número de parâmetros envolvidos ([HASAN; ZAKI, 2011](#)). Dessa forma, é normal que este problema seja sub-dividido em situações menores, estudadas por diferentes

áreas, sendo que uma destas busca compreender como ocorre a ligação entre dois nós da rede. Este problema, em específico, é chamado de predição de ligações, e pode ser descrito como: dado um grafo G , em um momento t , o objetivo é prever, com precisão, as arestas (A), que serão adicionadas à rede durante o intervalo t e um momento futuro t' . A partir da resolução deste problema, é possível construir um sistema de avaliação específico para analisar a evolução de determinadas redes (HU et al., 2012). A dificuldade do problema reside no dinamismo típico destes objetos, cujos nós e ligações mudam constantemente (PEREZ-CERVANTES, 2015).

O problema da predição de ligações é apresentado na Figura 3. No momento $t - 1$ pode-se perceber a presença de quatro vértices, e duas arestas, conectando os nós $V2$ e $V4$, e, $V1$ e $V3$. Já no instante t , uma nova ligação é realizada, estabelecendo uma conexão entre os nós $V1$ e $V2$. No momento $t + 1$, pode-se perceber que ocorreu mais um relacionamento, ligando, assim, os nós $V4$ e $V1$. O objetivo, portanto, é utilizar o conhecimento obtido nos instantes $t - 1$, t , e, $t + 1$ para prever as arestas presentes em um tempo futuro n .

Figura 3 – Problema da predição de links



Fonte: Autoria Própria

Seja $x, y \in V$ nós do grafo $G(V, E)$ e $l^{(x,y)}$ a existência ou não de uma aresta entre os nós (x,y) (TANG, 2017). Uma ligação é definida por:

$$l^{(x,y)} = \begin{cases} 1 & \text{se } (x,y) \in E \\ 0 & \text{se } (x,y) \notin E \end{cases} \quad (1)$$

Sendo assim, obtêm-se o valor 1 quando a ligação existir no grafo e zero em caso negativo. Esse é um problema típico de classificação binária e muitas técnicas de aprendizado de máquinas podem ser usados para resolvê-lo (WANG et al., 2015).

Existem diversas aplicações para a predição de ligações, como prever quais páginas da Internet serão visitadas em um determinado dia dado um histórico de navegação, os locais para os quais um indivíduo irá viajar em determinado mês dado o seu histórico de

viagens, entre outros (ACAR; DUNLAVY; KOLDA, 2009). Liben-Nowell e Kleinberg (2003) também destacam que essa área de pesquisa pode ser utilizada para sugerir interações promissoras ou colaborações que ainda não foram descobertas em uma determinada organização. Já na área da segurança, é possível utilizá-la para identificar grupos de terroristas ou criminosos (HASAN; ZAKI, 2011).

Diversas abordagens diferentes são utilizadas para resolução deste problema. Alguns utilizam métodos de caminhada aleatória ou algoritmos para grafos espectrais (YANG et al., 2011). De acordo com Hasan e Zaki (2011), as abordagens existentes podem ser classificadas em diversos grupos: algoritmos que calculam a similaridade entre um par de nós, para que seja possível a utilização de um método de aprendizado supervisionado; outros utilizam modelos probabilísticos; e também existem os modelos baseados em formulações de álgebra linear.

Os primeiros trabalhos que buscavam realizar a predição de ligações utilizaram a similaridade entre os nós para calcular a proximidade entre eles, a fim de prever se estes, deveriam possuir uma ligação entre si (ADAMIC; ADAR, 2003). Essa abordagem utiliza do conceito de homofilia, que descreve que pessoas com interesses similares possuem mais chance de serem amigos (NEWMAN, 2018). A partir disso, é possível inferir informações topológicas da rede, e até mesmo calcular a similaridade entre um nó e todos os outros nós da rede.

Uma segunda abordagem, utiliza métodos probabilísticos para inferir a existência de ligações ou não. Diversos algoritmos diferentes podem ser utilizados para este fim (OLIVEIRA et al., 2014). A principal vantagem do uso desses modelos é que eles não dependem de um conjunto de dados específico, sendo mais genéricos, o que resulta numa aplicação mais vasta para diferentes cenários (PEREZ-CERVANTES, 2015).

Também é possível utilizar métodos baseados em álgebra linear para resolução do problema de predição de ligações. Nesse caso, o conjunto de dados é transformado em matrizes e tensores, que representam uma dimensão acima das matrizes. Conforme explicado por Acar, Dunlavy e Kolda (2009), este método possibilita levar a temporalidade em consideração, onde cada unidade de tempo pode ser transformada em uma matriz. Dessa forma, o conjunto de matrizes pode ser representado como um tensor. Nesse momento, técnicas relacionadas a álgebra linear podem ser utilizadas para se descobrir padrões em conjuntos de dados de diversas dimensões (DUNLAVY; KOLDA; ACAR, 2011).

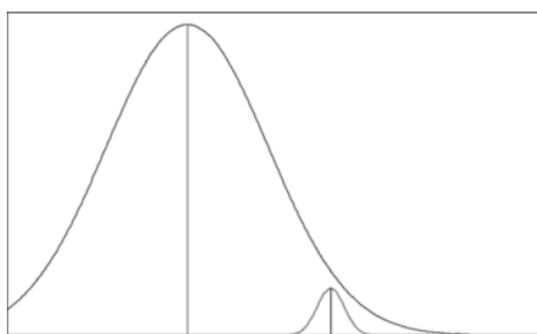
De acordo com Menon e Elkan (2011) e Lichtnwalter e Chawla (2012), existem três grandes desafios ao se realizar a predição de ligações: o primeiro deles é relacionado às informações topológicas do grafo, onde normalmente é necessário possuir informações adicionais a respeito dos nós. Já o segundo desafio está relacionado ao conjunto de dados normalmente utilizado para realizar a predição de ligações, que apresenta um extremo desbalanceamento de classes. Isso significa que o número de arestas presentes é, na

maioria das vezes, significativamente menor do que o número de arestas ausentes. Por fim, o terceiro desafio apresentado pelos, diz respeito à eficiência dos modelos computacionais propostos, uma vez que é necessário lidar com um grande número de nós e arestas. Por se utilizar do conjunto de currículos cadastrados na Plataforma Lattes, o primeiro desafio é resolvido a partir da utilização de extratores de informação específicos para tanto. Porém, ainda é necessário lidar com as outras duas situações. Ao longo do [Capítulo 4](#) serão abordadas as estratégias utilizadas para tornar o processo computacional eficiente, enquanto a questão do balanceamento de classes será melhor retratada abaixo.

3.2.1 Balanceamento de Classes

O número de ligações possíveis em um grafo é quadraticamente relacionado ao número de nós. No entanto, o número de ligações reais representa apenas uma pequena fração deste número ([HASAN; ZAKI, 2011](#)). De acordo com [Menon e Elkan \(2011\)](#), esta situação interfere nos resultados devido a duas razões: (i) com menos exemplos de uma determinada classe, é mais difícil inferir padrões confiáveis; (ii) os modelos treinados são enviesados em direção à classe predominante. Uma técnica tradicional para superar o desbalanceamento das classes é chamada de sob-amostragem, que consiste em reduzir o número de amostras da classe determinante, de forma randômica, igualando assim o número de componentes para ambos os casos.

Figura 4 – Desbalanceamento entre as classes



Fonte: [Rattigan e Jensen \(2005\)](#)

Diversos estudos apresentam o problema mais a fundo, como, por exemplo, [Rattigan e Jensen \(2005\)](#), que destacam o impacto do desbalanceamento de classes na precisão dos algoritmos. A [Figura 4](#) apresenta a grande diferença entre as arestas presentes e ausentes, onde o menor valor representa o número real de arestas existentes na rede, e o valor mais alto a quantidade de arestas ausentes. Nesses casos, os classificadores padrão tendem a ser sobrecarregados pela classe dominante, gerando resultados incorretos ([CHAWLA; JAPKOWICZ; KOTCZ, 2004](#)).

A partir das considerações aqui apresentadas, foi possível compreender melhor o problema da predição de ligações, bem como as diferentes abordagens que podem ser utilizadas ao estudá-lo. Dessa forma, a próxima seção apresenta o conjunto de técnicas escolhidas para tanto.

3.3 Aprendizado de Máquinas

Diversos autores apresentam diferentes definições a respeito do termo Aprendizado de Máquinas, ou *Machine Learning* (ML), em inglês. De acordo com Choudhary e Gianey (2017), ML concede aos sistemas a habilidade de aprender automaticamente e ganhar experiência sem a necessidade de serem explicitamente programados. Nesse contexto, Mohri, Rostamizadeh e Talwalkar (2018) explicam que, de uma forma geral, ML pode ser entendido como um método computacional que utiliza algum tipo de experiência para melhorar sua performance ou fazer predições melhores. Dessa forma, as experiências se referem a informações antigas. Também é possível entender ML como o processo de se adaptar a novas circunstâncias e detectar e extrapolar padrões (RUSSELL; NORVIG, 2016).

A partir da ampla utilização de técnicas relacionadas ao aprendizado de máquinas, Domingos (2012) destaca que a próxima onda de inovações será guiada pelas descobertas neste campo. Neste contexto, tarefas que envolvem reconhecimento, diagnóstico, planejamento, controle de robôs, predição, entre outros, já utilizam algumas destas técnicas (NILSSON, 1996). Entre os problemas que podem ser resolvidos a partir da utilização de ML, destacam-se:

- **Classificação:** Pode ser considerado um dos problemas mais frequentemente estudados e abordados em ML, e consiste em identificar a qual classe uma determinada observação pertence (DOMINGOS, 2012). Pode-se citar diversos exemplos, como, por exemplo, identificar se um *e-mail* é um *spam*; classificar um usuário de uma rede social como amigo, ou não; entre outros;
- **Regressão:** Enquanto o resultado de um algoritmo de classificação consiste em um número finito de categorias, o resultado da regressão consiste em um ou mais valores contínuos (BISHOP, 2006). Nesse caso, o objetivo é estimar um valor $y \in \mathbb{R}$ dado um padrão x (ALPAYDIN, 2014). Um exemplo de aplicação da regressão consiste em estimar o valor de um imóvel baseado no padrão dos imóveis da vizinhança.
- **Ranqueamento:** Tem como objetivo melhorar a ordem pelo qual os resultados são retornados a um determinado usuário, melhorando a relevância e eficiência destes (RICHARDSON; PRAKASH; BRILL, 2006). Como, por exemplo, títulos a serem assistidos em uma plataforma de vídeos, ou páginas da internet em uma plataforma de buscas.
- **Agrupamento:** Consiste na detectar potenciais grupos dentre as observações analisa-

das (RUSSELL; NORVIG, 2016). Alguns exemplos incluem agrupar usuários baseado a partir do seu padrão de compra, ou agrupar plantas de acordo com suas espécies;

- **Detecção de anomalias:** Consiste na tarefa de identificar comportamentos não esperados em um conjunto de dados (CAN; ALATAS, 2019). O caso mais comum para sua utilização é a identificação de fraudes ao conceder crédito.

Baseado no tipo de problema a ser resolvido, deve-se considerar o cenário nos quais é possível identificar algum deles. Estes cenários se diferem no que diz respeito aos tipos de dados, à ordem, aos métodos e às formas de avaliação dos algoritmos de aprendizado de máquinas (MOHRI; ROSTAMIZADEH; TALWALKAR, 2018). De acordo com Duda, Hart e Stork (2012), podem ser classificados três diferentes grupos:

- **Supervisionado:** O modelo recebe um conjunto de rótulos já identificados e realiza a predição para os dados ainda não vistos (MOHRI; ROSTAMIZADEH; TALWALKAR, 2018). Nesse caso, busca-se encontrar uma hipótese que explica o conjunto de dados já apresentado, e prediz quando um dado possui uma grande probabilidade de, também, se adequar aos rótulos ainda não conhecidos (NILSSON, 1996). A maioria dos problemas discutidos se encaixa nessa categoria.
- **Não Supervisionado:** Nesse caso, não existem categorias ou rótulos para o conjunto de dados. Na maioria das vezes, o objetivo é conseguir agrupar as observações da forma correta, de acordo com sua similaridade (NILSSON, 1996).
- **Por reforço:** Os algoritmos desse tipo estão preocupados com o problema de encontrar possíveis ações a serem tomadas, dada uma situação específica, onde o objetivo é maximizar a recompensa final (BISHOP, 2006).

Para que um algoritmo possa realizar a classificação de uma nova observação, uma série de passos se faz necessária. Conforme Duda, Hart e Stork (2012) descrevem, inicialmente um conjunto de dados deve ser obtido, para que, em seguida, seja processado. Após esse passo é possível realizar a extração das características, e então a classificação pode ser realizada.

3.3.1 Algoritmos de Classificação

Conforme citado anteriormente, os algoritmos de classificação utilizados quando se deseja identificar a qual classe uma determinada observação pertence. No caso de co-autorias de artigos científicos, existem duas possíveis classes: os autores irão escrever um artigo em conjunto ou não. Desta forma, podemos retratar essa situação em específico como uma classificação binária. Levando estes fatores em consideração, são analisados os principais algoritmos que podem ser utilizados neste tipo de problema.

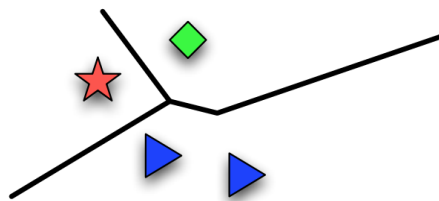
3.3.1.1 Método Geral

Ao se aplicar o aprendizado de máquinas em problemas de classificação, o objetivo é encontrar a qual categoria uma determinada observação pertence. Nesse contexto, pode-se utilizar um método padrão para qualquer algoritmo a ser apresentado. Suponha que existem duas distribuições de probabilidade $p(X|0)$ e $p(X|1)$, e dado um padrão, X , deseja-se utilizar um método estatístico para determinar se o mesmo pertence a categoria 0 ou 1 (NILSSON, 1996). O trabalho do classificador, nesse caso, é usar um padrão já aprendido e determinar, utilizando como base uma função de erro, a qual categoria uma nova observação deve pertencer (MOHRI; ROSTAMIZADEH; TALWALKAR, 2018).

3.3.1.2 K Vizinhos Mais Próximos:

Uma classe de métodos de aprendizado está relacionada a estatística, sendo chamados de métodos de vizinhos mais próximos, ou métodos baseados em memória (NILSSON, 1996). Dado um conjunto de treino t , contendo m padrões rotulados, um procedimento de vizinhos mais próximos decide que um novo rótulo, X , pertence à mesma categoria que os vizinhos mais próximos em t (NILSSON, 1996). Mais precisamente, um método de vizinhos mais próximos aloca uma nova observação, X , à categoria na qual a maioria dos seus vizinhos está contida (ALPAYDIN, 2014; NILSSON, 1996). Esse processo pode ser melhor compreendido a partir da Figura 5, onde a observação seria rotulada de acordo com a sua proximidade com o triângulo, estrela ou quadrado, sendo assim, uma dessas três opções (ALPAYDIN, 2014).

Figura 5 – Representação de um classificador de vizinhos mais próximos.



Fonte: Alpaydin (2014)

Alguns fatores se fazem importantes nesse momento, como, por exemplo, a quantidade de categorias existentes para uma nova observação ser alocada. Portanto, caso os dados possuam ruídos, as predições realizadas também possuirão ruídos, interferindo na qualidade dos resultados (ALPAYDIN, 2014).

