UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

UM MODELO PARA AVALIAÇÃO DE RELEVÂNCIA CIENTÍFICA BASEADO EM MÉTRICAS DE ANÁLISE DE REDES SOCIAIS

AYSLÂNYA JERONIMO WANDERLEY

JOÃO PESSOA-PB
Março-2015

AYSLÂNYA JERONIMO WANDERLEY

UM MODELO PARA AVALIAÇÃO DE RELEVÂNCIA CIENTÍFICA BASEADO EM MÉTRICAS DE ANÁLISE DE REDES SOCIAIS

DISSERTAÇÃO APRESENTADA AO CENTRO DE INFORMÁTICA DA UNIVERSIDADE FEDERAL DA PARAÍBA, COMO REQUISITO PARCIAL PARA OBTENÇÃO DO TÍTULO DE MESTRE EM INFORMÁTICA (SISTEMAS DE COMPUTAÇÃO).

Orientador: Prof. Dr. Alexandre Nóbrega Duarte

JOÃO PESSOA-PB
Março-2015

W245u Wanderley, Ayslânya Jeronimo.
Um modelo para avaliação de relevância científica baseado em métricas de análise de redes sociais / Ayslânya Jeronimo Wanderley.- João Pessoa, 2015.
102f. : il.
Orientador: Alexandre Nóbrega Duarte
Dissertação (Mestrado) - UFPB/CI
1. Informática. 2. Sistemas de computação. 3. Ciência dos Dados. 4. Pesquisadores - avaliação. 5. Regressão Logística. 6. Redes sociais - análise.

UFPB/BC CDU: 004(043)

Ata da Sessão Pública de Defesa de Dissertação de Mestrado de **Ayslânya Jerônimo Wanderley**, candidata ao título de Mestre em Informática na Área de Sistemas de Computação, realizada em 30 de março de 2015.

Aos trinta dias do mês de março do ano de dois mil e quinze, às quinze horas, no Centro de Informática (Unidade Mangabeira) da Universidade Federal da Paraíba, reuniram-se os membros da Banca Examinadora constituída para examinar a candidata ao grau de Mestre em Informática, na área de "*Sistemas de Computação*", na linha de pesquisa "*Computação Distribuída*", a **Sra. Ayslânya Jerônimo Wanderley**. A comissão examinadora foi composta pelos professores doutores: **Alexandre Nóbrega Duarte (PPGI - UFPB)**, orientador e presidente da Banca, **Anand Subramanian (UFPB),** examinador interno e **Nazareno Ferreira de Andrade (UFCG),** examinador externo ao Programa. Dando início aos trabalhos, o professor Alexandre Nóbrega Duarte cumprimentou os presentes, comunicou aos mesmos a finalidade da reunião e passou a palavra à candidata para que a mesma fizesse, oralmente, a exposição do trabalho de dissertação intitulado "*Um modelo para avaliação de relevância científica baseado em métricas de Análise de Redes Sociais*". Concluída a exposição, a candidata foi arguida pela Banca Examinadora que emitiu o seguinte parecer: "*aprovada*". Eu, Nadja Rayssa Soares de Almeida, Secretária do Programa de Pós Graduação em Informática - PPGI, lavrei a presente ata que vai assinada por mim e pelos membros da Banca Examinadora. João Pessoa, 30 de março de 2015.

Nadja Rayssa Soares de Almeida

Profº Alexandre Nóbrega Duarte
Orientador (PPGI-UFPB)

Profº Anand Subramanian
Examinador Interno (PPGI-UFPB)

Profª Nazareno Ferreira de Andrade
Examinadora Externa (UFCG)

Agradecimentos

Primeiramente quero agradecer a Deus, Divino Pai Eterno, que sempre foi minha fortaleza nos momentos de angústia e desânimo durante essa caminhada. A Sua mão me guiou até aqui, abrindo as porta certas, no momento certo.

Ao meu orientador, Dr. Alexandre Nóbrega Duarte, por toda sua dedicação, suas ideias e contribuições que foram de extrema relevância para o desenvolvimento deste trabalho. Todo esse tempo de convivência me fez perceber não só a sua genialidade, mas também a sua competência como docente e orientador;

Aos meus pais, Vilma e Damião, que sempre incentivaram minha caminhada em busca do conhecimento, e nunca mediram esforços para me ajudar a alcançar meus sonhos;

A minha tia Maria do Céu Jeronimo (*in memorian*), que apesar de não estar ao meu lado fisicamente comemorando essa conquista, sempre estará presente dentro do meu coração, me dando forças para seguir em frente apesar de tudo;

Ao meu noivo Jhonatan, pelo amor, compreensão e companheirismo em todos os momentos em que estive ausente, sobretudo pelo apoio nas horas mais difíceis desta caminhada;

A minha amiga, quase irmã, Angélica Félix. Nessa caminhada não dividimos apenas o apartamento, dividimos sonhos, angústias, noites em claro estudando para as provas de Estrutura de Dados e também muitas conquistas. Que nossa parceria perdure por toda a nossa vida;

Aos colegas de Mestrado, em especial a galera do LabSNA, por todo o conhecimento compartilhado durante essa caminhada;

Aos professores do PPGI, pelos valorosos conhecimentos compartilhados ao longo desse período;

Aos companheiros de pesquisa, Mateus Prestes e Felipe Crispim, por todo o auxílio prestado no desenvolvimento deste trabalho;

Por fim, quero agradecer ao meu orientador da graduação e amigo, professor Pablo Ribeiro Suárez, por todo o incentivo e auxílio prestado, sem os quais essa conquista não seria possível.

Resumo

A tarefa de avaliar a relevância científica de um pesquisador nem sempre é trivial. Geralmente esse processo é baseado em índices que consideram a produção e o impacto do mesmo em sua área de pesquisa. Entretanto, a literatura aponta que tais indicadores tomados isoladamente são insuficientes uma vez que desconsideram os padrões de relação nos quais os pesquisadores se inserem. Além disso, muitos trabalhos já comprovaram que as relações de colaboração exercem forte impacto sobre a relevância de um pesquisador. Nesse contexto, entende-se que a modelagem e análise dessas relações pode ajudar a construir novos indicadores que complementem o processo de avaliação vigente. Sendo assim, o objetivo deste trabalho foi especificar um modelo estatístico que permite avaliar a relevância científica de um pesquisador, definida pela detenção de bolsa de produtividade do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), baseado em métricas aplicadas às suas redes de colaboração científica. Para tanto foram aplicadas métricas de Análise de Redes Sociais (ARS) às redes de colaboração de 1592 docentes vinculados aos Programas de Pós-Graduação na área de Ciência da Computação que posteriormente serviram como base para construção de um modelo de Regressão Logística utilizando a técnica de validação cruzada *10-fold* estratificada. O modelo proposto apresentou resultados bastante animadores e demonstrou que as métricas de ARS que mais influenciam na avaliação de relevância de um pesquisador são a Centralidade de Intermediação, o Grau Ponderado, o *PageRank* e o Coeficiente de Agrupamento Local, tendo as duas primeiras influência positiva e as duas últimas influência negativa. Isso demonstra que pesquisadores que desempenham um papel de intermediador dentro da rede e que costumam manter relacionamentos fortes com seus colaboradores são mais propensos a serem contemplados com bolsas de produtividade, enquanto que aqueles pesquisadores que possuem uma rede mais coesa e costumam colaborar com pesquisadores que já são líderes na sua área têm menor probabilidade de serem bolsistas.

Palavras-chave: Ciência dos Dados, Avaliação de Pesquisadores, Regressão Logística, Análise de Redes Sociais.

Abstract

The task of assessing the scientific relevance of a researcher is not always trivial. Generally, this process is based on indices that consider the production and the impact of it in their area of research. However, the literature indicates that such indicators taken separately are insufficient, since they ignore the standards of relationship in which researchers are inserted. In addition, many studies have proven that collaborative relationships have a serious impact on the relevance of a researcher. In this context, it is understood that the modeling and analysis of these relationships can help building new indicators that complement the current evaluation process. Thus, this work aimed to specify a statistical model which allows for assessing the scientific relevance of a researcher, defined by the detention of productivity grant from the National Council for Scientific and Technological Development (Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq), based on metrics applied to their scientific collaboration networks. Therefore, we applied metrics of Social Network Analysis (SNA) to collaborative networks of 1592 professors connected with Postgraduate Program in Computer Science area that later served as the basis for construction of a logistic regression model using the stratified 10-fold cross-validation technique. The proposed model produced very encouraging results and demonstrated that the SNA metrics that most influence in assessing the relevance of a researcher are the Betweenness Centrality, Weighted Degree, PageRank and Local Clustering Coefficient, having the first two positive influence and the last two negative influence. This shows that researchers who play an intermediary role within the network and usually maintain strong relationships with its collaborators are more likely to be contemplated with productivity grants, while those researchers with a more cohesive network and often collaborate with researchers who are already leaders in their field are less likely to be a scholarship student.

Keywords: Data Science, Researchers Assessment, Logistic Regression, Social Network Analysis.

Conteúdo

Lista de Símbolos

ACC : *Acurácia*

ARS : *Análise de Redes Sociais*

AUC : *Área sob a curva ROC*

AUT : *Autoridade*

CAL : *Coeficiente de Agrupamento Local*

CAPES : *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior*

CG : *Centralidade de Grau*

CI : *Centralidade de Intermediação*

CNPq : *Conselho Nacional de Desenvolvimento Científico e Tecnológico*

CP : *Centralidade de Proximidade*

ESP : *Especificidade*

FAP : *Fundações de Amparo à Pesquisa*

GML : *Graph Modeling Language*

GP : *Grau Ponderado*

LabSNA : *Laboratório de Social Networks Analysis*

PPGCCs : *Programas de Pós-Graduação em Ciência da Computação*

PPGs : *Programas de Pós-Graduação*

PR : *PageRank*

ROC : *Receiver Operating Characteristics*

SENS : *Sensibilidade*

VPN : *Verdadeiros Preditivo Negativo*

VPP : *Verdadeiros Preditivo Positivo*

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Este capítulo é destinado a descrever os principais aspectos que motivaram a realização deste trabalho (Seção 1.1), bem como o problema de pesquisa tratado nesta dissertação (Seção 1.2). Além disso, na Seção 1.3 são apresentados os objetivos do trabalho em questão e, por fim, a Seção 1.4 apresenta a organização dos demais capítulos deste documento.

1.1 Motivação

A avaliação da relevância de um pesquisador é de grande importância para comunidade científica. Entretanto, esse processo avaliativo ainda se configura como um grande desafio devido à análise subjetiva empregada no mesmo (DIGIAMPIETRI et al., 2014). Tal processo se fundamenta em critérios quantitativos (número de produção) e qualitativos (qualidade da pesquisa) para mensurar o mérito científico de um pesquisador.

Apesar desse modelo de avaliação ser amplamente utilizado por órgãos de fomento à pesquisa, ele possui diversas críticas por parte até dos próprios pesquisadores. Sendo assim, para garantir a meritocracia na distribuição de fundos de pesquisa, buscam-se novas abordagens que possam complementar a precisão das avaliações da relevância de um pesquisador.

Além dessa demanda por novos critérios de avaliação, observa-se um envolvimento cada vez maior de cientista em atividades de pesquisas colaborativas, motivado pelas características cada vez mais interdisciplinares, complexas e caras da ciência moderna (DIGIAMPIETRI et al., 2014). Outro fator de motivação, são os inúmeros trabalhos (LEE; BOZEMAN, 2005; ABBASI; ALTMANN; HOSSAIN, 2011; LIAO, 2011; ARAUJO et al., 2014) que

apontam o impacto significativo da colaboração científica nos índices de desempenho de um pesquisador.

Mais recentemente, o enorme desenvolvimento da metodologia de Análise de Redes Sociais (ARS) tem possibilitado a representação dessas relações de colaboração a partir de grafos sociais, que podem ser analisados mediante a aplicação das métricas desenvolvidas nessa metodologia. Isso permite o desenvolvimento de novas técnicas de avaliação do mérito científico baseadas no perfil de colaboração demonstrado pelo pesquisador.

Tais técnicas utilizam a análise de padrões encontrados nas redes de colaboração científica para determinar a relevância de um pesquisador. Nesse contexto, as pesquisas desenvolvidas nessa temática são caracterizadas como projetos de Ciência dos Dados, uma vez que demandam por conhecimentos do domínio estudado, de Ciência da Computação, Matemática e Estatística.

Diante dessa perspectiva, observam-se diversos trabalhos na literatura que usam a metodologia adotada na Ciência dos Dados para estabelecer relações entre as dinâmicas sociais das redes de colaboração e a relevância de um pesquisador no meio científico (MCCARTY et al., 2013; WAINER; VIEIRA, 2013b; CIMENLER; REEVES; SKVORETZ, 2014; BORDONS et al., 2015). Esses trabalhos evidenciam que a maneira como os pesquisadores colaboram dentro da rede, bem como as métricas de ARS aplicadas às suas redes são fatores determinantes para a avaliação do seu impacto científico.

Levando em consideração esse aspecto, este trabalho verificou a influência que as métricas de ARS aplicadas às redes de colaboração científica dos docentes vinculados aos programas de Pós-Graduação na área de Ciência da Computação exercem sobre a relevância científica dos mesmos para utilizá-las em um modelo estatístico capaz de avaliar o mérito científico de um pesquisador.

Esse modelo pode ser de grande utilidade na medida em que contribui significativamente para a melhor compreensão da importância das relações de colaboração para os propósitos da política científica, principalmente quanto aos objetivos de avaliação da relevância de um pesquisador.

1.2 Problema de Pesquisa

No Brasil, a avaliação individual dos pesquisadores é realizada pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) com a finalidade de contemplá-los com bolsas de produtividade, sendo os detentores dessas bolsas considerados pesquisadores de grande relevância científica.

Levando em conta esse fato, pesquisas vêm sendo realizadas no intuito de verificar se o perfil de colaboração dos membros desse grupo tem alguma influência sobre sua relevância científica, entre essas pesquisas, destaca-se a de (ARAUJO et al., 2014) que demonstrou que o perfil de colaboração dos bolsistas de produtividade em pesquisa do CNPq revela um forte impacto sobre a sua produtividade.

Levando em conta esse aspecto, este estudo aplicou métricas de ARS à rede de colaboração dos docentes vinculados ao quadro permanente dos Programas de Pós-Graduação na área de Ciência da Computação, com o intuito de responder a seguinte questão de pesquisa:

É possível identificar se um pesquisador é bolsista de produtividade através de sua interação em uma rede de colaboração científica?

A hipótese proposta para este estudo é que existe uma relação significativa entre algumas métricas de Análise de Redes Sociais e a relevância científica de um pesquisador, caracterizada pela detenção de bolsas de produtividade do CNPq. Sendo assim, do ponto de vista estatístico, seria possível estimar a probabilidade de influência das variáveis independentes estudadas sobre o mérito científico dos pesquisadores vinculados aos programas de Pós-Graduação na área de Ciência da Computação.

1.3 Objetivos

O objetivo geral desta dissertação é a construção e validação de um modelo capaz de avaliar a relevância científica de um pesquisador em termos de detenção de bolsa de produtividade a partir de métricas de ARS aplicadas as suas redes de colaboração científica.

Diante disso, para se alcançar o objetivo geral foram definidos os seguintes objetivos específicos:

1. Identificar quais as métricas de ARS, adotadas sob uma perspectiva individual, apre-

sentam relações com o mérito científico de um pesquisador;

2. Investigar a influência de cada métrica estudada sobre a relevância científica de um pesquisador;

3. Aplicar a técnica de Regressão Logística para modelar a associação das variáveis independentes (métricas de ARS) com a variável resposta, caracterizada pela detenção ou não detenção de bolsa de produtividade;

4. Analisar a capacidade de discriminação do modelo gerado na base de validação, a fim de verificar o desempenho.

1.4 Estrutura da Dissertação

Este documento está dividido em seis capítulos, incluindo este introdutório, que apresentou a motivação, o problema de pesquisa abordado e os objetivos deste trabalho. O restante do documento está organizado conforme descrito a seguir.

O Capítulo 2 contextualiza o problema tratado nesta dissertação, apresentando, inicialmente, os principais conceitos que norteiam os projetos de Ciência dos Dados, ressaltando a modelagem feita por Regressão Logística. Dando sequência, apresenta os fundamentos da Análise de Redes Sociais e apresenta a Plataforma Lattes, base de dados utilizada para gerar as redes desta pesquisa. Por fim, apresenta os principais aspectos considerados na avaliação do mérito científico no Brasil, enfatizando os critérios utilizados para concessão de bolsas de produtividade do CNPq.

O Capítulo 3 apresenta os trabalhos que já foram desenvolvidos acerca da temática tratada nesta pesquisa de modo que o leitor possa identificar o estágio atual do conhecimento referente ao tema.

O Capítulo 4 apresenta os detalhes da construção do modelo logístico bem como os testes realizados para verificar o ajuste do modelo aos dados.

O Capítulo 5 apresenta a avaliação dos modelos gerados no Capítulo 4, de modo a verificar a capacidade de discriminação dos mesmos, com o intuito de identificar qual modelo foi capaz de generalizar melhor os dados. Além disso, é apresentada uma análise do modelo final escolhido.

Por fim, o Capítulo 6 encerra o trabalho com uma análise dos resultados obtidos, apresentando as principais contribuições da pesquisa e suas limitações, e ainda as perspectivas para possíveis trabalhos futuros.

Capítulo 2

Fundamentação Teórica

O problema abordado nesta dissertação envolve uma série de conceitos básicos que precisam ser devidamente explicitados antes da apresentação dos resultados obtidos. Dessa forma, este capítulo apresenta uma breve fundamentação dos assuntos necessários para a compreensão desta pesquisa, dando ênfase aos pontos mais relevantes para a compreensão dos resultados. Esses assuntos foram subdivididos em seções, a fim de permitir uma melhor compreensão dos mesmos.

A Seção 2.1 apresenta uma visão geral sobre Ciência dos Dados, ressaltando os principais estágios envolvidos nos projetos desta área de pesquisa, dando ênfase ao modelo de regressão logística. A Seção 2.2 apresenta os conceitos básicos sobre Análise de Redes Sociais, enfatizando as Redes de Colaboração Científica e ainda, a Plataforma Lattes do CNPq, visto que os CVLattes dos pesquisadores foram a base para a construção das redes de colaboração científica deste trabalho. Por fim, a Seção 2.3 apresenta os principais aspectos relacionados à avaliação do mérito científico no Brasil, esclarecendo os critérios que envolvem a concessão de Bolsas de Produtividade no país.

2.1 Ciência dos Dados

O termo Ciência dos Dados (ou *Data Science*) é utilizado para definir uma área emergente da Ciência que se preocupa em extrair conhecimento a partir da coleta, preparação, análise, visualização, gestão e preservação de grandes quantidades de informação (STANTON, 2012). De forma sintetizada, Dhar (2013) define a mesma como o estudo da extração generalizável

de conhecimento a partir de dados.

Apesar de já existirem algumas iniciativas em instituições de ponta no exterior com foco em Ciência dos Dados, ela é considerada uma área recente no Brasil (PORTO; ZIVIANI, 2014). Por esse motivo, ainda existe divergência na literatura quanto a real abrangência do termo, que por muitas vezes é confundido com outras ciências, tais como, Ciência da Computação, Matemática e Estatística. Isso ocorre, em geral, porque a Ciência dos Dados envolve técnicas bastante semelhantes às utilizadas nas ciências citadas acima.

Todavia, Drew (2013) demonstra através de um diagrama de Venn (Figura 2.1) que os projetos realizados pela Ciência dos Dados envolvem três habilidades distintas: conhecimentos inerentes à Ciência da Computação (habilidades de *hackers*), conhecimentos matemáticos e estatísticos e domínio sobre uma área específica do conhecimento para que a análise possa ser interpretável.

Sendo assim, fica evidente que para se realizar um bom projeto de Ciência dos dados é preciso entender não só o domínio no qual foram gerados os dados, mas também as operações matemáticas realizadas durante a "aprendizagem"e a "melhoria"do modelo, e isso vai além da Estatística.

Figura 2.1: Diagrama de Venn com as habilidades necessárias para a Ciência dos Dados

Ainda acerca das habilidades necessárias, Porto e Ziviani (2014) destacam que apesar da Ciência dos Dados estar intimamente ligada a muitas das disciplinas tradicionais já estabelecidas, a mesma viabiliza uma nova área altamente interdisciplinar. É justamente a partir dessa interdisciplinaridade que é possível vislumbrar aplicações da mesma em diversos contextos de estudo.

Dessa forma, é possível vislumbrar como a Ciência dos Dados auxilia os pesquisadores a gerarem um conhecimento relevante a partir de grandes quantidades de dados encontrados durante sua pesquisa, sendo primordial para solucionar o problema tratado.

2.1.1 Estágios de um projeto de Ciência dos Dados

A Ciência dos Dados possui uma metodologia própria e sistemática para alcançar os objetivos a que se propõe. Zumel e Mount (2014) descrevem os seis estágios básicos que compõem um projeto dessa área de pesquisa:

1. **Definição do objetivo**

   Na primeira etapa de um projeto de Ciência dos Dados é preciso definir um objetivo mensurável e que possa ser quantificado. Neste estágio, deve-se aprender tudo sobre o contexto em que o problema está inserido.

2. **Coleta e armazenamento dos dados**

   Esta etapa abrange todo o processo de identificação dos dados que serão necessários para alcançar o objetivo traçado na primeira etapa, o processo de coleta desses dados, e o processo de armazenamento adequado para uma análise eficiente. Geralmente é o estágio mais demorado de todo o processo e também o mais importante.

3. **Modelagem**

   Esta etapa é onde o cientista de dados tenta extrair informações úteis a partir dos dados, a fim de atingir seus objetivos. Uma vez que muitos procedimentos de modelagem fazem suposições específicas sobre a distribuição e o relacionamento dos dados, haverá sobreposição e alternância entre a etapa de modelagem e o estágio de limpeza de dados. As tarefas de modelagem mais comuns são:

- Classificação - Decidir se algo pertence a uma categoria ou outra.
- *Scoring* - Prever ou estimar um valor numérico.
- *Ranking* - Aprender a ordenar itens por preferências.
- *Clustering* - Agrupar itens por similaridades
- Descoberta de relações - Encontrar correlações ou causas potenciais de efeitos observados nos dados.
- Caracterização - Gerar relatórios a partir dos dados.

Para cada uma dessas tarefas existem várias abordagens diferentes, sendo que o objetivo traçado e a qualidade dos dados disponíveis serão os norteadores da escolha do modelo certo para solucionar o problema que foi definido.

4. **Avaliação crítica do modelo**

   Na etapa de avaliação verifica-se se o modelo que foi gerado atende aos objetivos do projeto. Nela avalia-se o quanto o modelo é genérico, a sua precisão para resolução do problema, bem como se os resultados gerados pelo modelo fazem sentido no mundo real.

   Se nesta etapa for verificado que o modelo não atende aos pressupostos elencados acima será preciso retornar as etapas anteriores, sendo a partir da definição de objetivos mais realistas ou a partir da coleta de dados adicionais ou outros recursos necessários para alcançar seus objetivos originais.

5. **Apresentação e Documentação**

   Esta etapa é caracterizada pela apresentação dos resultados obtidos a partir do modelo final, seja através de um relatório ou de um artigo científico. Além disso, é importante também documentar o modelo implantado para fins de execução e manutenção do mesmo.

6. **Implantação e manutenção do modelo**

   Finalmente, com a etapa de implantação, o modelo é colocado à prova através de testes que devem verificar a sua precisão para resolver o problema encontrado no mundo real.

Nessa etapa pode surgir a necessidade de alguns ajustes no modelo final, de forma a manter sua acurácia.

Ainda de acordo com Zumel e Mount (2014) a execução das seis etapas envolvidas no projeto são regidas por um ciclo de vida básico que pode ser observado na Figura 2.2.

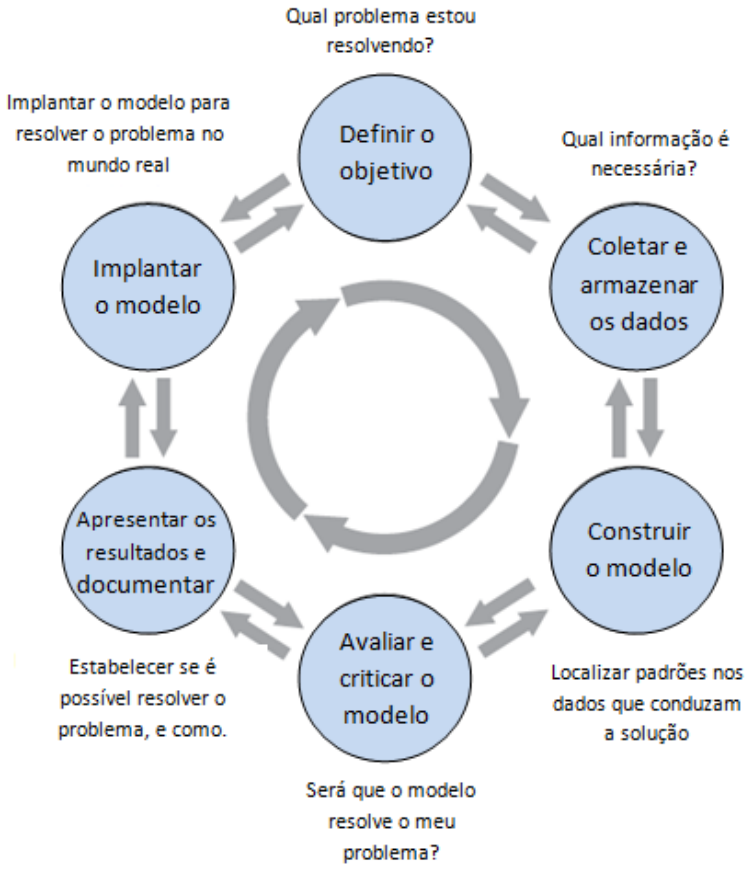


Figura 2.2: Ciclo de vida de um projeto de Ciência dos Dados

Através da figura é possível verificar que o encerramento de uma etapa não é definitivo, pois durante o projeto podem ocorrer diversas situações que façam retomar um estágio que havia sido concluído.

