UNIVERSIDADE ESTADUAL DE MONTES CLAROS
Centro de Ciências Exatas e Tecnológicas
Programa de Pós Graduação em Modelagem Computacional e Sistemas

Hugo Andrei Mendes da Silva

UM MODELO PARA PREDIÇÃO DE DESEMPENHO DE PESQUISADORES NA GRANDE ÁREA DE CONHECIMENTO CIÊNCIA DA COMPUTAÇÃO

Montes Claros – MG
Agosto de 2016

Hugo Andrei Mendes da Silva

UM MODELO PARA PREDIÇÃO DE DESEMPENHO DE PESQUISADORES NA GRANDE ÁREA DE CONHECIMENTO CIÊNCIA DA COMPUTAÇÃO

Dissertação apresentada no Programa de Pós Graduação em Modelagem Computacional e Sistemas, da Universidade Estadual de Montes Claros como exigência para obtenção do grau de Mestre em Modelagem Computacional e Sistemas.

Orientador: **Profº. Dr. Renê Rodrigues Veloso**

Montes Claros – MG

Agosto de 2016

Hugo Andrei Mendes da Silva

UM MODELO PARA PREDIÇÃO DE DESEMPENHO DE PESQUISADORES NA GRANDE ÁREA DE CONHECIMENTO CIÊNCIA DA COMPUTAÇÃO

Dissertação apresentada no Programa de Pós Graduação em Modelagem Computacional e Sistemas, da Universidade Estadual de Montes Claros como exigência para obtenção do grau de Mestre em Modelagem Computacional e Sistemas.

Montes Claros, 11 de Agosto de 2016.

Orientador: _____

Profº. Dr. Renê Rodrigues Veloso

Universidade Estadual de Montes Claros

Coorientador: _____

Profº. Dr. Marcos Flávio Silveira Vasconcelos D'Angelo

Universidade Estadual de Montes Claros

Membros:

_____

Profº. Dr. Antônio Wilson Vieira

Universidade Estadual de Montes Claros

_____

Profº. Dr. João Batista Mendes

Universidade Estadual de Montes Claros

Montes Claros – MG

Agosto de 2016

S586m

Silva, Hugo Andrei Mendes da.
Um modelo para predição de desempenho de pesquisadores na grande área de conhecimento Ciência da Computação [manuscrito] / Hugo Andrei Mendes da Silva. – 2016.
74 f. : il.

Bibliografia: f. 66-68.
Dissertação (mestrado) - Universidade Estadual de Montes Claros - Unimontes, Programa de Pós-Graduação em Modelagem Computacional e Sistemas/PPGMCS, 2016.

Orientador: Prof. Dr. Renê Rodrigues Veloso.
Coorientador: Prof. Dr. Marcos Flávio Silveira Vasconcelos D'Ângelo.

1. Pesquisadores – Ciência da Computação – Produtividade. 2. Mineração de dados. 3. Classificação. I. Veloso, Renê Rodrigues. II. D'Ângelo, Marcos Flávio Silveira Vasconcelos. III. Universidade Estadual de Montes Claros. IV. Título.

Catalogação Biblioteca Central Professor Antônio Jorge

RELATÓRIO DE DEFESA DE DISSERTAÇÃO

MESTRADO

Universidade Estadual de Montes Claros - Unimontes
Pró-Reitoria de Ensino/Coordenadoria de Pós-Graduação
Programa de Pós-Graduação em Modelagem Computacional e Sistemas - PPGMCS

1-Indentificação do Aluno

Nome: Hugo Andrei Mendes da Silva	Matrícula: 100001681
Linha de Pesquisa: Inteligência Computacional, Otimização e suas Aplicações	

2- Sessão de Qualificação

Título:

UM MODELO PARA PREDIÇÃO DE DESEMPENHO DE PESQUISADORES NA GRANDE ÁREA DE CONHECIMENTO CIÊNCIA DA COMPUTAÇÃO.

3- Comissão Examinadora

Nome	Função	Assinatura
Renê Rodrigues Veloso	Orientador (a)	
Marcos Flávio S.V. D'Ângelo	Coorientador(a)	
Antônio Wilson Vieira	Examinador(a)	
João Batista Mendes	Examinador(a)	

4- Resultado

A comissão Examinadora, em **11/08/2016** após Defesa de Dissertação e arguição do(a) candidato(a), decidiu:

(X) pela aprovação da Dissertação
( ) pela reprovação da Dissertação
( ) pela revisão de forma, indicando o prazo de 30 dias para apresentação definitiva.
( ) pela reformulação da Dissertação, indicando o prazo de _____ dias para nova versão.

Preencher somente em caso de revisão de forma:	**Preencher somente em caso de revisão de reformulação:**
( ) O(a) aluno(a) apresentou a revisão de forma e a Dissertação foi aprovada.	( ) O(a) aluno(a) apresentou a reformulação e a Dissertação foi aprovada.
( ) O(a) aluno(a) apresentou a revisão de forma e a Dissertação foi reprovada.	( ) O(a) aluno(a) apresentou a reformulação e a Dissertação foi reprovada.
( ) O(a) aluno(a) não apresentou a revisão da forma.	( ) O(a) aluno(a) não apresentou a reformulação.

Autenticação
Orientador(a) Comissão Examinadora

11/08/2016
Data

Assinatura

Autenticação
Coordenador

Prof. Nilton Alves Maia
Coordenador do PPGMCS
UNIMONTES
Masp 1046657-1

11/08/2016
Data

Assinatura

Aos meus familiares:
Mariana, José, Gislene, Fernanda, Lucas, Laíse, Rodrigo e Carlos Eduardo.