Também é importante levar em consideração a medida utilizada para calcular a distância entre as observações. A mais simples delas (para atributos numéricos) é a

distância Euclidiana, e pode ser encontrada na [Equação \(2\)](#), onde x e y representam duas amostras do conjunto de dados.

$$distancia = \sqrt{\sum_{j=1}^n (x_j - y_j)^2} \quad (2)$$

Na [Equação \(2\)](#), os valores x e y representam os pontos em um plano cartesiano, e, após realizar o somatório de todas as distâncias elevadas ao quadrado, é feito o cálculo da raiz quadrada. Esse processo garante que os números positivos permaneçam inalterados, enquanto os números negativos são substituídos pelos seus valores absolutos.

3.3.1.3 Regressão Logística:

A Regressão Logística é uma abordagem matemática que pode ser utilizada para explicar a relação entre variáveis dependentes e independentes ([KLEINBAUM et al., 2002](#)). Esse método apresenta a mesma base da regressão linear, um modelo que utiliza situações de predição de valores, mas, para calcular a probabilidade condicional da observação pertencer a determinada categoria ([MOHRI; ROSTAMIZADEH; TALWALKAR, 2018](#)). O ponto mais importante deste algoritmo é função logística, apresentada no formato de uma curva em S, responsável por converter um valor qualquer do mundo real, em um valor entre 0 e 1 ([CHOUDHARY; GIANEY, 2017](#)).

Para tanto, a [Equação \(3\)](#) é apresentada. É importante compreender que o resultado irá variar entre 0 e 1, e este representará a probabilidade da observação pertencer a determinada categoria. A [Figura 6](#) apresenta o exemplo de uma curva formada a partir da utilização da [Equação \(3\)](#), também chamada de Função Sigmóide.

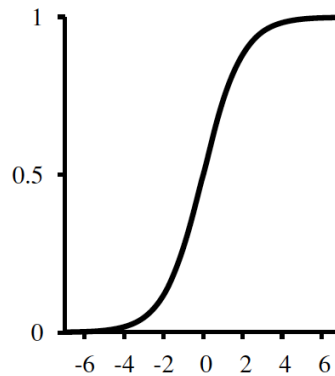
$$f(x) = \frac{1}{1 + e^{-x}} \quad (3)$$

Conforme pode ser observado na [Equação \(3\)](#), o elemento $1 + e^{-x}$ garante que resultado da equação irá se aproximar de 1, caso o valor x seja positivo, e irá se aproximar de 0, caso o valor x seja negativo.

3.3.1.4 Florestas Aleatórias:

Apresentado inicialmente por [Breiman \(2001a\)](#), o classificador Florestas Aleatórias consiste em um conjunto de diferentes classificadores estruturados no tipo árvore, onde cada unidade vota na classe mais popular para uma determinada entrada de dados. Cada nó da árvore, ou cada decisão, é baseada em um sub conjunto aleatório de atributos disponíveis, levando a resultados mais precisos e robustos ([GUNS; ROUSSEAU, 2014](#)). Ao fim, a predição é feita levando em consideração a maioria dos votos obtidos pelas árvores de decisão ([CHEN; LIN, 2006](#)).

Figura 6 – Função Sigmóide



Fonte: [Russell e Norvig \(2016\)](#)

[Choudhary e Gianey \(2017\)](#) detalham o passo a passo do algoritmo Florestas Aleatórias:

1. Usando os dados disponíveis, retire n amostras;
2. Para cada amostra, crie uma árvore com as seguintes alterações: selecione um conjunto randômico de preditores para cada nó, e escolha a melhor divisão entre os parâmetros;
3. Estime o resultado desejado através da acumulação das predições das n árvores criadas anteriormente.

Além de ser utilizado como classificador, o algoritmo Florestas Aleatórias também pode ser utilizado para encontrar os atributos mais relevantes no conjunto de dados.

3.3.1.5 Naive Bayes

Uma categoria específica dos métodos de aprendizado de máquinas está relacionada à probabilidade. O caso a ser abordado é o teorema de Bayes. Através desse método é possível identificar as causas de um determinado efeito, analisando apenas o efeito ([RUSSELL; NORVIG, 2016](#)). Este atesta que a presença, ou ausência, de uma característica específica de uma classe não esta relacionada presença, ou ausência, de qualquer outra característica ([LIU et al., 2011](#)).

Para tanto, a [Equação \(4\)](#) é utilizada. Desta forma, é possível identificar a probabilidade de A, dado B, possuindo apenas os outros termos da equação. No caso do algoritmo Naive Bayes, esta fórmula é utilizada com o objetivo de identificar qual a probabilidade de uma nova observação da base de dados ser pertencente a uma determinada classe.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (4)$$

3.4 Resumo do Capítulo

Neste capítulo foram apresentadas as fundamentações teóricas necessárias para a compreensão do restante do trabalho. Inicialmente, foram descritos os conceitos relacionados às redes sociais, bem como as suas propriedades. Posteriormente, o problema da predição de ligações foi demonstrado, abordando também o problema do desbalanceamento de classes. Por fim, uma pequena explicação a respeito dos algoritmos de aprendizado de máquinas, que serão utilizados no trabalho, foi realizada. A partir destes tópicos será possível compreender a metodologia utilizada no próximo capítulo.

4 Metodologia

Para que seja possível atingir os objetivos propostos neste trabalho, alguns passos se fazem necessários. Dessa forma, neste capítulo serão destacados os métodos utilizados para que a predição de futuras ligações seja realizada. Para tanto, oito áreas do conhecimento foram selecionadas, chegando a um total de 71.010 pesquisadores que possuem seus currículos cadastrados na Plataforma Lattes, responsáveis por 5.823.173 de colaborações científicas entre os anos 1960 e 2019. Inicialmente, a [Seção 4.1](#) apresenta o fluxo de trabalho utilizado, onde serão abordadas todas as etapas necessárias para alcançar o objetivo proposto. A partir deste fluxo de trabalho, as outras seções apresentam as etapas realizadas em detalhes. O presente capítulo conclui-se com um resumo, abordando as principais contribuições geradas.

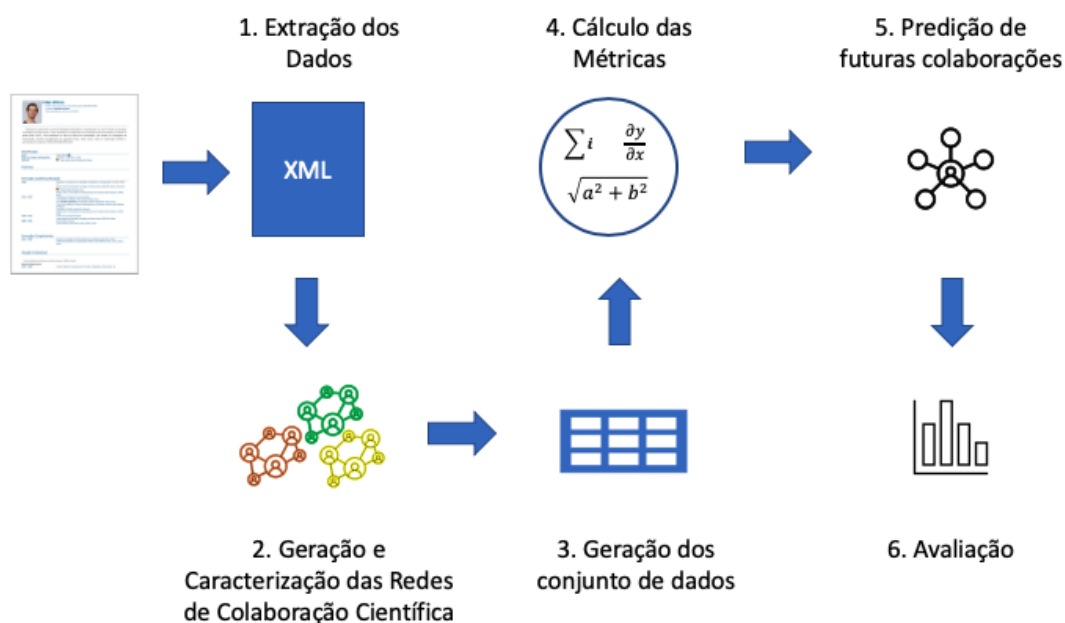
4.1 Método Proposto

Inicialmente, é importante destacar a importância da Plataforma Lattes como repositório de currículos de pesquisadores no Brasil. Conforme [Lane \(2010\)](#) afirma, esta representa uma das fontes de dados científicos mais relevantes e reconhecidos mundialmente. Dentre todos os pesquisadores que possuem seus currículos cadastrados na plataforma, foram escolhidos para análise apenas os que possuem o doutorado concluído. Conforme explicado por [Dias \(2016\)](#), estes representam apenas 5,38% do total de indivíduos presentes na plataforma, apesar de serem responsáveis por 74,51% dos artigos publicados em periódico e 64,67% dos artigos publicados em anais de congresso, o que contribui para a riqueza deste estudo. Além disso, os indivíduos com o doutorado concluído possuem, em seus currículos, uma data de atualização recente, e, em geral, possuem algum tipo de publicação registrada.

Ainda a respeito do conjunto de dados selecionado para desenvolvimento do trabalho, foram escolhidas oito áreas do conhecimento especificadas pelo CNPq, são elas: Agronomia, Biologia, Ciência da Computação, Ciência da Informação, Educação, Engenharia Elétrica, Medicina e, por fim, Linguística, Letras e Artes. Com exceção da Ciência da Informação, escolhida inicialmente por proximidade com o grupo de pesquisa pelo qual o trabalho está sendo desenvolvido, todas as demais foram escolhidas por serem as que possuem a mais vasta produção científica nas suas respectivas grandes áreas.

Conforme apresentado na [Figura 7](#), após a definição do conjunto de dados a ser trabalhado – na figura representado pela imagem de um currículo cadastrado na Plataforma Lattes – é realizada a primeira etapa do método proposto, responsável por realizar a extração destes currículos. Essa etapa será detalhada na [Seção 4.2](#), mas já é possível destacar

Figura 7 – Método Proposto



Fonte: Autoria própria

que o processo é realizado utilizando o *framework* desenvolvido por Dias (2016). Neste momento são identificados e separados os currículos dos indivíduos com título de doutor pertencentes às áreas especificadas acima, e, em um segundo momento, suas produções científicas são divididas por períodos previamente estabelecidos. Os períodos para divisão estabelecidos são: 1960 (data da primeira publicação cadastrada no repositório) a 2000, 2001 a 2010, e, por fim, 2011 a 2019 (último ano considerado, tendo em vista que a coleta foi realizada em 2020).

Já com os dados selecionados e distribuídos em períodos, as redes de colaboração científica são caracterizadas. Nesse momento, utilizando o método proposto em Dias et al. (2014), as redes são geradas utilizando as publicações registradas em cada currículo, considerando o período e áreas específicas, conforme explicado na Seção 4.3. A partir das redes de colaboração científica, é possível identificar, e gerar o conjunto de dados no formato correto a ser utilizado no restante do trabalho.

A Seção 4.4 descreve este processo em detalhes, sendo este, responsável pela redução de dimensionalidade do conjunto de informações, o que também soluciona o problema do balanceamento de classes, fazendo com que as métricas, geradas na próxima etapa, sejam calculadas apenas para a porção da rede a ser utilizada nas predições. Posteriormente, a partir do conjunto previamente filtrado, é necessário realizar o cálculo de algumas métricas a serem utilizadas como atributos para predição de colaborações. Esse processo, bem como a definição dos atributos a serem utilizados é descrito na Seção 4.5.

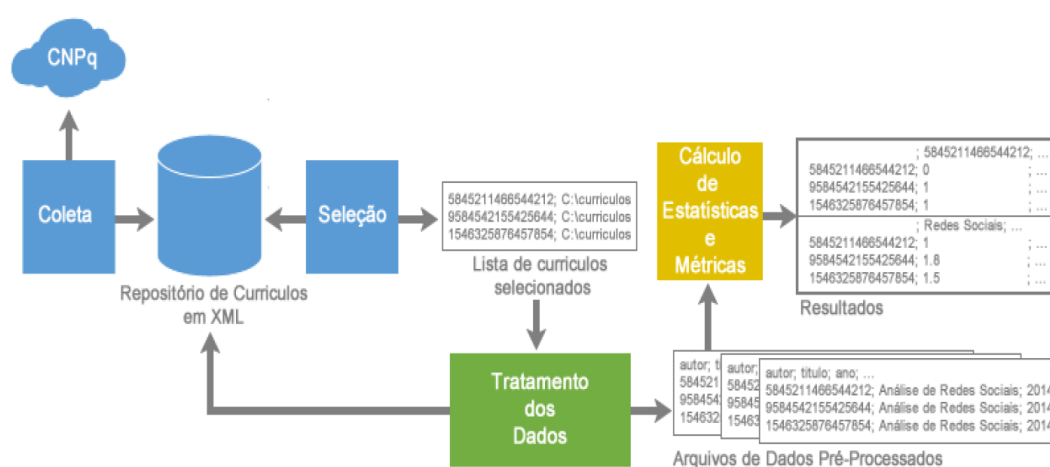
Por fim, todo o processo feito anteriormente é utilizado como entrada em alguns

algoritmos de aprendizado de máquinas. As técnicas usadas são explicadas na [Seção 4.6](#). Após utilizar os algoritmos de aprendizado de máquinas, é necessário realizar a avaliação dos resultados, e explicar quais técnicas foram utilizadas para tanto, o que é detalhado na [Seção 4.7](#).

4.2 Extração dos Dados

Como o objetivo da predição de ligações é determinar se uma nova ligação será formada em um próximo instante de tempo, pode-se compreendê-lo como uma classificação binária. Dessa forma, é natural utilizar modelos de aprendizado de máquinas desenvolvidos especificamente para tais situações (TANG, 2017). Estes dependem fortemente do pré-processamento de dados, onde a forma de extração e processamento irá influenciar nas predições realizadas (GOLINKO; SONDERMAN; ZHU, 2017). Portanto, esta seção tem como objetivo apresentar os métodos utilizados para realizar a extração dos currículos dos pesquisadores cadastrados na Plataforma Lattes.

Figura 8 – LattesDataXplorer



Fonte: Dias (2016)

Para tanto, o LattesDataXplorer (DIAS, 2016), um *framework* para extração e tratamento de currículos cadastrados na Plataforma Lattes, foi utilizado. Conforme apresentado na [Figura 8](#), para que os dados curriculares sejam extraídos com um baixo custo computacional, algumas etapas se fazem necessárias. Inicialmente, é necessário identificar a URL (*Uniform Resource Locator*) do currículo a ser extraído, para que as informações de cada pesquisador possam ser acessadas individualmente. As URL's são compostas pelo endereço de busca na Plataforma Lattes, acrescido do código que representa cada um dos

indivíduos. A partir disso, um arquivo de identificação é gerado, sendo este armazenado localmente o que garante que os currículos possam ser extraídos de forma sequencial.