2.1.2 Regressão Logística Múltipla

Na maioria dos problemas tratados pela Ciência dos Dados é de grande interesse verificar se duas ou mais variáveis estão relacionadas de alguma forma, estabelecendo essa relação através de um modelo matemático. Este tipo de modelagem é denominado regressão, e ajuda

a entender como o comportamento de uma ou mais variáveis pode mudar o comportamento de outra.

Dentre as técnicas utilizadas pelos cientistas dos dados destaca-se o modelo de regressão logística, que é uma técnica estatística cujo objetivo é modelar, a partir de um conjunto de observações, a relação "logística" entre uma variável resposta dicotômica e uma série de variáveis explicativas numéricas (contínuas, discretas) e/ou categóricas (CABRAL, 2013).

A regressão logística é a principal representante de uma classe de modelos conhecidos como modelos lineares generalizados, uma vez que é um caso particular destes, usado em situações onde a variável resposta é binária. Sendo assim, ele é utilizado, normalmente, para resolver problemas em que a variável resposta apresenta como possíveis realizações um atributo e não uma mensuração.

No caso do problema proposto nesta dissertação, a variável resposta (ou variável dependente) indica se um pesquisador é ou não bolsista de produtividade do CNPq. Essa variável será denominada de **STATUS_PESQ** e assume apenas dois valores possíveis:

- 0 - para representar o pesquisador que não possui bolsa de produtividade.
- 1 - para representar o pesquisador que possui bolsa de produtividade.

De acordo com Hair, Anderson e Tatham (2009), para definir uma relação delimitada por zero e um, a regressão logística usa uma função assumida entre as variáveis independentes e a variável dependente que lembra uma curva em forma de "S", ou seja, uma função sigmoide, como pode ser observada na Figura 2.3.

Há dois tipos de regressão logística: a regressão logística simples, onde a variável resposta é explicada por uma única variável independente (explicativa) e a regressão logística múltipla onde a variável resposta é explicada por duas ou mais variáveis independentes (explicativas), como o caso tratado neste trabalho.

Segundo Favero (2009), quando a variável dependente Y assumir apenas dois possíveis estados (1 ou 0) e houver um conjunto de p variáveis independentes X_1 , X_2 , ... , X_p, o modelo de regressão logística é caracterizado como múltiplo e é dado pela Equação (2.1):

$$P(evento) = \frac{1}{1 + e^{-(Z)}} \tag{2.1}$$

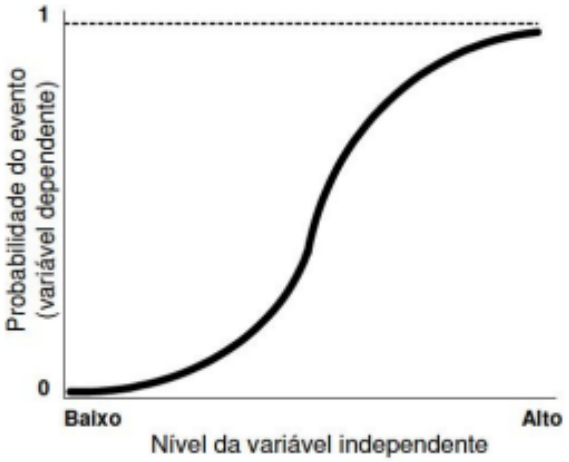


Figura 2.3: Forma da relação logística entre variáveis dependente e independente

onde, $Z = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$ e $B^T = (\beta_1, ..., \beta_k)$ é o vetor de parâmetros a ser estimado.

É importante destacar, conforme relata Silva (2011), que independentemente do número de variáveis usadas para definir o modelo de regressão logística, o foco sempre será distinguir dois grupos de indivíduos, consoante apresentem ou não determinadas características.

A grande utilização desse modelo na análise de variáveis de saída dicotômicas é justificada por dois motivos principais: sob o ponto de vista matemático, é razoavelmente flexível e fácil de ser utilizado; permite uma interpretação de resultados bastante rica e direta (HOSMER; LEMESHOW, 2000). Esses argumentos justificam o uso deste modelo para tratar o problema desta dissertação em detrimento de outros modelos alternativos apresentados na literatura como as redes neurais e a análise discriminante.

2.1.2.1 Métodos de seleção de variáveis

Na etapa de construção do modelo é necessário identificar quais as variáveis de entrada são importantes para a classificação da variável de saída. Na regressão logística é importante minimizar o número de variáveis para que o modelo resultante seja mais facilmente generalizado e mais estável numericamente. Assim, se utilizam procedimentos que de acordo com regras de decisão, adicionam ou removem variáveis do modelo.

Nesse sentido, Brocco (2006) ressalta que os métodos de seleção de modelos mais conhecidos são: o *backward*, o *forward* e o *stepwise*. Tais métodos são utilizados para selecionar quais variáveis mais influenciam o conjunto de saída podendo, assim, diminuir o número de variáveis a compor a equação de regressão (ALVES, 2013).

O método *backward* caracteriza-se por incorporar todas as variáveis e iterar através de um conjunto de etapas eliminando uma variável por vez. Se em uma determinada etapa não houver eliminação de alguma variável, o processo é então interrompido e as variáveis restantes definem o modelo final. Já o método *forward* caracteriza-se por considerar a variável de maior coeficiente de correlação amostral observado com a variável resposta. A cada etapa, uma variável pode vir a ser incorporada (KARAM, 2006).

O método *stepwise* (passo a passo) é considerado uma combinação dos dois métodos anteriores, pois realiza o processo de inclusão ou exclusão iterativa de variáveis do modelo, baseado em critérios tais com a estatística G e o teste Wald (KARAM, 2006). Nesse método as variáveis são selecionadas a cada passo, de acordo com critérios que otimizem o modelo, reduzindo a variância e evitando problemas de multicolinearidade (GONÇALVES; GOUVÊA; MANTOVANI, 2013).

Tal método é considerado mais robusto pelos pesquisadores. Por esse motivo será adotado neste trabalho. De acordo com Alves, Lotufo e Lopes (2013), o método *stepwise* realiza os seguintes procedimentos para seleção de variáveis:

1. Escolhe-se a variável X_k que possui o maior coeficiente de correlação para entrar no modelo.

2. Uma variável X_i entra no modelo, então se o coeficiente de correlação for maior que o anterior X_i permanece no modelo, caso contrário X_i sai do modelo.

3. X_i sai do modelo e se o coeficiente de correlação for menor que o anterior, X_i fica no modelo, caso contrário, X_i permanece fora do modelo. Este passo é repetido até que não tenha mais X_i para sair do modelo. Terminada esta etapa retorna-se ao passo 2 e este passo continua até que não tenham mais variáveis para entrar no modelo.

2.1.2.2 Estimação dos parâmetros

Na regressão logística um dos métodos mais utilizado para obter as estimativas para o vetor $\beta = (\beta_0, \beta_1, ..., \beta_p)$ dos parâmetros do modelo e a matriz de covariâncias é o método da máxima verossimilhança.

Tal método permite obter valores para os parâmetros desconhecidos, que maximizam a probabilidade de obter o conjunto de observações (HOSMER; LEMESHOW, 2000). Ele consiste, basicamente, em encontrar o valor de β que maximize o logaritmo neperiano da função de verossimilhança dado pela Equação (2.2):

$$l(\beta) = \sum_{i=1}^{n} [y_i x_i^T \beta - \ln(1 + exp(x_i^T \beta)] \tag{2.2}$$

Para tanto, é necessário derivar $l(\beta)$ em relação a cada um dos parâmetro $(\beta_0, \beta_1, ..., \beta_p)$ utilizando um processo iterativo, obtendo-se a Equação (2.3) para cada um dos parâmetros:

$$\frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} [y_i x_{ij} - \frac{exp(x_i^T \beta)}{1 + exp(x_i^T \beta)} x_{ij}] = \sum_{i=1}^{n} [y_i - \pi_i] x_{ij} \tag{2.3}$$

Assim, os valores de $\beta_0, \ldots, \beta_j$ que maximizam essa função são obtidos igualando-se as $p + 1$ equações a zero, sendo esse máximo garantido pelo fato da função logaritmo ser estritamente crescente.

$$\frac{\partial \beta}{\partial \beta_0} = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \widehat{\pi}_i = 0 \tag{2.4}$$

$$\frac{\partial \beta}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij} y_i - \sum_{i=1}^{n} x_{ij} \widehat{\pi}_i = 0 \tag{2.5}$$

As Equações (2.4) e (2.5) são solucionadas a partir de processos iterativos como o Newton-Raphson e o escore de Fisher, devido a não linearidade das mesmas.

2.1.2.3 Teste de Significância dos Coeficientes

Os métodos de inferência são baseados na teoria de máxima verossimilhança. Conforme a literatura atual, os testes estatísticos mais utilizados para verificar a significância dos coefici-

entes encontrados são o teste da razão de verossimilhança e o teste de *Wald*. Ambos os testes são descritos abaixo.

1. Teste da razão de verossimilhança

   O teste da razão de verossimilhança (LR) consiste em comparar os valores observados da variável resposta com os valores preditos obtidos dos modelos com e sem a variável em questão. A comparação desses valores é baseada no log da verossimilhança. A estatística desse teste é dada pela Equação (2.6) :

   $$LR = -2ln(L_s) + 2ln(L_c) \tag{2.6}$$

   em que L_s é a verossimilhança do modelo sem a covariável e L_c é a verossimilhança do modelo com a covariável.

   No caso da regressão logística múltipla, temos o interesse em saber se pelo menos uma variável é significativa para o modelo. Sob a hipótese nula, os p coeficientes são iguais a zero, assim, a estatística LR tem distribuição qui-quadrado com p graus de liberdade. Nesse caso, L_c é a verossimilhança do modelo com as p variáveis explicativas e L_s é a verossimilhança do modelo apenas com o intercepto (ACTION, 2014)

2. Teste de *Wald*

   O teste de *Wald* (W) averigua se uma determinada variável independente apresenta uma relação estatisticamente significativa com a variável dependente (CABRAL, 2013). Ele é comumente adotado para testar hipóteses relativas a um único parâmetro β_j, para $j = 1, 2, ..., p$.

   Sendo assim, suponha que um modelo contenha três variáveis explicativas X_1, X_2 e X_3 com coeficientes dados respectivamente por β_1, β_2 e β_3. A estatística de *Wald* é usada para testar a hipótese nula $\beta_1 = 0$ na presença de β_2 e β_3. Caso não existam evidências para rejeitar esta hipótese, conclui-se que a variável X_1 não é necessária no modelo na presença de X_2 e X_3.

   Considerando-se a hipótese nula $H_0 : \beta = \beta_0$, a estatística para este teste é dada pela

Equação 2.7:

$$W = \frac{(\widehat{\beta} - \beta_0)^2}{\widehat{Var}(\widehat{\beta})} \tag{2.7}$$

que sob a hipótese nula tem distribuição qui-quadrado com p graus de liberdade. Valores de W superiores ao valor tabelado da distribuição qui-quadrado indicam que a covariável associada a β é importante para explicar a variação da resposta (FERREIRA, 2007).

O teste de *Wald*, todavia, se comporta de maneira inesperada em determinadas situações, rejeitando a hipótese nula quando o coeficiente é significativo (HAUCK; DONNER, 1977). Sendo assim, os autores aconselham que os coeficientes identificados pelo teste de Wald como sendo estatisticamente não significativos, sejam testados novamente pelo teste da razão de verossimilhança. Se acontecer dos resultados divergirem é aconselhável confiar no teste da razão de verossimilhança.

2.1.2.4 Interpretação dos Coeficientes

Uma vez que os coeficientes foram estimados e realizaram-se os testes de significância dos mesmos, é necessário interpretar os valores obtidos. No caso da regressão logística o coeficiente da variável independente representa a taxa de mudança ou a inclinação da função *logit* para cada incremento unitário na unidade de medida da variável, considerando-se fixas as demais variáveis do modelo (MENDONÇA, 2008). Em outras palavras, o coeficiente da regressão logística indica o quanto a probabilidade de ocorrência de um evento aumenta com o aumento de uma unidade na variável independente.

Entretanto, como a principal suposição da regressão logística é que o logaritmo da razão entre as probabilidades de ocorrência e não ocorrência do evento é linear, é comum interpretar os coeficientes a partir dos *odds ratio* - OR, ou razão de chances. A *odds ratio* é calculada através da exponenciação do parâmetro estimado, sendo que se o seu valor for igual a 1, entende-se que a variável independente não contribui para a explicação da variável resposta, enquanto que valores maiores que 1 indicam que a variável independente aumentam as chances de ocorrência do evento e valores menores que 1 indicam que a variável independente diminue as chances de ocorrência do evento.

2.1.2.5 Medidas de Qualidade de Ajuste do Modelo

Em qualquer modelo de regressão, é necessário proceder à análise dos resíduos para validação da qualidade do modelo estimado. Sendo assim, são empregadas medidas para detectar diferenças significativas entre os valores observados e os valores estimados, sendo que as mais utilizadas são baseadas nos resíduos de *Pearson* e nos resíduos *Deviance* (CABRAL, 2013).

O resíduo de Pearson para o *j*-ésimo individuo é definido pela Equação (2.8):

$$r(y_j\widehat{\pi}_j) = r_j = \frac{y_j\widehat{\pi}_j}{\sqrt{\widehat{\pi}_j(1-\widehat{\pi}_j)}} \quad j = 1, 2, ..., n \tag{2.8}$$

A estatística de teste global baseada nos resíduos de *Pearson* designa-se pela estatística de qui-quadrado de *Pearson* e é calculada através da Equação (2.9):

$$\chi^2 = \sum_{j=1}^{n} r(y_j\widehat{\pi}_j)^2 \tag{2.9}$$

Já uma estatística alternativa é obtida à custa dos resíduos *Deviance*, que para o j-ésimo indivíduo é definida pela Equação 2.10:

$$d(y_j\widehat{\pi}_j) = d_j = \pm\left\{2[y_j ln(\frac{y_j}{\widehat{\pi}_j}) + (1-y_j)ln(\frac{1-y_j}{1-\widehat{\pi}_j})]\right\}^{1/2} \tag{2.10}$$

Assim, a estatística a utilizar é:

$$D = \sum_{j=1}^{n} d(y_j\widehat{\pi}_j)^2 \tag{2.11}$$

Há ainda o teste de *Hosmer-Lemeshow* (HOSMER; LEMESHOW, 2000), que associa os dados às suas probabilidades estimadas, da mais baixa a mais alta, e então faz um teste qui-quadrado para determinar se as frequências estimadas estão próximas das frequências observadas. A finalidade deste teste é verificar se existem diferenças significativas entre as classificações realizadas pelo modelo e a realidade observada.

Assim, a hipótese a ser testada é:

$$\begin{cases} H_0 : o_j = e_j, \quad \forall j = 1, ..., g \\ H_1 : \exists j | o_j \neq e_j, \quad j = 1, ..., g \end{cases}$$

A estatística deste teste sob a hipótese nula é dada pela Equação (2.12):

$$\chi_{HL}{}^2 = \sum_{j=1}^{g} \frac{(o_j - e_j)^2}{e_j(1 - \frac{e_j}{n_j})} = \sum_{j=1}^{g} \frac{(o_j - \bar{p}_j)^2}{n_j \bar{p}_j (1 - \bar{p}_j)} \sim \chi^2{}_{g-2} \tag{2.12}$$

onde n_j é o número de observações pertencentes ao grupo j, verificando-se $n = \sum_{j=1}^{g} n_j$, o_j é a frequência de sucesso observada no grupo j, onde $o_j = \sum_{j=1}^{n_j} y_{ij}$ e y_{ij} é a i-ésima observação do grupo j, e_j é a frequência esperada de sucesso no grupo j, onde $e_j = n_j \bar{p}_j$ e $\bar{p}_j = \dfrac{\sum_{j=1}^{n_j} \hat{p}_j}{n_j}$ e $\hat{p}_j$ é a probabilidade predita correspondente à i-ésima observação do grupo j.

2.1.2.6 Avaliação do Poder de Discriminação do Modelo

Como o objetivo do modelo logístico é predizer o valor de uma variável binária, classificando o indivíduo em um dos dois grupos, a avaliação da capacidade preditiva do modelo construído é essencial. Para tanto se faz necessário a aplicação do modelo em um conjunto de teste diferente daquele que foi utilizado para o ajuste do modelo.

De acordo com Brocco (2006), para se classificar um indivíduo em dois grupos, com base na probabilidade da resposta predita, é necessário identificar um valor limiar, conhecido como ponto de corte. Geralmente 0,5 é um valor razoável, entretanto, se os dois grupos não podem ser classificados como simétricos, um valor diferente de 0,5 deve ser considerado. Uma maneira de se determinar este valor é através da curva ROC (*Receiver Operating Characteristics*), que permite avaliar a capacidade preditiva de um modelo usando o ponto de corte escolhido.

Esta análise é feita por meio de um método gráfico simples e robusto, que permite estudar a variação da sensibilidade e especificidade, para diferentes valores de corte. Conforme Brocco (2006), o melhor ponto de corte produz valores para sensibilidade e especificidade que se localiza no "ombro" da curva, ou próximo dela, ou seja, no ponto mais à esquerda e superior possível. Na Figura 2.4, verifica-se três graus de discriminação possíveis fornecidos pelas curvas ROC (BRAGA, 2000).

A partir da figura, infere-se que quanto mais a curva estiver distante da diagonal principal, melhor o desempenho de modelo associado a ela, já que a linha diagonal representa um modelo de classificação aleatória. Esse fato sugere que quanto maior for a área entre a curva

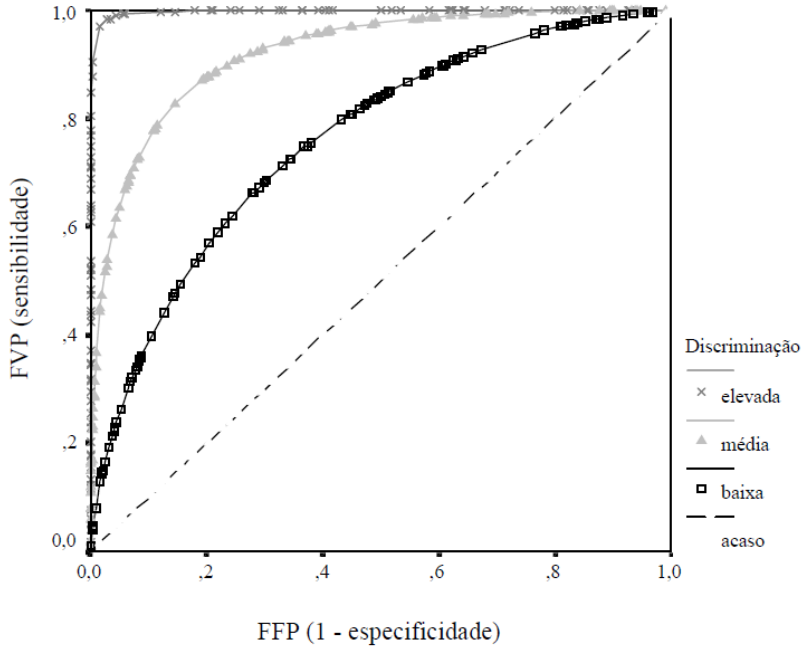


Figura 2.4: Curvas ROC representativas de três graus de capacidade de discriminação

ROC produzida e a diagonal principal, melhor o desempenho global do modelo (OLIVEIRA, 2011). Acerca desse aspecto, Hosmer e Lemeshow (2000) apresentam uma regra geral para avaliação do resultado da área sob a Curva ROC:

- Se a área for igual a 0,5: não há discriminação;
- área no intervalo entre 0,7 e 0,8: discriminação aceitável;
- área no intervalo entre 0,8 e 0,9: excelente discriminação;
- área acima de 0,9: excepcional discriminação.

Após a determinação do ponto de corte, é importante avaliar o poder de discriminação do modelo, isto é, avaliar a capacidade preditiva de um modelo após a classificação das observações em um dos dois grupos.

Para tanto é necessário criar a matriz de classificação, que descreve uma tabulação cruzada entre a classificação predita através de um único ponto de corte e a condição real e conhecida de cada indivíduo, onde a diagonal principal representa as classificações corretas e valores fora dessa diagonal correspondem a erros de classificação (BROCCO, 2006). Um exemplo da matriz de classificação pode ser visto na Tabela 2.1.

Tabela 2.1: Exemplo de Matriz de Classificação

		Valor Observado	
Valor Predito		Y = 0	Y = 1
	Y = 0	VN (verdadeiro negativo)	FN (falso negativo)
	Y = 1	FP (falso positivo)	VP (verdadeiro positivo)

A partir dos valores observados nessa matriz são calculadas as seguintes métricas de desempenho, descritas em (ACTION, 2014):

- Acurácia: é definida pela proporção de predições corretas, considerando-se apenas os acertos totais, sem diferenciar os positivos dos negativos. É dada pela Equação (2.13):

$$ACC = \frac{(VP + VN)}{(P + N)} \tag{2.13}$$

- Sensibilidade: é definida pela proporção de verdadeiros positivos, ou seja, avalia a capacidade do modelo classificar um indivíduo como evento 1 dado que realmente ele é 1. É calculada através da Equação (2.14):

$$SENS = \frac{(VP)}{(VP + FN)} \tag{2.14}$$

- Especificidade: é definida pela proporção de verdadeiros negativos, ou seja, ela avalia a capacidade do modelo predizer um indivíduo como evento 0 dado que ele realmente é 0. É calculada através da Equação 2.15:

$$ESPEC = \frac{(VN)}{(VN + FP)} \tag{2.15}$$

- Verdadeiro Preditivo Positivo: é a proporção de verdadeiros positivos em relação a todas as predições positivas, sendo calculada pela Equação (2.16):

$$VPP = \frac{(VP)}{(VP + FP)} \tag{2.16}$$

- Verdadeiro Preditivo Negativo: proporção de verdadeiros negativos em relação a todas

as predições negativas, sendo calculada pela Equação (2.17):

$$VPN = \frac{(VN)}{(VN + FN)} \tag{2.17}$$

2.1.3 Validação Cruzada

O modelo de regressão logística busca por padrões encontrados nos dados que possam generalizar situações que possivelmente ocorrerão no futuro. Dessa forma, é importante que a validação do modelo gerado seja feita em um conjunto diferente do que foi usado para o ajuste do modelo, sendo fundamental uma partição criteriosa do conjunto de dados, de modo a torná-los representativos.

Nesse contexto, um dos métodos estatísticos mais utilizados para divisão dos dados entre conjunto de treinamento e conjunto de teste é a validação cruzada (do inglês, *cross validation*), que segundo Herrera et al. (2004) consiste em particionar o conjunto de dados em subconjuntos mutuamente exclusivos, gerando assim um subconjunto para a estimação dos parâmetros do modelo (dados de treinamento) e o restante dos subconjuntos (dados de teste) é empregado na validação do modelo.

A partir do método original da validação cruzada se derivam outras técnicas, uma delas é conhecida como *multifold-cross-validation* ou *k-fold cross validation*. Nessa técnica, o conjunto de dados original de N exemplos é dividido em K subconjuntos, onde $K > 1$. O treinamento do modelo ocorre em todos os subconjuntos exceto em um, e o erro de validação é medido testando-o sobre o subconjunto deixado de fora. O procedimento descrito é repetido por K tentativas, cada vez utilizando um subconjunto de validação diferente (HAYKIN, 1999). O percentual de classificações corretas é acumulado para todas as observações da amostra, indicando a precisão global do modelo.

Essa abordagem apresenta como vantagem a utilização de um grande conjunto de dados durante o treinamento, além de gerar conjuntos mutuamente exclusivos que cobrem todo o conjunto de dados (TAN et al., 2009). Tal técnica é utilizada, normalmente, quando o número de observações é pequeno.

2.2 Análise de Redes Sociais

As redes sociais podem ser definidas como um conjunto de nós interligados por arestas, onde esses nós representam pessoas (denominados de atores) e as arestas correspondem aos laços gerados pelas interações sociais existentes entre elas (RECUERO, 2005). Complementando essa definição, Marteleto (2001) considera que as redes sociais são representações de um conjunto de participantes autônomos, unindo ideias e recursos em torno de valores e interesses compartilhados.

A partir da modelagem dessas interações sociais através de grafos, surgiu uma nova metodologia de pesquisa, denominada de Análise de Redes Sociais (ARS) cujo foco é a interação que ocorre entre as diversas entidades que compõem uma rede social, bem como os padrões e implicações dessas interações (WASSERMAN; FAUST, 1994).