Agradecimentos

Os meus sinceros agradecimentos aos que me acompanharam e contribuíram de alguma forma durante esses dois anos de curso. Em especial, agradeço:

A Deus, pela vida e por tudo.

Ao meu orientador Prof. Renê Veloso, pela amizade, ensinamentos e compreensão diante todas as dificuldades que tive no cruso.

Ao meu coorientador Prof. Marcos Flávio, por estar sempre disposto a ajudar e pelo exemplo de pesquisador.

Aos demais professores e membros do PPGMCS, que sempre bem humorados e solícitos, mostram dedicação em seus ensinamentos e atividades.

Aos colegas do PPGMCS, pela amizade e apoio durante o curso.

Aos meus pais, José e Gislene, que sempre fizeram questão de estar ao meu lado e mostram diariamente que o exemplo é a melhor forma de ensinar e motivar um filho.

À minha irmã Fernanda que, de sua forma alegre e amorosa, torna a vida de todos nós muito melhor.

À minha irmã e orientadora Laíse, que define em pessoa o conceito da palavra dedicação. Te agradeço todo auxílio que me deu na reta final desse projeto.

Todos meus familiares, que me apoiam e incentivam sempre.

Em especial, a minha esposa Mariana, pelo amor e companheirismo diário; Desde a prova de seleção até o último ponto final dessa dissertação, a sua paciência e apoio constante foram fundamentais.

Resumo

Avaliar o desempenho científico de pesquisadores nem sempre é um trabalho simples. É comum esse processo ser realizado através de índices utilizados globalmente. Entretanto, esse tipo de avaliação deve levar em consideração a área de atuação dos pesquisadores e o contexto em que eles trabalham, visto que realidades distintas necessitam de avaliações específicas, o que muitas vezes não acontece com esses indicadores globais, classificados como cienciométricos.

Diferentes trabalhos têm sido realizados para avaliação da produtividade científica de instituições de ensino. Mesmo assim, não é comum encontrar estudos que avaliam o desempenho do pesquisador em relação ao grupo de pesquisadores que trabalham na mesma área. Tais estudos são importantes para direcionar o pesquisador e indicar o seu potencial. Para tanto, esse trabalho tem o objetivo de fazer uma análise da produtividade de pesquisadores da grande área Ciência da Computação, que são cadastrados na plataforma Lattes, e desenvolver um modelo preditivo do desempenho de pesquisadores nessa área.

Para isso, foram coletados dados dos pesquisadores, dos quais foram construídos atributos a serem considerados nos modelos preditivos. Esses atributos foram analisados através de algoritmos conhecidos na mineração de dados e de comparações com avaliações feitas pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Tais análises destacaram informações relevantes para predizer o desempenho desses pesquisadores e, posteriormente, foi criado um modelo preditivo de classificação que obteve resultados relevantes, com acurácia média acima de 75%, o que pode auxiliar no processo de julgamento de bolsas pelas instituições de fomento à pesquisa, além de possibilitar o direcionamento da carreira acadêmica dos pesquisadores que trabalham na grande área Ciência da Computação.

Palavras-chave: Pesquisadores, Mineração de Dados, Classificação, Modelo.

Abstract

Assess the scientific performance of researchers is not always a simple task. Often this process is done with global indexes. However, such an assessment should take into consideration the area of expertise of the researchers and the context in which they work, since distinct realities require specific assessments. It often does not happen with these global indicators, classified like scientometric.

Different researches have been made to evaluate the scientific productivity of universities. However, it's not usual to find works that relate the researcher performance with his research group. Such studies are important to direct the researchers and indicate their potential. Therefore, this work aims to make an analysis of the productivity of researchers from Computer Science area that are registered in the Lattes Platform, and develop a predictive model of the performance of researchers in this area.

For this, the data of researchers were collected and used to construct attributes that were considered in the predictive models. These attributes were analyses using common data mining algorithms and by comparing them with assessments made by the *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq). The result of these analyses are relevant informations to predict the performance of these researchers. After that, was created a predictive classification model that obtained relevant results, with average accuracy above 75%. It can assist in the judging process by the research founding agencies, and to help in academic career of the Computer Science researchers.

Keywords: Researchers, Data Mining , Classification, Model.

Lista de Figuras

2.1 Critérios avaliados pelo CA-CC 22
2.2 Processo de KDD 24

3.1 Detalhamento da criação da base de dados 32
3.2 Detalhamento da criação da base de dados com mesmo potencial 34
3.3 Avaliação dos atributos segundo perfil dos pesquisadores 35
3.4 Avaliação da média dos atributos segundo perfil dos pesquisadores 36
3.5 Avaliação da média normalizada dos atributos segundo perfil dos pesquisadores 36
3.6 Avaliação da média normalizada dos atributos segundo grupos pesquisadores criados pelo algoritmo 38
3.7 Fluxo de atividades para seleção dos atributos 38
3.8 Análise estatística da base por atributo 40
3.9 Matriz de correlação dos 16 atributos normalizados 41
3.10 Avaliação da média normalizada dos atributos segundo grupos com 15 atributos 42
3.11 Avaliação das divergências entre os grupos e a classificação do CNPq . . . 43
3.12 Avaliação da classificação do CNPq com os atributos finais 44
3.13 Avaliação dos grupos criados pelo agrupamento com os atributos finais . . 44