A partir do arquivo gerado anteriormente e das URL's obtidas, cada currículo é acessado, a fim de gerar um novo arquivo contendo a identificação e a data de atualização do currículo. Esta identificação é composta por 16 dígitos e é única para cada indivíduo cadastrado na plataforma. Ao se iniciar a fase de extração, última etapa do processo, é realizado um teste. Neste momento, caso o identificador já exista na base de dados local, a data de atualização do currículo é comparada com a data da última versão presente no repositório local, e, caso sejam divergentes, a cópia local é deletada e a versão atualizada é obtida. Caso o identificador não esteja presente localmente, o currículo é extraído pela primeira vez. Através dessa sequência de passos, é possível extrair apenas os currículos que passaram por algum tipo de atualização, reduzindo assim o custo computacional.

Os currículos extraídos a partir dos passos citados acima são armazenados localmente em formato XML, e representam o objeto de origem dos dados deste trabalho. A partir destes arquivos, o custo computacional necessário para processamento dos dados é reduzido, uma vez que os mesmos estão armazenados localmente, sem depender do acesso à uma base de dados *on-line*. Neste momento, é realizada a filtragem dos currículos a serem utilizados no restante do trabalho, levando em consideração as áreas de atuação dos indivíduos, e as datas das publicações registradas na plataforma.

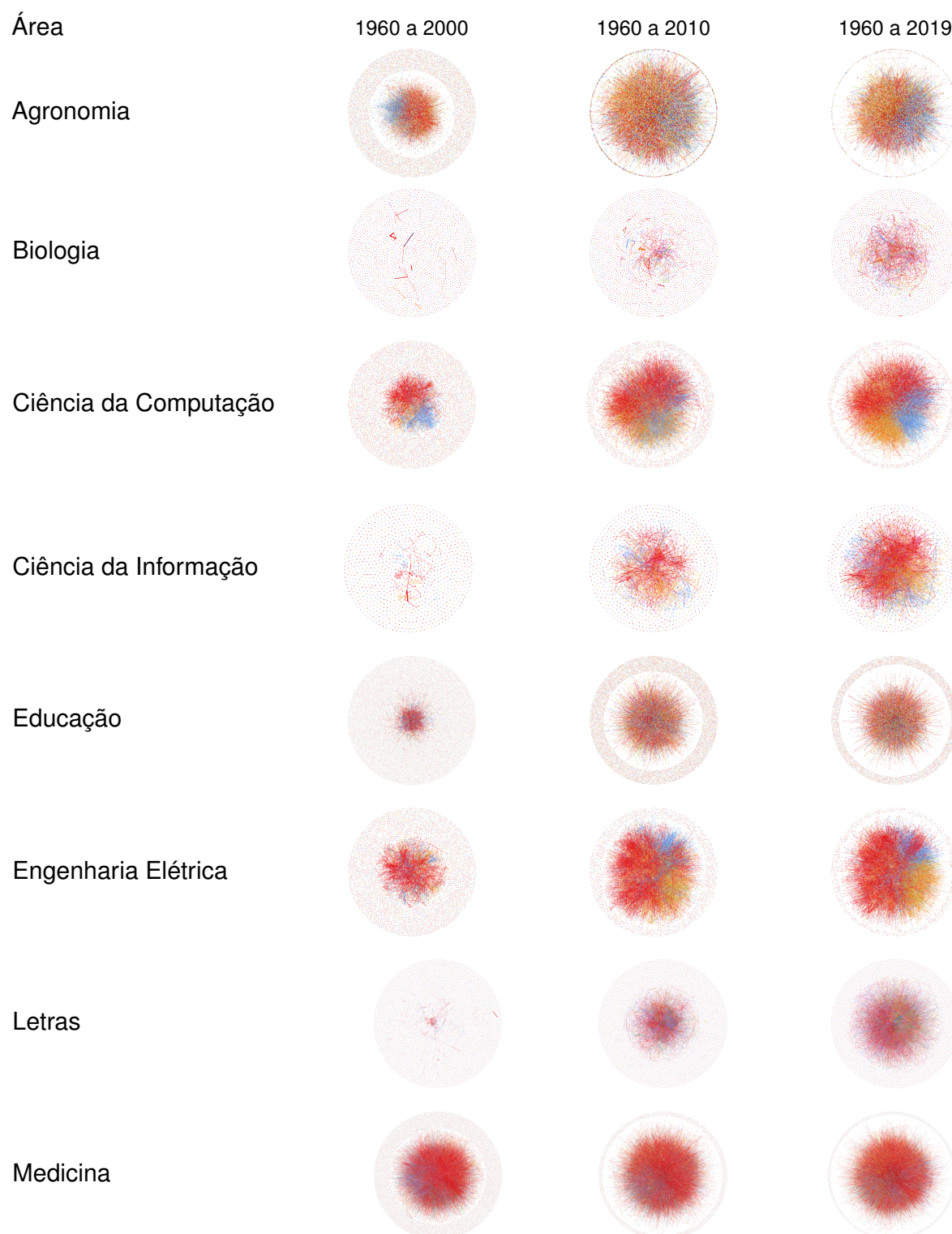
4.3 Caracterização das Redes de Colaboração Científica

Com os dados já extraídos e organizados, é necessário realizar a geração das redes de colaboração científica, definidas por [Newman \(2004\)](#) como a documentação de uma colaboração entre dois ou mais autores. Utilizando o *framework* demonstrado na seção anterior, [Dias et al. \(2014\)](#) apresentam um método para a construção destas redes utilizando um baixo custo computacional. Para tanto, os títulos dos trabalhos presentes nos currículos são transformados em uma chave única, e caso dois autores possuam a mesma chave em seu currículo, uma ligação entre os mesmos é estabelecida.

O trabalho aqui apresentado utiliza a mesma estratégia para criação das redes formadas pelos trabalhos realizados pelos indivíduos com doutorado concluído pertencentes às oito áreas citadas acima. Ao todo foram utilizados os dados de 71.010 pesquisadores, responsáveis por mais de 5.8 milhões de colaborações científicas.

Para melhor compreender os atributos a serem utilizados e as possíveis diferenças nos resultados dos algoritmos, é importante realizar uma análise das principais características encontradas. Ao todo, foram caracterizadas 24 redes, diferenciadas por área e período. A divisão por períodos é necessária, já que deve-se identificar os atributos responsáveis por haver uma aresta em determinado instante de tempo analisado. Estas redes são

Figura 9 – Evolução das Redes de Colaboração científica



Fonte: Autoria Própria

demonstradas visualmente na [Figura 9](#) e suas características são detalhadas na [Tabela 1](#).

O *software Gephi* foi utilizado para visualização, manipulação e exportação das imagens a partir do processo descrito anteriormente ([BASTIAN; HEYMANN; JACOMY](#),

Tabela 1 – Caracterização das Redes de Colaboração científica

Área	Período	Nós	Colaborações	Grau Médio	Densidade	Diâmetro	Caminho médio	Nós no componente Gigante	Nós Isolados
Agronomia	1960 a 2000	11.513	191.242	5,942	0,001	11	3,94	5.181	6.332
	2000 a 2010	11.513	844.700	30,606	0,001	11	3,29	10.449	1064
	2010 a 2019	11.513	467.010	46,609	0,004	9	3	10.979	534
Biologia	1960 a 2000	2.251	367	0,066	0	4	1,47	65	2.186
	2000 a 2010	2.251	3.446	0,717	0	17	6,76	538	1.713
	2010 a 2019	2.251	3.983	1,554	0,001	21	6,62	980	1.271
Ciência da Computação	1960 a 2000	7.134	20.194	1,305	0	12	5,335	2.012	5.122
	2000 a 2010	7.134	109.154	6,656	0,001	12	4,377	5.356	1.778
	2010 a 2019	7.134	108.577	11,415	0,001	11	3,906	6.157	977
Ciência da Informação	1960 a 2000	1.061	1.064	0,477	0	15	5,527	178	883
	2000 a 2010	1.061	6.084	3,668	0,003	10	4,067	613	448
	2010 a 2019	1.061	13.408	12,535	0,012	7	3,194	810	251
Educação	1960 a 2000	16.977	22.389	0,977	0	18	5,79	2.643	14.334
	2000 a 2010	16.977	417.792	29,176	0,002	12	3,732	10.319	6.658
	2010 a 2019	16.977	773.775	84,283	0,005	10	3,162	13.018	3.959
Engenharia Elétrica	1960 a 2000	5.665	27.731	1,684	0	14	5,458	1.876	3.789
	2000 a 2010	5.665	105.551	6,883	0,001	12	4,418	4.154	1.511
	2010 a 2019	5.665	89.596	11,324	0,002	10	3,945	4.787	878
Letras	1960 a 2000	8.527	1.408	0,209	0	15	5,738	524	8.003
	2000 a 2010	8.527	60.589	7,519	0,001	14	4,414	3.062	5.465
	2010 a 2019	8.527	176.343	30	0,004	12	3,655	4.989	3.538
Medicina	1960 a 2000	17.882	570.748	9,845	0,001	12	3,928	10.504	7.378
	2000 a 2010	17.882	1.233.731	32,03	0,002	9	3,325	15.504	2.378
	2010 a 2019	17.882	574.291	45,364	0,003	10	3,121	16.471	1.411

Fonte: Autoria Própria

2009). Inicialmente, a [Figura 9](#) apresenta as redes correspondentes à evolução da produção científica ao longo dos três períodos, onde as cores representam os estados do país nos quais os indivíduos analisados estão vinculados profissionalmente, que seguem o mesmo padrão para todas as áreas. Conforme explicado ao longo do trabalho, os nós representam pesquisadores, e as arestas representam trabalhos publicados em conjunto.

Inicialmente, a partir das informações apresentadas na [Tabela 1](#), pode-se perceber que houve, em geral, um grande aumento no número de colaborações científicas ao longo dos períodos estudados. Em média, as áreas analisadas tiveram um aumento de 38 vezes no número de colaborações ao se considerar o primeiro e último período. Algumas áreas merecem destaque dado o considerável aumento no número de publicações em conjunto. Um exemplo é a área de Linguística, Letras e Artes, que possuía 1.408 colaborações entre os anos de 1960 a 2000, e ao final do período de análise apresentou 176.343 trabalhos em conjunto, representando um aumento de 125 vezes. Em segundo lugar, a área da Educação saiu de 22.389 para 773.775 colaborações, o que representa um aumento de 34 vezes em comparação ao primeiro período analisado. Ainda analisando a quantidade de colaborações, a rede da Medicina apresenta o maior valor em um único momento, tendo 1.233.731 trabalhos publicados em conjunto entre 2000 e 2010, fazendo com que esta fosse a área responsável pela maior quantidade de trabalhos publicados em congressos e periódicos na análise aqui realizada.

Considerando a densidade do grafo, ou seja, a fração de arestas realmente presentes dentre todas as possíveis ([NEWMAN, 2018](#)) – fator extremamente relevante uma vez que a quantidade de arestas presentes ao longo dos períodos é um atributo crucial para compreensão da evolução da rede – pode-se perceber que existem comportamentos diferentes a depender da área sendo analisada. Ainda sobre a densidade, caso todos os pesquisadores estudados tivessem realizado, ao menos, uma publicação entre si, a densidade seria igual a 1. Dessa forma, é possível afirmar que, quanto mais próximo a 1, maior o número de arestas em comparação ao tamanho total da rede. A partir disso, pode-se observar que a Ciência da Informação, no terceiro período de tempo, possui a maior densidade encontrada ao longo do trabalho, sendo esta 0,012, demonstrando que esta é uma área do conhecimento onde os pesquisadores realizam mais trabalhos em colaboração. É possível observar a influência deste valor nos resultados ao longo das próximas etapas do trabalho.

Outros fatores que podem ser analisados estão relacionados à conexão entre os pesquisadores. Ao se observar o diâmetro da rede – que representa a distância máxima entre dois nós ([NEWMAN, 2018](#)) – espera-se que, com o aumento na quantidade de colaborações científicas, este número diminua, o que de fato ocorre na maioria das redes. Porém, a área da Biologia parte de uma densidade de 4, no primeiro período, indo para 17 em um segundo momento, e finaliza a análise com um diâmetro de 21. Este fator, que

representa a falta de trabalhos em conjunto pode ser observado também na quantidade de nós isolados na rede, onde também espera-se uma grande diminuição ao longo do tempo. Porém, aliado ao aumento do diâmetro, o número de nós isolados, entre os anos de 2010 a 2019, na área da Biologia representam 56% do total da rede.

As características da rede, brevemente aqui apresentadas, podem influenciar o comportamento dos algoritmos de predição, uma vez que uma parte dos atributos a serem utilizados se baseiam em informações topológicas. Conforme explicado no [Capítulo 3](#), uma outra situação que surge ao lidar com o problema da predição de ligações está relacionada ao custo computacional necessário para realizar o cálculo de tais atributos. Levando em consideração a área da Medicina, que possui 17.882 pesquisadores, uma simples matriz para representar os resultados obtidos pelos cálculos possuiria 159 milhões de colunas, e o mesmo número de linhas. Portanto, a próxima seção apresenta as técnicas utilizadas para realizar a geração do conjunto de dados a ser utilizado no trabalho, de forma a reduzir o custo computacional necessário para tanto.

4.4 Processamento do conjunto de dados

Neste momento, a partir das redes de colaboração científicas caracterizadas na etapa anterior, deverão ser calculadas e preparadas as informações necessárias para gerar um conjunto de dados a ser utilizado como entrada nos algoritmos de aprendizado de máquina. Os passos para atingir tal processo serão brevemente apresentados a seguir, e posteriormente serão abordados em detalhes.

- Inicialmente, os identificadores originais, obtidos pela Plataforma Lattes, devem ser transformados em uma nova numeração sequencial iniciando em 1. Esse processo é importante, uma vez que possibilita percorrer todos os pesquisadores de uma forma mais simples, sem a necessidade de criar uma nova estrutura de dados para tanto.
- A partir do passo anterior, um dicionário é criado, onde o identificador presente no currículo extraído da Plataforma Lattes corresponde a um novo número. Com isso, é possível realizar a atualização destes no arquivo que representa as arestas da rede.
- A partir da utilização dos novos identificadores, se inicia a etapa de codificação dos dados categóricos. Nesse momento, estado, cidade e instituição do indivíduo são convertidos em categorias, e não apenas em um texto simples.
- Após as categorias terem sido convertidas em uma matriz esparsa, que contém diversas colunas, é necessário realizar a redução de dimensionalidade do conjunto de dados. Dessa forma, o custo computacional para realização das predições também é reduzido.
- Posteriormente, é necessário realizar o cálculo dos atributos topológicos da rede. Para evitar realizar esse processo para arestas que não serão utilizadas, deve-se realizar

o balanceamento do conjunto de dados, isto é, fazer com que ambas as classes tenham o mesmo número de observações. Portanto, esta etapa consiste em identificar quantas arestas existem na rede, e selecionar aleatoriamente a mesma quantidade de arestas ausentes.

- Por fim, os atributos topológicos são calculados apenas para o conjunto de arestas definidos no passo anterior. Com isso, o custo computacional é reduzido consideravelmente.
- A última etapa consiste em unir o conjunto de dados que possui os atributos específicos do domínio aos atributos topológicos, gerados na etapa anterior.

A partir dos passos aqui apresentados, são abordadas em detalhes cada uma dessas etapas, bem como as técnicas utilizadas para tanto. Ao longo de todo o processo a linguagem Python 3 (ROSSUM; DRAKE, 2009) foi utilizada, juntamente com algumas bibliotecas específicas para pré-processamento das informações e utilização de algoritmos de aprendizado de máquina, que serão destacadas ao longo do texto.