Ainda segundo os autores, a Análise de Redes Sociais evoluiu a partir da combinação de ciências como Matemática, Antropologia e Sociologia. Por esse motivo a aplicação da metodologia de ARS permite descrever de forma quantitativa aspectos importantes dos fenômenos sociológicos.

Na ARS o estudo do comportamento, das escolhas e das orientações de um indivíduo está intimamente ligado ao conjunto de relações estabelecido pelo mesmo através de suas interações, e não somente pelos seus atributos individuais (MARTELETO, 2001).

De acordo com Guerra (2012), a ARS emergiu como um conjunto de métodos para a análise da estrutura social que permitem, especificamente, uma investigação dos aspectos relacionais dessas estruturas. A utilização destes métodos, portanto, depende da disponibilidade de relações ao invés de dados de atributos.

Sendo assim, a metodologia de ARS torna-se imprescindível quando se pretende verificar a influência da estrutura da rede na explicação dos fenômenos pesquisados, pois evidencia o impacto das interações em diversos aspectos sociais de um indivíduo.

Conforme a característica dos métodos utilizados é possível dividir a ARS em duas vertentes distintas: a análise global (*global network analysis*), que se preocupa em investigar as características da rede como um todo e a análise individual (*ego networks analysis*), onde a rede de um indivíduo é analisada com o intuito de obter informações sobre a quantidade de nós com o qual o mesmo se relaciona, sobre a sua posição dentro da rede, entre outros fato-

res (MENA-CHALCO; DIGIAMPIETRI; JR., 2012). Assim, cada uma dessas abordagens fornecem um conjunto de métricas específicas desenvolvidas dentro da própria metodologia.

Neste trabalho, a análise das redes é feita utilizando-se uma abordagem individual, dado que o interesse desta pesquisa é relacionar as métricas associadas a um pesquisador com seu mérito científico. Essa modalidade de análise permite enxergar a influência do grupo sobre um indivíduo, e ainda verificar o quanto suas conexões na rede podem influenciar em suas oportunidades e restrições na mesma(AZEVEDO, 2011). Sendo assim, as principais métricas utilizadas na abordagem de análise individual são expressas a seguir:

1. **Centralidade de Grau** (*Degree Centrality*): representa o número de ligações que um nó possui (CHELMIS; PRASANNA, 2011). A centralidade de grau de um nó i é expressa pela Equação (2.18):

$$C_G(i) = \sum_{j=1}^{n} a_{ij} \tag{2.18}$$

onde a_{ij} indica se existe ligação entre o nó i e o nó j (se existir, então $a_{ij} = 1$, caso contrário, $a_{ij} = 0$) e n representa o número de nós dentro da rede.

2. **Centralidade de Intermediação Normalizada** (*Normalized Betweenness Centrality*): representa a quantidade de vezes que um determinado nó aparece no caminho geodésico entre dois nós da rede, sendo expressa pela Equação (2.19):

$$C_I(i) = \frac{\sum_{j,k \wedge i \neq j \neq k} \frac{g_{jik}}{g_{jk}}}{\frac{(n-1)(n-2)}{2}} \tag{2.19}$$

onde n é o número de nós, g_{jk} é o número de caminhos mais curtos do nó j para o nó k, e g_{jik} é o número de caminhos mais curtos de nó j para o nó k que passam pelo nó i. Esta métrica permite analisar o potencial de comunicação de um ator dentro da rede (SILVA; MA; ZENG, 2008)

3. **Centralidade de Proximidade Normalizada** (*Normalized Closeness Centrality*): mede o comprimento médio dos caminhos mais curtos de um vértice para cada um dos outros vértices de um grafo. A centralidade de proximidade de um vértice i é

calculada pela Equação (2.20):

$$C_C(i) = \frac{n-1}{\sum_{j=1}^{n} e_{ij}} \tag{2.20}$$

onde n é o número de nós e e_{ij} é o número de arestas existentes no caminho mais curto do nó i para o nó j. Esta métrica indica a capacidade de alcance de um nó dentro da rede (CHELMIS; PRASANNA, 2011).

4. **Grau Ponderado** (*Weighted Degree*): Foi definido por Abbasi e Altmann (2011) como a soma de todos os pesos das arestas ligadas a um nó, sendo expressa pela Equação (2.21).

$$G_P(i) = \sum_{j=1}^{n} w_{ij} \tag{2.21}$$

onde n é o número de nós, w_{ij} representa o peso da aresta entre o nó i e o nó j, ou seja, representa a quantidade de vezes que os dois nós se relacionaram. Tal métrica evidencia a força das relações entre os pesquisadores.

5. **Autoridade** (*Authority*): o valor de autoridade é calculado somando-se a quantidade de *hubs* com o qual o nó em questão está conectado. Seu cálculo é feito através do Algoritmo HITS (KLEINBERG, 1999). A Equação (2.22) descreve como se obtém a autoridade de um nó dentro da rede, onde $h(j)$ representa o número de *hubs* ligado ao nó j.

$$A(i) = \sum_{j=1}^{n} h(j) \tag{2.22}$$

6. **Coeficiente de Agrupamento Local** (*Local Clustering Coefficient*): o coeficiente de agrupamento local foi introduzido por Watts e Strogatz (1998) para representar o quanto os vizinhos de um determinado nó estão conectados. Sendo assim, ele é definido como a razão entre as arestas existentes e as arestas possíveis entre os vizinhos

de um dado vértice. Em grafos não orientados, ele é definido pela Equação (2.23):

$$CC(i) = \frac{2|\{e_{jk} : v_j, v_k \in N_i, e_{jk} \in E\}|}{K_i(K_i - 1)} \tag{2.23}$$

onde CC_i é o coeficiente de agrupamento local, K_i é o número de vértices do grafo, j e k representam os vértices e j para k, e N_i são os vizinhos imediatamente ligados a i.

7. ***PageRank***: o *PageRank* (PAGE et al., 1998), se caracteriza como uma métrica de prestígio. Tal métrica atribui maior peso aos atores que se conectam com diferentes nós e com nós que já estão bem conectados. Em outras palavras, atores com *PageRank* elevado estão associados a nós populares dentro da rede, sendo assim, o objetivo dessa métrica é avaliar a qualidade das conexões e não só a quantidade, como acontece nas métricas de centralidade (KUMAR; JAN, 2014). Segundo Benevenuto, Almeida e Silva (2011), o *PageRank* (PR) de um nodo i, pode ser calculado através de Equação 2.24

$$PR(i) = (1 - d) + d \sum_{v \in S(i)} \frac{PR(v)}{Nv} \tag{2.24}$$

onde $S(i)$ é o conjunto de páginas que apontam para i, Nv denomina o número de arestas que saem do nodo v, e o parâmetro d é um fator que pode ter valor entre 0 e 1.

No contexto das redes de colaboração científica, as métricas mencionadas acima permitem auferir a popularidade, a posição e o prestígio de um pesquisador dentro de sua rede. Esse motivo justifica a escolha dessas métricas para o desenvolvimento deste trabalho, em detrimento de tantas outras disponibilizadas pela ARS. É importante ressaltar ainda que as métricas de centralidade de Intermediação e de Proximidade utilizadas nesta pesquisa são normalizadas a fim de permitir a comparação entre atores de redes diferentes.

2.2.1 Redes de Colaboração Científica

O processo de produção científica está cada vez mais pautado na ideia de colaboração. Segundo Sidone, Haddad e Mena-Chalco (2014) o aumento o perfil colaborativo é uma característica preponderante da ciência moderna, sendo independente da área, visto que 75% dos

artigos produzidos atualmente no mundo estão associados a autores de diferentes instituições.

A colaboração científica pode ser definida como o trabalho conjunto de pesquisadores para atingir um objetivo comum que consiste em produzir novos conhecimentos científicos (KATZ; MARTIN, 1997). Essa colaboração é motivada por diversos fatores, entre os quais se destacam: o desejo de aumentar a popularidade científica, a visibilidade e o reconhecimento pessoal; a possibilidade de maior divulgação da pesquisa e a possibilidade de aumento da produtividade (VANZ, 2010).

Segundo Katz e Martin (1997) são considerados colaboradores: (i) os pesquisadores que trabalham juntos ao longo de um projeto ou durante parte considerável dele; (ii) os pesquisadores que fazem frequentes e substanciais contribuições; (iii) os pesquisadores cujos nomes aparecem no projeto de pesquisa original e (iv) os responsáveis por um ou mais elementos da pesquisa.

Fica evidente que a produção científica colaborativa, em suas diferentes formas de publicação, acaba constituindo pontos de conexão que permitem sua caracterização como verdadeiras redes sociais. Essas conexões podem ser modeladas através de grafos, emergindo assim as chamadas redes de colaboração científica, também chamadas de redes sociais acadêmicas.

De acordo com Freire e Figueiredo (2011), este tipo de rede pode ser definido como uma rede social onde os pesquisadores são representados por nós e as relações de colaboração entre eles sejam elas publicações, orientações, participações em projetos, entre outras, formam arestas entre estes nós. Sendo assim, as redes de colaboração científica evidenciam a maneira como os pesquisadores interagem entre si, possibilitando uma análise mais profunda do impacto dessa interação sobre a produtividade científica.

Uma vez que essas redes também são caracterizadas como redes sociais, pois representam uma estrutura social composta por nós (pesquisadores) e arestas (relações de colaboração), observa-se um crescimento da aplicação da metodologia de ARS para compreender a dinâmica das relações de cooperação entre os pesquisadores.

Nesta pesquisa os grafos de colaboração científica utilizados para a análise são ponderados e não orientados, ou seja, suas arestas representam o número de trabalhos colaborativo entre dois pesquisadores distintos. Um exemplo de grafo de colaboração gerado neste traba-

lho pode ser observado na Figura 2.5.

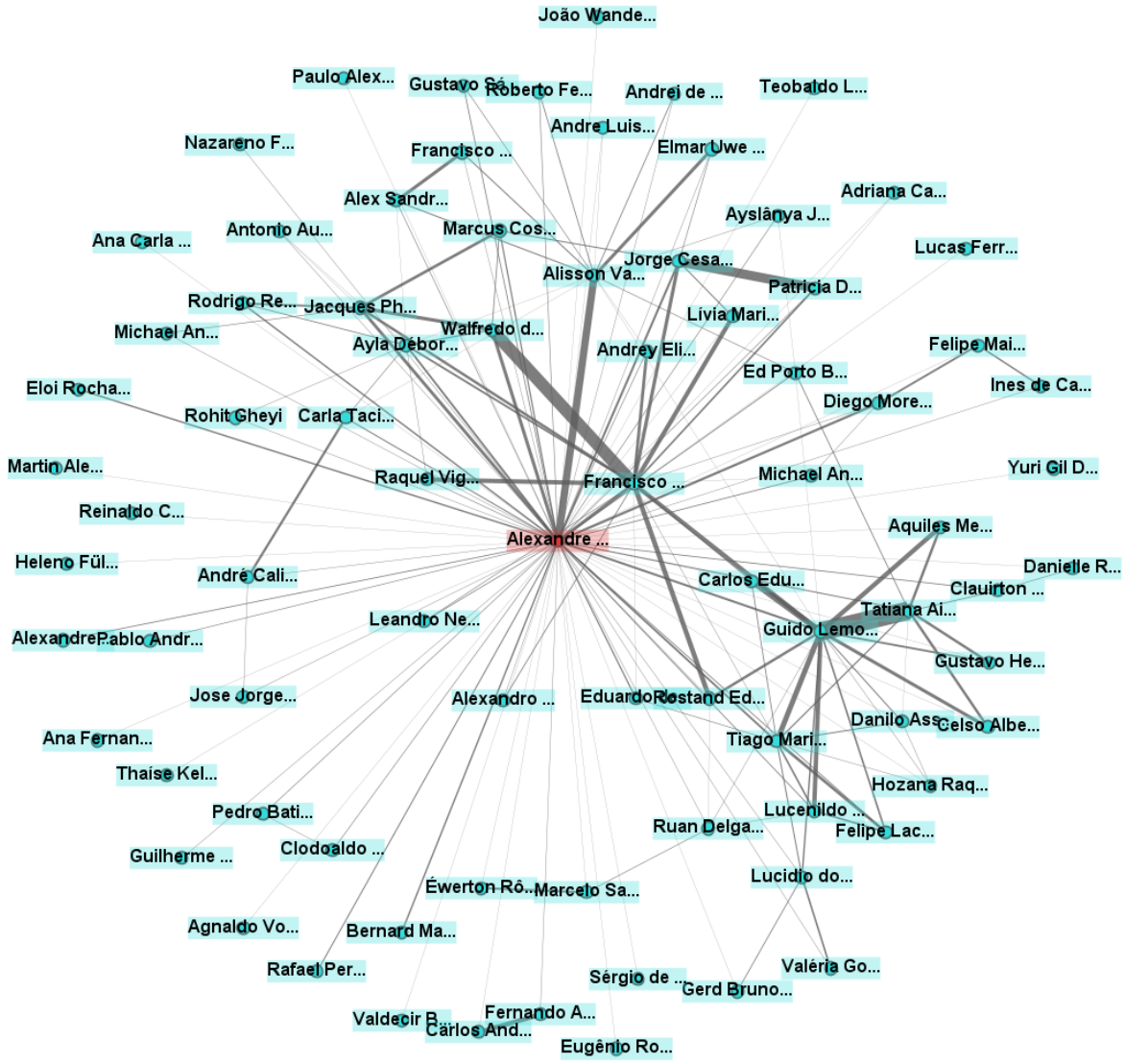


Figura 2.5: Exemplo de uma rede de colaboração científica representada por um grafo

É possível verificar através da figura a existência de arestas mais grossas, as quais indicam que existiram diversas colaborações entre os dois pesquisadores conectados, configurando assim um peso maior a essa ligação.

Em seu trabalho Newman (2003) destaca que em redes de colaboração científica ocorre uma interação local muito maior do que em outros tipos de redes, sendo que o processo no qual um pesquisador apresenta um colaborador à outro pesquisador é de fundamental importância para o crescimento da comunidade científica.

2.2.2 Plataforma Lattes

A Plataforma Lattes (PL) é um sistema de informação desenvolvido e implantado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) para gerenciar informações relacionadas a pesquisadores e instituições do país (CNPQ, 2014).

Criada em 1999, a partir de um projeto desenvolvido pelas Universidades Federal de Santa Catarina e Federal de Pernambuco em parceria com empresas privadas, a plataforma visa integrar e padronizar as informações dos currículos dos pesquisadores, de modo a facilitar a busca e seleção dos mesmos, bem como gerar estatísticas que orientem as políticas públicas de educação (GUERRA, 2012).

Segundo Mena-Chalco, Digiampietri e Jr. (2012), os currículos Lattes tornaram-se um padrão nacional utilizado na avaliação individual das atividades científicas, acadêmicas e profissionais dos pesquisadores, ressaltando que o número de cadastros na Plataforma Lattes já ultrapassa a marca dos quatro milhões. Dada essa magnitude, as agências de fomento, centros de pesquisas e departamentos das universidades utilizam a plataforma para extrair dados dos pesquisadores para análise de currículos.

De acordo com CNPq (2014), todos os bolsistas de pesquisa, de mestrado, de doutorado e de iniciação científica, orientadores credenciados e outros membros da comunidade ligados ao CNPq devem ter o currículo cadastrado na Plataforma. Sendo assim, cada pesquisador é responsável pelo preenchimento do seu currículo, sendo solicitadas dos mesmo as seguintes informações:

1. Dados pessoais do autor do currículo.
2. Dados da formação acadêmica e titulação.
3. Atuações profissionais.
4. Trabalhos em projetos de pesquisa.
5. Produções (bibliográficas, técnicas ou artísticas).
6. Apresentações em eventos.
7. Orientações em andamento ou concluídas.

Para cada currículo cadastrado na plataforma, o sistema gera um código de identificação único que compõe o *link* do endereço do currículo, conforme pode ser observado na Figura 2.6. Esse código é especialmente importante nessa pesquisa, pois é utilizado pelo *crawler*[1] para extrair as informações e gerar as redes dos pesquisadores.

Figura 2.6: Página com endereço do CV-Lattes, enfatizando o código de identificação.

Diante da existência de um identificador único, a Plataforma Lattes garante que pesquisadores que possuam nomes semelhantes sejam reconhecidos corretamente, se tornando assim um dos bancos de dados de pesquisadores mais robustos na atualidade (LANE, 2010).

Sendo assim, ao digitar o endereço da Plataforma Lattes *(http:/lattes.cnpq.br/)* seguindo do número de identificação do pesquisador é possível verificar todas as informações em sua página, incluindo publicações, participações em bancas, orientações de mestrado e doutorado, etc. Um exemplo das informações sobre publicações em anais de congresso contidas na página de um pesquisador pode ser vista na Figura 2.7.

Nota-se, através da figura, que diversas publicações foram feitas em conjunto com outros pesquisadores, sendo essas associações utilizadas para modelar a rede de colaboração do pesquisador associado ao currículo Lattes em questão.

É importante ressaltar que grande parte dos editais de financiamento de projetos feitos por instituições de amparo à pesquisa, como o próprio CNPq, utilizam os currículos Lattes dos pesquisadores como uma das formas de avaliação das propostas. Esse fato motiva os

[1] Aplicativo que acessa páginas web e recupera "dados"para uma futura análise.

Capítulos de livros publicados

1. BRASILEIRO, F. V. ; **DUARTE, A.** ; AQUINO, M. S. ; ALMEIDA, M. J. S. C. ; SOUZA, A. D. D. ; GUERRA, F. V. A. . ComuniNet - Comunicação Social na Internet: Webdifusão de um Serviço de Áudio e de Vídeo para a UFPB. In: Lúcia de Fátimaa Guerra Ferreira; Iraci Araújo Ferreira. (Org.). Prêmio Elo Cidadão. João Pessoa: Editora Universitária, 1999, v. 21, p. 173-181.

Trabalhos completos publicados em anais de congressos

1. doi BELE, J. C. S. ; WANDERLEY, A. ; ABREU, G. ; PORTO, E. ; BRITO, A. V. ; **DUARTE, A.** . Infography Usage in a Systematic Mapping About Social Network Analysis. In: 11th International FLINS Conference on Uncertainty Modeling in Knowledge Engineering and Decision Making, 2014, João Pessoa. Proceedings of the 11th International FLINS Conference on Decision Making and Soft Computing, 2014. p. 30-38.
2. doi FALCÃO, E. L. ; ARAÚJO, T. M. U. ; **DUARTE, A.** . Deaf Accessibility as a Service: uma Arquitetura Elástica e Tolerante a Falhas para o Sistema de Tradução VLIBRAS. In: XII Workshop de Computação em Clouds e Aplicações, 2014, Florianópolis. Anais do XXXII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos, 2014. p. 16-22.
3. doi BAPTISTA, V. M. P. S. ; BRITO, F. M. ; PEREIRA, J. ; ALMEIDA, F. ; LIMA, P. ; **DUARTE, A.** ; PORTO, E. ; GUIMARÃES, S. D. P. . Uma ferramenta para analisar mudanças na coesão entre parlamentares em votações nominais. In: III Brazilian Workshop on Social Network Analysis and Mining, 2014, Brasília. Anais do XXXIV Congresso da Sociedade Brasileira de Computação, 2014.
4. doi WANDERLEY, A. ; **DUARTE, A.** ; BRITO, A. V. ; PRESTES, M. A. ; CRISPIM, F. . Identificando correlações entre métricas de Análise de Redes Sociais e o h-index de pesquisadores de Ciência da Computação. In: III Brazilian Workshop on Social Network Analysis and Mining, 2014, Brasília. Anais do XXXIV Congresso da Sociedade Brasileira de Computação, 2014.
5. doi MENDONÇA-JUNIOR, M. L. ; BRITO, A. V. ; **DUARTE, A.** . Uma Ferramenta para Extração Semiautomática e Análise de Relevância de Artigos Científicos. In: III Brazilian Workshop on Social Network Analysis and Mining, 2014, Brasília. Anais do XXXIV Congresso da Sociedade Brasileira de Computação, 2014.
6. doi SOUZA, J. ; LYRA, D. ; CAVALCANTI, J. ; SIMÃO, R. ; FILHO, Z. ; **DUARTE, A.** ; BRITO, A. V. . Análise de redes de palavras baseada em títulos extraídos de um sistema de atendimento. In: III Brazilian Workshop on Social Network Analysis and Mining, 2014, Brasília. Anais do XXXIV Congresso da Sociedade Brasileira de Computação, 2014.
7. doi BRITO, F. E. ; Costa, R. E. O. ; **Duarte, A** . Sobre o Uso de Model Canvas em Planos de Gerenciamento de Dados para Curadoria Digital em Projetos de Pesquisa. In: VIII Brazilian e-Science Workshop, 2014, Brasília. Anais do XXXIV Congresso da Sociedade Brasileira de Computação, 2014.

Figura 2.7: Exemplo de página do CV-Lattes com informações sobre as publicações

pesquisadores a manter seus currículos com informações corretas e atualizadas, tornando a Plataforma Lattes uma fonte adequada para análise da produção científica brasileira (FARIAS; VARGAS; BORGES, 2012).

Apesar dessa vasta abrangência, a Plataforma Lattes ainda possui restrições quanto a recuperação de informações dos pesquisadores, pois não permite uma busca automatizada de um conjunto de currículos de pesquisadores de uma mesma área de pesquisa, por exemplo. Sendo assim, torna-se necessário o desenvolvimento de *crawlers* que realizem esse procedimento de forma automática, como o que é utilizado neste trabalho.

2.3 Mérito Científico no Brasil

A questão da avaliação do mérito científico de pesquisadores é tarefa diária de agências de fomento, de editores científicos, de gestores de política científica e dos próprios pesquisadores em todo o mundo. Mesmo assim, a busca por indicadores para avaliação da influência de um cientista ainda é um grande desafio para agências responsáveis por realizar esse processo.

No caso do Brasil os procedimentos de avaliação do mérito científico são realizados por várias instituições que atendem a diferentes objetivos. Sendo as principais delas: o CNPq

(Conselho Nacional de Desenvolvimento Científico e Tecnológico), a CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), as FAPs (Fundações de Amparo à Pesquisa), e as Universidades.

É de responsabilidade do CNPq o apoio à formação de recursos humanos para as atividades de investigação científica, a partir da concessão de bolsas e auxílios para a pesquisa (CAFE, 2012). Nesse sentido, o órgão, que é vinculado ao Ministério da Ciência e Tecnologia, realiza trienalmente a avaliação individual dos pesquisadores que pleiteiam bolsas de produtividade em pesquisa e em desenvolvimento tecnológico e extensão inovadora.

Na abordagem adotada pelo CNPq, a avaliação do pesquisador é realizada por outros cientistas que atuam na mesma área de pesquisa do avaliado, sendo assim denominada de "avaliação por pares"ou *peer review* (CAFE, 2012). Esse tipo de avaliação tem como objetivo preservar a autonomia da ciência através da utilização de mecanismos autorreguladores que evitam a necessidade de outros agentes interferirem nesse campo (BINOTTO; HOFF; SIQUEIRA, 2008).

O mérito científico de um pesquisador tem sido comumente auferido através de métricas de produção e impacto. As medidas de produção consideram o número de artigos publicados por um pesquisador durante sua carreira acadêmica, enquanto as métricas de impacto avaliam o quanto a produção do cientista teve importância para sua área de pesquisa (WAINER; VIEIRA, 2013a).

Sendo assim, é comum que as avaliações realizadas para a concessão de financiamento levem em consideração os resultados alcançados em termos de produtividade científica de um pesquisador. Entretanto, muitos pesquisadores se queixam de que as ferramentas usadas nesse processo focam mais na quantidade do que na qualidade.

Como todos os indicadores tomados de forma isolada, as análises de produção e impacto só refletem parcialmente a relevância da atividade científica, desconsiderando outros fatores que também são importantes, como por exemplo, as relações de colaboração científica estabelecida pelo pesquisador.

Segundo Martins e Ferreira (2013), os padrões de relação nos quais os pesquisadores se inserem, bem como a relação entre estes padrões e os atributos e comportamentos individuais são por muitas vezes excluídos do processo de avaliação da relevância de um pesquisador.

Em contrapartida, Spinak (1998) afirma que a atividade científica deve ser analisada e

interpretada dentro do contexto social em que está inserida, uma vez que as avaliações de desempenho científico devem levar em conta o contexto social, econômico e histórico da sociedade em que se aplicam, não podendo, assim, ser medida em escala absoluta.

Nesse sentido, nota-se um crescente esforço em adotar, durante o processo de avaliação, um conjunto variados de métricas que permitem expressar de forma mais fidedigna a relevância de um pesquisador. Diante desse contexto, muitas pesquisas vêm sendo desenvolvidas com o objetivo de relacionar os aspectos das interações sociais dos pesquisadores com sua produção.

2.3.1 Bolsas de Produtividade do CNPq

Com o intuito de fomentar a pesquisa no país, o CNPq concede diversas modalidades de bolsas, entre as quais se destacam as Bolsas de Produtividade em Pesquisa (PQ) e as Bolsas de Produtividade em Desenvolvimento Tecnológico e Extensão Inovadora (DT), que são regidas pela Resolução Normativa RN 016/2006 (CNPQ, 2014).

Tanto as Bolsas de Produtividade em Pesquisa quanto em Desenvolvimento Tecnológico e Extensão Inovadora são organizadas em níveis, em ordem crescente: 2, 1D, 1C, 1B, 1A, onde cada nível provê uma compensação salarial também crescente.