4.1 Processo de validação cruzada *10-fold* estratificada 47
4.2 Acurácia média do *k-Nearest Neighbors* com variação dos valores de K . . . 48
4.3 Acurácia Média do *Random Forest* com Diferentes Estimadores 49
4.4 Distribuição da Acurácia dos Classificadores - Validação Cruzada 52
4.5 Novos rótulos para classificação binária 54
4.6 Avaliação das divergências encontradas no Teste 1 56
4.7 Avaliação das divergências encontradas no Teste 2 57
4.8 Novo arupamento prévio 60

Lista de Tabelas

3.1 Avaliações de Bolsa pelo Comitê CA-CC por Ano 31
3.2 Amostra da Base Completa para Análise dos Atributos 33
3.3 Exemplo de Contagem dos Atributos 34
3.4 Resultado do agrupamento com 5 grupos - 59 atributos 37
3.5 Resultado de agrupamento com 5 Grupos - 15 atributos 42

4.1 Distribuição Anual da Amostra de Dados Completa 46
4.2 Avaliação com Diferentes Configurações do Algoritmo *Naive Bayes* 50
4.3 Avaliação do Modelo com Algoritmos *Support Vector Machines* 50
4.4 Modelos de Classificação com os Respectivos Parâmetros Ajustados 51
4.5 Resultado da Validação Cruzada para os Modelos - Acurácia (%) 51
4.6 Teste de Hipótese entre Modelo 2 e Demais Modelos 52
4.7 Resultado da Validação Cruzada para Classificação Binária (%) 54
4.8 Matriz de Confusão da Classificação no Teste 1 - Acurácia 52,99% 55
4.9 Matriz de Confusão da Classificação no Teste 2 - Acurácia 60,68% 55
4.10 Avaliações de Bolsa CA-CC por Ano - Teste de Predição 58
4.11 Exemplo de Contagem dos Atributos para Predição 58
4.12 Acurácia Média da Validação Cruzada para Predição - Acurácia (%) 59
4.13 Distribuição dos Pesquisadores em Novos Rótulos 61
4.14 Resultado da Validação Cruzada para Predição com Agrupamento - Acurácia (%) 62

A.1 Tabela com 15 Atributos Selecionados 69
A.2 Tabela com 16 Atributos 70
A.3 59 Atributos Iniciais e Informações Complementares 71
A.3 59 Atributos Iniciais e Informações Complementares 72
A.4 Tabela com Agrupamentos Ponderados 73
A.4 Tabela com Agrupamentos Ponderados 74

Acrônimos

1. **ARS** – *Análise de Redes Sociais.*
2. **CA** – *Comitê Assessor.*
3. **CA-CC** – *Comitê Assessor da Ciência da Computação.*
4. **CNPq** – *Conselho Nacional de Desenvolvimento Científico e Tecnológico.*
5. **CAPES** – *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior.*
6. **DTD** – *Document Type Definition.*
7. **XML** – *eXtensible Markup Language.*
8. **KDD** – *Knowledge-Discovery in Databases.*
9. **KNN** – *k-Nearest Neighbors.*
10. **MCTI** – *Ministério da Ciência, Tecnologia e Inovação.*
11. **PQ** – *Bolsa de Produtividade em Pesquisa.*
12. **SCI** – *Science Citation Index.*
13. **SNA** – *Social Network Analysis.*

Sumário

Agradecimentos **vi**

Resumo **vii**

Abstract **viii**

Lista de Figuras **ix**

Lista de Tabelas **x**

1 Introdução **14**
1.1 Caracterização do Problema e Objetivos 17
1.2 Metodologia 18
1.3 Contribuições 18
1.4 Dificuldades e Questões em Aberto 19
1.5 Organização da Dissertação 20

2 Fundamentação Teórica **21**
2.1 Bolsa de Produtividade em Pesquisa 21
2.2 Algoritmos em Mineração de Dados 23
2.3 Trabalhos Relacionados 27

3 Extração dos Atributos **30**
3.1 Criação do Banco de Dados com Pesquisadores 30
3.2 Análise Inicial dos Atributos 32
3.2.1 Agrupamentos Ponderados 39
3.3 Redução por Baixa Representatividade 39
3.4 Ponderação de Atributos 40
3.5 Redução por Correlação de Variáveis 41
3.6 Análises dos 15 Atributos Finais 41

4 Construção e Ajuste do Modelo **45**
4.1 Seleção da Amostra 45
4.2 Definição dos Modelos e Ajustes dos Parâmetros 46
4.2.1 Ajuste do *k-Nearest Neighbors* 48
4.2.2 Ajuste do *Random Forest* 49
4.2.3 Ajuste do *Naive Bayes* 50
4.2.4 Ajuste do *Support Vector Machines* 50
4.2.5 Parâmetros Ajustados para os Modelos 51
4.3 Comparação dos Modelos 51
4.3.1 Discussão 53
4.4 Análise Detalhada do Modelo 55
4.5 O Modelo de Predição 57
4.6 Ajuste do Modelo Preditivo 59
4.7 Considerações Finais 62

5 Conclusões e Trabalhos Futuros **64**
5.1 Trabalhos Futuros 64

Referências Bibliográficas **66**

Apêndice A Tabelas com Atributos **69**
A.1 Lista de 15 atributos Finais Utilizados no Modelo 69
A.2 Lista de 16 Atributos 70
A.3 Lista com os 59 Atributos Analisados Inicialmente 71
A.4 Agrupamento Ponderado 73

Capítulo 1

Introdução

A avaliação da capacidade produtiva dos pesquisadores é um fator importante para gestão de processos e alocação de recursos financeiros. Atualmente, existe um grande interesse em identificar o potencial de pesquisadores através de avaliações do desempenho científico, uma vez que esse tipo de avaliação oferece suporte ao recrutamento de pesquisadores em instituições de ensino superior e ao processo de concessão de financiamentos por órgãos de fomento, além de verificar e direcionar o desempenho da produção científica de pesquisadores [Abbasi et al., 2011].