Inicialmente, os identificadores vindos da Plataforma Lattes – que possuem 16 dígitos – devem ser transformados em uma numeração sequencial, uma vez que as estruturas de repetição das próximas etapas serão responsáveis por percorrer os nós da rede com base no identificador. Dessa forma, um dicionário é criado para armazenar a numeração original, juntamente com a nova. A partir disso, pode-se converter todas as arestas da rede para os novos identificadores. Após essa etapa, é possível dar prosseguimento à codificação das variáveis categóricas.

Atributos referentes ao domínio podem ser utilizados durante o processo de predição. Nesse caso, é necessário avaliar o conjunto de dados sendo utilizado e as técnicas necessárias para conversão destes em formatos corretos para sua utilização. Por se utilizar a Plataforma Lattes, diversas informações estão presentes em cada currículo dos pesquisadores, como, por exemplo: seu endereço profissional, orientações realizadas, participações em bancas, congressos nos quais foi realizada alguma publicação, instituições onde estudou, entre outras. Inicialmente, a quantidade de colaborações em conjunto que dois pesquisadores possuem ao longo de um determinado instante de tempo foi considerado um atributo a ser utilizado, chamado de peso. Dessa forma, é possível identificar colaboradores que já trabalham juntos constantemente, e possivelmente possuem uma maior influência nos próximos instantes de tempo.

Em um segundo momento, as informações de cidade, estado e instituição cadastrados no endereço profissional do pesquisador foram selecionadas e devem ser codificadas. Para tanto, o primeiro passo consiste na remoção de palavras que não interferem na categoria a ser apresentada, como, por exemplo, "do", "da", "dos", "das", chamadas de *stopwords*. Posteriormente, é necessário transformar os textos simples em números, ou seja, no lugar de "Belo Horizonte", o valor 5; para "São Paulo", o valor 13; e assim para cada informação

categórica. Porém, ao enviar tais valores para o algoritmo, pode ocorrer a interpretação errônea de que o valor 13 é maior do que o valor 5, fazendo entender que pesquisadores que possuem "São Paulo" em seus currículos receberiam um peso diferente dos de "Belo Horizonte". Para que essa situação não aconteça, é necessário transformar cada um desses números em colunas. Ao final, obtêm-se uma grande matriz esparsa, preenchida com diversos números 0, e poucos números 1, onde os valores 1 representam a cidade, estado, ou instituição dos pesquisadores. Concluindo esta etapa, o conjunto de dados enviado para o algoritmo passa a possuir milhares de colunas, a depender da quantidade de categorias possíveis.

Mesmo possuindo a característica de apresentar bons resultados a partir da utilização de grandes matrizes esparsas (que possuem varias colunas com os números zero e um), os atributos, ou seja, as colunas do conjunto de dados, utilizados nos algoritmos de predição representam o número de dimensões dos mesmos. Por exemplo, o algoritmo Vizinhos Mais Próximos, apresentado no [Capítulo 3](#), utiliza a proximidade para classificar a observação em uma determinada categoria. Dessa forma, quanto maior o número de dimensões utilizadas, maior o tempo para realizar a classificação das observações. Portanto, para se reduzir o custo computacional, é necessário realizar a redução de dimensionalidade dos atributos gerados anteriormente.

Neste momento, diversas técnicas podem ser utilizadas. A mais popular é chamada de Análise de Componentes Principais (PCA), e é responsável por condensar uma grande quantidade de variáveis em um conjunto menor, sem que haja perda de informações. Porém, este processo é recomendado para atributos numéricos, e não deve ser utilizado para dados categóricos, como é o caso deste trabalho, onde são utilizadas a cidade, estado e instituição de atuação profissional dos pesquisadores. Dessa forma, a análise de correspondências múltiplas (MCA) deve ser utilizada. Esta representa uma extensão da análise de correspondências (AC), e permite analisar padrões de relacionamento de diversas variáveis categóricas dependentes ([ABDI; VALENTIN, 2007](#)).

Neste processo, foram selecionados 10% do total de atributos identificados na seção anterior. No caso do conjunto de dados formado por Medicina, por exemplo, o conjunto total possuía mais de 9 mil colunas, compostas por dados categóricos, que foram condensados em, aproximadamente, 900, sem que houvesse perda considerável de informações, conforme explicado anteriormente.

Antes de se realizar o cálculo dos atributos topológicos, se faz necessário definir para quais arestas os mesmos serão calculados. Nessa etapa do trabalho, o problema do desbalanceamento de classes vêm a tona. Diversos autores ([ACAR; DUNLAVY; KOLDA, 2009](#); [HASAN; ZAKI, 2011](#); [GONG et al., 2011](#); [PEREZ-CERVANTES et al., 2013](#)) propõem técnicas e métodos para resolução deste desafio. Uma abordagem tradicional para superar o desbalanceamento das classes é chamada de sob-amostragem, a qual é utilizada neste

trabalho. Após este passo, o número de arestas presentes e ausentes é o mesmo. Com isso, é possível realizar o cálculo dos atributos topológicos para apenas uma pequena porção da rede, no lugar de realizar o processo para todas as possíveis arestas.

4.5 Cálculo das Métricas

Com os atributos relacionados ao domínio previamente estabelecidos e o balanceamento de classes realizado, se faz necessário calcular os atributos topológicos a serem utilizados no processo de predição. Para tanto, um conjunto básico de características, oriundas de outros trabalhos que abordaram esse tema foram selecionados ([ADAMIC; ADAR, 2003](#); [DIGIAMPIETRI et al., 2015](#); [MARUYAMA; DIGIAMPIETRI, 2019a](#); [HASAN; ZAKI, 2011](#); [CHEN; LI; HUANG, 2005](#); [LÜ; ZHOU, 2011](#)). A [Tabela 2](#) apresenta a notação padrão utilizada ao longo desta seção.

Tabela 2 – Notação utilizada

Símbolo	Significado
$\Gamma(x)$	Conjunto de Vizinhos de x
$\Gamma(y)$	Conjunto de Vizinhos de y
Letras em minúsculo	Nós na rede social
$ \Gamma(x) $	Número de vizinhos do nó x
$ \Gamma(y) $	Número de vizinhos do nó y

Fonte: [Tang \(2017\)](#)

De acordo com [Liben-Nowell e Kleinberg \(2003\)](#), a forma mais simples de realizar a predição de ligações é através da métrica vizinhos em comum, que pode ser entendida como a quantidade de nós em comum que dois nós específicos possuem. Ao utilizar esse atributo nas redes de colaboração científica, [Newman \(2010\)](#) destaca que indivíduos que nunca trabalharam juntos, mas possuem um colaborador em comum, têm uma probabilidade muito maior de colaborar no futuro. O atributo Vizinhos em Comum (VC) é mostrado na [Equação \(5\)](#), onde x e y representam vértices do grafo. O resultado, apresentado na equação por $VC(x,y)$, corresponde à quantidade de vizinhos que ambos os nós possuem em comum. Neste caso, o símbolo $\cap$ representa a interseção entre os vizinhos de x e os vizinhos de y .

$$VC(x,y) = |\Gamma(x) \cap \Gamma(y)| \quad (5)$$

Outra métrica que pode ser obtida utilizando as características estruturais da própria rede é chamada de Coeficiente de Jaccard (CJ), e mede a probabilidade de que ambos x e y possuam um vizinho v , escolhido aleatoriamente que x ou y possuam. [Hasan e Zaki](#)

(2011) explicam que, ao contrário do atributo Vizinhos em Comum, o CJ normaliza o número de vizinhos em comum, conforme apresentado na [Equação \(6\)](#).

$$CJ(x,y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \quad (6)$$

Similar à [Equação \(5\)](#), o resultado da [Equação \(6\)](#) também leva em consideração os vizinhos em comum dos dois nós. No caso do CJ, a quantidade de vizinhos em comum é dividida pela união total do conjunto, ou seja, o número total de vizinhos de x e de y , resultando em um valor que representa a similaridade do conjunto total analisado.

Com o objetivo de estabelecer a similaridade entre duas páginas, [Adamic e Adar \(2003\)](#) propuseram a métrica Adamic/Adar. Para que seja possível a sua utilização em algoritmos de predição de ligações, [Liben-Nowell e Kleinberg \(2007\)](#) customizaram a métrica conforme apresentado na [Equação \(7\)](#). Essa formulação atribui características mais raras com um peso maior ([POTGIETER et al., 2009](#)). Podemos entendê-la como o número de propriedades compartilhadas pelos nós, dividido pelo $\log$ da frequência das características.

$$Adamic/Adar(x,y) = \sum_{w \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log |\Gamma(w)|} \quad (7)$$

Logo, o resultado da [Equação \(7\)](#) é dado pela soma da divisão de 1 pelo $\log$ do número de vizinhos de w , onde w representa os vizinhos em comum de x e y . Um valor próximo a zero indica que os nós não são similares. Assim, quanto mais alto o valor, mais similares os nós são entre si.

Seguindo o mesmo raciocínio, a métrica Alocação de Recursos (RA) atribui peso na relação de dois nós, favorecendo as relações entre aqueles que possuem poucos relacionamentos ([DIGIAMPIETRI et al., 2015](#)), e pode ser encontrada na [Equação \(8\)](#).

$$RA(x,y) = \sum_{w \in \Gamma(x) \cap \Gamma(y)} \frac{1}{|\Gamma(w)|} \quad (8)$$

Considerando apenas o tamanho das vizinhanças dos nós, a métrica Atração Preferencial (PA) foi proposta, e é apresentada na [Equação \(9\)](#). Em suma, estabelece que a probabilidade de um novo relacionamento com outros vértices é baseada no grau do nó em questão ([HASAN; ZAKI, 2011](#)). Essa métrica não requer informações relacionadas à vizinhança de cada nó, fazendo com que possua um custo computacional mais baixo ([LÜ; ZHOU, 2011](#)). Neste caso, o resultado de PA é entendido como a multiplicação da quantidade de vizinhos que o nó x possui, pelo número de vizinhos que o nó y possui.

$$PA(x,y) = |\Gamma(x)| |\Gamma(y)| \quad (9)$$

O fato de que amigos de amigos podem criar uma ligação sugere que a distância entre os nós de uma rede podem influenciar na formação de novas ligações (HASAN; ZAKI, 2011). Dessa forma, a métrica Menor Caminho(MC) também pode ser utilizada para que a predição de ligações seja realizada. Podemos entendê-la como o caminho mínimo entre dois nós, ou seja, quanto menor a distância, mais similares os nós são (DIGIAMPIETRI et al., 2015; TANG, 2017).

$$MC(x,y) = Distancia(Menor(x,y)) \quad (10)$$

No caso do cálculo da menor distância entre os nós, diversos métodos podem ser utilizados. O resultado será o valor identificado através do método utilizado, que corresponde à quantidade de nós que foi necessário percorrer a partir de x para chegar à y . Quanto menor o valor, mais próximos os nós estão entre si.

Por fim, os oito conjunto de dados referentes à cada uma das áreas especificadas, contendo os pesquisadores, as ligações entre si e os atributos selecionados foram utilizados como entrada em um algoritmo de aprendizado de máquina. Cada linha do conjunto de dados é composta pelos seguintes itens: Vizinhos em Comum, Coeficiente de Jaccard, Adamic/Adar, Alocação de Recursos, Atração Preferencial, Menor Caminho, Cidade, Estado, Instituição, Peso, e por fim, a presença, ou não, de uma ligação. É importante salientar que as métricas correspondem aos cálculos apresentados anteriormente para os dois pesquisadores específicos. O objetivo, portanto, é responder a seguinte pergunta: dado este conjunto de atributos, uma nova aresta será gerada? Essa informação será enviada para o algoritmo de predição.

4.6 Predição de futuras colaborações

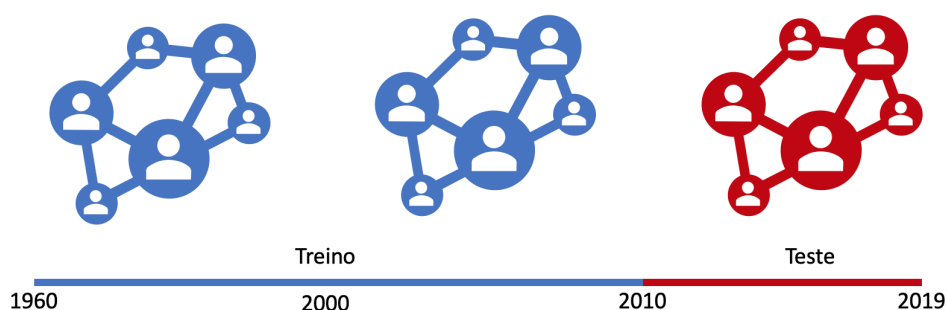
Esta etapa do processo consiste em utilizar o conjunto de dados gerado na etapa anterior como entrada em diversos algoritmos de aprendizado de máquinas para predição de futuras colaborações científicas. Todas as técnicas utilizadas foram explicadas no Capítulo 3, de modo que, os detalhes e parâmetros de cada serão apresentados nesta seção.

É importante salientar que todos os 8 conjuntos de dados passaram pelo processo, sem que houvesse alterações de parâmetro para cada um deles. No total, 4 diferentes algoritmos foram utilizados: Vizinhos Mais Próximos, Regressão Logística, Florestas Aleatórias e *Naive Bayes*.

Previamente à predição, é necessário separar os conjuntos de dados em dois subconjuntos, sendo um para treino dos classificadores, e outro para teste destes. Dessa forma, o teste é realizado em um conjunto de dados até então não utilizado anteriormente, e os resultados obtidos representam a performance do modelo em um cenário real. No caso

de problemas relacionados à predição de ligações, [Lichtnwalter e Chawla \(2012\)](#) destacam que ao se utilizar dados longitudinais, a separação entre treino e teste também deve levar esse fator em consideração. Normalmente, em aprendizado de máquina, o processo ocorre a partir de uma simples divisão percentual (entre 75% e 85% dos total), onde a maioria dos dados são utilizados para que o algoritmo aprenda os fatores que levam a uma determinada classe e, posteriormente, um novo conjunto de informações (entre 25% e 15% do total) ainda não visto pelos algoritmos é utilizado para realmente validar a eficácia da técnica utilizada. No caso deste trabalho, as duas primeiras redes serão utilizadas para treino, e a última rede para testes, conforme apresentado na [Figura 10](#).

Figura 10 – Divisão do Conjunto de Dados



Fonte: Autoria Própria

Nesse momento, a coluna referente à existência ou não de uma aresta em determinado período também é separada do restante dos dados. Normalmente, essa coluna é chamada de variável dependente, ou seja, seu valor depende do restante do conjunto de dados. Já as outras colunas se referem às chamadas variáveis independentes. A partir disso, as variáveis independentes passam pelo processo de normalização, que é necessário para manter os atributos dentro de um limite similar a todos, uma vez que, caso haja um valor fora do padrão, este pode influenciar o posicionamento do restante das classificações, impedindo os algoritmos de realizarem a aprendizagem ([DUDA; HART; STORK, 2012](#)).

Após esta etapa, os dados, já normalizados, são divididos entre treino e teste, e aplicados aos algoritmos especificados acima. Para tanto, a biblioteca scikit-learn ([PEDREGOSA et al., 2011](#)) foi utilizada. Esta possui todos os algoritmos necessários já implementados, bem como os métodos para normalização dos dados.