As bolsas PQ são concedidas para pesquisadores de todas as áreas do conhecimento com o objetivo de distinguir seu trabalho e valorizar sua produção. Entre os critérios para a concessão estão a produção científica, a participação na formação de recursos humanos e a efetiva contribuição para a área de pesquisa, além do requisito de ser doutor titulado a pelo três anos. O período de vigência da bolsa para pesquisador nível 1A é de 60 meses, níveis 1B, 1C e 1D de 48 meses e de 36 meses para os bolsistas da categoria 2. (CNPQ, 2014).

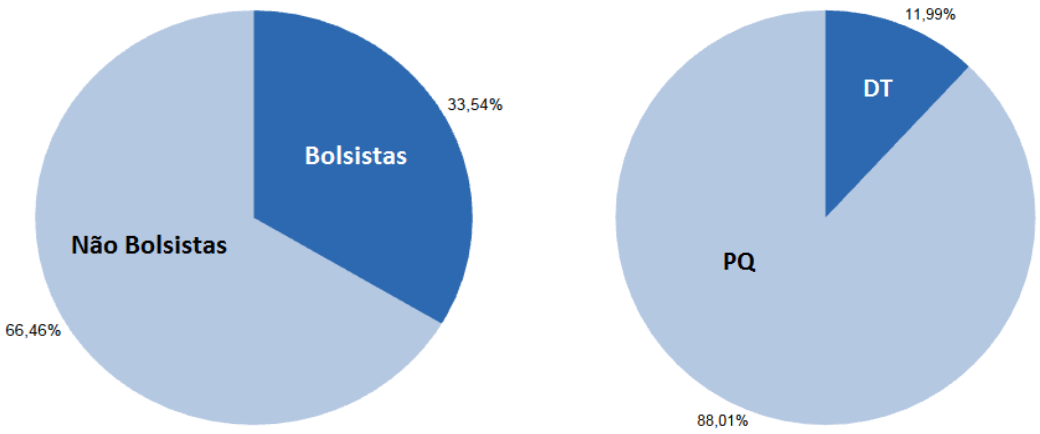
Já as bolsas DT tem como finalidade distinguir o pesquisador doutor, valorizando sua produção em desenvolvimento tecnológico e inovação. São requisitos para a obtenção desta bolsa ter um bom e crescente histórico de formação de recursos humanos, produção e transferência de tecnologia, e um projeto de pesquisa claramente inovador. O período de vigência dessa bolsa é semelhante a modalidade PQ de acordo com o nível (CNPQ, 2014). E assim como acontece com as bolsas PQ existe a exigência titulação há pelo menos três anos.

Cabe ressaltar, que os bolsistas de produtividade do CNPq, quaisquer que seja sua área, passam por um processo avaliativo extremamente competitivo, onde sua qualidade científica

é analisada e julgada pelos comitês assessores de suas respectivas áreas. De acordo com Wainer e Vieira (2013a), o número de bolsas de produtividade é fixo, por subárea. Dessa forma, só é possível atribuir uma nova bolsa de nível 1C a um pesquisador durante a avaliação, se outro pesquisador perder sua bolsa 1C, nesta mesma avaliação.

Os pesquisadores detentores destas bolsas são considerados a elite da comunidade em sua respectiva área do conhecimento, sendo a bolsa um critério para permitir o acesso a diversos canais de fomento, à coordenação de institutos nacionais de pesquisa (INCTs) e para avaliar os programas de pós- graduação pela CAPES (LIMA; VELHO; FARIA, 2012).

Por esses motivos, a distribuição das bolsas PQ e DT é extremamente limitada, atingindo uma parcela mínima de pesquisadores em cada área do conhecimento. Para ilustrar essa afirmação, a Figura 2.8(a) apresenta a porcentagem de pesquisadores bolsistas dentro do universo formado pelos docentes dos Programas de Pós-Graduação (PPGs) brasileiros na área de Ciência da Computação e a Figura 2.8(b) mostra a distribuição dos bolsistas nas modalidades PQ e DT.



(a) Porcentagem de bolsistas e não bolsistas (b) Porcentagem de PQs e DTs

Figura 2.8: Distribuição do número de bolsas PQs e DTs entre os docentes vinculados aos programas de Pós-Graduação na área de Ciência da Computação

A partir da Figura 2.8(a) observa-se que os bolsistas de produtividade nas modalidades PQ e DT correspondem a aproximadamente um terço da população de docentes. Em valores absolutos isso corresponde a 534 pesquisadores detentores de bolsa e 1508 pesquisadores não bolsistas. Já a Figura 2.8(b) expressa que no universo formado pelos 534 bolsistas, verifica-se uma proporção maior de bolsista PQ em relação aos bolsistas DT.

2.4 Considerações Finais

O objetivo deste capítulo foi fornecer um conjunto mínimo de conhecimento acerca dos assuntos abordados nesta dissertação. Nesse sentido, foram abordados inicialmente aspectos inerentes à Ciência dos Dados para uma melhor compreensão dos procedimentos utilizados nessa área da Ciência, com enfoque na modelagem de Regressão Logística Múltipla.

Além disso, foram apresentados conceitos e métricas referentes à Análise de Redes Sociais, com ênfase nas redes de colaboração científica e uma breve apresentação da Plataforma Lattes, fonte dos dados utilizados neste trabalho de Mestrado. Por fim, foram apresentados os principais aspectos referentes à avaliação do mérito científico no Brasil, priorizando os aspectos relevantes para a concessão de bolsas de produtividade.

No próximo capítulo serão apresentados os principais trabalhos relacionados à esta pesquisa, analisando os principais resultados encontrados nos mesmos, bem como um comparativo destes com esta dissertação.

Capítulo 3

Trabalhos Relacionados

Neste capítulo são descritos alguns trabalhos relacionados à pesquisa realizada nesta dissertação. Na Seção 3.1, são apresentadas pesquisas que investigam o impacto da colaboração científica entre pesquisadores na avaliação do mérito científico dos mesmos. A Seção 3.2 destina-se a descrever trabalhos que realizam predição de relevância científica de um pesquisador a partir de métricas de ARS. Já a Seção 3.3 apresenta uma breve discussão sobre os principais resultados encontrados pelos autores e sua contribuição para esta pesquisa.

3.1 Impacto das Colaborações no Mérito Científico

A caracterização das redes de colaboração científica, bem como o estudo do padrão de conectividade existente entre os pesquisadores atraíram diversos esforços de pesquisa na última década. Grande parte dos artigos focados nessa área busca investigar aspectos relacionados aos efeitos das relações de colaboração sobre a produtividade e o impacto de um pesquisador.

Lee e Bozeman (2005) apresentaram um estudo sobre o impacto da colaboração científica sobre a produtividade de um pesquisador. Os autores utilizaram como base de dados o *curriculum vitae* de 443 cientistas norte-americanos para gerar uma rede de coautoria entre os mesmos. A partir de duas abordagens distintas, a contagem total e a contagem fracionada de artigos publicados (alcançada dividindo o número de publicações pelo número de autores), os autores identificaram que em termos de contagem total a colaboração possui correlação significativa com a produção, já em termos de contagem fracionada o número de colaboradores não é um preditor de produtividade significativo.

Liao (2011) demonstrou através de seu trabalho que a colaboração científica produz efeitos significativos na qualidade de pesquisa. O autor investigou se existe relação entre a intensidade de colaboração e a diversidade de membros em coautoria com os índices de qualidade científica, representados em termos de citações, fator de impacto e quantidades de prêmios de pesquisa. Sendo assim, ele mensurou a intensidade de colaboração baseado na rede de coautoria em artigos científicos, utilizando o conceito de centralidade da ARS e a diversidade de membros dividindo o número de artigos produzidos em coautoria pelo número de colaboradores distintos.

No entendimento dos autores, os resultados demonstraram existir uma correlação positiva significativa entre a intensidade de colaboração e as três métricas de qualidade de pesquisa, enquanto que a diversidade de membros não produz efeitos sobre as mesmas.

Em uma pesquisa mais recente, Kumar e Jan (2013) analisaram a rede de colaboração em pesquisa na área de Negócios e Gestão da Malásia buscando demonstrar os efeitos das métricas de popularidade, posição e prestígio sobre a produtividade em pesquisa. Sendo a principal relevância deste artigo para essa pesquisa a constatação de que a popularidade de pesquisadores e a força e diversidade de suas conexões apresentam efeitos significativos sobre o seu desempenho.

Alves, Benevenuto e Laender (2013) desenvolveram um estudo interessante evidenciando que os pesquisadores influentes dentro das redes de colaboração são essenciais para a evolução das comunidades científicas. Em seu trabalho, os autores propuseram uma nova métrica capaz de evidenciar a importância de um pesquisador para a comunidade científica como um todo. A partir disso, foram analisadas as características topológicas dos membros pertencentes ao núcleo de comunidades científicas, verificando-se que os mesmos são responsáveis por aumentar o grau médio da rede e diminuir o agrupamento, pois se conectam com pequenos grupos distintos. Isso revela que os pesquisadores com altos índices de produção, geralmente pertencem a núcleos de comunidade científica.

No contexto brasileiro, Araujo et al. (2014) realizaram um estudo utilizando como base de dados a Plataforma Lattes. Nele os autores utilizaram os 2,7 milhões de currículos cadastrados nessa base de dados para gerar um grafo bipartido entre os autores e os artigos publicados em periódicos. A partir disso, foram feitas diversas análises com a rede gerada, sendo a principal contribuição desse estudo para o desenvolvimento desta pesquisa os resul-

tados encontrados na análise do subconjunto de pesquisadores bolsistas do CNPq.

Segundo os autores, as estatísticas de produtividade e colaboração científica dos pesquisadores contemplados com bolsas de estudo diferem dos pesquisadores regulares, ressaltando que eles apresentam maior probabilidade de manter contatos em outros grupos de pesquisa, apresentando assim uma rede com menos componentes gigantes. Com isso, eles concluíram que os bolsistas de produtividade são pesquisadores com alto grau de colaboração e produzem mais trabalhos.

Ainda considerando o contexto brasileiro, identifica-se o trabalho de Digiampietri et al. (2014), onde foi estudada a relação entre os Programas de Pós-Graduação em Ciência da Computação (PPGCCs), a partir dos trabalhos produzidos em colaboração por seus docentes, explorando características como produção bibliográfica e métricas de redes sociais. De acordo com os autores, a principal contribuição do trabalho é a apresentação de novas métricas e a exploração de relacionamentos para caracterizar a produtividade dos programas estudados.

A partir da análise da estrutura da rede os autores identificaram que os programas mais bem localizados topologicamente são mais produtivos. Embora esse resultado seja focado na colaboração entre os PPGCCs, ele reflete também a colaboração dos pesquisadores, uma vez que as arestas se formam através das conexões entre eles.

Sendo assim, os resultados apresentados nesse artigo tem importância significativa para este trabalho, uma vez que foram obtidos a partir de um conjunto de dados semelhante ao utilizado nesta dissertação.

3.2 Predição de Relevância Científica a partir de Métricas de ARS

Tomando por base as evidências de correlação significativa entre métricas de ARS e a relevância científica de um pesquisador, muitos autores têm concentrado esforços em desenvolver modelagens matemáticas que possam expressar essa relação de forma clara e objetiva. Na literatura são observados diversos trabalhos propondo modelos capazes de prever a relevância científica de um pesquisador a partir de métricas de ARS.

Um dos trabalhos pioneiros que representaram os efeitos da colaboração científica sobre

a produtividade de um pesquisador através de uma modelagem matemática foi o de Eaton et al. (1999). Nessa pesquisa, os autores analisaram a relação entre os aspectos estruturais de uma rede de coautoria com a produtividade dos autores, utilizando três níveis de análise: a rede total (formada por toda a amostra), a *macronetwork* (formada pelos autores que mais publicam frequentemente) e 20 *micronetworks* (agrupamentos de autores que publicam com frequência).

A partir do coeficiente de correlação de *Pearson*, os autores investigaram a relação entre métricas de ARS (centralidade de grau, centralidade de intermediação e centralidade de grau ponderado) e a quantidade de artigos publicados pelo autor, observando um grau de correlação superior a 0,70 entre essas métricas em todos os níveis de análise, exceto pela centralidade de intermediação, onde verificou-se correlação forte apenas na análise de *macronetwork*. Além disso, os autores apresentaram um modelo de regressão linear múltipla para os níveis de análise total e de *macronetwork*, e um modelo de regressão linear simples para a análise das *micronetworks*.

O modelo de regressão completo para o nível de análise total, contou com as seguintes variáveis explicativas: centralidade de grau, centralidade de grau ponderado, número de artigos com um único coautor, e conseguiu explicar 89% da variação do número de publicações. Já em relação à análise da *macronetwork* o modelo de regressão completo foi semelhante ao anterior acrescido apenas da variável centralidade de intermediação, sendo responsável por 91% da variação do número de publicações.

Em termos da análise de regressão simples, 19 das 20 *micronetworks* apresentam centralidade de grau responsável por 66,4% da variação do número de publicação, centralidade de grau ponderado responsável por 69,5% e a centralidade de intermediação responsável por 73,3% dessa variação. E ainda 14 *micronetworks* tem o número de artigos publicados por um único coautor sendo responsável por 51,3% da variação do número de publicações. Sendo assim, os autores concluíram que a produtividade está intimamente relacionada com a posição estrutural de um pesquisador em todos os três níveis de colaboração analisados.

Abbasi, Altmann e Hossain (2011) estudaram a correlação existente entre algumas métricas de ARS, coletadas a partir de uma rede de coautoria entre pesquisadores de 5 universidades e o desempenho científico dos mesmos mensurado em termos de citações. As métricas investigadas pelos autores foram centralidade de grau, centralidade de intermediação, cen-

tralidade de proximidade, centralidade de autovetor (todas normalizadas), eficiência e média dos laços fortes.

Aplicando-se a análise de correlação de posto de *Spearman* os autores observaram uma correlação significativa positiva entre as métricas de centralidade de grau, centralidade de autovetor, média laços fortes e eficiência e o *g-index*. A partir desses resultados, eles propuseram um modelo matemático, utilizando a análise de regressão múltipla de *Poisson*, para prever o impacto dessas métricas sobre o *g-index* de um pesquisador. Com isso, chegaram à conclusão que é possível prever o desempenho de um pesquisador a partir da análise de sua rede de colaboração.

No trabalho de McCarty et al. (2013) é realizada uma investigação sobre as características da rede de coautoria de 238 autores do Web of Science com o intuito de gerar um modelo preditivo, tendo como variável resposta o *h-index* e como variáveis explicativas métricas que refletem o comportamento colaborativo de um autor, a estrutura de colaboração e as características dos coautores, todas elas encontradas a partir da geração de redes egocêntricas para cada um dos pesquisadores. Dessa forma, o modelo de regressão linear multivariado final contou com quatro variáveis explicativas:

- Netsize: Número de autores que compõem a rede;
- Hierarchy: O quanto os coautores são intermediados por um único autor;
- MeanTie: Número médio de artigos publicados entre os co-autores;
- MeanAlterH: Média do *h-index* dos co-autores.

O modelo proposto apresentou $R^2 = 0,69$ e 59% da variação do *h-index* foi explicada pelo tamanho da rede. Com base nesses resultados, os autores sugerem que o impacto científico aumenta à medida que o pesquisador colabora com um maior número de coautores, dando preferência a autores que já possuem alto impacto científico.

Outro estudo bastante relevante sobre predição com métricas de ARS foi realizado por Cimenler, Reeves e Skvoretz (2014). Os autores estenderam a pesquisa de Abbasi, Altmann e Hossain (2011) utilizando uma base de dados mais ricos gerando um modelo bivariado de regressão de *Poisson* associando o *h-index* dos pesquisadores com métricas de ARS. As métricas utilizadas pelos autores foram a centralidade de grau, a centralidade de proximidade, a

centralidade de intermediação, a centralidade de autovetor, a média dos laços fortes, o coeficiente de eficiência de Burt e o coeficiente de agrupamento local, extraídos de quatro tipos de redes distintas: rede de comunicação, rede de publicação conjunta, rede de propostas de concessão conjuntas e rede de patentes conjuntas.

Os resultados obtidos pelos autores a partir da regressão indicaram que o grau de centralidade foi estatisticamente significativo e teve um impacto positivo em todas as redes, exceto na rede de comunicação. A centralidade de proximidade e de autovetor foram estatisticamente significativos e tiveram impacto positivo sobre o *h-index* em todas as redes. A centralidade de intermediação teve um impacto positivo significativo apenas para a rede de publicações conjuntas. A média dos laços fortes foi estatisticamente significativa, e teve um impacto positivo apenas para a rede de publicações conjuntas e patentes. O coeficiente de eficiência teve um impacto positivo significativo apenas para a rede de patentes. E por fim, o coeficiente de agrupamento local foi estatisticamente significativo e teve um impacto positivo apenas para a rede de publicações conjuntas e propostas de concessão.

Sarigol et al. (2014) apresentaram em seu trabalho uma abordagem diferente das anteriores para realizar predição do sucesso científico baseado em redes de coautoria. Os autores utilizaram métodos de aprendizado de máquina com base em métricas de posição de autores em redes de coautoria no momento da publicação para prever se um artigo será muito citado cinco anos mais tarde.

A abordagem utilizada pelos autores considera como variáveis preditoras a centralidade de grau, a centralidade de autovetor, a centralidade de intermediação e centralidade *k-core* dos autores. Dessa forma, eles avaliaram como a posição de um autor na sua rede de coautoria no momento da publicação de um artigo, pode influenciar no impacto dessa publicação no futuro.

Os resultados encontrados pelos autores indicam que o método proposto permitiu uma predição precisa do futuro sucesso de um pesquisador em termos de citações, evidenciando forte relação entre a posição de autores em redes de colaboração científica e seu sucesso futuro em termos de citações.

Em um trabalho mais recente, Bordons et al. (2015) analisou as redes de colaboração de três áreas distintas (Nanociência, Farmacologia e Estatística) explorando a relação entre o desempenho individual dos pesquisadores mensurado através do *g-index* e a posição do

cientista na rede de coautoria.

Os autores utilizaram o modelo de regressão múltipla de *Poisson* para explorar até que ponto existe relação entre o *g-index* de cientistas e sua posição nas redes de coautoria. As métricas utilizadas foram centralidade de grau, centralidade de intermediação, centralidade de proximidade, centralidade de autovetor, coeficiente de agrupamento e média dos laços fortes.

O modelo ajustado obtido na pesquisa foi considerado significativo nas três áreas analisadas obtendo $R^2 = 0,652$ em Farmacologia, $R^2 = 0,573$ em Nanociência $R^2 = 0,195$ em Estatística, sendo que as variáveis que mostram uma relação mais forte com o *g-index* são a média de laços fortes e o grau normalizado.

Outro resultado bastante interessante desse trabalho foi a constatação que o *g-index* sofre maior influência das métricas relativas a posição de um pesquisador na rede em campos de pesquisa experimentais (Nanociência e Farmacologia) do que em campo de pesquisas teóricos (Estatística).

3.3 Considerações Finais

Observando os trabalhos apresentados na Seção 3.1, percebeu-se que as redes de colaboração têm desempenhado papel significativo nos índices de produção e impacto de um pesquisador. Por conseguinte, alguns resultados interessantes dos artigos mencionados foram utilizados para subsidiar esta pesquisa.

Analisando essa vasta literatura, encontraram-se aspectos importantes utilizados no desenvolvimento deste trabalho, como por exemplo, uma forte tendência de utilização de métricas relacionadas com a posição de um pesquisador dentro da rede para mensurar seu impacto científico (LIAO, 2011; KUMAR; JAN, 2013).

Além disso, os resultados propostos por Araujo et al. (2014) foram utilizados para considerar como fator de relevância neste trabalho o fato do pesquisador possuir bolsa de produtividade do CNPq.

Já a partir da análise dos trabalhos reportados na seção 3.2, verifica-se um crescente esforço de pesquisa na busca pela predição do mérito científico de um pesquisador a partir de métricas encontradas em suas redes de colaboração científica.

Com o objetivo de proporcionar uma melhor visualização desses trabalhos e do enquadramento desta dissertação em comparação aos mesmos, a Tabela 3.1 apresenta uma classificação destes em termos de perspectiva utilizada na análise da rede, de métricas abordadas, de fator de relevância considerado e de modelagem utilizada para predição.

Tabela 3.1: Comparação entre os trabalhos relacionados e enquadramento da dissertação

Trabalho	Modelagem da rede	Métricas Abordadas	Fator de Relevância	Modelagem Preditiva
[Eaton et. al, 1999]	Global	Métricas de centralidade	*h-index*	Regressão Linear Múltipla e Simples
[Abbasi et. al 2011]	Global	Métricas de centralidade, posição e prestígio	*g-index*	Regressão de *Poisson*
[Mccarty, 2013]	Egocêntrica	Métricas de estrutura, posição e características dos coautores	*h-index*	Regressão Linear Múltipla
[Cimenler et al. 2014]	Global	Métricas de centralidade, posição e prestígio	*h-index*	Regressão de *Poisson*
[Sarigöl et al. 2014]	Egocêntrica	Métricas de centralidade, posição e prestígio	Número de citações	Aprendizagem de Máquina
[Bordons et al. 2015]	Global	Métricas de centralidade, posição e prestígio	*g-index*	Regressão de *Poisson*
Esta Dissertação	Egocêntrica	Métricas de centralidade, prestígio e posição	Detenção de bolsa de produtividade	Regressão Logística Múltipla

Dos seis trabalhos relacionados à predição de relevância científica, cinco utilizaram um modelo de regressão para associar às métricas de ARS com os índices de relevância (*h-index* e *g-index*). Levando-se em consideração esse cenário, optou-se por utilizar nesta dissertação um modelo de regressão, porém a abordagem utilizada é a Regressão Logística Múltipla, devido à natureza dicotômica da variável resposta estudada, diferentemente do que se observa

nos trabalhos estudados.

Uma característica preponderante que distingue essa pesquisa dos trabalhos relatados neste capítulo é a análise voltada ao impacto das métricas de ARS no tocante a detenção de bolsas de produtividade do CNPq e não apenas a índices de desempenho. Essa abordagem pode fornecer grandes contribuições aos comitês de avaliação científica do país, pois permite uma melhor compreensão das diferenças de colaboração entre bolsistas e não bolsistas.

Capítulo 4

Construção e Ajuste do Modelo

Neste capítulo são apresentadas as principais etapas envolvidas na construção e ajuste do modelo de classificação proposto neste trabalho. A Figura 4.1 apresenta de maneira sucinta a metodologia aplicada para se chegar ao modelo de regressão logística final.

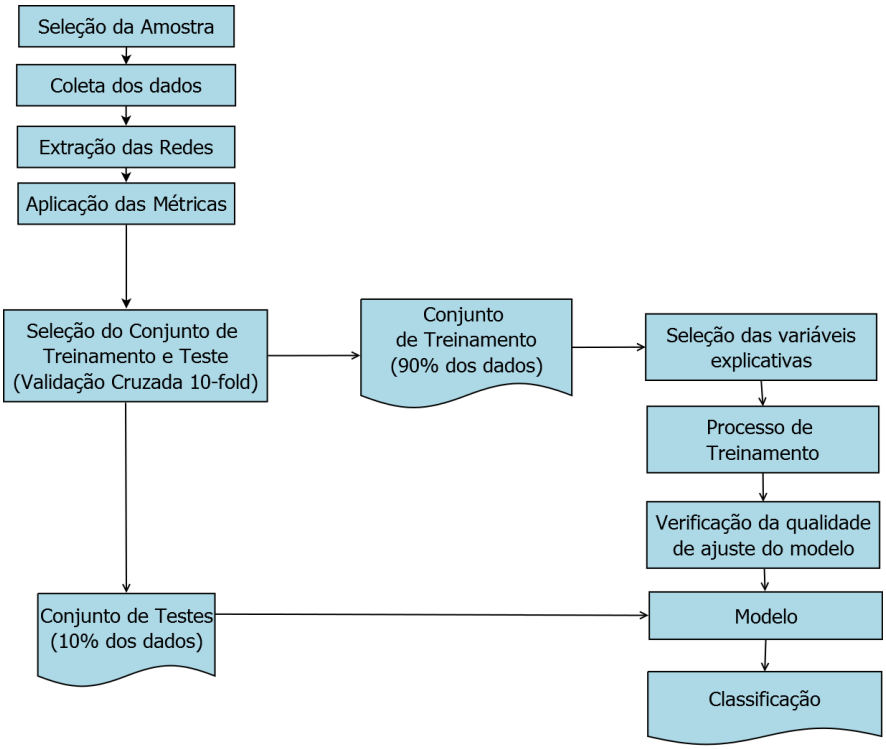


Figura 4.1: Fluxograma da metodologia aplicada

4.1 Seleção da Amostra e Coleta dos Dados

A amostra utilizada na realização desta pesquisa é formada pelos docentes pertencentes ao quadro permanente dos Programas de Pós-Graduação (PPGs) na área de Ciência da Computação reconhecidos pela CAPES e relacionados no site [1] da mesma.

Cabe ressaltar que foram considerados todos os Programas de Pós-Graduação pertencentes à área de Ciência da Computação e não apenas aqueles com nomenclatura de Ciência da Computação. Essa estratégia foi adotada com a finalidade de tornar a base de dados mais robusta para construção do modelo de regressão.