Diante dos desafios na construção de indicadores efetivos para avaliação da performance de pesquisa e predição da capacidade produtiva científica, surge a possibilidade de estudos para a criação de modelos computacionais voltados para avaliação e comparação do desempenho científico, visto que identificar cientistas em potencial se torna uma tarefa relativamente complexa, mas com resultados e aplicações relevantes na gestão da produtividade científica.

Assim, espera-se que os critérios dessas avaliações, normalmente relacionados com indicadores científicos, sejam corretos e consistentes, uma vez que servem como parâmetros para o desenvolvimento da ciência, tecnologia e inovação no Brasil. Mugnaini et al. [2004] afirma que as atividades relacionadas com a produção de indicadores quantitativos vêm se fortalecendo no país na última década, com o reconhecimento da necessidade, por parte do estado e da comunidade científica, de dispor de ferramentas para auxiliar na avaliação de atividades relacionadas ao desenvolvimento científico e tecnológico no país .

Portanto, pesquisas voltadas para o aperfeiçoamento dos índices científicos brasileiros, contexto dessa dissertação de mestrado, tem uma relação direta com o crescimento da produtividade acadêmica nacional e, por consequência, impactam no desenvolvimento da ciência no Brasil.

Para possibilitar tais estudos, o conhecimento de disciplinas ligadas à Ciência da Informação se faz necessário [Borko, 1968]. Como é o caso da Cienciometria , que é definida como o estudo dos aspectos quantitativos da ciência enquanto disciplina ou atividade econômica, sendo um segmento da Sociologia da Ciência e aplicada no desenvolvimento de políticas científicas, bem como envolve estudos quantitativos das atividades científicas, incluindo as publicações. Da mesma forma a Infometria, que se dedica ao estudo dos aspectos quantitativos da informação em qualquer formato, referente a qualquer grupo social [Macias-Chapula, 1998].

Macias-Chapula [1998] afirma que a análise dos dados informétricos e cienciométricos oferece informações sobre a dinâmica científica de um país, bem como sobre sua participação na ciência e na tecnologia mundial. Isso mostra a importância da criação e aperfeiçoamento dos indicadores cienciométricos, índices científicos que quantificam as características produtivas como o fator de impacto das publicações no meio científico.

Alguns dos indicadores cienciométricos bastante utilizados pela comunidade acadêmica mundial são:

- ***Fator de impacto***, também chamado de SCI (*Science Citation Index*), que oferece uma forma de comparar a importância relativa dos artigos de um periódico em relação aos artigos de outros periódicos do mesmo campo [Mugnaini, 2006];

- ***Índice h:*** Do inglês *h-index*, utilizado para avaliar a produtividade e o impacto das pesquisas dos cientistas, que é determinado pelo número de artigos publicados com a quantidade de citações maiores ou iguais ao número de artigos [Thomaz et al., 2011];

- ***Índice g:*** Do inglês *g-index*, que consiste em um método proposto para suprir algumas deficiências do *h-index* que desconsidera um número excessivo de citações a um determinado artigo [Abbasi et al., 2011].

Esses indicadores apresentam um índice numérico baseado no histórico dos artigos publicados de cada pesquisador, o que permite uma possível comparação em relação a outros pesquisadores, mas não oferece um modelo para análise e predição de sua capacidade produtiva para tentar identificar qual o potencial de um determinado cientista em relação a outros cientistas. Outro aspecto importante é que, normalmente, essas métricas somente levam em consideração as publicações de artigos e suas citações e não avaliam outras atividades que são comuns na carreira acadêmica de pesquisadores no Brasil, como é o caso dos prêmios, trabalhos em eventos e orientações em trabalhos e pesquisas.

Grande parte dos índices avaliativos conhecidos avaliam o pesquisador através de indicadores globais que variam consideravelmente entre diferentes áreas de pesquisa, o que mostra a necessidade de avaliação de modelos específicos em cada área [De Lima, 2014]. De Lima [2014] afirma que as avaliações de performance devem ser feitas levando em consideração a área e subárea de pesquisa, visto que existem diferentes aspectos abordados em cada uma e propõe, em seu trabalho, um ranking interno em áreas como a ciência da computação, levando em consideração a subárea de pesquisa do pesquisador em análise.

Os trabalhos que relacionam a Cienciometria e índices de pesquisa no Brasil são voltados para uma análise de grupos de pesquisas ou instituições, mas não fornecem direcionamento para um único pesquisador. Garcia [2015], em seu trabalho de mestrado, fez uma análise da produção científica brasileira sob vários aspectos e identificou várias formas de colaboração entre países da América Latina, mas não entrou em detalhes para viabilizar um modelo com foco no pesquisador, fato que poderia direcionar o cientista em quais atributos de sua carreira ele pode investir para obter mais resultados.

Mazloumian et al. [2011] indicam, no entanto, que não há um consenso a respeito da melhor forma de se identificar a capacidade de um pesquisador em relação à média daqueles de sua área de pesquisa, ao mesmo tempo em que é possível identificar novos talentos a partir de picos de citações advindas de suas publicações. Isso implica que, somente quantificar os valores produtivos de cada pesquisador não é suficiente, outras informações podem ser necessárias para essa avaliação, o que dificulta o processo de predição.