Para o algoritmo Vizinhos Mais Próximos, foram utilizados 6 vizinhos, a métrica minkowski (definida por padrão), e a distância euclidiana. No algoritmo Florestas Aleatórias foram utilizadas 10 árvores de decisão, com o critério entropia. Por fim, os algoritmos Regressão Logística e *Naive Bayes* não sofreram alterações, e foram utilizados com todos os parâmetros no padrão da biblioteca importada.

4.7 Avaliação

Uma etapa muito importante dos métodos de aprendizado de máquina é a avaliação dos resultados, uma vez que cada tipo de problema demanda um tipo de avaliação. Para os problemas de classificação, como é o caso deste trabalho, é observado o número de predições certas e erradas realizadas. Neste contexto, é comum utilizar a chamada matriz de confusão (RON; FOSTER, 1998), exibida na Tabela 3, que apresenta o número de instâncias que foram corretamente e erroneamente classificadas (CHOUDHARY; GIANEY, 2017). Através dela, pode-se classificar as predições realizadas de quatro formas diferentes:

- **Verdadeiro Positivo (VP):** observações positivas realmente preditas como positivas;
- **Falso Positivo (FP):** observações negativas incorretamente preditas como positivas;
- **Falso Negativo (FN):** observações preditas como negativas incorretamente;
- **Verdadeiro Negativo (VN):** observações corretamente preditas como negativas;

Tabela 3 – Matriz de Confusão

		Predito	
		0	1
Atual	0	VN	FP
	1	FN	VP

Fonte: Autoria Própria

Porém, tal análise pode levar em consideração diversos fatores e correlações entre estas variáveis. Dentre as diferentes abordagens existentes, as utilizadas em situações de classificação são: precisão, revogação e F1. Além disso, para problemas de predição de ligações, diversos autores recomendam a utilização da métrica área sob a curva (AUC), a qual será utilizada como base.

Cada uma das métricas, descritas acima, possui características próprias. A precisão, apresentada na Equação (11) (LICHTNWALTER; CHAWLA, 2012), tem como objetivo responder a seguinte pergunta: "de todos os valores preditos positivos, quantos realmente estão corretos?". Uma alta precisão está relacionada a poucos falsos positivos. Já considerando todos os valores positivos, a revogação tem como objetivo saber quantos destes foram realmente preditos, e é apresentada na Equação (12) (HAN, 2009). Já a métrica F1 leva em consideração a precisão e a revogação, fazendo uma média ponderada das duas anteriores. Por último, a área sob a curva, ou *area under the curve* (AUC), em inglês, é utilizada para exibir o desempenho de um modelo de classificação ao longo de todo o processo de aprendizagem. Esta é a métrica mais recomendada para medir, de forma justa, o desempenho em situações de predição de ligações. Neste caso, o valor varia entre 0 e 1, onde valores acima de 0,5 indicam que o método avaliado alcança uma performance maior

do que uma predição aleatória (BÜTÜN; KAYA, 2019).

$$Precisao = \frac{VP}{VP + FP} \quad (11)$$

$$Revocacao = \frac{VP}{VP + FN} \quad (12)$$

Na Equação (11), o resultado é dado pelo número de predições positivas classificadas corretamente (VP), dividido pela soma dos verdadeiros positivos (VP) aos falsos positivos (FP), ou seja, a quantidade de observações que foram classificadas como um, e na verdade deveriam ter sido classificadas como zero. Já a Equação (12) também realiza a divisão dos verdadeiros positivos (VP), porém a divisão é realizada levando em consideração a soma dos verdadeiros positivos (VP) aos falsos negativos (FN), que corresponde às observações classificadas erroneamente como zero, quando na verdade seriam um.

4.8 Resumo do capítulo

Neste capítulo foi apresentado o método para realizar a predição de futuras colaborações científicas. Em primeiro momento é apresentado o fluxo de trabalho contendo todas as etapas necessárias para tanto. Ao longo do capítulo cada processo é apresentado e explicado em detalhes. As etapas descritas abordam a extração dos dados, geração das redes de colaboração científica, geração do conjunto de dados, passando por um procedimento para balanceamento das classes. Posteriormente, são calculadas as métricas apenas de uma fração do conjunto total de dados, tendo como objetivo realizar a predição de futuras colaborações utilizando 4 diferentes algoritmos, e, ao final, avaliar o desempenho destes. O próximo capítulo apresentará os resultados obtidos a partir da metodologia aqui apresentada.

5 Análise e Discussão dos Resultados

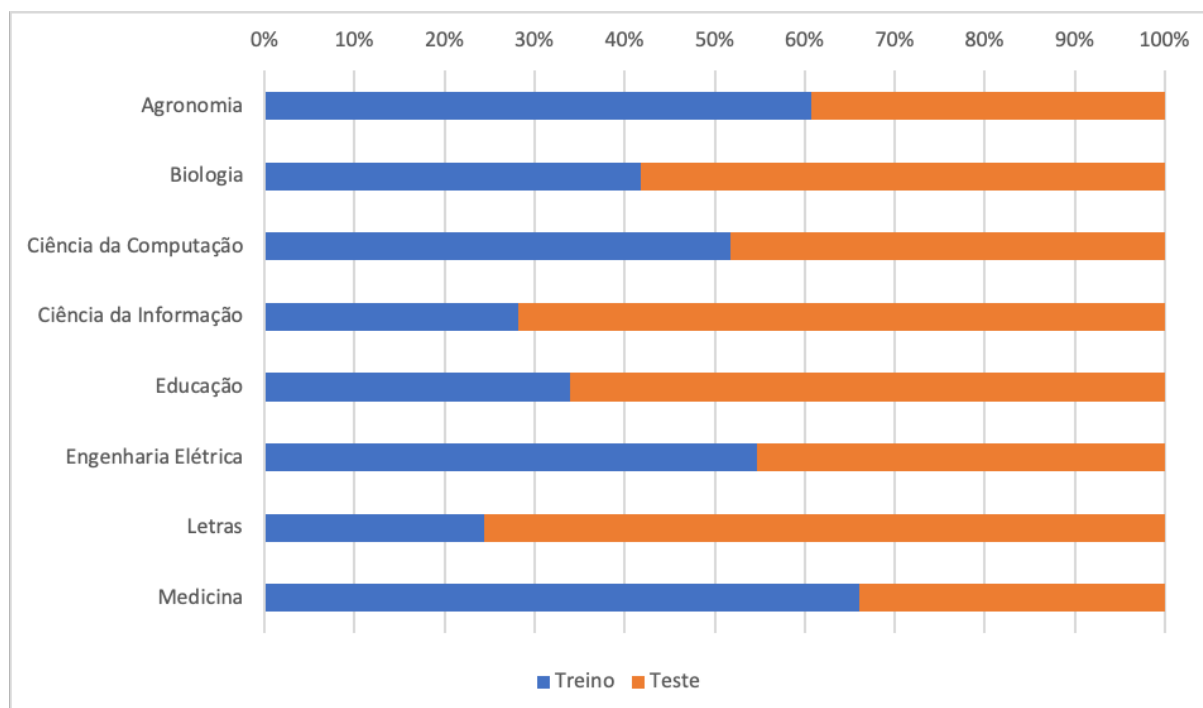
Este capítulo tem como objetivo apresentar os resultados encontrados, bem como discuti-los. Inicialmente, a [Seção 5.1](#) apresenta a caracterização dos dados utilizados nos algoritmos de predição. Posteriormente, [Seção 5.2](#) exibe os resultados encontrados, onde num primeiro momento é feita uma descrição geral das descobertas realizadas, e, em seguida, cada um deles é abordado em mais detalhes.

5.1 Caracterização dos dados

A partir da aplicação da metodologia previamente demonstrada no conjunto de dados selecionado para o trabalho aqui proposto, uma série de resultados foram obtidos. A partir destes, é possível compreender quais fatores podem influenciar no surgimento de uma nova colaboração científica, bem como características inerentes às próprias áreas do conhecimento que possuem a capacidade de influenciar na co-autoria de artigos. Dessa forma, conforme destacado na última seção do [Capítulo 4](#), uma série de métricas pode ser utilizada para avaliar a performance do aprendizado conduzido. Além das técnicas aplicadas para predição e avaliação, o próprio conjunto de dados pode influenciar na qualidade dos resultados encontrados, sendo este o primeiro fator a ser analisado. Como o objetivo da utilização do aprendizado de máquina é generalizar um tipo de padrão de comportamento, de forma a compreender, no caso de um problema de classificação, a qual classe uma determinada observação pertence, é essencial que haja dados suficientes para tanto. Portanto, uma análise do tamanho dos conjuntos analisados neste trabalho foi conduzida, e é apresentada a seguir.

Desta forma, o tamanho do conjunto de dados representa uma característica inerente às áreas sendo analisadas, e pode influenciar o comportamento dos algoritmos, uma vez que este conjunto deve ser dividido duas partes, uma para treino, e outra para testes, conforme já abordado no [Capítulo 4](#). Como o problema aqui retratado utiliza de dados em um formato longitudinal, ou seja, que leva em consideração o fator tempo, não se pode realizar uma simples divisão baseada no tamanho total do conjunto, como ocorre na maioria das vezes. Portanto, foi realizada segmentação considerando os períodos das redes criadas, onde as redes dos anos de 1960 a 2000, e 2000 a 2010 foram utilizadas para treino dos algoritmos, e o período de 2011 a 2019 foi utilizado para testes. A partir disto, a [Figura 11](#) apresenta a proporção entre os dados a serem utilizados para cada etapa nas diferentes áreas analisadas. Considerando a [Tabela 4](#), cabe salientar que, normalmente, é utilizado algo em torno de 15% a 25% para testes, e o restante para treino dos modelos. Porém, no caso das redes aqui estudadas, mesmo selecionando um período de 1960 à 2010 –

Figura 11 – Proporção entre dados de treino e dados de teste.



Fonte: Autoria Própria

correspondente a 50 anos – para treinamento, na maioria das áreas, o período de testes, que compreende apenas os anos de 2011 a 2019, representou uma parcela muito maior dos dados.

Considerando as diferentes áreas analisadas, e tomando como base a [Tabela 4](#), é possível perceber que existe uma grande diferença entre a produção científica nos períodos estabelecidos. Em alguns casos, como, por exemplo, na área de Linguística, Letras e Artes, apenas 24% dos dados estão disponíveis para treinamento dos modelos, enquanto que, para a área da Medicina, 66% poderão ser utilizados. Isso significa que os algoritmos possuirão uma parcela menor de informações para realizar o aprendizado, e uma quantidade maior para avaliar se realmente foi possível ou não encontrar um padrão de comportamento nas observações estudadas. Neste caso, o número de observações corresponde à quantidade total de arestas presentes e ausentes. Este fator poderá influenciar a qualidade das predições realizadas, e, consequentemente, os resultados encontrados.

Além da proporção entre treino e teste – que pode influenciar a capacidade dos modelos de compreenderem os padrões de comportamento, e, consequentemente, conseguirem generalizar o que influencia o surgimento de uma nova colaboração científica – o próprio tamanho do conjunto de dados também é um fator crucial neste processo. Analisando a [Figura 12](#), é possível perceber a grande diferença na quantidade de publicações realizadas no período analisado. Enquanto a área da Educação possui mais de 1,4 milhão

Tabela 4 – Fatias do conjunto de dados

Áreas	Observações	Qtd. Treino	% Treino	Qtd. Teste	%Teste
Agronomia	625.168	379.186	60,65%	245.982	39,35%
Biologia	4.052	1.692	41,76%	2.360	58,24%
Ciência da Computação	99.668	51.620	51,79%	48.048	48,21%
Ciência da Informação	14.770	4.152	28,11%	10.618	71,89%
Educação	1.474.540	499.512	33,88%	975.028	66,12%
Engenharia Elétrica	79.404	43.368	54,62%	36.036	45,38%
Linguística, Letras e Artes	264.732	64.496	24,36%	200.236	75,64%
Medicina	956.006	631.020	66,01%	324.986	33,99%

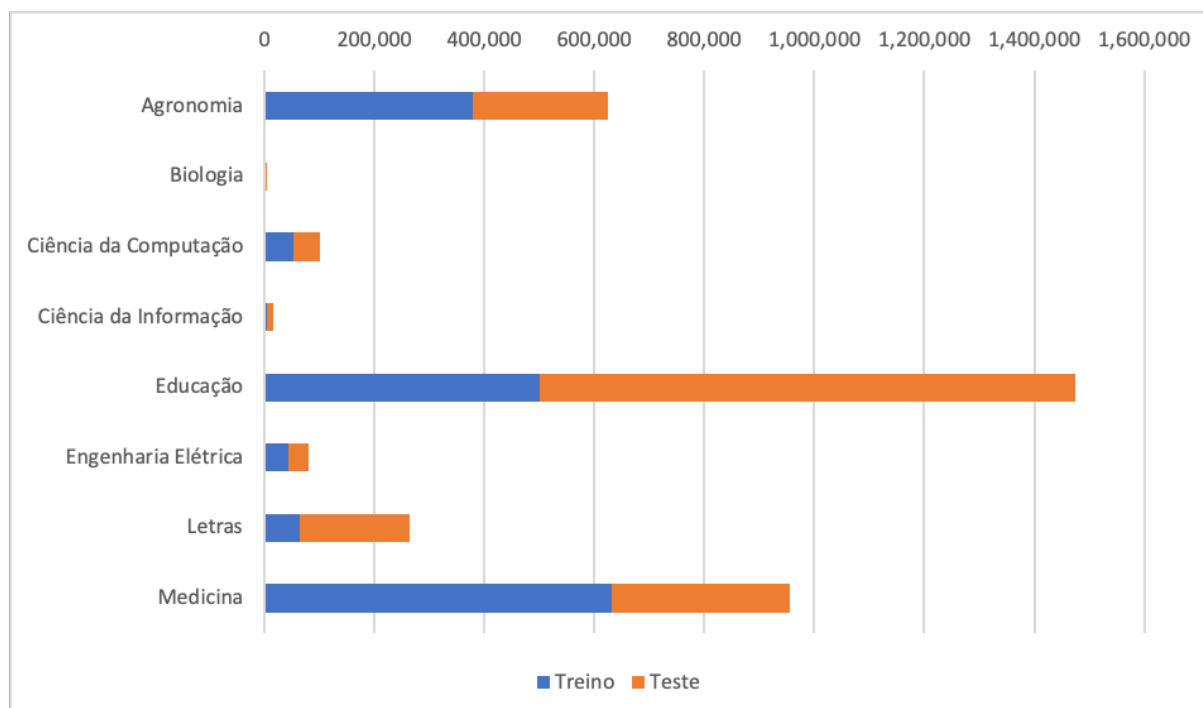
Fonte: Autoria Própria

de observações a serem analisadas, a Biologia possui apenas 4.052. Este pode ser um fator determinante, uma vez que, com um conjunto de dados menor, os modelos podem apresentar dificuldades em identificar fatores que influenciam o surgimento de colaborações científicas. Será possível observar os efeitos das diferenças aqui apresentadas quando a performance dos algoritmos for apresentada ao longo deste capítulo.