A lista dos docentes foi obtida a partir da página *web* dos respectivos programas. Esta coleta foi feita de forma manual pois não havia padrão nas estruturas das páginas, impossibilitando a criação de coletores automáticos por expressão regular. A coleta dos 73 programas resultou numa lista com 1592 docentes, sendo 534 bolsistas de produtividade e 1058 não bolsistas.

Essa lista foi utilizada para a obtenção das redes de colaboração utilizadas nesse trabalho. Para tanto, utilizou-se os *IDs* dos respectivos currículos dos docentes como entrada para geração automática das redes mediante as informações contidas na plataforma Lattes com uso de um programa denominado *Lattescrawler* [2] , desenvolvido no Laboratório de *Social Networks Analysis* (LabSNA) da Universidade Federal da Paraíba.

É importante destacar que a abordagem de ARS adotada nesta pesquisa é egocêntrica. Dessa forma, foi preciso extrair uma rede de colaboração para cada um dos pesquisadores estudados, sendo necessária, portanto, a automatização do processo a partir de um script desenvolvido na linguagem PowerShell. Tais redes são representadas a partir de um arquivo no formato gml (*Graph Modeling Language*).

Logo, cada uma das redes foi gerada a partir das relações de colaboração expressas pelo docente usado como semente em seu currículo Lattes, onde considerou-se as publicações em coautoria, as participações em projetos de pesquisas e as orientações de Mestrado e Doutorado. Sendo assim, cada rede foi formada pelo docente pesquisado e seus colaboradores

[1] Disponível em: http://conteudoweb.capes.gov.br/conteudoweb/ProjetoRelacaoCursosServlet?acao=pesquisarIescodigoArea=10300007descricaoArea=descricaoAreaConhecimento=CI%CANCIA+DA+COMPUTA%C7%C3OdescricaoAreaAvaliacao=CI%CANCIA+DA+COMPUTA%C7%C3O

[2] Disponível em: https://github.com/marcilioLemos/LABSNA/tree/master/LattesCrawler/src/br/ufpb/ci/labsna/lattescrawler

diretos.

Concluída a etapa anterior, utilizou-se a ferramenta Gephi (BASTIAN; HEYMANN; JACOMY, 2009) para geração dos grafos e aplicação das métricas de ARS. Essa ferramenta permite exportar os dados gerados em um arquivo no formato *csv* (*comma-separated values*). Esse processo também foi realizado de forma automática a partir de um script desenvolvido em Java, que aplicou as métricas de ARS e posteriormente salvou os dados em um arquivo no formato *csv*, sendo gerados, portanto, 1592 arquivos que foram agrupados em um único banco de dados, de onde foram extraídos apenas os dados dos pesquisadores que serviram de semente para geração das redes, ou seja, os docentes permanentes dos programas apresentados anteriormente, formando assim a base de dados utilizada para a construção do modelo.

Para um melhor entendimento do processo descrito acima, a Figura 4.2 mostra o exemplo de uma rede gerada na ferramenta Gephi de um pesquisador de nível 1A com as respectivas métricas de ARS aplicadas. Já na Figura 4.3 observa-se a rede de um pesquisador que não possui bolsa também com suas respectivas métricas.

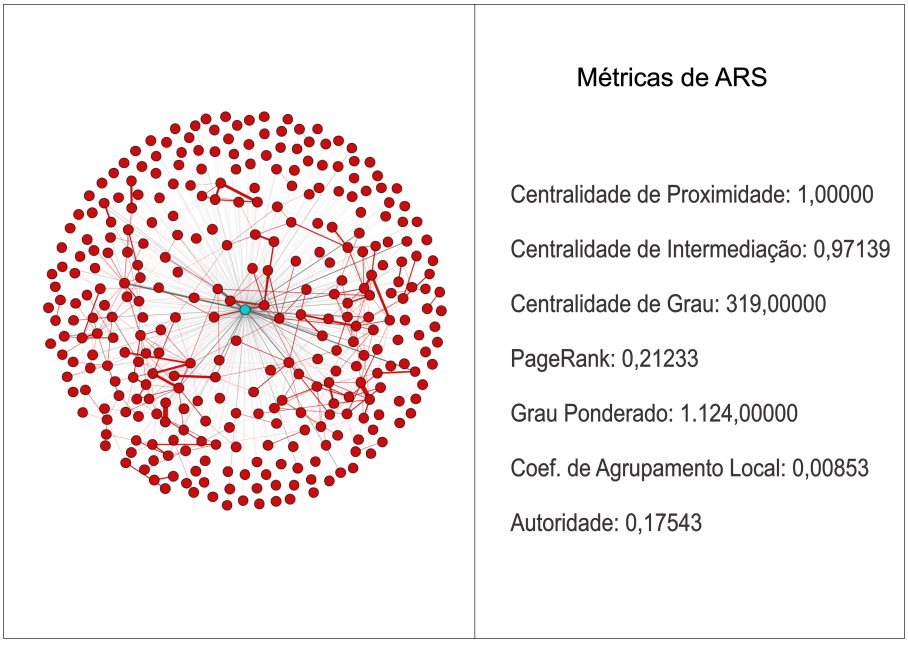


Figura 4.2: Rede de um pesquisador PQ nível 1A com suas métricas

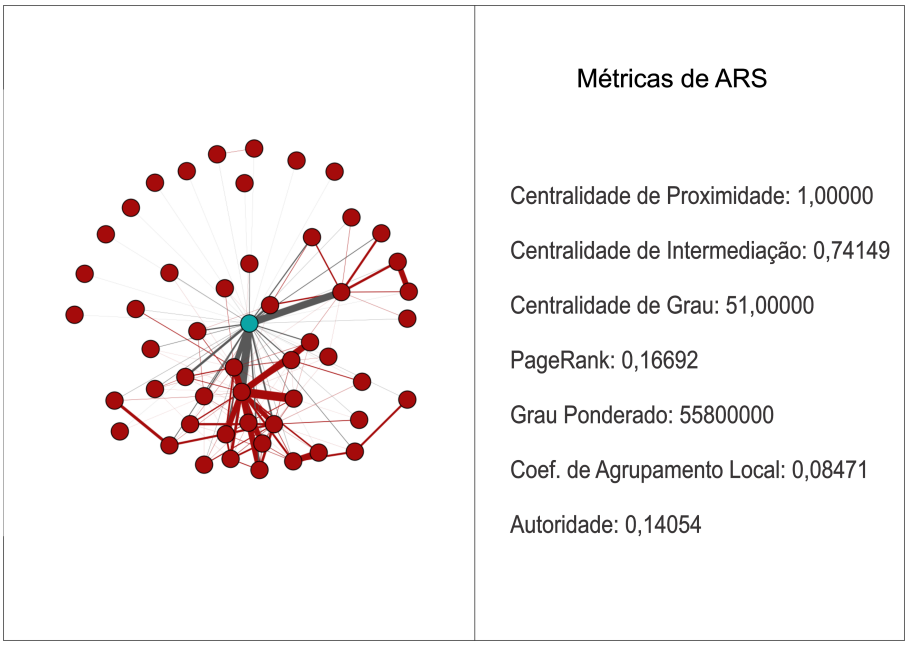


Figura 4.3: Rede de um pesquisador não bolsista com suas métricas

4.2 Divisão da Base de Dados

Antes de iniciar a construção do modelo preditivo é necessário dividir a base de dados em dois conjuntos distintos: o conjunto de treinamento, que será usado para a construção e ajuste do modelo e o conjunto de teste que será usado para validar o modelo construído. Em bases de dados de tamanho reduzido, como é o caso desta pesquisa, é comum utilizar técnicas de amostragem de dados, sendo um dos métodos amplamente utilizado a validação cruzada (*cross validation*), pois permite obter melhores estimativas de acurácia (WITTEN; FRANK, 2005).

Dessa forma, esta pesquisa utilizou a técnica de validação cruzada 10-*fold* estratificada (*stratified 10-fold cross validation*), que divide a base de dados em 10 subconjuntos de tamanhos equivalentes sem sobreposição e mantém a proporção das classes em cada subconjunto (COVÕES, 2010).

Sendo assim, a construção do modelo é realizada utilizando 9 dos 10 subconjuntos e validada no subconjunto restante, sendo esse processo repetido 10 vezes, onde em cada etapa um subconjunto diferente é deixado de fora para validação e os demais são utilizados para a construção do modelo, conforme é ilustrado na Figura 4.4. Nesse caso, a acurácia final do modelo é obtida pela média das acurácia de cada um dos modelos parciais.

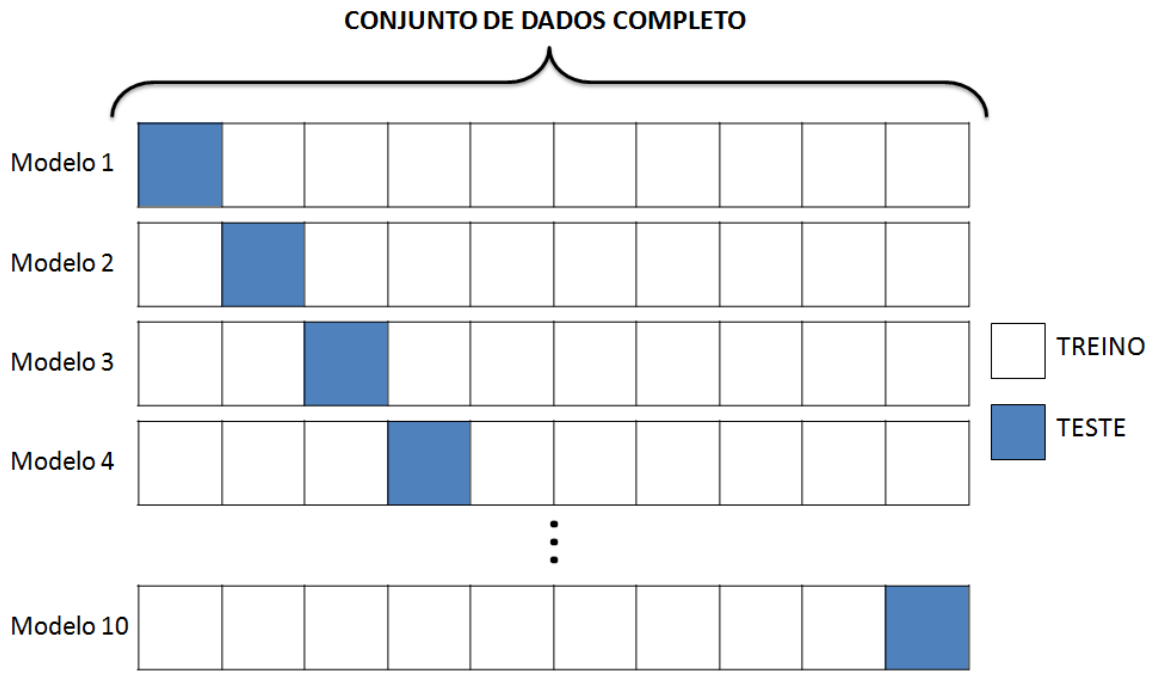


Figura 4.4: Processo de validação cruzada 10-*fold* estratificada

No caso do trabalho em questão, a base de dados original, formada por 1592 observações, foi dividida em 10 subconjuntos, sendo oito com 159 observações e dois com 160 observações e em cada um desses subconjuntos foram mantidas as proporções das classes da base de dados originais, ou seja, 33,54% de pesquisadores bolsistas e 66,46% de pesquisadores não bolsistas.

4.3 Construção do Modelo

Dado que a finalidade desta pesquisa é a construção de um modelo matemático capaz de classificar a relevância de um pesquisador a partir das métricas de ARS aplicadas em suas redes de colaboração, bem como analisar os aspectos colaborativos que mais influenciam nessa relevância, optou-se por utilizar a modelagem de Regressão Logística, com o intuito de estabelecer uma relação entre a relevância científica de um pesquisador (variável resposta) e as métricas de ARS (variáveis explicativas). Sendo assim, as subseções a seguir demonstram o processo de construção desse modelo. Todo o processo de análise estatística e tratamento dos dados foi feito utilizando o software Action (Equipe Estatcamp, 2014) e o software

estatístico R (R Core Team, 2013), versão 3.0.2.

4.3.1 Variável Resposta

Para a construção do modelo, definiu-se como variável resposta capaz de expressar a relevância científica de um pesquisador o fato do mesmo possuir ou não bolsa de produtividade do CNPq, considerando que os seus detentores são considerados a elite da ciência brasileira. Tal variável é denominada nesta pesquisa de **STATUS_PESQ**, e a seguinte codificação é utilizada para torná-la dicotômica:

- 0 para representar o pesquisador que não possui bolsa de produtividade;
- 1 para representar o pesquisador que possui bolsa de produtividade.

Como foi visto anteriormente, a distribuição dessas duas classes na base de dados se dá da seguinte forma: 1058 (66,46%) representantes da classe 0 (Não Bolsista) e 534 (33,54%) representantes da classe 1 (Bolsista de Produtividade). Apesar dessa distribuição de classes não ser totalmente igual, ou seja, não possuir 50% representantes de cada classe, ainda é válida para a construção do modelo logístico, pois o tamanho das desigualdades das amostras é reduzido em termos percentuais, conforme sugere (MANDALA, 2003).

4.3.2 Variáveis Explicativas

Visando explicar a probabilidade de um pesquisador ser classificado como bolsista de produtividade a partir de suas interações sociais, foram selecionadas sete métricas de ARS utilizadas na abordagem de análise egocêntrica: Centralidade de Grau, Centralidade de Intermediação, Centralidade de Proximidade, Grau Ponderado, Autoridade, Coeficiente de Agrupamento Local e *PageRank*. Todas as métricas foram calculadas de maneira individual para cada pesquisador da amostra estudada. A Tabela 4.1 apresenta a codificação dessas variáveis para construção do modelo, bem como sua descrição e natureza estatística. Através da descrição verifica-se o que as variáveis representam em termos práticos, bem como que todas elas possuem natureza contínua.

Para uma melhor compreensão de como se dá a distribuição dessas variáveis nas duas classes representadas na variável resposta, é apresentada na Tabela 4.2 uma análise descritiva

Tabela 4.1: Descrição das Variáveis Explicativas

Variável	Cód	Descrição	Natureza
Centralidade de Grau	CG	Reflete a quantidade de colaboradores de um pesquisador	Contínua
Centralidade de Intermediação	CI	Reflete o potencial de comunicação de um pesquisador dentro da rede	Contínua
Centralidade de Proximidade	CP	Reflete a capacidade de alcance de um pesquisador dentro da rede	Contínua
Coeficiente de Agrupamento Local	CAL	Reflete o quanto o pesquisador colabora com grupos coesos	Contínua
Grau Ponderado	GP	Reflete a força das relações estabelecidas pelo pesquisador	Contínua
PageRank	PR	Reflete qual o prestígio de um pesquisador, a partir da qualidade de suas conexões	Contínua
Autoridade	AUT	Reflete a tendência de um pesquisador colaborar com *hubs*	Contínua

Tabela 4.2: Estatística descritiva das variáveis explicativas por classes

	STATUS_PESQ	Média	Desvio Padrão	Erro padrão da Média
CP	Bolsista	1,00	0,00	0,00
	Não Bolsista	0,98	0,126	0,004
CI	Bolsista	0,82	0,13	0,01
	Não Bolsista	0,70	0,32	0,01
GP	Bolsista	327,57	276,29	11,96
	Não Bolsista	109,68	140,65	4,32
AUT	Bolsista	0,20	0,06	0,00
	Não Bolsista	0,27	0,10	0,00
PR	Bolsista	0,24	0,07	0,00
	Não Bolsista	0,33	0,13	0,00
CAL	Bolsista	0,09	0,08	0,00
	Não Bolsista	0,18	0,24	0,01
CG	Bolsista	41,52	33,96	1,47
	Não Bolsista	15,99	19,03	0,59

das variáveis independentes em relação aos dois grupos representados na variável resposta.

Considerando que o objetivo da pesquisa é identificar entre as variáveis estudadas aquelas que mais contribuem para classificação de um pesquisador como bolsista ou não bolsista de produtividade, é interessante primeiro investigar aquelas variáveis que apresentam grande diferença entre as duas classes, pois elas podem ser boas candidatas a preditoras. Nesse sentido, a Figura 4.5 apresenta gráficos que ilustram a relação das variáveis independentes com o status do pesquisador, onde é possível notar claramente que existem muitas variáveis a serem consideradas no modelo.

Através da figura, percebe-se que os bolsistas de produtividade tendem a ter as métricas de grau ponderado, centralidade de intermediação e centralidade de grau, superiores às dos pesquisadores não bolsistas. Já em relação às métricas de PageRank, Autoridade e Coeficiente de Agrupamento Local, a situação se inverte, pois os bolsistas, em geral, possuem índices menores do que os não bolsistas.

4.3.3 Processo de seleção de variáveis

Antes da construção de um modelo logístico é necessário garantir que todas as variáveis explicativas contribuem de forma significativa para classificar a variável resposta. Para garantir isso, é preciso eliminar as variáveis que são fortemente correlacionadas, permanecendo

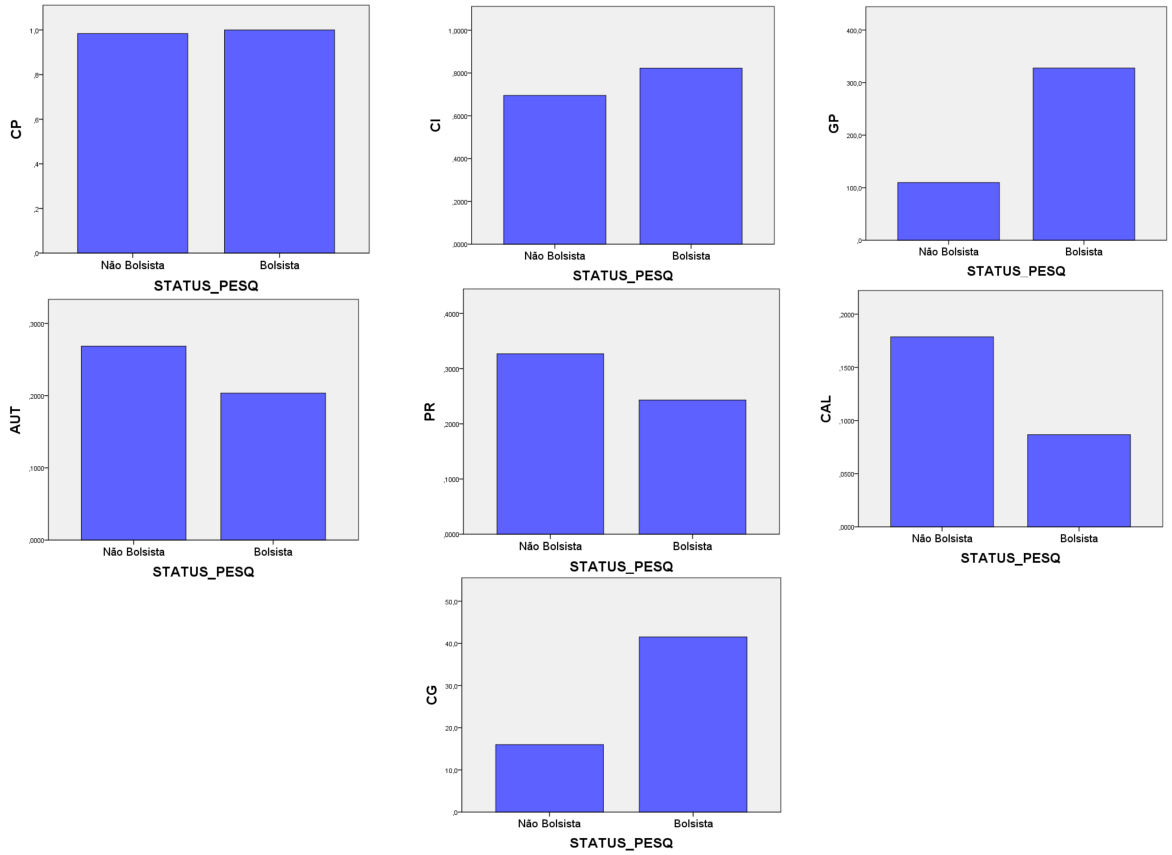


Figura 4.5: Gráficos das relações entre as variáveis explicativas e a variável resposta

com apenas uma delas no modelo. Tendo em vista isso a Tabela 4.3 apresenta a matriz de correlação de Spearman das variáveis estudadas.

Tabela 4.3: Matriz de Correlação de Spearman

	STATUS	CP	CI	GP	AUT	PR	CAL	CG
STATUS	1							
CP	0,07380	1						
CI	0,10077	0,16701	1					
GP	0,52726	0,17802	0,11428	1				
AUT	-0,37783	0,17808	0,35653	-0,70054	1			
PR	-0,36174	-0,17803	0,40682	-0,69294	0,92235	1		
CAL	-0,11442	0,15827	-0,70887	-0,06535	-0,30432	-0,47453	1	
CG	0,52210	0,178089	0,290551	0,903071	-0,65323	-0,61917	-0,22209	1

A partir da apresentação da matriz de correlação verifica-se que as variáveis PR e AUT são fortemente correlacionadas entre si, pois apresentam um coeficiente de correlação de 0,92235 a um nível de significância de 5%, o que também acontece com as variáveis CG e GP que apresentam coeficiente de correlação de 0,903071. Isso demonstra que nem todas

as variáveis devem ser consideradas no modelo, sendo necessário deixar de fora as variáveis que se correlacionam entre si.

Visando garantir um modelo parcimonioso com o menor número de variáveis possível, tornou-se necessário utilizar um processo automático de seleção de variáveis, a fim de garantir que somente aquelas que possuem contribuição significativa para a classificação do pesquisador irão permanecer no modelo.

Dessa forma, optou-se pelo método *stepwise*, utilizando a abordagem *forward stepwise*, no qual o modelo inicial não contém nenhuma variável explicativa, e estas são adicionadas etapa a etapa, tendo como critério de parada o momento em que nenhuma nova variável acrescenta informação ao modelo e todas as contidas nele são significantes (ASSUNÇÃO, 2012). Os níveis de significância definidos neste trabalho foram de 0,05 para a entrada e 0,1 para a saída.

Sendo assim, a Tabela 4.4 apresenta as variáveis selecionadas na etapa final desse processo para os 10 modelos criados, seguidas dos respectivos índices "-2 Log Likelihood" do modelo inicial (sem as variáveis) e do modelo final escolhido. A apresentação das variáveis de cada modelo se dá pela ordem de entrada da mesma em cada etapa do processo.

Tabela 4.4: Variáveis selecionadas para entrar em cada modelo

Modelo	Variáveis que entraram no modelo	-2LL (inicial)	-2LL
1	CI + GP + AUT + CG	1485,122	1375,054
2	CI + GP + PR + CAL	1481,162	1361,708
3	CI + GP + PR + CAL	1496,968	1358,93
4	CI + GP + PR + CAL	1487,76	1354,228
5	CI + GP + PR + CAL	1466,852	1336,462
6	CI + GP + PR + CAL	1494,407	1356,804
7	CI + GP + PR + CAL + CG	1477,593	1340,039
8	CI + GP + PR + CAL	1477,772	1346,583
9	CI + GP + PR + CAL + CG	1477,546	1347,548
10	CI + GP + PR + CAL	1488,464	1358,775

Como pode ser observado através da Tabela, as variáveis escolhidas pelo método *forward stepwise* para compor o modelo colaboram para mudanças significativas no -2 log-verossimilhança. Diante disso, concluiu-se que as mesmas desempenham um papel importante na classificação da relevância de um pesquisador.

4.3.4 Estimação dos coeficientes

A partir da definição das variáveis que entrariam na construção dos modelos procedeu-se a etapa de verificação de multicolinearidade entre as variáveis preditivas. A multicolinearidade se caracteriza quando duas ou mais variáveis independentes estão fortemente correlacionadas, fazendo com que os estimadores dos coeficientes da regressão logística apresentem considerável incerteza (WALTER et al., 2010).

Dessa forma, em cada modelo foi aplicado o Fator de Inflação da Variância (VIF) que realiza um diagnóstico de colinearidade nos dados. Nesse caso, as variáveis com VIF superior a cinco são consideradas fortemente correlacionadas, sendo necessário excluir uma delas do modelo.

Terminado o processo anterior, prosseguiu-se com a estimação e inferência dos coeficientes, que foi feita a partir do método de máxima verossimilhança e do teste de *Wald*. Sendo assim, a aplicação do Fator de Inflação da Variância e os coeficientes estimados de cada modelo são apresentados a seguir:

1. Modelo 1

   Aplicando o teste de colinearidade ao modelo 1 verificou-se que não havia multicolinearidade entre as variáveis independentes, conforme pode ser visto na Tabela 4.5

Tabela 4.5: Diagnóstico de colinearidade para o modelo 1

	CI	GP	AUT	CG
VIF	2.85	2.59	3.31	3.06

   Uma vez que os VIFs de todas as variáveis são inferiores a 5, considera-se todas elas na construção do modelo. Dessa forma, a Tabela 4.6 apresenta os coeficientes estimados do modelo 1, com seus respectivos índices de significância e razões de chances.