A criação de modelos preditivos somente com informações de sistemas brasileiros é importante para um gerenciamento interno da inovação. A plataforma Lattes[1], sistema de integração em uma base dados das informações de pesquisa e currículos dos pesquisadores, é uma ferramenta reconhecida nacionalmente e serve como base de parâmetro para o fomento em pesquisas nacionais [CNPq, 2016]. Mena-Chalco & Júnior [2013] afirmam que a plataforma Lattes é uma ferramenta de padrão nacional entre os pesquisadores e foi utilizada como fonte de dados no projeto ScripLattes[2], sistema para auxílio na análise de produção científica brasileira que mostra dados estatísticos sobre a produtividade dos pesquisadores cadastrados na plataforma Lattes. [Mena-Chalco & Cesar-Jr, 2016].

Diante do exposto, esta pesquisa propõe a utilização de conhecimentos da Ciência da Informação, Ciência da Computação e da Modelagem Computacional para desenvolver, por meio da mineração de dados, um modelo preditivo para avaliação do potencial

[1]Plataforma Lattes: `http://lattes.cnpq.br/`
[2]ScripLattes: `http://scriptlattes.sourceforge.net/`

de pesquisadores na grande área Ciência da Computação que estão cadastrados na plataforma Lattes.

Com isso, o presente trabalho tem como principal contribuição uma metodologia para direcionar os pesquisadores brasileiros no sentido de identificar em quais aspectos o mesmo pode evoluir, impactando na sua carreira produtiva e, consequentemente, na produção do seu grupo de pesquisa. Além disso, apresenta-se como ferramenta útil na identificação de pesquisadores com potencial para as instituições com programas de pós-graduação, bem como para o direcionamento de recursos por parte das agências de fomento.

1.1 Caracterização do Problema e Objetivos

Com o tema do trabalho proposto, o problema geral envolvido com esse projeto está relacionado aos desafios na avaliação da produtividade científica de pesquisadores brasileiros através de informações disponíveis em sistemas nacionais.

Sendo assim, surgem algumas questões que procuram ser respondidas pelo presente estudo, sendo elas: (a) Quais as características a serem extraídas dos currículos da plataforma Lattes que viabilizem a identificação de desempenho de produtividade de pesquisadores em Ciência da Copmutação? (b) Como é possível predizer o potencial de trabalho de um pesquisador em uma etapa futura de sua carreira tomando os dados fornecidos por ele na plataforma Lattes? (c) Somente os dados constantes na base de currículos Lattes são suficientes para fornecer um bom indicador do perfil do pesquisador?

Norteado por tais questões, o objetivo geral deste trabalho consiste no desenvolvimento de um modelo para predição de desempenho de pesquisadores cadastrados na base de currículos Lattes na grande área de conhecimento Ciência da Computação.

Por conseguinte, os objetivos específicos do trabalho foram:

- Identificar as informações disponíveis que são relevantes, a partir da base Lattes, que caracterizem o potencial de produção científica individual de pesquisadores em Ciência da Computação;
- Analisar a necessidade de informações disponíveis em outros sistemas, além da plataforma Lattes, que auxiliem na avaliação do potencial dos pesquisadores;
- Construir um modelo preditivo para avaliar o potencial produtivo de pesquisadores em diferentes estágios da carreira;

- Ajustar do modelo preditivo desenvolvido para comparação de pesquisadores em diferentes realidades produtivas no Brasil.

1.2 Metodologia

Para conseguir alcançar os objetivos apresentados e desenvolver o modelo preditivo proposto nesta pesquisa, o trabalho foi dividido em quatro etapas. A primeira parte do projeto foi uma revisão bibliográfica de fontes de informações relacionadas à Cienciometria e à extração de conhecimento de bases de dados, da língua inglesa, *Knowledge-Discovery in Databases* (KDD).

A segunda parte do trabalho consistiu na obtenção dos dados, em formatação *eXtensible Markup Language* (XML), de currículos cadastrados na plataforma Lattes de pesquisadores com destaque produtivo, estudo da sua estrutura com o auxílio do *Document Type Definition* (DTD), seleção de atributos com potencial de caracterização de perfis de pesquisadores e extração dos dados avaliados como relevantes para análises.

Na terceira etapa, a massa de dados extraída foi explorada com técnicas de KDD, para seleção de atributos relevantes que são considerados no modelo. Nesta fase, os resultados obtidos foram avaliados e comparados para direcionar os principais atributos contidos na base que determinam a produção científica de pesquisadores, com intuito de fornecer dados relevantes para o desenvolvimento de um modelo preditivo de potencial.

Por fim, na quarta etapa, foram realizados testes em diferentes tipos de modelos e um deles foi selecionado como resultado do trabalho. Além disso, o modelo final foi aplicado a perfis de pesquisadores que se enquadrem em diferentes classificações, para comparação e análise dos resultados obtidos. Adicionalmente, utilizando somente os atributos selecionados, foram efetuadas verificações nos perfis de pesquisadores a fim de testar a efetividade da predição proposta.