5.2 Resultados obtidos

Na seção anterior foram discutidos, separadamente, a influência da proporção entre dados de treino e teste, e também, a quantidade de observações disponíveis por área. Porém, é necessário analisar estes dois fatores em conjunto. Em primeiro lugar, é importante observar que algoritmos de aprendizado de máquinas ganharam força nas últimas décadas em decorrência do aumento do poder computacional disponível e do grande volume de dados sendo gerados a todo instante. Portanto, esses algoritmos foram pensados de forma a tirar vantagem de grandes volumes de informações para realizar as tarefas às quais foram desenvolvidos. Dessa forma, espera-se que, quanto maior o conjunto de dados, melhores serão os resultados apresentados. Neste mesmo contexto, espera-se que existam mais observações para treino do que para testes, de forma a fazer com que os modelos possam ser treinados em uma parcela maior do que a que será utilizada para testes.

Figura 12 – Quantidade de observações para treino e teste



Fonte: Autoria Própria

Estes dois fatores devem ser levados em consideração nas áreas estudadas no presente trabalho. Ao compararmos a área da Ciência da Informação, e Biologia, através da [Tabela 4](#), podemos perceber a diferença em alguns destes fatores. Enquanto a primeira possui um total de 14.770 observações a serem utilizadas, a segunda possui apenas 4.052. Porém, ao compararmos a proporção dos dados de treino e testes, podemos observar que a Biologia possui 41,76% dos dados disponíveis para treinamento dos modelos, enquanto a Ciência da Informação possui apenas 28,11% para o mesmo processo. Estes fatores devem ser levados em consideração ao se analisar os resultados apresentados em sequência.

Levando estes fatores em consideração, a [Tabela 5](#) apresenta o resultado global dos algoritmos, isto é, a média dos resultados encontrados levando em consideração todo o conjunto de áreas e os algoritmos utilizados em cada uma delas. É possível perceber que, mesmo se tratando de uma média, os modelos treinados apresentaram um comportamento melhor do que um classificador randômico. Utilizando como um valor base a taxa de 50% de acerto, e a área sob a curva (AUC) como métrica ideal para medir a performance de um algoritmo utilizado no problema de predição de ligações, pode-se observar que os algoritmos aplicados conseguiram compreender alguns dos padrões que influenciam o surgimento ou não de colaborações científicas. Ainda observando as métricas apresentadas na [Tabela 5](#), pode-se perceber que a precisão encontrada foi mais alta que as métricas revocação e F1, o que é normal em um modelo de classificação, uma vez que ao aumentar

a precisão, algumas predições podem ser feitas de forma equivocada.

Tabela 5 – Resultado Global

AUC	Precisão	Revocação	F1
0.88	0.88	0.86	0.86

Fonte: Autoria Própria

Mesmo apresentando um bom resultado, a média das predições ainda fica um pouco abaixo de valores apresentados na literatura. Nos próximos parágrafos, a partir da apresentação dos resultados em detalhes, serão analisadas as possíveis causas desta diferença. No momento, é importante salientar que os algoritmos conseguiram realizar a predição de futuras colaborações científicas com sucesso. Neste caso, considera-se um sucesso valores da métrica área sob a curva (AUC) acima de 50%. Esta taxa é importante, pois representa um acerto aleatório, ou seja, as mesmas chances de uma colaboração surgir ou não. Também é importante destacar que na análise aqui apresentada – após filtragem por área e nível de escolaridade – foram utilizados 5.823.173 colaborações científicas, publicadas por um total de 71.010 doutores com currículos cadastrados na Plataforma Lates, tornando este trabalho inédito devido a sua representatividade. Portanto, a diversidade entre as diferentes áreas também pode influenciar no desempenho dos algoritmos.

A partir desta análise inicial, pode-se voltar a atenção à [Tabela 6](#), que apresenta os resultados detalhados por área, algoritmo e métrica utilizados. Inicialmente, percebe-se que grande parte das técnicas aplicadas obteve um resultado acima de 90%, com exceção ao algoritmo *Naive Bayes*. Também é possível perceber que, conforme observado anteriormente, existe uma pequena diferença entre precisão e revocação. Porém, em grande parte dos resultados, os valores são similares, o que representa que os modelos conseguiram realizar as predições corretas e se mantiveram sensíveis ao problema apresentado. As únicas diferenças neste quesito se referem, novamente, ao algoritmo *Naive Bayes*, portanto, que será abordado em detalhes.

Partindo para uma análise mais específica a respeito da performance dos algoritmos, destacam-se os resultados apresentados pelo *Naive Bayes*, que, em diversas áreas, ficam próximos à faixa de 50% de acertos. Ao se calcular a média apenas desta técnica, uma AUC de 0.64 é obtida, valor não adequado para possíveis aplicações práticas. Neste caso, é destacada a importância de se utilizar diversos algoritmos diferentes, que possuam diferentes abordagens, para que seja feita uma comparação entre estes. Ainda abordando a baixa performance deste algoritmo, [Rish et al. \(2001\)](#) destacam que neste método se considera que as variáveis utilizadas são independentes umas das outras. Ao se considerar os atributos topológicos, é possível observar que estes não apresentam dependências entre si, enquanto os atributos relacionados ao domínio sim, uma vez que a instituição do pesquisador está relacionada à sua cidade, e consequentemente, ao seu estado de

Tabela 6 – Resultados obtidos nas áreas selecionadas

Área	Algoritmo	AUC	Precisão	Revocação	F1
Agronomia	K Vizinhos Mais Próximos	0.95	0.90	0.90	0.90
	Regressão Logística	0.98	0.97	0.97	0.97
	Florestas Aleatórias	0.99	0.99	0.99	0.99
	<i>Naive Bayes</i>	0.50	0.25	0.50	0.50
Biologia	K Vizinhos Mais Próximos	0.80	0.75	0.71	0.71
	Regressão Logística	0.94	0.90	0.89	0.89
	Florestas Aleatórias	0.98	0.96	0.96	0.96
	<i>Naive Bayes</i>	0.59	0.66	0.58	0.58
Ciência da Computação	K Vizinhos Mais Próximos	0.96	0.92	0.92	0.92
	Regressão Logística	0.96	0.95	0.94	0.94
	Florestas Aleatórias	0.99	0.98	0.98	0.98
	<i>Naive Bayes</i>	0.98	0.98	0.98	0.98
Ciência da Informação	K Vizinhos Mais Próximos	0.91	0.86	0.85	0.85
	Regressão Logística	0.96	0.92	0.92	0.92
	Florestas Aleatórias	0.99	0.98	0.98	0.98
	<i>Naive Bayes</i>	0.53	0.73	0.53	0.53
Educação	K Vizinhos Mais Próximos	0.96	0.93	0.92	0.92
	Regressão Logística	0.99	0.98	0.98	0.98
	Florestas Aleatórias	0.99	0.99	0.99	0.99
	<i>Naive Bayes</i>	0.56	0.75	0.56	0.56
Engenharia Elétrica	K Vizinhos Mais Próximos	0.96	0.91	0.91	0.91
	Regressão Logística	0.97	0.96	0.95	0.95
	Florestas Aleatórias	0.99	0.99	0.99	0.99
	<i>Naive Bayes</i>	0.74	0.79	0.70	0.67
Medicina	K Vizinhos Mais Próximos	0.98	0.95	0.95	0.95
	Regressão Logística	0.98	0.97	0.97	0.97
	Florestas Aleatórias	0.99	0.99	0.99	0.99
	<i>Naive Bayes</i>	0.64	0.74	0.63	0.63
Linguística, Letras e Artes	K Vizinhos Mais Próximos	0.95	0.90	0.89	0.89
	Regressão Logística	0.99	0.98	0.98	0.98
	Florestas Aleatórias	0.99	0.99	0.99	0.99
	<i>Naive Bayes</i>	0.55	0.76	0.55	0.55

Fonte: Autoria Própria

atuação profissional. Dessa forma, acredita-se que esse fator possa ter contribuído para os resultados apresentados. Por outro lado, a abordagem sendo discutida apresentou resultado de 0.98 para a área de Ciência da Informação, destacando que esta técnica pode trazer bons resultados a depender do conjunto de dados utilizado.

Posteriormente, analisando o resultado apresentado pelos outros algoritmos selecionados, é possível perceber que o seu comportamento foi similar na maioria das áreas. Com exceção à Biologia, onde a técnica K Vizinhos Mais Próximos apresentou resultados de 80% de predições corretas, enquanto em todas as outras áreas os resultados do restante dos algoritmos (desconsiderando o Naive Bayes) apresentou resultados acima de 94% de

acerto. O fato de três, das quatro técnicas, selecionadas terem apresentados resultados com altas taxas de acerto pode ser explicado por [Breiman \(2001b\)](#), que apresenta o conceito chamado de "Multiplicidade de bons modelos". Neste caso, é possível dizer que, a partir da utilização de um conjunto de dados que passou por um processo de limpeza, extração das variáveis relevantes e organização, diversas técnicas alcançarão boas performances.

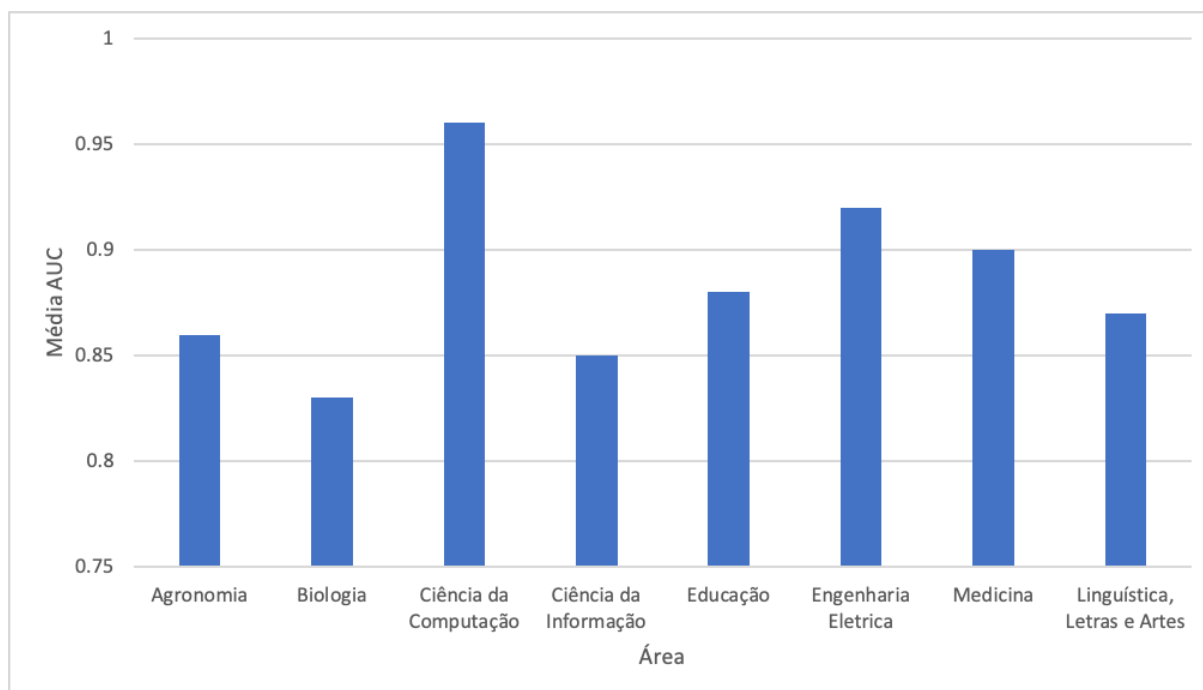
Ainda considerando os três algoritmos que possuíram os melhores resultados, a técnica Florestas Aleatórias apresenta os melhores resultados em todas as áreas do conhecimento, sendo esta considerada a melhor opção para predição de futuras colaborações científicas, de acordo com os estudos aqui realizados. Em segundo lugar, a Regressão Logística também apresenta ótimos resultados, muitas vezes similares às Florestas Aleatórias. Por fim, o algoritmo K Vizinhos Mais Próximos também apresenta resultados acima dos encontrados na literatura, porém abaixo dos citados acima.

Ao se comparar os resultados apresentados na [Tabela 6](#) às características de cada rede, apresentadas no [Capítulo 4](#), pode-se fazer uma análise levando em consideração os fatores topológicos citados anteriormente. Inicialmente, a área da Medicina – citada como maior produtora de artigos científicos e, conseqüentemente, maior conjunto de dados disponível para treino e testes – apresentou o terceiro melhor resultado em relação ao percentual de predições realizadas corretamente, conforme apresentado na [Figura 13](#). Já considerando a densidade do grafo, onde a área de Ciência da Informação foi destacada por possuir o maior valor, e também por possuir a menor quantidade de autores, esta área apresentou uma das piores performances, possuindo resultados melhores apenas do que a Biologia. Esta – cujo destaque foi relacionado ao alto diâmetro da rede, fator que demonstra a pouca colaboração científica entre os pesquisadores da área em específico – apresentou um desempenho médio de 83% de acertos, o mais baixo na análise conduzida. Ainda com relação aos resultados por área, destacam-se os resultados alcançados na Ciência da Informação, fortemente impulsionados pelo bom comportamento do algoritmo *Naive Bayes* em específico.

Os fatores destacados acima deixam claro que a predição de colaborações científicas possui um grande potencial para compreender a evolução deste tipo de rede. Porém, o comportamento intrínseco destas deve ser levado em consideração, uma vez que pode influenciar o comportamento dos modelos. Dessa forma, também fica claro que áreas onde os cientistas possuem o costume de publicar trabalhos entre si, e possuem menos pesquisadores isolados, se destacam ao estimar futuras colaborações científicas. Ainda é importante citar que características topológicas da rede permitem realizar um outro tipo de análise, relacionado à importância de cada atributo desta categoria no processo de predição.