Tabela 4.6: Coeficientes estimados do Modelo 1

	Estimativa	Desvio Padrão	LR	P-Valor(LR)	W	P-Valor(W)	OR
(Intercepto)	-1.7019	0.3515	453.78	> 0.001	-4.84	> 0.0001	0.182
CI	4.9880	0.6799	80.40	> 0.001	7.34	> 0.0001	146.642
GP	0.0033	0.0006	35.68	> 0.001	5.54	> 0.0001	1.003
AUT	-14.4638	1.8562	76.11	> 0.001	-7.79	> 0.0001	> 0.001
CG	-0.0095	0.0046	3.88	0.049	-2.05	0.0401	0.991

Infere-se pelos dados apresentados que a Centralidade de Intermediação e o Grau Ponderado influenciam positivamente na relevância de um pesquisador, pois apresentam coeficientes positivos. Enquanto isso, a Autoridade e a Centralidade de Grau tem impacto negativo, já que seus coeficientes são negativos.

Já em relação aos testes de inferência, todos os coeficientes obtiveram um índice de significância inferior a 5%, que foi o nível de confiança estabelecido nesta pesquisa, tanto no teste da razão de máxima verossimilhança (LR) quanto no teste de *Wald*.

2. Modelo 2

O diagnóstico de colinearidade para o modelo 2 é apresentado na Tabela 4.7, onde mais uma vez verificou-se que os VIFs de todas as variáveis são inferiores à cinco, portanto todas elas permanecem no modelo.

Tabela 4.7: Diagnóstico de colinearidade para o modelo 2

	CI	GP	PR	CAL
VIF	4.69	1.80	2.36	4.26

Sendo assim, os coeficientes estimados desse modelo são apresentados na Tabela 4.8 seguidos dos testes de inferência estatística e das razões de chance.

Tabela 4.8: Coeficientes estimados do Modelo 2

	Estimativa	Desvio Padrão	LR	P-Valor(LR)	W	P-Valor(W)	OR
(Intercepto)	0.7671	0.8973	467.14	> 0.001	0.85	0.3926	2.153
CI	1.9584	0.9606	6.70	0.010	2.04	0.0415	7.086
GP	0.0022	0.0005	21.96	> 0.001	4.35	> 0.0001	1.002
PR	-10.4900	1.2641	87.37	> 0.001	-8.30	> 0.0001	> 0.001
CAL	-4.2454	1.4312	7.44	0.006	-2.97	0.0030	0.014

Observando os dados apresentados, nota-se que a Centralidade de Intermediação e o Grau Ponderado mais uma vez têm influencia positiva na relevância de um pesquisador, já o *PageRank* e o Coeficiente de Agrupamento Local têm impacto negativo na variável resposta.

De acordo com os testes de razão de máxima verossimilhança (LR) e de *Wald* (W), constata-se que todos os coeficientes são significativos a um nível de 5%. Nota-se que, no teste de *Wald*, o intercepto possui um p-valor superior a 0,05 apenas, o que pode ser ignorado dado que em amostras pequenas quando ocorre divergência sugere-se confiar

no teste da razão de máxima verossimilhança, já que o teste de *Wald* muitas vezes apresenta comportamento estranho.

3. Modelo 3

   De maneira similar, aplicou-se o teste de colinearidade às variáveis selecionadas no modelo 3, onde os resultados encontrados são demonstrados na Tabela 4.9.

Tabela 4.9: Diagnóstico de colinearidade para o modelo 3

	CI	GP	PR	CAL
VIF	3.84	1.83	2.33	3.81

Nota-se que todas as variáveis selecionadas para o modelo não possuem colinearidade forte, visto que seus VIFs são inferiores a cinco. Dessa forma, as quatro variáveis permaneceram no modelo e a estimação dos seus coeficientes, seguido dos testes de inferência, bem como das razões de chances são apresentados na Tabela 4.10.

Tabela 4.10: Coeficientes estimados do Modelo 3

	Estimativa	Desvio Padrão	LR	P-Valor(LR)	W	P-Valor(W)	OR
(Intercepto)	1.5388	0.8581	469.91	> 0.000	1.79	0.0729	4.658
CI	1.6194	0.8659	5.25	0.022	1.87	0.0615	5.049
GP	0.0017	0.0005	12.94	> 0.001	3.40	0.0007	1.002
PR	-11.4135	1.2434	103.60	> 0.001	-9.18	> 0.0001	> 0.001
CAL	-5.5092	1.4093	12.49	> 0.001	-3.91	0.0001	0.004

A partir da análise dos coeficientes do modelo 3 notou-se que tanto a Centralidade de Intermediação quanto o Grau Ponderado apresentam um impacto positivo sobre o status do pesquisador, enquanto que o PageRank e o Coeficiente de Agrupamento Local apresentam um impacto negativo.

Em relação aos testes de inferência verifica-se que o teste de *Wald* aplicado ao intercepto e ao coeficiente da variável CI apresenta significância superior a 0,05, porém o teste de máxima verossimilhança garante que todos os coeficientes são significativos a um nível de confiança de 5%, logo, foi considerado o resultado obtido pelo teste de máxima verossimilhança.

4. Modelo 4

   A Tabela 4.11 apresenta o teste de multicolinearidade aplicado ao modelo 4, onde verifica-se os valores dos VIFs de cada uma das variáveis selecionadas.

Tabela 4.11: Diagnóstico de colinearidade para o modelo 4

	CI	GP	PR	CAL
1	4.45	1.80	2.32	4.15

A partir da observação dos dados apresentados na tabela, conclui-se que todas as variáveis devem ser consideradas no modelo, tendo em vista que os seus VIFs são todos inferiores a cinco, denotando que não existe colinearidade entre as mesmas. Nesse caso, os valores dos coeficientes juntamente com os seus testes de significância e razão de chances podem ser vistos na Tabela 4.12.

Tabela 4.12: Coeficientes estimados do Modelo 4

	Estimativa	Desvio Padrão	LR	P-Valor(LR)	W	P-Valor(W)	OR
(Intercepto)	1.4013	0.8880	474.62	> 0.000	1.58	0.1146	4.058
CI	1.6611	0.9355	4.70	0.030	1.78	0.0758	5.266
GP	0.0018	0.0005	15.53	> 0.001	3.73	0.0002	1.002
PR	-11.4820	1.2715	104.64	> 0.001	-9.03	0.> 0001	> 0.001
CAL	-4.9014	1.4270	9.75	0.002	-3.43	0.0006	0.007

De modo similar aos outros modelos apresentados verifica-se uma influência positiva da Centralidade de Intermediação e do Grau Ponderado no status do pesquisador, enquanto que o *PageRank* e o Coeficiente de Agrupamento Local apresentam um impacto negativo sobre o mesmo.

Além disso, o testes da razão de máxima verossimilhança indica que todos os coeficientes são significativos a um nível de 5%. Sendo assim, embora o teste de *Wald* tenha rejeitado a hipótese nula para o intercepto e para o coeficiente da variável CI, esses resultados não foram considerados, já que o teste de máxima verossimilhança é mais confiável.

5. Modelo 5

Prosseguindo a análise aplicou-se o teste de colinearidade ao modelo 5, cujos resultados estão expressos na Tabela 4.13.

Tabela 4.13: Diagnóstico de colinearidade para o modelo 5

	CI	GP	PR	CAL
VIF	3.26	1.83	2.31	3.44

Os resultados apresentados na tabela demonstram que as variáveis não estão correlacionadas, pois apresentam VIF inferior a 5. Diante disso, todas elas foram utilizadas para construção do modelo, sendo apresentados na Tabela 4.14 os coeficientes estimados das mesmas, seguidos dos testes de significância e das razões de chances.

Tabela 4.14: Coeficientes estimados do Modelo 5

	Estimativa	Desvio Padrão	LR	P-Valor(LR)	W	P-Valor(W)	OR
(Intercepto)	1.6955	0.8322	492.39	> 0.001	2.04	0.0416	5.447
CI	1.2877	0.8086	3.49	0.062	1.59	0.1113	3.624
GP	0.0020	0.0005	17.56	> 0.001	3.94	0.0001	1.002
PR	-11.0790	1.2383	97.44	> 0.001	-8.95	> 0.0001	0.000
CAL	-6.1016	1.3997	15.66	> 0.001	-4.36	> 0.0001	0.002

A partir da análise dos coeficientes estimados no modelo 5 verifica-se um impacto positivo das variáveis Centralidade de Grau e Grau Ponderado na explicação da relevância de um pesquisador, em contrapartida, as variáveis PageRank e Coeficiente de Agrupamento Local influenciam de maneira negativa a variável resposta.

Já os testes de inferência indicam que o coeficiente da variável Centralidade de Intermediação não é significativo a um nível de 5%. Como esse resultado foi obtido tanto pelo teste de *Wald*, quanto pelo teste de máxima verossimilhança, considera-se que a retirada da variável não traz prejuízos ao modelo.

6. Modelo 6

A aplicação do teste de colinearidade no modelo 6 pode ser observada na Tabela 4.15.

Tabela 4.15: Diagnóstico de colinearidade para o modelo 6

	CI	GP	PR	CAL
VIF	4.97	1.78	2.37	4.51

Diante dos resultados apresentados na Tabela 4.15, verificou-se que todas as variáveis possuem VIFs inferiores a cinco, o que as credenciam a entrarem no modelo. Nesse caso, a Tabela 4.16 apresenta os coeficientes estimados de cada uma das variáveis, bem como os testes de significância e a razão de chances.

A análise dos coeficientes do modelo 6 demonstram mais uma vez o impacto positivo das métricas Centralidade de Intermediação e Grau Ponderado no status do pesqui-

Tabela 4.16: Coeficientes estimados do Modelo 6

	Estimativa	Desvio Padrão	LR	P-Valor(LR)	W	P-Valor(W)	OR
(Intercepto)	1.3600	0.9402	472.05	> 0.001	1.45	0.1480	3.895
CI	1.8466	1.0086	5.17	0.023	1.83	0.0671	6.338
GP	0.0017	0.0005	14.21	> 0.001	3.55	0.0004	1.002
PR	-11.8016	1.2883	109.41	> 0.001	-9.16	> 0.0001	> 0.001
CAL	-4.9428	1.5118	8.94	0.003	-3.27	0.0011	0.007

sador. Já as variáveis *PageRank* e Coeficiente de Agrupamento Local exercem uma influência negativa.

Observa-se ainda que todos os coeficientes são significativos a um nível de 5% no teste de máxima verossimilhança, logo desconsidera-se os testes de *Wald* que, por ventura, não atingiram o nível de significância estabelecido.

7. Modelo 7

O diagnóstico de colinearidade para o modelo 7 é apresentado na Tabela 4.17, onde observa-se o VIF de cada uma das variáveis selecionadas para o modelo.

Tabela 4.17: Diagnóstico de colinearidade para o modelo 7

	CI	GP	PR	CAL	CG
VIF	4.84	2.87	2.75	4.32	3.02

Os resultados do teste de colinearidade demonstram que todas as variáveis apresentaram VIF inferior a cinco, e portanto, não existe correlação forte entre as mesmas, sendo possível utilizá-las para a construção do 4.18 os coeficientes do modelo, com seus respectivos testes de significância e razão de chances.

Tabela 4.18: Coeficientes estimados do Modelo 7

	Estimativa	Desvio Padrão	LR	P-Valor(LR)	W	P-Valor(W)	OR
(Intercepto)	1.3269	0.9492	487.45	> 0.001	1.40	0.1621	3.768
CI	2.1637	1.0280	7.65	0.006	2.10	0.0353	8.702
GP	0.0026	0.0006	20.37	> 0.001	4.23	> 0.0001	1.003
PR	-11.9713	1.3637	97.88	> 0.001	-8.78	> 0.0001	> 0.001
CAL	-5.7108	1.5764	11.07	0.001	-3.62	0.0003	0.003
CG	-0.0089	0.0045	3.74	0.053	-2.00	0.0450	0.991

Analisando os coeficientes do modelo 7, verificou-se que, assim como aconteceu nos outros modelos apresentados, as variáveis que apresentam impacto positivo sobre a relevância de um pesquisador são a Centralidade de Intermediação e o Grau Ponderado,

em contrapartida, as variáveis que apresentam impacto negativo são o *PageRank*, o Coeficiente de Agrupamento Local e a Centralidade de Grau.

Já os resultados obtidos pelo teste de *Wald* e pelo teste da razão de máxima verossimilhança indicam que o coeficiente da variável Centralidade de Grau não é significativo a um nível de 5%, sendo assim a variável independente CG não produz efeito significativo sobre a variável resposta.

8. Modelo 8

Os resultados obtidos a partir do teste de colinearidade aplicado ao modelo 8 são apresentados na Tabela 4.19.

Tabela 4.19: Diagnóstico de colinearidade para o modelo 8

	CI	GP	PR	CAL
VIF	4.32	1.84	2.38	4.11

Analisando a tabela verifica-se que todas as variáveis devem ser consideradas no modelo, já que não há correlações fortes entre elas. Na Tabela 4.20, são apresentados, portanto, os coeficientes estimados para cada uma dessas variáveis, bem como os testes de significância estatística e as razões de chance de cada uma delas.

Tabela 4.20: Coeficientes estimados do Modelo 8

	Estimativa	Desvio Padrão	LR	P-Valor(LR)	W	P-Valor(W)	OR
(Intercepto)	1.2607	0.9168	480,91	> 0,001	1.38	0.1691	3,526
CI	1.7929	0.9542	5.50	0.019	1.88	0.0602	6.007
GP	0.0019	0.0005	16.21	> 0.001	3.77	0.0002	1.002
PR	-11.2611	1.2741	98.13	> 0.001	-8.84	> 0.0001	>0.001
CAL	-5.1496	1.4848	9.92	> 0.002	-3.47	0.0005	0.006

Nota-se que as variáveis que apresentam impacto positivo no status do pesquisador são a Centralidade de Grau e o Grau Ponderado, seguindo o mesmo padrão dos modelos já apresentados, enquanto o PageRank e o Coeficiente de Agrupamento Local impactam negativamente a explicação da variável resposta.

Além disso, o teste de razão de máxima verossimilhança indica que todos os coeficientes estimados são significativos a um nível de 5%. Apesar do teste de *Wald* divergir quanto a significância do intercepto e do coeficiente da variável CI, esse resultado será ignorado, uma vez que o teste de razão de máxima verossimilhança é mais confiável.

9. Modelo 9

   A aplicação do diagnóstico de colinearidade ao modelo 9 gerou os resultados apresentados na Tabela 4.21.

Tabela 4.21: Diagnóstico de colinearidade para o modelo 9

	CI	GP	PR	CAL	CG
VIF	5.98	2.85	2.98	5.05	3.28

   Nota-se a partir da análise da tabela, que o modelo 9 apresenta duas variáveis fortemente correlacionadas, portanto, foi necessária a exclusão de uma delas para realizar um novo ajuste no modelo. Dessa forma, procedeu-se com a exclusão da variável Coeficiente de Agrupamento Local, já que a Centralidade de Intermediação teve impacto mais significativos nos outros modelos apresentados. Sendo assim, a Tabela 4.22 apresenta os coeficientes estimados das variáveis que permaneceram no modelo, com seus respectivos testes de significância e razão de chances.

Tabela 4.22: Coeficientes estimados do Modelo 9

	Estimativa	Desvio Padrão	LR	P-Valor(LR)	W	P-Valor(W)	OR
(Intercepto)	-1.9112	0.3880	470.76	0.000	-4.93	0.0000	0.148
CI	5.6586	0.7683	86.64	> 0.000	7.37	> 0.0001	286.621
GP	0.0032	0.0006	29.46	> 0.000	5.07	> 0.0001	1.003
PR	-13.1310	1.5636	89.48	> 0.000	-8.40	> 0.0001	> 0.001
CG	-0.0103	0.0052	3.77	0.052	-1.98	0.0482	0.990

   Observa-se claramente a influência positiva das variáveis Centralidade de Intermediação e Grau Ponderado na explicação da variável resposta. Já as variáveis PageRank e Centralidade de Grau tem impacto negativo sobre a mesma.

   Quanto à inferência estatística dos coeficientes, tanto o teste da razão de verossimilhança quanto o teste de *Wald* garantiram que todos os coeficientes são significativos a um nível de 5%.

10. Modelo 10

    Os resultados do teste de colinearidade das variáveis selecionadas no modelo 10 estão expostos na Tabela 4.23.

    A partir da tabela apresentada verifica-se que a variável Centralidade de Intermediação possui um VIF superior à 5, entretanto como foi a única variável a apresentar

Tabela 4.23: Diagnóstico de colinearidade para o modelo 10

	CI	GP	PR	CAL
VIF	5.53	1.80	2.50	4.80

valor ligeiramente superior a 5, optou-se por mantê-la no modelo. Diante disso, os coeficiente estimados com seus respectivos teste de significância e razão de chances estão expostos na Tabela 4.24.

Tabela 4.24: Coeficientes estimados do Modelo 10

	Estimativa	Desvio Padrão	LR	P-Valor(LR)	W	P-Valor(W)	OR
(Intercepto)	0.7146	0.9846	467.89	>0.001	0.73	0.4680	2.042
CI	2.2961	1.0806	7.74	0.005	2.12	0.0336	9.938
GP	0.0020	0.0005	17.41	> 0.001	3.92	0.0001	1.002
PR	-11.1409	1.3021	95.63	> 0.001	-8.56	> 0.0001	> 0.001
CAL	-4.2853	1.5574	6.50	0.011	-2.75	0.0059	0.014

Os coeficientes estimados apresentados na tabela demonstram que a Centralidade de Intermediação e do Grau Ponderado afetam positivamente o status do pesquisador e as variáveis PageRank e Coeficiente de Agrupamento Local tem impacto negativo sobre o mesmo.

Em relação a significância estatística dos coeficientes, embora o teste de *Wald* divergir quanto a significância do intercepto, esse resultado será ignorado devido o fato do mesmo ter sido significativo no teste da razão de máxima verossimilhança.

4.3.5 Testes de qualidade de ajuste

Visando verificar se os modelos estimados são realmente eficazes, buscou-se aferir o ajuste das variáveis dentro dos mesmos através de três testes estatísticos: o teste de *Pearson*, o teste *Deviance* e o *Teste Hosmer-Lemeshow*. Os testes de *Pearson* e *Deviance* verificam se o modelo conseguiu ajustar os dados para se alcançar estimativas seguras, sendo assim, quanto maior for o nível de significância $(p - valor)$ melhor é o ajuste dos dados no modelo. Na Tabela 4.25, pode ser verificado os resultados do teste de *Pearson* aplicado aos dez modelos. Já na Tabela 4.26 observam-se os resultados do teste *Deviance*.

A partir da observação das tabelas apresentadas, nota-se que todos os modelos obtiveram valores de p altos para os dois testes o que indica que os modelos estimados conseguiram

Tabela 4.25: Resultados do Teste de *Pearson*

Modelo	Qui-Quadrado	GL	p-valor
1	1354,91	1428	0,916
2	1368,06	1428	0,870
3	1381,5	1428	0,807
4	1358,84	1428	0,904
5	1407,41	1428	0,646
6	1362,29	1428	0,892
7	1351,14	1427	0,924
8	1384,46	1428	0,791
9	1353,8	1427	0,916
10	1377,52	1427	0,822

Tabela 4.26: Resultados do Teste *Deviance*

Modelo	Estatística Deviance	GL	p-valor
1	1375,06	1428	0,839
2	1361,69	1428	0,894
3	1358,92	1428	0,903
4	1354,21	1428	0,918
5	1336,44	1428	0,958
6	1356,79	1428	0,910
7	1340,02	1427	0,951
8	1346,56	1428	0,938
9	1355,88	1427	0,910
10	1358,76	1427	0,900

ajustar bem os dados.

Outro teste importante para verificar o ajuste dos modelos, é o *Hosmer-Lemeshow*. Esse teste compara as frequências observadas com as frequências esperadas a fim de averiguar se os modelos realizam predições com segurança. Os resultados desse teste podem ser vistos na Tabela 4.27.

Assim como os dois testes anteriores, os resultados mostrados na Tabela 4.27 demonstram que todos os modelos apresentaram valores de p superiores ou iguais a $0,05$ que foi o nível de significância estabelecido nesta pesquisa. Dessa forma, os modelos estimados também apresentaram um bom ajuste dos dados segundo o teste *Hosmer-Lemeshow*.

Tabela 4.27: Resultados do teste Hosmer-Lemeshow

Modelo	Qui-Quadrado	GL	p-valor
1	8,620	8	0,375
2	8,027	8	0,431
3	9,705	8	0,286
4	8,204	8	0,413
5	12,176	8	0,143
6	7,089	8	0,527
7	9,451	8	0,305
8	11,754	8	0,162
9	15,73	8	0,053
10	12,725	8	0,121

4.4 Considerações Finais

Nas seções anteriores foram apresentadas as etapas envolvidas na construção e ajuste dos modelos de regressão logística. A partir da utilização do método de validação cruzada 10-*fold* foram obtidos 10 modelos, utilizando-se as bases de treinamento, que apresentaram um ajuste de qualidade satisfatório.

Diante disso, o próximo capítulo será dedicado a verificar o poder de classificação dos modelos parciais obtidos, aplicando-os nas amostras de validação. A partir desse processo será possível avaliar a capacidade de generalização de cada modelo e verificar aquele que foi mais preciso nas classificações.

Capítulo 5

Resultados e Discussões

No capítulo anterior foram apresentados os modelos gerados no processo de validação cruzada. Dessa forma, este capítulo descreve os experimentos realizados para testar a capacidade de discriminação desses modelos a fim de verificar qual deles possui maior capacidade de generalização.

Sendo assim, a Seção 5.1 demonstra a capacidade preditiva dos modelos ajustados, realizando uma comparação dos modelos a partir da análise da área sob a curva ROC (AUC). Além disso, apresenta as matrizes de classificação geradas a partir da aplicação dos modelos tanto na base de construção quanto na base de validação, realizando uma avaliação a partir do cálculo das métricas de desempenho apresentadas na Seção 2.1.2.6. Já a Seção 5.2 discute os resultados encontrados a partir dos experimentos, apresentando uma análise acerca do modelo final obtido através dos testes.

5.1 Avaliação da Capacidade de Discriminação dos Modelos

A avaliação dos modelos gerados foi feita a partir da área sob a curva ROC e da análise da matriz de classificação. A matriz de classificação apresenta uma comparação entre a classificação realizada pelos modelos desenvolvidos e a classificação original das observações da amostra. Para tanto, é necessária a definição de um ponto de corte que seja considerado pelo modelo no momento de classificar o indivíduo em um dos dois grupos representados na

variável resposta.

A escolha de um ponto de corte satisfatório deve ser baseada em uma combinação ótima tanto da sensibilidade quanto da especificidade, supondo que a classificação de um indivíduo como evento dado que ele é não evento (falso positivo) e a classificação de um indivíduo como não evento dado que ele é evento (falso negativo) traz prejuízos equivalentes para o pesquisador (Equipe Estatcamp, 2014).

Para visualizar o ponto de corte que possibilita essa combinação ótima é comum utilizar a curva ROC, que permite estudar a variação da sensibilidade e especificidade. Sendo assim, o ponto de corte definido nesta pesquisa levou em consideração a média dos pontos de corte obtidos nas curvas ROC dos 10 modelos gerados a partir da base de treinamento que mais se aproximavam do canto superior esquerdo do gráfico. Logo, o ponto de corte adotado para validação dos modelos foi de $0,336$.

5.1.1 Avaliação a partir da área sob a curva ROC (AUC)

Além de servir de parâmetro para a escolha do melhor ponto de corte a ser adotado na validação do modelo, a curva ROC também é uma ferramenta amplamente utilizada para avaliar a habilidade do modelo em classificar corretamente os indivíduos em dois subgrupos.

Nesse caso, utiliza-se a área sob a curva ROC para verificar o poder de discriminação do modelo, e sua interpretação deve ser feita a partir dos intervalos da área obtida, conforme foi descrito na Subseção 2.1.2.6.

Levando em consideração isso, foi realizada uma comparação entre as curvas ROC geradas para cada um dos modelos obtidos, utilizando os resultados encontrados na base de validação, de modo a verificar qual deles tem maior capacidade de discriminação. A comparação gráfica entre as curvas dos modelos, bem como a AUC de cada um deles pode ser observada na Figura 5.1.

A análise da figura permite concluir que os modelos têm uma capacidade de discriminação excelente uma vez que nove dos dez modelos apresentados possuem a área sob a curva ROC superior a 0,80, obtendo uma média de área de 0,83. Observa-se ainda, que apenas o modelo 5 apresentou uma AUC inferior a 0,80, porém, ainda assim o seu poder de discriminação é considerado aceitável de acordo com a classificação proposta por hosmer2000applied.