1.3 Contribuições

As principais contribuições desta dissertação são:

- Desenvolvimento de uma metodologia computacional em mineração de dados para auxiliar programas de pós-graduação no sentido de identificar o potencial de pesquisadores;

- Desenvolvimento de metodologias para análise da base de currículos Lattes;

- Auxiliar na formação profissional do pesquisador a fim de permitir a identificação de pontos de melhoria em seu perfil e o direcionamento da sua atuação;
- Ferramenta de suporte para agências de fomento de pesquisa na identificação de equipes com maior potencial de avanço científico para aplicação de recursos;
- Algoritmos que auxiliam na identificação e seleção de pesquisadores para autoria em revistas científicas.

1.4 Dificuldades e Questões em Aberto

Apesar de conseguir gerar resultados satisfatórios, diversas dificuldades foram encontradas no desenvolvimento do trabalho, além de deixar alguns pontos em aberto que não eram objetivos do trabalho, mas que podem gerar bons resultados para a pesquisa como um todo.

A primeira e maior dificuldade encontrada no desenvolvimento do projeto foi extrair os arquivos da plataforma Lattes com os dados dos pesquisadore. Foram extraídos manualmente mais de 700 arquivos do sistema devido um bloqueio de acessos automáticos (*CAPTCHA*), implantado pouco tempo antes do início da pesquisa, e que é solicitado duas vezes para cada arquivo extraído, o que impactou fortemente na pesquisa.

Outro problema encontrado durante a pesquisa foi identificar quais pesquisadores seriam a referência da análise, sendo utilizados pesquisadores com bolsas de financiamento em pesquisa. O grande impacto na pesquisa foi gerado pelo fato que essa lista de pesquisadores não é oferecida em uma base estruturada pela instituição que oferece a bolsa. O que inicialmente seria feito por consultas manuais, foi evitado devido a existência de uma base de dados abertos com todos os pagamentos realizados, o que permitiu identificar os pagamentos de bolsas de produtividade de pesquisa e, por consequência, listar todos os pesquisadores que seriam utilizados como referência na área Ciência da Computação.

O principal ponto em aberto da dissertação está relacionado em identificar novos atributos, mesmo que a partir de outros sistemas ou metodologias, para adicionar ao modelo preditivo. Além disso, é possível fazer novos experimentos, como o vinculado ao fato que foram poucos os casos de pesquisadores reduziram de nível nas bolsas de financiamento. Possivelmente o fato de um pesquisador ter sido financiado anteriormente, interfere no processo de avaliação de novas bolsas.

1.5 Organização da Dissertação

Basicamente, essa pesquisa pode ser dividida em duas etapas. Sendo a primeira a geração da base de dados e extração dos atributos e a segunda a criação e validação do modelo preditivo. Para uma documentação didática dessas etapas, essa dissertação conta com mais quatro capítulos. No capítulo 2 é feita uma fundamentação teórica onde são apresentados trabalhos relacionados e fundamentos necessários para uma compreensão da pesquisa como um todo. No capítulo 3 é apresentado todo processo para escolha dos atributos necessários para uma comparação entre pesquisadores e quais critérios adotados nos algoritmos utilizados. No capítulo 4 é apresentada uma proposta de modelo para predição do potencial através de algoritmos de classificação e como é possível utilizar os conhecimentos da pesquisa para avaliação de pesquisadores na Ciência da Computação. Finalmente, no capítulo 5 são feitas as conclusões do trabalho e apresentados todos os ganhos obtidos, além da proposta de trabalhos futuros que permitam uma evolução da pesquisa em questão.

Capítulo 2

Fundamentação Teórica

Nesse capítulo é feita uma apresentação de conteúdos fundamentais para uma boa compreensão do trabalho, além de alguns trabalhos relacionados ao projeto que auxiliaram no direcionamento do mesmo.

2.1 Bolsa de Produtividade em Pesquisa

O Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), agência do Ministério da Ciência, Tecnologia e Inovação (MCTI), tem como principais atribuições fomentar a pesquisa científica e tecnológica, além de incentivar a formação de pesquisadores brasileiros visando contribuir para o desenvolvimento nacional [CNPq, 2016]. Dentre as possibilidades de financiamento via CNPq, a Bolsa de Produtividade em Pesquisa (PQ), atualmente é atribuída para pesquisadores doutores e possui um foco voltado para a qualidade das publicações do cientista, o que permite, com o auxílio da classificação utilizada nessa bolsa, entender o processo de avaliação de pesquisadores no Brasil [Wainer & Vieira, 2013].

As bolsas de produtividade e pesquisa são requisitadas pelos pesquisadores de alta produtividade no Brasil, mas somente quando existe a disponibilidade do recurso, são destinadas aos doutores que atendem aos critérios exigidos pelo respectivo Comitê Assessor (CA) de sua área de pesquisa. Além disso, as bolsas PQs são organizadas em ordem crescente de importância: PQs 2, 1D, 1C, 1B e 1A, ou seja, apesar de todos os níveis ter grande produção científica em relação aos demais cientistas, os pesquisadores classificados como PQ 1A são considerados os mais produtivos e os PQs 2 são os menos produtivos dentre os doutores classificados.

No caso da Ciência da Computação, o Comitê Assessor da Ciência da Computação (CA-CC), assim como outras grandes áreas, possui um critério de julgamento

próprio dos seus pesquisadores [1]. Nesse critério de avaliação existem itens necessários e classificatórios que consistem em variáveis quantitativas e qualitativas que são considerados para classificação do candidato à bolsa e pode ser observado de forma resumida na Figura 2.1.