Dessa forma, levando em consideração os atributos topológicos das redes caracterizadas anteriormente, a [Figura 14](#) apresenta importância de cada um deles com base na sua

Figura 13 – Média da métrica área sob a curva (AUC) para todos os algoritmos

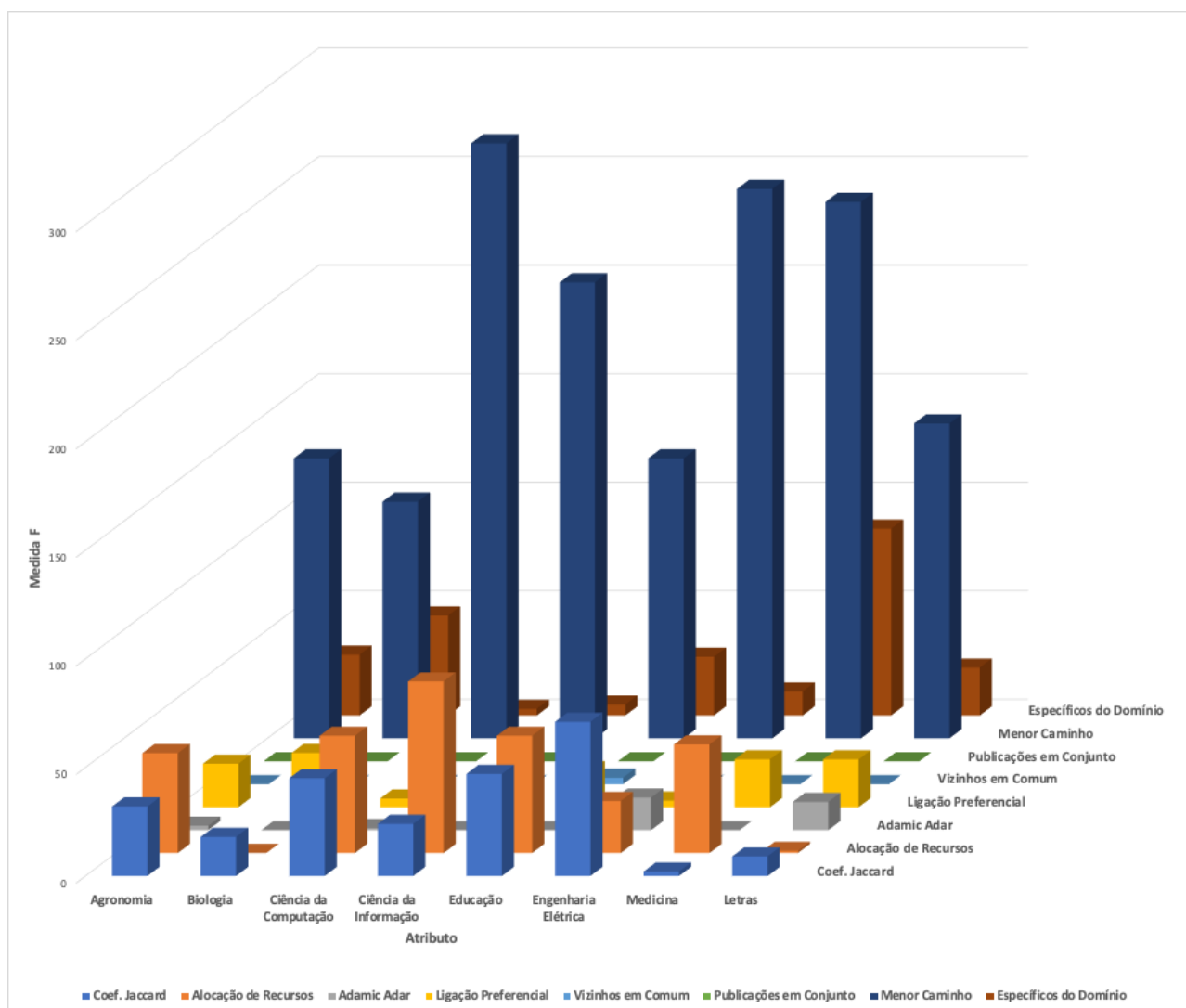


Fonte: Autoria Própria

influência na predição. Para tanto, foi utilizada a biblioteca XGBoost, desenvolvida por [Chen e Guestrin \(2016\)](#), que atribui pontos para cada variável ao longo do processo de aprendizado. Portanto, foram separados os 7 atributos referentes à topologia das redes, e um único atributo referente às especificidades do domínio. Neste caso, os atributos referentes à instituição, cidade e estado de atuação profissional dos pesquisadores foram agrupados. Através da análise dessa imagem, pode-se perceber que, para todas as áreas estudadas, o atributo Menor Caminho é o maior responsável pela predição de futuras colaborações científicas realizadas no conjunto selecionado, independente da área do conhecimento. Em segundo lugar, a métrica Alocação de Recursos se destaca como atributo mais importante ao longo da predição, seguida pelo Coeficiente de Jaccard, Atributos Específicos do Domínio, Ligação preferencial, Adamic/Adar, Vizinhos em Comum, e por fim a quantidade de colaborações realizadas em conjunto. É importante destacar que os dois últimos atributos apresentaram uma influência nas predições próxima a zero. Este comportamento se demonstra, até então, diferente da maioria dos referenciais teóricos aqui estudados, onde, na maior parte das vezes, o atributo mais relevante é o Vizinhos em Comum.

Ao se comparar os resultados apresentados na [Figura 14](#) às características de cada área do conhecimento analisada, percebe-se que houve algumas diferenças, como, por exemplo, a importância dos atributos específicos do domínio para a Medicina, demonstrando que a localização dos pesquisadores é um fator mais importante nessa área do que no restante. Já analisando a rede da Biologia, citada anteriormente como a que possui a pior

Figura 14 – Importância de cada atributo



Fonte: Autoria Própria

performance na predição de futuras colaborações científicas, se destaca a métrica Menor Caminho, que possui a menor influência entre todas as áreas, comprovando o fato de que, por se tratar de uma rede mais desconexa onde os nós continuam isolados, é mais difícil realizar conexões novas, impactando assim na estimativa de futuras colaborações científicas.

A partir dos resultados e considerações aqui apresentados, podemos concluir que a predição de futuras colaborações científicas utilizando os currículos dos doutores cadastrados na Plataforma Lattes se mostrou eficiente, e uma média de 88% das predições foram realizadas de forma correta. Ao se considerar a média do melhor preditor, para todas as áreas do conhecimento analisadas, uma média de 99% de acertos foi encontrada para a métrica AUC. Tal resultado apresenta valores similares aos encontrados na literatura.

Porém foram utilizados 5.8 milhões de trabalhos em conjunto, de um total de mais de 71 mil pesquisadores, ultrapassando os valores encontrados na literatura que normalmente analisam os pesquisadores de um determinado programa de pós graduação, ou área, em específico, tornando este um trabalho inédito devido a sua representatividade.

5.3 Resumo do Capítulo

Neste capítulo foram apresentados os resultados alcançados a partir da aplicação da metodologia citada anteriormente, no contexto do conjunto de dados formado pelos doutores com currículos cadastrados na Plataforma Lattes, em 8 diferentes áreas do conhecimento. Inicialmente foram demonstradas as diferenças entre os conjuntos de dados referentes às áreas especificadas e como este fator pode influenciar o comportamento dos algoritmos. Posteriormente, também foi apresentado o resultado médio, considerando todas as técnicas utilizadas, para o conjunto completo dos dados, o qual foi de 88% de precisão. Em seguida, uma análise considerando as especificidades de cada rede, juntamente com os detalhes dos resultados encontrados, foi realizada. Nesta etapa foi destacado o algoritmo Florestas Aleatórias, que alcançou uma média de 99% de acertos. Ao final do capítulo, foi apresentada uma consideração a respeito dos atributos utilizados para a predição, onde o Menor Caminho foi destacado como fator mais influente ao se realizar a estimativa de futuras colaborações científicas.

6 Conclusão

Com a adoção de métodos presentes na análise de redes sociais, se torna possível a explicação de diversos fenômenos. A partir das formulações geradas com base nestes métodos, um novo problema foi proposto por [Liben-Nowell e Kleinberg \(2003\)](#), chamado de predição de ligações, que consiste em estimar futuras conexões entre nós em uma determinada rede. Neste contexto, o presente trabalho tem como objetivo realizar a predição de futuras colaborações científicas, utilizando a Plataforma Lattes como fonte de dados. Através desta, redes de colaboração científica, nas quais os autores representam os nós, e os trabalhos publicados em conjunto representam as arestas, foram geradas. Para tanto, foram utilizados os dados de pesquisadores com o doutorado concluído pertencentes à oito áreas do conhecimento, sendo elas: Agronomia, Biologia, Ciência da Computação, Ciência da Informação, Educação, Engenharia Elétrica, Medicina, e, por fim, Linguística, Letras e Artes.

Para que seja possível realizar a predição de futuras colaborações científicas, algoritmos da área do aprendizado de máquinas, um ramo da inteligência artificial, foram utilizados. Estes foram escolhidos com base em suas abordagens, sendo a primeira técnica, K Vizinhos Mais Próximos, relacionada à proximidade entre os vizinhos (ou observações). Em segundo lugar foi utilizada a abordagem da Regressão Logística, que utiliza os mesmos princípios da regressão linear. posteriormente foi utilizada uma técnica que realiza uma combinação de diversas outras, chamada Florestas Aleatórias, e, por último, *Naive Bayes*, que possui uma abordagem baseada em probabilidades. Dessa forma, foi possível verificar o comportamento dos resultados a partir de quatro diferentes perspectivas.

Neste contexto, a partir da metodologia apresentada no [Capítulo 4](#), o conjunto de dados foi estruturado de forma a ser utilizado pelos algoritmos previamente definidos. Para tanto, divisões longitudinais foram realizadas, de forma a treinar os modelos, e, posteriormente testá-los. No total, foram utilizados 5.8 milhões de colaborações científicas, publicadas em conjunto por um total de 71.010 autores. A partir destas informações, a média de acertos, considerando todas as áreas e algoritmos foi de 88%, onde a técnica Florestas Aleatórias alcançou uma média de 99% para todas as áreas, e é a melhor abordagem no estudo aqui realizado. Neste contexto, o atributo Menor Caminho demonstrou ser o fator mais relevante para que uma colaboração científica possa ocorrer.

A partir da estratégia aqui apresentada, é possível replicar a metodologia para diferentes conjuntos de dados, ou até mesmo para uma nova fração de autores pertencentes à Plataforma Lattes. As descobertas constatadas deixam claro que fatores inerentes às próprias áreas do conhecimento podem influenciar futuras colaborações científicas.

Com relação aos desafios encontrados, o grande desbalanceamento das classes possui a capacidade de influenciar a qualidade dos modelos de forma considerável. É necessário que mais estudos e análises sejam conduzidos para verificar se existem possíveis alternativas à técnica de selecionar apenas um conjunto similar de arestas ausentes do conjunto de dados, de forma a melhor representar o modelo real. Também é importante analisar os diferentes momentos de tempo para que as porções de dados de treino e teste sejam mais próximas às normalmente aplicadas nos algoritmos.

Apesar dos desafios encontrados e da grande diferença entre as características das redes, os modelos treinados alcançaram uma boa performance. Neste caso, nas condições aqui apresentadas, a técnica Florestas Aleatórias poderia ser aplicada em outros cenários ou bases de dados, de forma a estimar futuras colaborações científicas de um determinado programa de pós graduação, ou até mesmo uma universidade.

Ao longo do trabalho, diversas considerações foram realizadas e apresentadas. Através de diversas etapas, uma série de contribuições para a literatura foram realizadas, dentre as principais, se destacam:

- Uma estratégia para a predição de futuras colaborações científicas a partir da aplicação de algoritmos de aprendizado de máquina;
- Um comparativo entre diferentes características das redes de colaboração científica referentes às áreas do conhecimento analisadas;
- Identificação do atributo que, dentro da análise realizada, mais influencia o surgimento de novas colaborações científicas.

Desta forma, é possível concluir que a estratégia apresentada é viável dado o conjunto de dados utilizado. Neste contexto, a técnica pode ser reaplicada, principalmente em grandes conjuntos de dados. A estratégia aqui apresentada não é restritiva com relação ao tamanho do conjunto, e os resultados são efetivos mesmo ao se utilizar uma base com centenas de milhares de observações.

6.1 Trabalhos Futuros

Através das contribuições e resultados apresentados é possível estabelecer alguns futuros trabalhos para continuidade dos estudos na área. Inicialmente, dando prosseguimento aos métodos relacionados a Inteligência Artificial, pode-se realizar o desenvolvimento de uma ferramenta para sugestão de colaborações científicas. Neste caso, o pesquisador recebe uma lista com os perfis mais similares ao dele, juntamente informações sobre os pesquisadores sugeridos. Esta ferramenta também pode ser utilizada para sugestão de pesquisadores em um tópico do conhecimento em específico. Fazendo com que os pesquisadores se aproximem de especialistas em diferentes áreas e que possuem algum vínculo entre si.

Outra possibilidade, que vai de encontro à proposta anterior se diz respeito a criação de uma interface de consulta e visualização de colaborações recomendadas. Para tanto, uma aplicação *web* deve ser desenvolvida permitindo que os pesquisadores vinculem o seu currículo, e solicitem colaborações, ou até mesmo, a análise de suas redes de colaboração. É importante deixar claro que para que o sistema retorne os resultados em tempo real, serão necessárias algumas alterações nos algoritmos de aprendizado de máquinas. Nesta aplicação também podem ser apresentados relatórios possibilitando o download e exportação para outras ferramentas.

Por fim, outras possibilidades de trabalhos futuros envolvem: extração de outras bases de dados a fim de aumentar o conjunto de características disponíveis; outros métodos para o balanceamento de classes, que considerem o contexto dos pesquisadores sem a necessidade de escolher aleatoriamente um grupo de arestas ausentes; utilização de novos algoritmos, ou até mesmo o aperfeiçoamento dos que já foram utilizados. A partir destas contribuições, a área da predição de colaborações científicas continuará a evoluir.

Referências

ABBASI, A.; ALTMANN, J.; HWANG, J. Evaluating scholars based on their academic collaboration activities: two indices, the rc-index and the cc-index, for quantifying collaboration activities of researchers and scientific communities. **Scientometrics**, Springer, v. 83, n. 1, p. 1–13, 2010. Citado na página 12.

ABDI, H.; VALENTIN, D. Multiple correspondence analysis. **Encyclopedia of measurement and statistics**, v. 2, n. 4, p. 651–657, 2007. Citado na página 38.

ACAR, E.; DUNLAVY, D. M.; KOLDA, T. G. Link prediction on evolving data using matrix and tensor factorizations. In: IEEE. **Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on**. [S.l.], 2009. p. 262–269. Citado 3 vezes nas páginas 1, 21 e 38.

ADAMIC, L. A.; ADAR, E. Friends and neighbors on the web. **Social networks**, Elsevier, v. 25, n. 3, p. 211–230, 2003. Citado 6 vezes nas páginas 2, 8, 9, 21, 39 e 40.

ALPAYDIN, E. **Introduction to machine learning**. [S.l.: s.n.], 2014. Citado 2 vezes nas páginas 23 e 25.

BARABÁSI, A.-L.; ALBERT, R. Emergence of scaling in random networks. **science**, American Association for the Advancement of Science, v. 286, n. 5439, p. 509–512, 1999. Citado 2 vezes nas páginas 1 e 6.

BASTIAN, M.; HEYMANN, S.; JACOMY, M. Gephi: an open source software for exploring and manipulating networks. In: **Third international AAAI conference on weblogs and social media**. [S.l.: s.n.], 2009. Citado na página 35.

BISHOP, C. M. **Pattern recognition and machine learning**. [S.l.]: springer, 2006. Citado 2 vezes nas páginas 23 e 24.

BONDY, J. A.; MURTY, U. S. R. et al. **Graph theory with applications**. [S.l.]: Citeseer, 1976. v. 290. Citado na página 16.

BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, n. 1, p. 5–32, 2001. Citado na página 26.

BREIMAN, L. Statistical modeling: The two cultures (with comments and a rejoinder by the author). **Statistical science**, Institute of Mathematical Statistics, v. 16, n. 3, p. 199–231, 2001. Citado na página 51.

BÜTÜN, E.; KAYA, M. A pattern based supervised link prediction in directed complex networks. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 525, p. 1136–1145, 2019. Citado na página 44.

CAN, U.; ALATAS, B. A new direction in social network analysis: Online social network analysis problems and applications. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 535, p. 122372, 2019. Citado na página 24.

CHAWLA, N. V.; JAPKOWICZ, N.; KOTCZ, A. Special issue on learning from imbalanced data sets. **ACM Sigkdd Explorations Newsletter**, ACM, v. 6, n. 1, p. 1–6, 2004. Citado na página 22.

CHEN, H.; LI, X.; HUANG, Z. Link prediction approach to collaborative filtering. In: IEEE. **Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'05)**. [S.l.], 2005. p. 141–142. Citado na página 39.