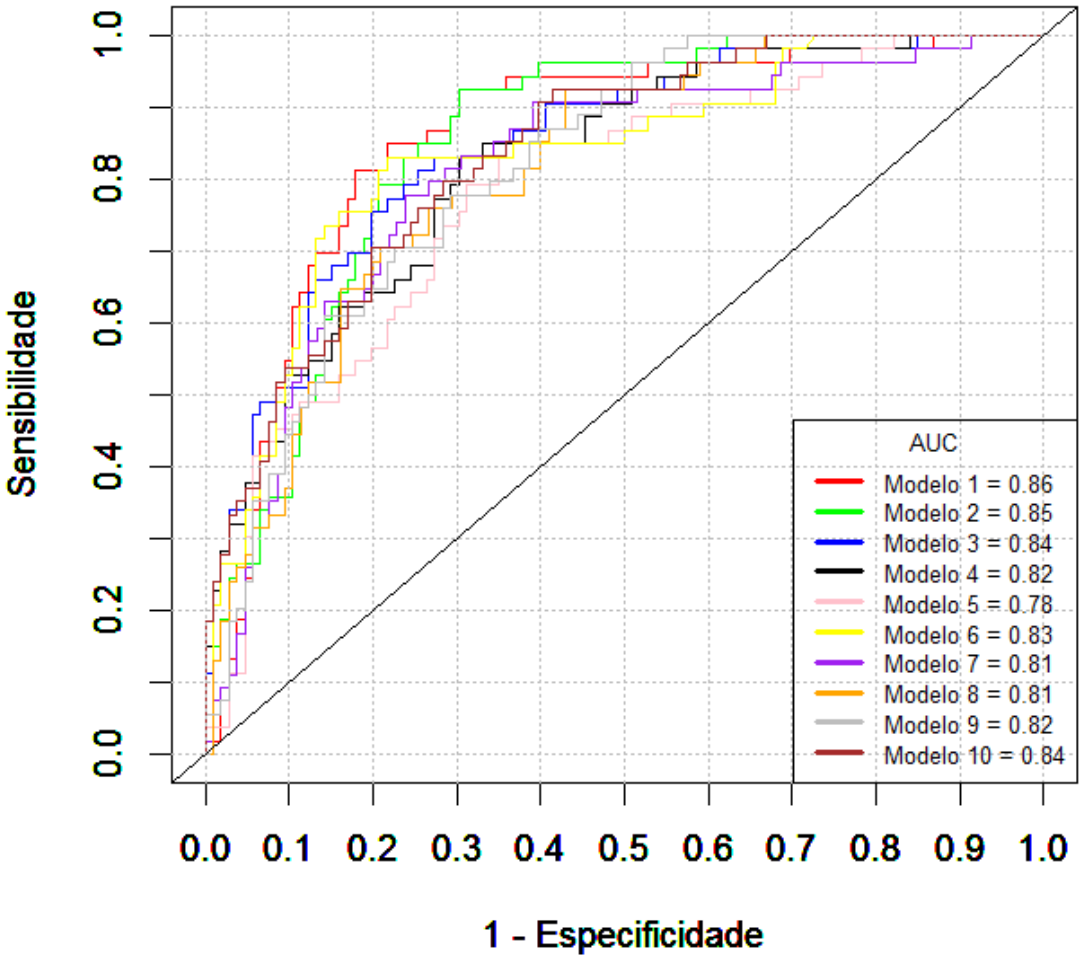


Figura 5.1: Comparação da capacidade de discriminação dos modelos ajustados a partir das curvas ROC

Ainda a partir da figura, podemos inferir que os possíveis candidatos a modelo final são os modelos 1 e 2, já que são os modelos que apresentam áreas sob a curva ROC mais elevadas. Entretanto, para corroborar essa suposição é necessária uma análise das matrizes de classificação tanto da base de construção quanto da base de validação, que permitem avaliar o poder de classificação dos modelos gerados.

5.1.2 Avaliação a partir da matriz de Classificação

A avaliação a partir da matriz de classificação permite identificar a quantidade de instâncias classificadas corretamente pelo modelo, bem como as instâncias que o modelo classificou como sendo de um grupo, mas que na verdade pertenciam ao outro grupo. Sendo assim, as matrizes de classificação para cada um dos modelos gerados são apresentadas a seguir:

1. Matriz de Classificação do Modelo 1

   A comparação entre as classificações feitas pelo modelo 1 e as observações reais nas bases de treinamento e de validação pode ser verificada a partir da matriz de Classificação apresentada na Tabela 5.1. Nela é possível observar, no lado direito, o número de classificações corretas e incorretas realizadas na base de treinamento e no lado esquerdo, o número de classificações corretas e incorretas realizadas na base de validação.

Tabela 5.1: Matriz de Classificação do Modelo 1

Base de Treinamento				Base de Validação			
		Valor Observado				Valor Observado	
Valor Predito		Não Bolsista	Bolsista	Valor Predito		Não Bolsista	Bolsista
	Não Bolsista	689	110		Não Bolsista	80	8
	Bolsista	263	371		Bolsista	26	45

De acordo com os dados de saída apresentados na matriz de Classificação gerada a partir da aplicação do modelo 1 e levando em consideração a classificação dos pesquisadores utilizados na pesquisa, foram obtidos os seguintes resultados:

- **Base de Construção:**
  - Dos 952 pesquisadores não bolsistas o modelo gerado classificou 689 pesquisadores como não bolsistas e 263 como bolsistas;
  - Dos 481 pesquisadores bolsistas o modelo classificou 371 pesquisadores como bolsistas e 110 como não bolsistas.
- **Base de Validação:**
  - Dos 106 pesquisadores não bolsistas o modelo gerado classificou 80 pesquisadores como não bolsistas e 26 como bolsistas;
  - Dos 53 pesquisadores bolsistas o modelo classificou 45 pesquisadores como bolsistas e 8 como não bolsistas.

2. Matriz de Classificação do Modelo 2

   Na Tabela 5.2 são apresentadas as classificação obtidas pelo modelo 2 na base de construção e na base de validação, onde mais uma vez verificamos no lado esquerdo o número de classificação corretas e incorretas na base de construção e no lado direito as classificações corretas e incorretas na base de validação.

Tabela 5.2: Matriz de Classificação do Modelo 2

Base de Treinamento				Base de Validação			
		Valor Observado				Valor Observado	
		Não Bolsista	Bolsista			Não Bolsista	Bolsista
Valor Predito	Não Bolsista	692	108	Valor Predito	Não Bolsista	81	9
	Bolsista	260	373		Bolsista	25	44

   A análise dos dados apresentados na Tabela 5.2 demonstram os resultados obtidos na classificação realizada pelo modelo 2. Sendo assim, levando em consideração o status real do pesquisador e a classificação feita pelo modelo obteve os seguintes resultados:

   - **Base de Construção:**
     - Dos 952 pesquisadores não bolsistas o modelo gerado classificou 692 pesquisadores como não bolsistas e 260 como bolsistas;
     - Dos 481 pesquisadores bolsistas o modelo classificou 373 pesquisadores como bolsistas e 108 como não bolsistas.
   - **Base de Validação:**
     - Dos 106 pesquisadores não bolsistas o modelo classificou 81 pesquisadores como não bolsistas e 25 como bolsistas;
     - Dos 53 pesquisadores bolsistas o modelo classificou 45 pesquisadores como bolsistas e 8 como não bolsistas.

3. Matriz de Classificação do Modelo 3

   Prosseguindo com a validação dos modelos, observa-se na tabela 5.3 as classificações realizadas pelo modelo 3 tanto na base de construção quanto na base de validação.

   De acordo com o que é apresentado na Tabela 5.3 e considerando a classificação real do pesquisador pelo CNPq, o modelo 3 obteve os seguintes resultados:

   - **Base de Construção:**

Tabela 5.3: Matriz de Classificação do Modelo 3

Base de Treinamento				Base de Validação			
		Valor Observado				Valor Observado	
		Não Bolsista	Bolsista			Não Bolsista	Bolsista
Valor Predito	Não Bolsista	692	102	Valor Predito	Não Bolsista	72	9
	Bolsista	260	379		Bolsista	34	44

  - Dos 952 pesquisadores não bolsistas o modelo gerado classificou 692 pesquisadores como não bolsistas e 260 como bolsistas;
  - Dos 481 pesquisadores bolsistas o modelo classificou 379 pesquisadores como bolsistas e 102 como não bolsistas.

- **Base de Validação:**
  - Dos 106 pesquisadores não bolsistas o modelo classificou 72 pesquisadores como não bolsistas e 34 como bolsistas;
  - Dos 53 pesquisadores bolsistas o modelo classificou 44 pesquisadores como bolsistas e 9 como não bolsistas.

4. Matriz de Classificação do Modelo 4

   O experimento de validação do modelo 4 é apresentado na Tabela 5.4. Nela é possível verificar a quantidade de classificações corretas e incorretas tanto da base de construção quanto na base de validação.

Tabela 5.4: Matriz de Classificação do Modelo 4

Base de Treinamento				Base de Validação			
		Valor Observado				Valor Observado	
		Não Bolsista	Bolsista			Não Bolsista	Bolsista
Valor Predito	Não Bolsista	691	101	Valor Predito	Não Bolsista	79	17
	Bolsista	261	380		Bolsista	27	36

   De acordo com os dados de saída apresentados na matriz de Classificação gerada a partir da aplicação do modelo 4 e levando em consideração a classificação dos pesquisadores utilizados na pesquisa, foram obtidos os seguintes resultados:

   - **Base de Construção:**
     - Dos 952 pesquisadores não bolsistas, o modelo gerado classificou 691 pesquisadores como não bolsistas e 261 como bolsistas;
     - Dos 481 pesquisadores bolsistas o modelo classificou 380 pesquisadores como bolsistas e 101 como não bolsistas.

- **Base de Validação:**
  - Dos 106 pesquisadores não bolsistas, o modelo classificou 79 pesquisadores como não bolsistas e 27 como bolsistas;
  - Dos 53 pesquisadores bolsistas o modelo classificou 36 pesquisadores como bolsistas e 17 como não bolsistas.

5. Matriz de Classificação do Modelo 5

Na Tabela 5.5 é apresentada a matriz de Classificação gerada a partir da aplicação do modelo 5 nos dados de construção e nos dados de validação, onde observa-se do lado esquerdo a quantidade de classificações corretas e incorretas na base de construção e no lado direito as classificações corretas e incorretas na base de validação.

Tabela 5.5: Matriz de Classificação do Modelo 5

Base de Treinamento				Base de Validação			
		Valor Observado				Valor Observado	
		Não Bolsista	Bolsista			Não Bolsista	Bolsista
Valor Predito	Não Bolsista	697	100	Valor Predito	Não Bolsista	76	15
	Bolsista	255	381		Bolsista	30	38

De acordo com os dados de saída apresentados na matriz de Classificação gerada a partir da aplicação do modelo 5 e levando em consideração a classificação dos pesquisadores utilizados na pesquisa, foram obtidos os seguintes resultados:

- **Base de Construção:**
  - Dos 952 pesquisadores não bolsistas, o modelo gerado classificou 697 pesquisadores como não bolsistas e 255 como bolsistas;
  - Dos 481 pesquisadores bolsistas o modelo classificou 381 pesquisadores como bolsistas e 100 como não bolsistas.
- **Base de Validação:**
  - Dos 106 pesquisadores não bolsistas, o modelo classificou 76 pesquisadores como não bolsistas e 30 como bolsistas;
  - Dos 53 pesquisadores bolsistas o modelo classificou 38 pesquisadores como bolsistas e 15 como não bolsistas.

6. Matriz de Classificação do Modelo 6

   A matriz de Classificação gerada a partir da aplicação do modelo 6 nos dados de construção e nos dados de validação pode ser observada na Tabela 5.6, sendo possível visualizar no lado esquerdo a quantidade de classificações corretas e incorretas feitas na base de construção e no lado esquerdo a quantidade de classificações corretas e incorretas feitas na base de validação.

Tabela 5.6: Matriz de Classificação do Modelo 6

Base de Construção				Base de Validação			
		Valor Observado				Valor Observado	
Valor Predito		Não Bolsista	Bolsista	Valor Predito		Não Bolsista	Bolsista
	Não Bolsista	687	106		Não Bolsista	84	12
	Bolsista	265	375		Bolsista	22	41

   De acordo com os dados de saída apresentados na matriz de Classificação e levando em consideração a classificação dos pesquisadores utilizados na pesquisa, foram obtidos os seguintes resultados:

   - **Base de Construção:**
     - Dos 952 pesquisadores não bolsistas, o modelo gerado classificou 687 pesquisadores como não bolsistas e 265 como bolsistas;
     - Dos 481 pesquisadores bolsistas o modelo classificou 375 pesquisadores como bolsistas e 106 como não bolsistas.
   - **Base de Validação:**
     - Dos 106 pesquisadores não bolsistas, o modelo classificou 84 pesquisadores como não bolsistas e 22 como bolsistas;
     - Dos 53 pesquisadores bolsistas o modelo classificou 41 pesquisadores como bolsistas e 12 como não bolsistas.

7. Matriz de Classificação do Modelo 7

   Na Tabela 5.7 é apresentada a matriz de Classificação gerada a partir da aplicação do modelo 5 nos dados de construção e nos dados de validação, onde observa-se no lado esquerdo a quantidade de classificações corretas e incorretas na base de construção e no lado direito as classificações corretas e incorretas na base de validação.

Tabela 5.7: Matriz de Classificação do Modelo 7

Base de Construção				Base de Validação			
		Valor Observado				Valor Observado	
		Não Bolsista	Bolsista			Não Bolsista	Bolsista
Valor Predito	Não Bolsista	694	102	Valor Predito	Não Bolsista	71	9
	Bolsista	259	378		Bolsista	34	45

De acordo com os dados de saída apresentados na matriz de Classificação e levando em consideração a classificação dos pesquisadores utilizados na pesquisa, verifica-se que o modelo obteve os seguintes resultados:

- **Base de Construção:**
  - Dos 953 pesquisadores não bolsistas, o modelo gerado classificou 694 pesquisadores como não bolsistas e 259 como bolsistas;
  - Dos 480 pesquisadores bolsistas o modelo classificou 378 pesquisadores como bolsistas e 102 como não bolsistas.
- **Base de Validação:**
  - Dos 105 pesquisadores não bolsistas, o modelo classificou 71 pesquisadores como não bolsistas e 34 como bolsistas;
  - Dos 54 pesquisadores bolsistas o modelo classificou 45 pesquisadores como bolsistas e 9 como não bolsistas.

8. Matriz de Classificação do Modelo 8

   O experimento de validação do modelo 8 a partir da matriz de Classificação é apresentado na Tabela 5.8. Nela é possível verificar a quantidade de classificações corretas e incorretas tanto da base de construção quanto na base de validação.

Tabela 5.8: Matriz de Classificação do Modelo 8

Base de Construção				Base de Validação			
		Valor Observado				Valor Observado	
		Não Bolsista	Bolsista			Não Bolsista	Bolsista
Valor Predito	Não Bolsista	696	103	Valor Predito	Não Bolsista	76	13
	Bolsista	257	377		Bolsista	29	41

De acordo com os dados de saída apresentados na matriz de Classificação e levando em consideração a classificação dos pesquisadores utilizados na pesquisa, verifica-se que o modelo obteve os seguintes resultados:

- **Base de Construção:**
  - Dos 953 pesquisadores não bolsistas, o modelo gerado classificou 696 pesquisadores como não bolsistas e 257 como bolsistas;
  - Dos 480 pesquisadores bolsistas o modelo classificou 377 pesquisadores como bolsistas e 103 como não bolsistas.
- **Base de Validação:**
  - Dos 105 pesquisadores não bolsistas, o modelo classificou 76 pesquisadores como não bolsistas e 29 como bolsistas;
  - Dos 54 pesquisadores bolsistas o modelo classificou 41 pesquisadores como bolsistas e 13 como não bolsistas.

9. Matriz de Classificação do Modelo 9

   Na Tabela 5.9 é apresentada a matriz de Classificação gerada a partir da aplicação do modelo 5 nos dados de construção e nos dados de validação, onde verifica-se a quantidade de classificações corretas e incorretas tanto da base de construção (lado direito) quanto na base de validação (lado esquerdo).

Tabela 5.9: Matriz de Classificação do Modelo 9

Base de Construção				Base de Validação			
		Valor Observado				Valor Observado	
		Não Bolsista	Bolsista			Não Bolsista	Bolsista
Valor Predito	Não Bolsista	693	111	Valor Predito	Não Bolsista	75	12
	Bolsista	259	369		Bolsista	31	42

   De acordo com os dados de saída apresentados na matriz de Classificação e levando em consideração a classificação dos pesquisadores utilizados na pesquisa, verifica-se que o modelo 9 obteve os seguintes resultados:

   - **Base de Construção:**
     - Dos 952 pesquisadores não bolsistas, o modelo gerado classificou 693 pesquisadores como não bolsistas e 259 como bolsistas;
     - Dos 480 pesquisadores bolsistas o modelo classificou 369 pesquisadores como bolsistas e 111 como não bolsistas.
   - **Base de Validação:**

  - Dos 106 pesquisadores não bolsistas, o modelo classificou 75 pesquisadores como não bolsistas e 31 como bolsistas;
  - Dos 54 pesquisadores bolsistas o modelo classificou 41 pesquisadores como bolsistas e 13 como não bolsistas.

10. Matriz de Classificação do Modelo 10

    Na Tabela 5.10 é apresentada a matriz de Classificação gerada a partir da aplicação do modelo 10 nos dados de construção e nos dados de validação, onde verifica-se a quantidade de classificações corretas e incorretas tanto da base de construção (lado direito) quanto na base de validação (lado esquerdo).

Tabela 5.10: Matriz de Classificação do Modelo 10

Base de Construção				Base de Validação			
		Valor Observado				Valor Observado	
		Não Bolsista	Bolsista			Não Bolsista	Bolsista
Valor Predito	Não Bolsista	694	104	Valor Predito	Não Bolsista	77	12
	Bolsista	258	376		Bolsista	29	42

De acordo com os dados de saída apresentados na matriz de classificação e levando em consideração a classificação dos pesquisadores utilizados na pesquisa, verifica-se que o modelo 9 obteve os seguintes resultados:

- **Base de Construção:**
  - Dos 952 pesquisadores não bolsistas, o modelo gerado classificou 694 pesquisadores como não bolsistas e 258 como bolsistas;
  - Dos 480 pesquisadores bolsistas o modelo classificou 376 pesquisadores como bolsistas e 104 como não bolsistas.
- **Base de Validação:**
  - Dos 106 pesquisadores não bolsistas, o modelo classificou 77 pesquisadores como não bolsistas e 29 como bolsistas;
  - Dos 54 pesquisadores bolsistas o modelo classificou 42 pesquisadores como bolsistas e 12 como não bolsistas.

Diante das matrizes expostas anteriormente, foi possível verificar o desempenho dos modelos obtidos, bem como o desempenho global do modelo, mediante a aplicação das métricas descritas na Subseção 2.1.2.6.

Na Tabela 5.11 podem ser visualizados o percentual de acurácia (ACC), de sensibilidade (SENS), de especificidade (ESP), de verdadeiros preditivos positivos (VPP) e de verdadeiros preditivos negativos (VPN) em relação aos resultados encontrados para a classificação dos pesquisadores em bolsistas e não bolsistas em cada um dos modelos gerados na base de construção. Já na Tabela 5.12 verifica-se as métricas de desempenho ao aplicar o modelo aos dados de validação, que não fizeram parte da construção do modelo.

Tabela 5.11: Métricas de desempenho dos modelos parciais na base de construção

	Base de Construção				
Modelo Parcial	ACC	SENS	ESP	VPP	VPN
1	73,90%	77,10%	72,30%	58,50%	86,20%
2	74,30%	77,50%	72,70%	58,90%	86,50%
3	74,70%	78,70%	72,60%	59,30%	87,10%
4	74,70%	79,90%	72,50%	59,20%	87,20%
5	75,20%	79,20%	73,20%	59,90%	87,40%
6	74,10%	77,90%	72,10%	58,50%	86,60%
7	74,80%	78,70%	72,80%	59,30%	87,10%
8	74,80%	78,50%	73,00%	59,40%	87,10%
9	74,10%	76,80%	72,70%	58,70%	86,10%
10	74,70%	78,30%	72,80%	59,30%	86,90%
Média	74,53%	78,26%	72,67%	59,10%	86,82%

Tabela 5.12: Métricas de desempenho dos modelos parciais na base de validação

	Base de Validação				
Modelo Parcial	ACC	SENS	ESP	VPP	VPN
1	78,60%	84,90%	75,40%	63,30%	90,90%
2	78,60%	83,00%	76,40%	63,70%	90,00%
3	72,90%	83,00%	67,90%	56,40%	88,80%
4	72,30%	67,90%	74,50%	57,10%	82,20%
5	71,70%	71,70%	71,70%	55,80%	83,50%
6	78,60%	77,30%	79,20%	65,00%	87,50%
7	72,90%	83,30%	67,60%	56,90%	88,70%
8	73,50%	75,90%	72,30%	58,50%	85,30%
9	73,10%	77,70%	70,70%	57,50%	86,20%
10	74,30%	77,70%	72,60%	59,10%	86,50%
Média	74,65%	78,24%	72,83%	59,33%	86,96%

A partir das duas tabelas observa-se que a média de acurácia dos modelos parciais construídos foi de 74,53% na base de construção e 74,65% na base de validação, demonstrando que o modelo generaliza bem os dados, uma vez que o desempenho foi similar nas duas

bases. Além disso, os modelos parciais obtiveram boas taxas de sensibilidade (78,26% na base de construção e 78,24% na base de validação) e especificidade (72,67% na base de construção e 72,83% na base de validação). Isso denota que, de modo geral, o modelo em si conseguiu classificar corretamente tanto os bolsistas como os não bolsistas, em ambas as bases.

Ainda observando os resultados apresentados, compreende-se que o modelo com maior capacidade de generalização é o modelo parcial 2, pois conseguiu apresentar um dos melhores desempenhos em relação a acurácia na base de construção, alcançando resultados ainda melhores na base de validação, o que indica que o modelo conseguiu generalizar bem os dados, fazendo classificações corretas nos dados que não foram utilizados durante o ajuste. Além disso, o modelo parcial 2 alcançou altas taxas de sensibilidade, especificidade, verdadeiros preditivos positivos e verdadeiro positivo negativo nos dados que não foram utilizados para construção do modelo (base de validação).

5.2 Análise dos Resultados e Discussão

Com base nos experimentos realizados na seção anterior, constatou-se que o modelo final a ser adotado na classificação de um pesquisador como bolsista ou não bolsista de produtividade a partir das métricas aplicadas a sua rede de colaboração foi o modelo 2, apresentado na Tabela 4.8.

Sendo assim, aplicando-se na Equação (2.1) os coeficientes obtidos na Tabela 4.8 verificamos que a probabilidade de um pesquisador ser bolsista de produtividade é dada pela Equação 5.1:

$$P(SerBolsista) = \frac{1}{1 + e^{-(0{,}7671+1{,}9584CI+0{,}0022GP-10{,}4900PR-4{,}2454CAL)}} \tag{5.1}$$

A partir da interpretação dos coeficientes estimados apresentados na equação, entende-se que as métricas de Centralidade de Intermediação e de Grau Ponderado exercem efeitos positivos na variável resposta, ou seja, quanto maior o valor desses indicadores, maior será a probabilidade de um pesquisador ser bolsista de produtividade. Em contraposição, as métricas de *PageRank* e Coeficiente de Agrupamento Local exercem efeitos negativos na variável resposta, indicando que os pesquisadores que possuem valores altos para essas métricas di-

minuem a probabilidade de serem bolsistas de produtividade.

Para uma análise mais completa do modelo apresentado, optou-se pela interpretação dos coeficientes a partir das razões de chances (do inglês odds ratio - OR) apresentadas na Tabela 4.8, em vez da interpretação dos próprios coeficientes estimados. Isso é possível porque a principal suposição da regressão logística é que o logaritmo da razão entre as probabilidades de ocorrência e não ocorrência do evento é linear.

A análise da razão de chances igual a 7,086 para o coeficiente da variável Centralidade de Intermediação indica que o aumento de uma unidade nessa variável aumenta em 7,086 vezes as chances de um pesquisador ser bolsista de produtividade. Em termos práticos, isso sugere que a relevância científica de um pesquisador tende a crescer à medida que ele assume um papel de intermediador dentro da rede, possibilitando a comunicação entre grupos distintos.

Prosseguindo com a análise das razões de chances dos coeficientes do modelo 2, observa-se que a variável Grau Ponderado possui razão de chance igual a 1,002, isso implica que a cada unidade de Grau Ponderado incrementada pelo pesquisador, ele consegue aumentar em 1,002 vezes as chances de ser bolsista. Nesse caso, entende-se que pesquisadores que colaboram frequentemente com os mesmos pesquisadores demonstram maior relevância científica do que aqueles que colaboram com muitos pesquisadores diferentes.

Já em relação a variável *PageRank*, verifica-se que a OR para cada unidade de incremento é menor do que 0,0001. Como a OR é menor que 1, a probabilidade de um pesquisador ser bolsista tende a diminuir quando há um aumento da sua métrica de PageRank, ou seja, a cada incremento de unidade na variável PageRank, o pesquisador diminui em 99,99% (1 - 0,0001) as chances de se tornar bolsista. Diante disso, infere-se que à medida que o pesquisador aumenta o seu grau de colaboração com pesquisadores de prestígio, as suas chances de ser bolsista de produtividade diminuem.

Essa característica peculiar pode ser explicada pelo fato de que os pesquisadores de grande relevância científica geralmente possuem vários orientandos de iniciação científica, mestrado e doutorado e, portanto, tendem a ter mais trabalhos em conjunto com eles, do que com pesquisadores já renomados, conferindo-lhes assim um PageRank baixo.

No caso da interpretação da OR igual a 0,0143 da variável Coeficiente de Agrupamento Local, verifica-se que quando o pesquisador incrementa em uma unidade o seu Coeficiente de Agrupamento Local, ele reduz em 98,57% (1 – 0,0143) as chances de ser bolsista de pro-

dutividade. Logo, pode-se inferir que quanto mais coeso for o grupo que o pesquisador está inserido menor é a sua relevância. Esse fato corrobora os resultados encontrados na pesquisa de (ALVES; BENEVENUTO; LAENDER, 2013) que demonstraram que pesquisadores de grande relevância são responsáveis por diminuir o agrupamento de suas redes científicas.