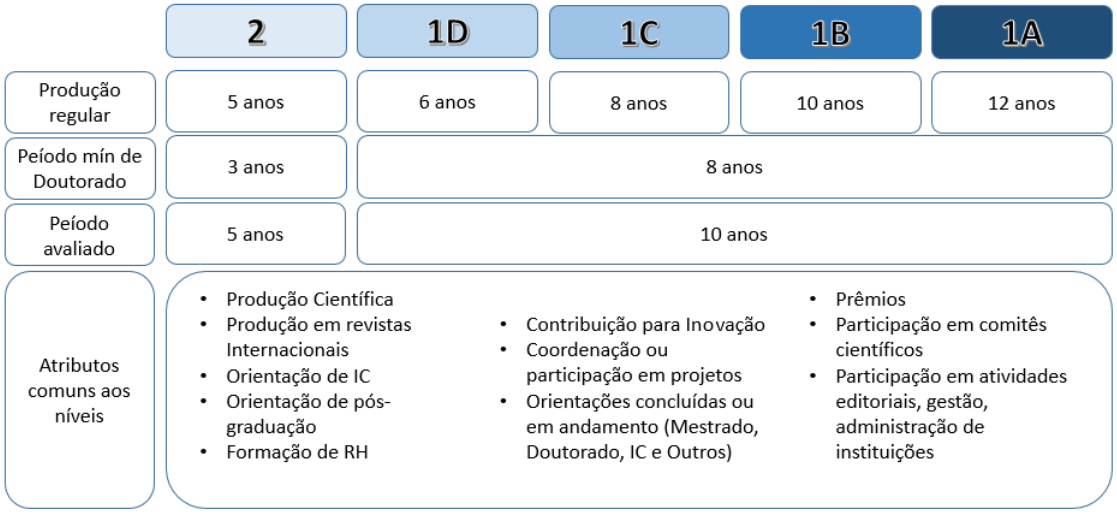

Figura 2.1: Critérios avaliados pelo CA-CC

É possível observar que o único atributo com exigência mínima é o tempo de doutorado, 3 anos para iniciar a bolsa como PQ 2 e, posteriormente, 8 anos para iniciar na categoria como PQ 1. Os demais atributos avaliados são comparados e analisados para direcionar a classificação do pesquisador dentro de cada nível.

O fato de um pesquisador atender todos os critérios exigidos para conseguir uma bolsa de produtividade, ou entrar em outra categoria caso já seja um pesquisador PQ, não implica no fato de alcançar a solicitação, é preciso existir a vaga para o nível do pesquisador em questão. Caso não exista a vaga, o cientista é alocado em um nível inferior ou não recebe a bolsa solicitada.

Para avaliar as solicitações de bolsa feitas pelos pesquisadores em destaque e classificá-las de acordo com o critério de julgamento, o CA-CC analisa os dados desses cientistas cadastrados na plataforma Lattes, ferramenta gerenciada pelo CNPq que tem como objetivo o cadastro das atividades relacionadas à carreira dos pesquisadores em geral no Brasil.

Ao avaliar os estudos cienciométricos existentes, é possível observar que em uma análise sobre o potencial de um determinado pesquisador, inicialmente é preciso avaliar alguns pesquisadores que são considerados como destaque entre os pares de pesquisa

[1] Disponível em: `http://cnpq.br/web/guest/view/-/journal_content/56_INSTANCE_0oED/10157/49290`

e os critérios de classificação usados para referenciá-los como destaque, isso permite obter um parâmetro de comparação e assim iniciar os trabalhos em relação aos dados desses cientistas.

2.2 Algoritmos em Mineração de Dados

A extração de conhecimento em base de dados, do inglês knowledge discovery in databases (kDD), consiste na execução de vários processos computacionais e funcionais, dentre eles a mineração de dados, para obtenção de conheicmento a partir de bases de dados estruturadas. Essa técnica ganhou força com o crescimento das bases e vem sendo utilizada para extração de conheicmentos implícitos nos dados atualmente.

A mineração de dados consiste em tentar obter conhecimento através de estudos em bases de dados. O termo mineração é utilizado para fazer uma analogia ao processo de extração de minerais, que tenta obter esse mineral em pesquisas geológicas que, nesse caso está relacionado ao conhecimento [Han et al., 2011].

O processo para obter o conhecimento em banco de dados, ou aprendizado de máquina, existe através da utilização de algoritmos computacionais que, com auxílio de técnicas de mineração de dados, tratam a informação em bases de dados para obter conclusões sobre o conhecimento existente [Da Costa Côrtes et al., 2002]. Esses algoritmos são agrupados em etapas que seguem um processo cíclico pois, ao terminar as etapas da mineração, é comum reiniciar o ciclo para ajustar cada algoritmo e tentar melhorar os resultados encontrados, conforme é possível observar na Figura 2.2.

Han et al. [2011] afirmam que as etapas da mineração compreendem em: limpeza dos dados, integração dos dados, transformação dos dados, aplicação de mineração de dados, evolução e avaliação dos resultados e por fim, apresentação do conhecimento ou modelo desenvolvido.

É possível agrupar todas as etapas que antecedem a mineração de dados em um processo chamado de pré-processamento dos dados, ou seja, todos os ajustes realizados antes de aplicar os demais conceitos[Han et al., 2011].

Cada base de dados utilizada em processos de mineração de dados pode ser separada entre atributos, conjunto de informações que podem ser extraídas da base, e rótulos, variáveis de interesse na mineração que normalmente estão relacionadas com alguma classificação prévia pelo processo em análise [Witten & Frank, 2005]. Os rótulos, ou classes, consistem nas variáveis desejadas da mineração, no caso da base de pesquisadores, os rótulos consistem nas classificações de bolsa PQ de cada pesquisador.