CHEN, T.; GUESTRIN, C. XGBoost: A scalable tree boosting system. In: **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York, NY, USA: ACM, 2016. (KDD '16), p. 785–794. ISBN 978-1-4503-4232-2. Disponível em: <<http://doi.acm.org/10.1145/2939672.2939785>>. Citado na página 52.

CHEN, Y.-W.; LIN, C.-J. Combining svms with various feature selection strategies. In: **Feature extraction**. [S.l.]: Springer, 2006. p. 315–324. Citado na página 26.

CHOUDHARY, R.; GIANEY, H. K. Comprehensive review on supervised machine learning algorithms. In: IEEE. **2017 International Conference on Machine Learning and Data Science (MLDS)**. [S.l.], 2017. p. 37–43. Citado 4 vezes nas páginas 23, 26, 27 e 43.

CLAUSET, A.; MOORE, C.; NEWMAN, M. E. Hierarchical structure and the prediction of missing links in networks. **Nature**, Nature Publishing Group, v. 453, n. 7191, p. 98, 2008. Citado na página 3.

DAUD, A. et al. Using machine learning techniques for rising star prediction in co-author network. **Scientometrics**, Springer, v. 102, n. 2, p. 1687–1711, 2015. Citado na página 11.

DIAS, T. Um estudo da produção científica brasileira a partir de dados da plataforma lattes. 181p. **Programa de Pós-Graduação em Modelagem Matemática e Computacional, Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte (Doutorado)**, 2016. Citado 3 vezes nas páginas 29, 30 e 31.

DIAS, T. M. et al. Modelagem e caracterização de redes científicas: um estudo sobre a plataforma lattes. In: **BRASNAM-II Brazilian Workshop on Social Network Analysis and Mining**. [S.l.: s.n.], 2013. p. 10–20. Citado 2 vezes nas páginas 2 e 14.

DIAS, T. M.; MOREIRA, T. H.; DIAS, P. M. Caracterização e análise das redes de colaboração científica dos bolsistas de produtividade em pesquisa do cnpq. In: SBC. **7º Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2018)**. [S.l.], 2018. Citado na página 14.

DIAS, T. M. R. et al. Identificação e caracterização de redes científicas de dados curriculares. **iSys-Revista Brasileira de Sistemas de Informação**, v. 7, n. 3, p. 5–18, 2014. Citado 5 vezes nas páginas 1, 2, 14, 30 e 32.

DIGIAMPIETRI, L. et al. Um sistema de predição de relacionamentos em redes sociais. In: **Brazilian Symposium on Information Systems**. [S.l.: s.n.], 2015. v. 11. Citado 5 vezes nas páginas 4, 10, 39, 40 e 41.

DIGIAMPIETRI, L. et al. Minerando e caracterizando dados de currículos lattes. In: **Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)**. [S.l.: s.n.], 2012. Citado na página 13.

- DOMINGOS, P. A few useful things to know about machine learning. **Communications of the ACM**, ACM New York, NY, USA, v. 55, n. 10, p. 78–87, 2012. Citado na página 23.
- DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern classification**. [S.l.]: John Wiley & Sons, 2012. Citado 2 vezes nas páginas 24 e 42.
- DUNLAVY, D. M.; KOLDA, T. G.; ACAR, E. Temporal link prediction using matrix and tensor factorizations. **ACM Transactions on Knowledge Discovery from Data (TKDD)**, ACM New York, NY, USA, v. 5, n. 2, p. 1–27, 2011. Citado na página 21.
- GOLINKO, E.; SONDERMAN, T.; ZHU, X. Cnfl: Categorical to numerical feature learning for clustering and classification. In: IEEE. **2017 IEEE Second International Conference on Data Science in Cyberspace (DSC)**. [S.l.], 2017. p. 585–594. Citado na página 31.
- GONG, N. Z. et al. Jointly predicting links and inferring attributes using a social-attribute network (san). **arXiv preprint arXiv:1112.3265**, 2011. Citado 2 vezes nas páginas 10 e 38.
- GRANOVETTER, M. S. The strength of weak ties. In: **Social networks**. [S.l.]: Elsevier, 1977. p. 347–367. Citado na página 9.
- GUNS, R.; ROUSSEAU, R. Recommending research collaborations using link prediction and random forest classifiers. **Scientometrics**, Springer, v. 101, n. 2, p. 1461–1473, 2014. Citado na página 26.
- HAN, J. Mining heterogeneous information networks by exploring the power of links. In: SPRINGER. **International Conference on Discovery Science**. [S.l.], 2009. p. 13–30. Citado 2 vezes nas páginas 7 e 43.
- HASAN, M. A. et al. Link prediction using supervised learning. In: **SDM06: workshop on link analysis, counter-terrorism and security**. [S.l.: s.n.], 2006. Citado 2 vezes nas páginas 2 e 9.
- HASAN, M. A.; ZAKI, M. J. A survey of link prediction in social networks. In: **Social network data analytics**. [S.l.]: Springer, 2011. p. 243–275. Citado 9 vezes nas páginas 1, 12, 19, 21, 22, 38, 39, 40 e 41.
- HU, K. et al. Link prediction in complex networks by multi degree preferential-attachment indices. **arXiv preprint arXiv:1211.1790**, 2012. Citado 2 vezes nas páginas 10 e 20.
- KLEINBAUM, D. G. et al. **Logistic regression**. [S.l.]: Springer, 2002. Citado na página 26.
- KLEINBERG, J. M. Navigation in a small world. **Nature**, Nature Publishing Group, v. 406, n. 6798, p. 845–845, 2000. Citado na página 19.
- KREBS, V. E. Mapping networks of terrorist cells. **Connections**, v. 24, n. 3, p. 43–52, 2002. Citado na página 8.
- LANE, J. Let's make science metrics more scientific. **Nature**, Nature Publishing Group, v. 464, n. 7288, p. 488, 2010. Citado 2 vezes nas páginas 2 e 29.
- LI, X. et al. Communities detection in social network based on local edge centrality. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 531, p. 121552, 2019. Citado na página 8.

- LIBEN-NOWELL, D.; KLEINBERG, J. The link-prediction problem for social networks. In: **Conference on Information and Knowledge Management (CIKM'03)**. [S.l.: s.n.], 2003. p. 556–559. Citado 7 vezes nas páginas [1](#), [8](#), [9](#), [12](#), [21](#), [39](#) e [55](#).
- LIBEN-NOWELL, D.; KLEINBERG, J. The link-prediction problem for social networks. **Journal of the American society for information science and technology**, Wiley Online Library, v. 58, n. 7, p. 1019–1031, 2007. Citado na página [40](#).
- LICHTNWALTER, R.; CHAWLA, N. V. Link prediction: fair and effective evaluation. In: IEEE COMPUTER SOCIETY. **Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)**. [S.l.], 2012. p. 376–383. Citado 4 vezes nas páginas [10](#), [21](#), [42](#) e [43](#).
- LIU, Z. et al. Link prediction in complex networks: A local naïve bayes model. **EPL (Europhysics Letters)**, IOP Publishing, v. 96, n. 4, p. 48007, 2011. Citado 2 vezes nas páginas [1](#) e [27](#).
- LÜ, L.; ZHOU, T. Role of weak ties in link prediction of complex networks. In: ACM. **Proceedings of the 1st ACM international workshop on Complex networks meet information & knowledge management**. [S.l.], 2009. p. 55–58. Citado na página [9](#).
- LÜ, L.; ZHOU, T. Link prediction in complex networks: A survey. **Physica A: statistical mechanics and its applications**, Elsevier, v. 390, n. 6, p. 1150–1170, 2011. Citado 4 vezes nas páginas [3](#), [11](#), [39](#) e [40](#).
- MARUYAMA, W. T.; DIGIAMPIETRI, L. A. Co-authorship prediction in academic social network. In: SBC. **Anais do V Workshop Brasileiro de Análise de Redes Sociais e Mineração**. [S.l.], 2019. p. 79–90. Citado 2 vezes nas páginas [2](#) e [39](#).
- MARUYAMA, W. T.; DIGIAMPIETRI, L. A. Predição de relacionamentos em redes sociais, uma revisao sistemática. In: SBC. **Anais do V Brazilian Workshop on Social Network Analysis and Mining**. [S.l.], 2019. p. 145–156. Citado na página [11](#).
- MCPHERSON, M.; SMITH-LOVIN, L.; COOK, J. M. Birds of a feather: Homophily in social networks. **Annual review of sociology**, Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, v. 27, n. 1, p. 415–444, 2001. Citado na página [19](#).
- MENA-CHALCO, J. P.; JUNIOR, R. M. C. Scriptlattes: an open-source knowledge extraction system from the lattes platform. **Journal of the Brazilian Computer Society**, Springer, v. 15, n. 4, p. 31–39, 2009. Citado na página [13](#).
- MENON, A. K.; ELKAN, C. Link prediction via matrix factorization. In: SPRINGER. **Joint european conference on machine learning and knowledge discovery in databases**. [S.l.], 2011. p. 437–452. Citado 2 vezes nas páginas [21](#) e [22](#).
- MILGRAM, S. The small world problem. **Psychology today**, New York, v. 2, n. 1, p. 60–67, 1967. Citado na página [19](#).
- MOHRI, M.; ROSTAMIZADEH, A.; TALWALKAR, A. **Foundations of machine learning**. [S.l.]: MIT press, 2018. Citado 4 vezes nas páginas [23](#), [24](#), [25](#) e [26](#).
- NEWMAN, M. **Networks: An Introduction Oxford Univ.** [S.l.]: Press, 2010. Citado na página [39](#).

NEWMAN, M. **Networks**. [S.l.]: Oxford university press, 2018. Citado 4 vezes nas páginas 3, 19, 21 e 35.

NEWMAN, M. E. The structure of scientific collaboration networks. **Proceedings of the national academy of sciences**, National Acad Sciences, v. 98, n. 2, p. 404–409, 2001. Citado 4 vezes nas páginas 1, 12, 18 e 19.

NEWMAN, M. E. Mixing patterns in networks. **Physical Review E**, APS, v. 67, n. 2, p. 026126, 2003. Citado 2 vezes nas páginas 1 e 6.

NEWMAN, M. E. Coauthorship networks and patterns of scientific collaboration. **Proceedings of the national academy of sciences**, National Acad Sciences, v. 101, n. suppl 1, p. 5200–5205, 2004. Citado 5 vezes nas páginas 1, 12, 18, 19 e 32.

NEWMAN, M. E.; PARK, J. Why social networks are different from other types of networks. **Physical Review E**, APS, v. 68, n. 3, p. 036122, 2003. Citado 3 vezes nas páginas 1, 6 e 18.

NEWMAN, M. E.; WATTS, D. J. Renormalization group analysis of the small-world network model. **Physics Letters A**, Elsevier, v. 263, n. 4-6, p. 341–346, 1999. Citado na página 6.

NILSSON, N. J. **Introduction to machine learning: An early draft of a proposed textbook**. [S.l.]: USA; Stanford University, 1996. Citado 3 vezes nas páginas 23, 24 e 25.

OLIVEIRA, M. A. d. et al. Predição semântica de unlinks em redes evolutivas utilizando lógica de descrição probabilística. 2014. Citado na página 21.

PARAND, F.-A.; RAHIMI, H.; GORZIN, M. Combining fuzzy logic and eigenvector centrality measure in social network analysis. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 459, p. 24–31, 2016. Citado na página 7.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in python. **Journal of machine learning research**, v. 12, n. Oct, p. 2825–2830, 2011. Citado na página 42.

PEREZ-CERVANTES, E. **Análise de redes de colaboração científica: uma abordagem baseada em grafos relacionais com atributos**. Tese (Doutorado) — Universidade de São Paulo, 2015. Citado 4 vezes nas páginas 13, 18, 20 e 21.

PEREZ-CERVANTES, E. et al. Using link prediction to estimate the collaborative influence of researchers. In: IEEE. **eScience (eScience), 2013 IEEE 9th International Conference on**. [S.l.], 2013. p. 293–300. Citado 3 vezes nas páginas 2, 13 e 38.

POTGIETER, A. et al. Temporality in link prediction: Understanding social complexity. **Emergence: Complexity & Organization (E: CO)**, Citeseer, v. 11, n. 1, p. 69–83, 2009. Citado 2 vezes nas páginas 16 e 40.

QIAO, J. et al. Bridging the gap between different social networks. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 410, p. 535–549, 2014. Citado na página 7.

RATTIGAN, M. J.; JENSEN, D. The case for anomalous link discovery. **Acm Sigkdd Explorations Newsletter**, ACM, v. 7, n. 2, p. 41–47, 2005. Citado na página 22.

RICHARDSON, M.; PRAKASH, A.; BRILL, E. Beyond pagerank: machine learning for static ranking. In: **Proceedings of the 15th international conference on World Wide Web**. [S.l.: s.n.], 2006. p. 707–715. Citado na página 23.

- RISH, I. et al. An empirical study of the naive bayes classifier. In: . [S.l.: s.n.], 2001. Citado na página 49.
- RON, K.; FOSTER, P. Special issue on applications of machine learning and the knowledge discovery process. **Journal of Machine Learning**, v. 30, p. 271–274, 1998. Citado na página 43.
- ROSSUM, G. V.; DRAKE, F. L. **Python 3 Reference Manual**. Scotts Valley, CA: CreateSpace, 2009. ISBN 1441412697. Citado na página 37.
- RUSSELL, S. J.; NORVIG, P. **Artificial intelligence: a modern approach**. [S.l.]: Malaysia; Pearson Education Limited,, 2016. Citado 3 vezes nas páginas 23, 24 e 27.
- STATTNER, E.; COLLARD, M. Descriptive modeling of social networks. **Procedia Computer Science**, Elsevier, v. 52, p. 226–233, 2015. Citado na página 7.
- STRÖELE, V.; ZIMBRÃO, G.; SOUZA, J. M. Análise de redes sociais científicas: modelagem multi-relacional. In: **I Brazilian Workshop Soc Netw Anal Min**. [S.l.: s.n.], 2012. Citado na página 17.
- TANG, F. **Link-Prediction and its Application in Online Social Networks**. Tese (Doutorado) — Victoria University, 2017. Citado 5 vezes nas páginas 3, 20, 31, 39 e 41.
- WANG, P. et al. Link prediction in social networks: the state-of-the-art. **Science China Information Sciences**, Springer, v. 58, n. 1, p. 1–38, 2015. Citado na página 20.
- WANG, S.; GROTH, P. Measuring the dynamic bi-directional influence between content and social networks. In: SPRINGER. **International Semantic Web Conference**. [S.l.], 2010. p. 814–829. Citado na página 7.
- WASSERMAN, S.; FAUST, K. **Social network analysis: Methods and applications**. [S.l.]: Cambridge university press, 1994. v. 8. Citado na página 16.
- WU, J. et al. General link prediction with influential node identification. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 523, p. 996–1007, 2019. Citado na página 11.
- YANG, S.-H. et al. Like like alike: joint friendship and interest propagation in social networks. In: ACM. **Proceedings of the 20th international conference on World wide web**. [S.l.], 2011. p. 537–546. Citado 4 vezes nas páginas 9, 16, 19 e 21.
- ZHOU, T.; LÜ, L.; ZHANG, Y.-C. Predicting missing links via local information. **The European Physical Journal B**, Springer, v. 71, n. 4, p. 623–630, 2009. Citado 2 vezes nas páginas 1 e 9.