5.2.1 Análise Qualitativa dos Resultados

Os resultados alcançados nesta pesquisa demonstraram a existência de uma relação significativa entre a maneira como os pesquisadores colaboram e sua avaliação de desempenho científico, representada pela detenção de bolsas de produtividade.

Entende-se que o modelo apresentado neste trabalho é especialmente importante para comunidade científica pelo fato de evidenciar como as relações de colaboração podem ter influência sobre o desempenho de um pesquisador, abrindo um leque de possibilidades para novas discussões acerca deste tema. Como foi apresentado no modelo é importante para os pesquisadores estabelecerem relações com diversos grupos de pesquisa, favorecendo assim uma alta produtividade. Além disso, manter conexões fortes pode garantir que o pesquisador alcance bons índices de publicação, uma vez que a pesquisa torna-se mais sólida a medida que é mantida ao longo dos anos.

De maneira geral, reconhece-se que a avaliação do CNPq considera a produtividade de um pesquisador como parâmetro primordial no momento da distribuição das bolsas. No entanto, buscou-se entender se essa produtividade está associada a fatores sociais, representados pelas relações que são estabelecidas pelos pesquisadores durante sua trajetória científica.

Através disso identificou-se que as métricas aplicadas as redes de colaboração de um pesquisador apresentaram correlação significativa com a classificação do pesquisador, uma vez que foram capazes de classificá-lo como bolsista de produtividade de maneira satisfatória através do modelo de Regressão Logística.

Portanto, conclui-se que esse modelo pode ser utilizado pela comunidade científica para a compreensão da influência das relações de colaboração sobre a produtividade de um pesquisador e consequentemente como isso pode influenciar a avaliação do CNPq.

5.3 Considerações Finais

O objetivo deste capítulo foi apresentar os resultados obtidos com os testes de avaliação da capacidade de discriminação dos modelos parciais gerados no Capitulo 4, além de fornecer uma análise acerca do modelo que melhor generalizou os dados na base de validação.

A partir dos resultados apresentados na Seção 5.1, inferiu-se que o modelo com maior número de acertos foi o modelo 2, o qual apresentou as melhores métricas de desempenho tanto na base de construção quanto na base de validação, além de obter uma das melhores áreas sob a curva ROC e ainda apresentar o conjunto de variáveis independentes que mais se repetiu nos modelos parciais.

Diante disso, a Seção 5.2 apresentou uma análise acerca das variáveis e dos coeficientes definidos no modelo 2, de modo a discutir os efeitos de cada uma das métricas na relevância de um pesquisador.

Levando em consideração esses resultados, o próximo capítulo apresenta as conclusões obtidas com este trabalho, bem como as principais contribuições, limitações e trabalhos futuros.

Capítulo 6

Conclusão e Perspectivas

Este capítulo apresenta as considerações finais a respeito do trabalho desenvolvido, as contribuições e limitações da pesquisa, bem como as frentes de trabalhos futuros.

O processo de avaliação de pesquisadores para concessão de bolsas de produtividade é totalmente baseado em informações sobre produção e impacto. Em contrapartida a essa abordagem, observa-se através de diversas pesquisas que as interações sociais que ocorrem entre os pesquisadores são fatores decisivos na avaliação de sua relevância. Sendo assim, é de suma importância a utilização de métricas alternativas baseadas nas informações de colaboração durante o processo de avaliação.

Levando em consideração essa necessidade, este trabalho apresentou um modelo de regressão logística que associa a relevância de um pesquisador às métricas de Análise de Redes Sociais aplicadas as suas redes de colaboração científica, avaliando sua relevância a partir da classificação como bolsista ou não bolsista de produtividade.

Para tanto se utilizou como amostra a base de dados formada pelos docentes dos programas de Pós-Graduação na área de Ciência da Computação, que conta tanto com pesquisadores bolsistas quanto não bolsistas. A partir dela foram extraídas redes de colaboração mediante as informações disponíveis no CV Lattes desses pesquisadores e aplicadas métricas de ARS que serviram de base para a construção do modelo, gerado a partir da aplicação da técnica de regressão logística, com validação cruzada *10-fold* estratificada.

Os resultados apresentados demonstram que o modelo final obtido apresentou um bom desempenho na classificação dos pesquisadores tanto na base de construção quanto na base de validação, atingindo nesta uma acurácia de 78,60%, sensibilidade de 83,00% e especifi-

cidade de 76,40%. Esses resultados podem ser considerados excelentes tendo em vista que nem todos os pesquisadores que possuem um perfil de bolsista de produtividade submetem projetos para concorrer às bolsas, ou ainda não são contemplados como bolsistas devido ao número restrito de bolsas oferecidas pelo CNPq.

Esse resultado reforça a hipótese inicial desta pesquisa, a qual considerava que existiam relações significativas entre as métricas de ARS e a relevância de um pesquisador a ponto de ser possível usá-las em um modelo de regressão logística capaz de avaliar a relevância de um pesquisador mediante sua classificação em termos de detenção de bolsa de produtividade.

Diante do modelo obtido, foi possível inferir ainda que as métricas de ARS que exercem maior influência sobre a relevância científica de um pesquisador, são a Centralidade de Intermediação, o Grau Ponderado, o *PageRank* e o Coeficiente de Agrupamento Local. Esse resultado revela algumas nuances da colaboração científica que interferem diretamente na avaliação do mérito científico de um pesquisador.

O fato da Centralidade de Intermediação apresentar efeitos positivos sobre o status do pesquisador aumentando as chances dele se tornar bolsista, reflete que os cientistas com alto poder de comunicação dentro da rede, ou seja, aqueles que assumem o papel de intermediador entre diversos grupos distintos tendem a ser mais relevantes do que aqueles que não apresentam relações com diversos grupos.

Outro fator a ser analisado é o efeito também positivo do Grau Ponderado sobre a relevância do pesquisador. Isso indica que a força dos laços estabelecidos por um pesquisador também é preponderante na avaliação da sua relevância. Sendo assim, compreende-se que pesquisadores que colaboram diversas vezes com os mesmos cientistas geralmente apresentam um desempenho melhor do que aqueles que estabelecem laços fracos com seus colaboradores.

Observou-se ainda que quanto maior a métrica de *PageRank* de um pesquisador menor são as chances dele ser classificado como bolsista. Como essa métrica reflete a qualidade das relações estabelecidas pelo pesquisador, entende-se que os pesquisadores com maior relevância costumam colaborar, em geral, com pesquisadores menos relevantes dentro da rede, em vez de estabelecer relações com os líderes da sua área de pesquisa.

Outra característica importante apontada pelo modelo gerado foi que os pesquisadores de maior relevância possuem um Coeficiente de Agrupamento Local baixo indicando que

os seus colaboradores não formam um grupo coeso, ou seja, não costumam colaborar entre si. Essa constatação reforça a ideia de que pesquisadores relevantes colaboram com diversos grupos distintos que não interagem entre si, sendo ele o responsável por intermediar a comunicação entre esses grupos.

6.1 Contribuições

Diante do exposto, acredita-se que a principal contribuição deste trabalho consiste no fato do mesmo apresentar uma estratégia eficaz para a avaliação da relevância de um pesquisador a partir do seu perfil de colaboração científica, apresentando como vantagem o retorno de resultados imparciais e instantâneos, reduzindo os esforços e tornando o processo menos subjetivo.

Cita-se ainda como contribuição a possibilidade de entender o impacto da dinâmica das relações sociais dos pesquisadores sobre sua relevância científica, permitindo aos pesquisadores avaliarem as melhores estratégias para estabelecerem suas conexões no meio acadêmico a fim de maximizar seu mérito científico.

Além disso, o modelo apresentado pode servir de complemento durante a avaliação dos pesquisadores pelos órgãos de fomento, uma vez que já foi comprovado por diversas pesquisas que a colaboração científica tem forte influência no desempenho de um pesquisador, e, portanto deve ser levada em consideração no contexto avaliativo.

6.2 Limitações

Este estudo apresenta algumas limitações, sendo a primeira delas o nível da amostra utilizada, que foi restrita apenas aos pesquisadores da área de Ciência da Computação, o que limita o poder de generalização do modelo para pesquisadores de outras áreas sem as devidas investigações.

Além disso, entende-se como outra possível limitação o fato do modelo apresentado não considerar os aspectos temporais das relações, que podem contribuir para um melhor entendimento do impacto dessas métricas na relevância do pesquisador ao longo de sua trajetória científica.

Apesar destas limitações entende-se que os resultados apresentados nesta pesquisa foram satisfatórios, pois permitiram um melhor entendimento acerca da influência das relações de colaboração sobre a avaliação da relevância científica de um pesquisador.

6.3 Trabalhos Futuros

O modelo apresentado neste trabalho representa um primeiro esforço na tentativa de entender como o perfil de colaboração de um pesquisador impacta na avaliação de seu mérito científico e apesar dos resultados obtidos terem sido bastante satisfatórios ainda há muito trabalho a ser realizado nessa área de pesquisa.

Compreende-se que os questionamentos levantados nesta pesquisa possui outros fatores que também devem ser considerados, acarretando possíveis desdobramentos deste trabalho com a incorporação tanto de novas técnicas como de novos dados que possam agregar valor aos resultados aqui apresentados.

Diante dessas perspectivas, almeja-se como trabalho futuro generalizar o modelo apresentado, uma vez que o mesmo está adaptado apenas à realidade brasileira devido à fonte de dados utilizada, permitindo assim a sua utilização por pesquisadores de vários países.

Outro direcionamento de trabalho futuro consiste em utilizar novas técnicas que permitam a incorporação de aspectos temporais com o intuito de investigar a influência das colaboração ao longo de toda a trajetória do pesquisador, visando assim prever sua relevância em longo prazo. Outro direcionamento possível seria a aplicação de Análise Discriminante ou Árvores de Decisão e a incorporação de métricas de produção e impacto ao modelo com o intuito de possibilitar uma comparação com as métricas de ARS em termos de explicação para a classificação de um pesquisador.

Além disso, pretende-se estender o arcabouço desenvolvido nesta pesquisa para avaliação de pesquisadores de programas de outras áreas, a fim de verificar se os resultados encontrados serão mantidos ou se haverá divergências.

Ainda como trabalho futuro, pretende-se disponibilizar o modelo gerado para comunidade científica a partir de uma aplicação web que permita a avaliação automática da relevância de um pesquisador mediante a submissão do *link* do seu currículo Lattes.

Bibliografia

ABBASI, A.; ALTMANN, J. On the correlation between research performance and social network analysis measures applied to research collaboration networks. In: *System Sciences (HICSS), 2011 44th Hawaii International Conference on*. [S.l.: s.n.], 2011. p. 1–10. ISSN 1530-1605.

ABBASI, A.; ALTMANN, J.; HOSSAIN, L. Identifying the effects of co-authorship networks on the performance of scholars: A correlation and regression analysis of performance measures and social network analysis measures. *Journal of Informetrics*, v. 5, n. 4, p. 594–607, out. 2011. ISSN 17511577. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1751157711000630>.

ACTION: Portal action: Predição. 2014. Disponível em: <http://www.portalaction.com.br/989-predicao>. Acesso em: 12 nov.2014.

ALVES, B. L.; BENEVENUTO, F.; LAENDER, A. H. The role of research leaders on the evolution of scientific communities. In: *Proceedings of the 22Nd International Conference on World Wide Web Companion*. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2013. (WWW '13 Companion), p. 649–656. ISBN 978-1-4503-2038-2.

ALVES, M. F. *Previsão de Demanda de Cargas Eletricas por Seleção de Variaveis Stepwise e Redes Neurais Artificiais*. Dissertação (Mestrado) — Universidade Estadual Paulista, 2013.

ALVES, M. F.; LOTUFO, A. D. P.; LOPES, M. L. M. Seleção de variaveis stepwise aplicadas em redes neurais artificiais para previsão de demanda de cargas eletricas. In: *Anais do XI Simpósio Brasileiro de Automação Inteligente*. [S.l.: s.n.], 2013.

ARAUJO, E. B. et al. Collaboration networks from a large cv database: Dynamics, topology and bonus impact. *PLoS ONE*, Public Library of Science, v. 9, n. 3, p. e90537, 03 2014. Disponível em: <http://dx.doi.org/10.1371%2Fjournal.pone.0090537>.

ASSUNÇÃO, F. *Estratégias para tratamento de variáveis com dados faltantes durante o desenvolvimento de modelos preditivos*. Dissertação (Mestrado) — Instituto de Matemática e Estatística, Universidade de São Paulo, 2012.

AZEVEDO, T. B. de. *Análise de Redes Sociais através de softwares de ARS e de Data Mining: um estudo de caso em turmas de graduação*. 2011.

BASTIAN, M.; HEYMANN, S.; JACOMY, M. *Gephi: An Open Source Software for Exploring and Manipulating Networks*. 2009. Disponível em: <http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154>.

BENEVENUTO, F.; ALMEIDA, J.; SILVA, A. Explorando redes sociais online: Coleta e análise de grandes bases de dados de redes sociais online. In: *Jornadas de Atualização em Informática (JAI)*. [S.l.: s.n.], 2011. p. 11–57.

BINOTTO, E.; HOFF, D. N.; SIQUEIRA, E. S. Peer review e a qualificação da produção científica: limites e avanços. In: *XXXII Encontro da ANPAD*. [S.l.: s.n.], 2008.

BORDONS, M. et al. The relationship between the research performance of scientists and their position in co-authorship networks in three fields. *Journal of Informetrics*, v. 9, n. 1, p. 135 – 144, 2015. ISSN 1751-1577. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1751157714001138>.

BRAGA, A. C. da S. *Curvas ROC: aspectos funcionais e aplicações*. Tese (Doutorado) — Universidade do Minho, 2000.

BROCCO, J. B. *Ponderação de Modelos com Aplicação em Regressão Logistica Binaria*. Dissertação (Mestrado) — Departamento de Estatística, Universidade Federal de São Carlos, 2006.

CABRAL, C. I. S. *Aplicação do Modelo de Regressão Logistica num Estudo de Mercado*. Dissertação (Mestrado) — Departamento de Estatística e Investigação Operacional, Universidade de Lisboa, 2013.

CAFE, A. L. da P. *A Produção Científica do Campo da Sociologia Brasileira face aos critérios de avaliação do Cnpq e da Capes: 2007-2009*. Dissertação (Mestrado) — Universidade Federal da Bahia, 2012.

CHELMIS, C.; PRASANNA, V. K. Social networking analysis: A state of the art and the effect of semantics. In: *SocialCom/PASSAT*. [S.l.]: IEEE, 2011. p. 531–536. ISBN 978-1-4577-1931-8.

CIMENLER, O.; REEVES, K. A.; SKVORETZ, J. A regression analysis of researchers' social network metrics on their citation performance in a college of engineering. *Journal of Informetrics*, v. 8, n. 3, p. 667 – 682, 2014. ISSN 1751-1577. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1751157714000571>.

CNPQ: Critérios de julgamento dos comitês de assessoramento. 2014. Disponível em: <http://www.cnpq.br/web/guest/criterios-de-julgamento>. Acesso em: 12 out.2014.

COVÕES, T. F. *Seleção de atributos via agrupamento*. Dissertação (Mestrado) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2010.

DHAR, V. Data science and prediction. *Commun. ACM*, ACM, New York, NY, USA, v. 56, n. 12, p. 64–73, dez. 2013. ISSN 0001-0782.

DIGIAMPIETRI, L. A. et al. Brax-ray: An x-ray of the brazilian computer science graduate programs. *PLoS ONE*, Public Library of Science, v. 9, n. 4, p. e94541, 04 2014. Disponível em: <http://dx.doi.org/10.1371%2Fjournal.pone.0094541>.

DREW: The data science venn diagram. 2013. Disponível em: <http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>. Acesso em: 20 out.2014.

EATON, J. P. et al. Structural analysis of co-author relationships and author productivity in selected outlets for consumer behavior research. *Journal of Consumer Psychology*, v. 8, n. 1, p. 39 – 59, 1999. ISSN 1057-7408. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1057740899703438>.

Equipe Estatcamp. *Software Action*. São Carlos - SP, Brasil, 2014. Disponível em: <http://www.portalaction.com.br/>.

FARIAS, L. R. de; VARGAS, A. P.; BORGES, E. N. Um sistema para análise de redes de pesquisa baseado na plataforma lattes. In: *Anais da VIII Escola Regional de Banco de Dados*. [S.l.: s.n.], 2012.

FERREIRA, J. M. *Analise de sobrevivência: uma visão de risco comportamental na utilização de cartão de credito*. Dissertação (Mestrado) — Departamento de Estatística e Informática, Universidade Federal Rural de Pernambuco, 2007.

FREIRE, V. P.; FIGUEIREDO, D. R. Ranking in collaboration networks using a group based metric. *Journal of the Brazilian Computer Society*, Springer-Verlag, v. 17, n. 4, p. 255–266, 2011. ISSN 0104-6500. Disponível em: <http://dx.doi.org/10.1007/s13173-011-0041-7>.

FAVERO, L. P. L. *Analise de dados: modelagem multivariada para tomada de decisões*. [S.l.]: Elsevier, 2009.

GONÇALVES, E. B.; GOUVÊA, M. A.; MANTOVANI, D. M. N. Analise de risco de credito com o uso de regressão logistica. *Revista Contemporânea de Contabilidade*, v. 10, n. 20, p. 139–160, 2013.

GUERRA, G. N. *Modelo de Reputação e Ontologia Aplicados à Rede Social Científica do ObserveUnB*. Dissertação (Mestrado) — Departamento de Ciência da Computação, Universidade de Brasília, 2012.

HAIR, J.; ANDERSON, R.; TATHAM, R. *Analise Multivariada de Dados*. 6. ed. [S.l.]: Bookman, 2009. ISBN 9788536304823.

HAUCK, W. W.; DONNER, A. Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association*, v. 72, n. 360a, p. 851–853, 1977.

HAYKIN, S. *Redes Neurais, Princípios e prática*. 2. ed. [S.l.]: Bookman, 1999.

HERRERA, F. et al. Un estudio empirico preliminar sobre los tests estadisticos más habituales en el aprendizaje automático. Tendencias de la Mineria de Datos en Espana, Red Espanola de Mineria de Datos y Aprendizaje (TIC2002-11124-E), 2004.

HOSMER, D.; LEMESHOW, S. *Applied Logistic Regression*. [S.l.]: Wiley, 2000. 392 p. ISBN 9780471654025.

KARAM, K. de A. *Regressão Logistica : Um modelo de Risco de Cancelamento de Clientes*. Dissertação (Mestrado) — Pontifícia Universidade Católica do Rio de Janeiro, 2006.

KATZ, J. S.; MARTIN, B. R. What is research collaboration? *Research Policy*, v. 26, 1997.

KLEINBERG, J. M. Authoritative sources in a hyperlinked environment. *J. ACM*, ACM, New York, NY, USA, v. 46, n. 5, p. 604–632, set. 1999. ISSN 0004-5411. Disponível em: <http://doi.acm.org/10.1145/324133.324140>.

KUMAR, S.; JAN, J. Mapping research collaborations in the business and management field in malaysia, 1980–2010. *Scientometrics*, Springer Netherlands, v. 97, n. 3, p. 491–517, 2013. ISSN 0138-9130. Disponível em: <http://dx.doi.org/10.1007/s11192-013-0994-8>.

KUMAR, S.; JAN, J. Research collaboration networks of two oic nations: comparative study between turkey and malaysia in the field of 'energy fuels', 2009–2011. *Scientometrics*, Springer Netherlands, v. 98, n. 1, p. 387–414, 2014. ISSN 0138-9130. Disponível em: <http://dx.doi.org/10.1007/s11192-013-1059-8>.

LANE, J. Let's make science metrics more scientifiic. *Nature*, v. 454, n. 25, p. 488–489, 2010.

LEE, S.; BOZEMAN, B. The impact of research collaboration on scientific productivity. *Social Studies of Science*, v. 35, n. 5, p. 673–702, 2005.

LIAO, C. How to improve research quality? examining the impacts of collaboration intensity and member diversity in collaboration networks. *Scientometrics*, Springer Netherlands, v. 86, n. 3, p. 747–761, 2011. ISSN 0138-9130.

LIMA, R. A. de; VELHO, L. M. L. S.; FARIA, L. I. L. de. Bibliometria e "avaliação" da atividade científica: um estudo sobre o índice h. *Perspectivas em ciencia da informacao*, v. 17, p. 3–17, 2012. ISSN 1413-9936.

MANDALA, G. S. *Introdução à Econometria*. [S.l.]: LTC, 2003. ISBN 9788521613862.

MARTELETO, R. M. Análise de redes sociais: aplicação nos estudos de transferência da informação. *Ciência da informação*, SciELO Brasil, v. 30, n. 1, p. 71–81, 2001.

MARTINS, D. L.; FERREIRA, S. M. S. P. Mapeamento e avaliação da produção científica da universidade de são paulo com foco na estrutura e dinâmica de suas redes de colaboração científica. *Liinc em Revista*, v. 9, n. 1, 2013.

MCCARTY, C. et al. Predicting author h-index using characteristics of the co-author network. *Scientometrics*, Springer Netherlands, v. 96, n. 2, p. 467–483, 2013. ISSN 0138-9130.

MENA-CHALCO, J.; DIGIAMPIETRI, L.; JR., R. C. Caracterizando as redes de coautoria de currículos lattes. In: *CSBC 2012 - BraSNAM ()*. [S.l.: s.n.], 2012.

MENDONÇA, T. S. *Modelos de regressão logística clássica, Bayesiana e redes neurais para Credit Scoring*. Dissertação (Mestrado) — Departamento de Estatística, Universidade Federal de São Carlos, 2008.

NEWMAN, M. The structure and function of complex networks. *SIAM review*, JSTOR, v. 45, p. 167–256, 2003. ISSN 0036-1445.

OLIVEIRA, L. dA S. *Seleção de Covariaveis para Ajuste de Regressão Logistica na Analise da Abundância de Invertebrados edaficos em Diferentes Agroecossistemas*. Dissertação (Mestrado) — Universidade Federal de Viçosa, 2011.

PAGE, L. et al. The pagerank citation ranking: Bringing order to the web. In: *Proceedings of the 7th International World Wide Web Conference*. Brisbane, Australia: [s.n.], 1998. p. 161–172. Disponível em: <citeseer.nj.nec.com/page98pagerank.html>.

PORTO, F. A. M.; ZIVIANI, A. Ciência de dados. In: *Anais III Seminário de Grandes Desafios da computação*. [S.l.: s.n.], 2014.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2013. Disponível em: <http://www.R-project.org/>.

RECUERO, R. Redes sociais na internet: Considerações iniciais. *Revista E-Compós*, v. 10, 2005.

SARIGOL, E. et al. Predicting scientific success based on coauthorship networks. *CoRR*, p. –1–1, 2014.

SIDONE, O. J. G.; HADDAD, E. A.; MENA-CHALCO, J. Padrões de colaboração científica no brasil: O espaço importa? In: *Anais do XLI Encontro Nacional de Economia [Proceedings of the 41th Encontro Brasileiro de Economia*. [S.l.: s.n.], 2014.

SILVA, A. C. da. *Analise estatistica de inqueritos online*. Dissertação (Mestrado) — Escola de Ciências, Universidade do Minho, 2011.

SILVA, M. R.; MA, H.; ZENG, A. P. Centrality, network capacity, and modularity as parameters to analyze the core-periphery structure in metabolic networks. In: *Proceedings of the IEEE*. [S.l.: s.n.], 2008.

SPINAK, E. Indicadores cienciométricos. *Ci. Inf*, SciELO Brasil, v. 27, n. 2, p. 141–148, 1998.

STANTON, J. *An Introduction to Data Science*. [S.l.]: jsresearch.net, 2012. 196 p.

TAN, P. et al. *Introdução ao datamining: mineração de dados*. [S.l.]: Ciencia Moderna, 2009. ISBN 9788573937619.

VANZ, I. R. C. S. Samile Andrea de S. Colaboração científica: revisão teórico-conceitual. *Perspectivas em Ciência da Informação*, v. 2, n. 15, 2010.

WAINER, J.; VIEIRA, P. Avaliação de bolsas de produtividade do cnpq e medidas bibliométricas: correlações para todas as grandes áreas. *Perspectivas em Ciência da Informação*, v. 18, n. 2, p. 60–78, 2013.

WAINER, J.; VIEIRA, P. Correlations between bibliometrics and peer evaluation for all disciplines: the evaluation of brazilian scientists. *Scientometrics*, Springer Netherlands, v. 96, n. 2, p. 395–410, 2013. ISSN 0138-9130.

WALTER, S. A. et al. Lealdade de estudantes: Um modelo de regressão logistica. *Revista de Administração FACES Journal*, Universidade FUMEC, v. 10, n. 4, p. 129–151, 2010.

WASSERMAN, S.; FAUST, K. *Social Network Analysis: Methods and Applications*. [S.l.]: Cambridge University Press, 1994. (Structural Analysis in the Social Sciences). ISBN 9780521387071.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of'small-world'networks. *Nature*, v. 393, n. 6684, p. 409–10, 1998.

WITTEN, I. H.; FRANK, E. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005. ISBN 0120884070.

ZUMEL, N.; MOUNT, J. *Practical Data Science with R*. [S.l.]: Manning Publications Company, 2014. ISBN 9781617291562.