Após o pré-processamento dos dados, são realizados os testes com os algoritmos

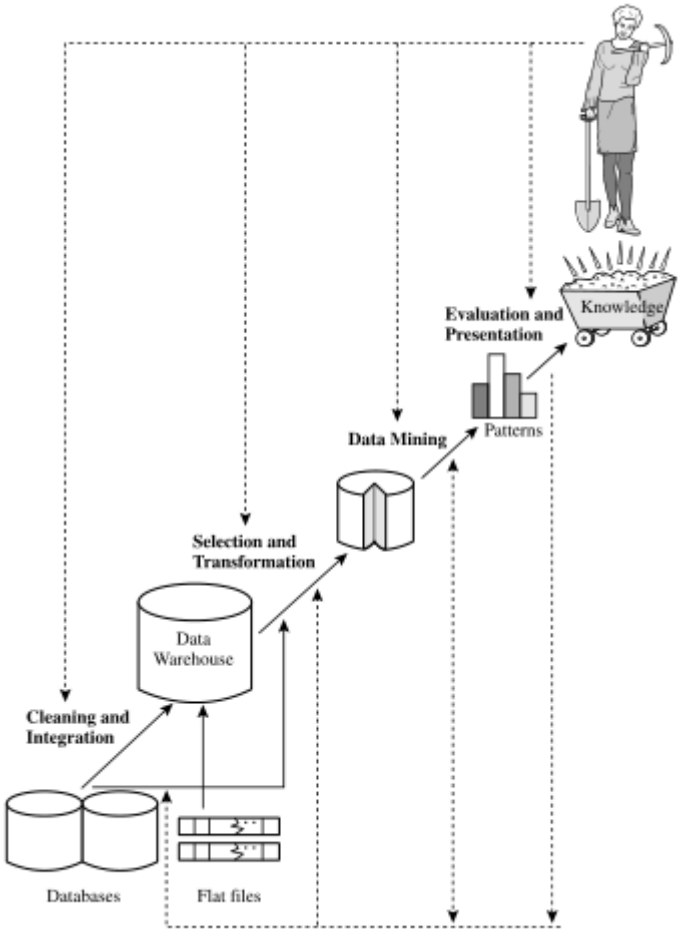

Figura 2.2: Processo de KDD

Fonte: [Han et al., 2011, p. 6]

de mineração de dados, sendo as técnicas mais conhecidas o agrupamento e a classificação. O agrupamento consiste em algoritmos que geram grupos nos dados e que não necessitam dos rótulos nos dados conhecidos previamente, chamados de algoritmos não supervisionados. A classificação, são algoritmos que classificam registros não conhecidos e, em sua maioria, utilizam rótulo dos dados previamente conhecidos como uma base de treinamento, chamados de algoritmos supervisionados.

Os algoritmos de agrupamento utilizam de técnicas computacionais para criar grupos entre os registros através de medidas de similaridade entre os dados da base, considerando a coesão interna entre o grupo e o isolamento externo com registros de outros grupos. O método mais comum de agrupamento é o *K-means* e suas variações, que utilizam medidas numéricas para criação de grupos [Witten & Frank, 2005].

O algoritmo de mineração de dados *K-means*, em português K-médias, separa

a base de dados em K grupos (K definido previamente) que são representados por centróides, ponto médio entre as distâncias de cada grupo, motivo do nome atribuído ao algoritmo. Normalmente, utiliza-se o critério de erro quadrado médio (equação 2.1) como medida de distância entre os pontos [Zaki & Meira Jr, 2014]. Em cada grupo de registro criado pelo algoritmo, é calculado um centróide C_i que representa o ponto médio entre todos os pontos presentes no grupo. O Erro médio quadrado consiste na diferença entre todas as dimensões (atributos da base) e o ponto médio elevado ao quadrado, conforme pode ser observado na equação 2.1.

$$E = \sum_{i=1}^{k} \sum_{p \epsilon C_i} |p - m_i|^2 \tag{2.1}$$

O processo de geração dos grupos acontecem através de iterações que iniciam em K grupos esolhidos aleatoriamente e, a cada iteração, os pontos da base são distribuídos entre os grupos existentes. O processo é repetido até que algum critério de parada seja alcançado, como por exêmplo não existir mais alterações entre os grupos criados. Já a escolha da quantidade de grupos a ser criada pode ser definida por conhecimento funcional (número de grupos já é definido pelo processo), método que foi utilizado nessa pesquisa, ou por testes e variação de vários valores de K, a quantidade de grupos que der resultados de coesão interna e menores erros médios são escolhidos.

O processo de classificação consiste em tarefas de predição de rótulos para registros que ainda não estão classificados, ou seja, baseado no histórico de acontecimentos em uma base de treinamento, o algoritmo tenta prever qual a classe, ainda desconhecida, de um novo registro [Harrington, 2012].

Os algoritmos de classificação utilizados no desenvolvimento deste trabalho foram:

- ***k-Nearest Neighbors***

  O Método dos K vizinhos mais próximos identifica quais são os k registros mais próximos através de alguma medida de distância (como exemplo a distância euclidiana na equação 2.2), e, baseado no rótulo dos vizinhos, o algoritmo classifica o registro desconhecido por maioria das classes, conforme a tendência da vizinhança [Harrington, 2012].

  A distância euclidiana é calculada através da raiz quadrada da soma de todas as distâncias entre os pontos, sendo a distância entre dois pontos a soma das diferenças entre todas as dimensões, ou atributos da base, conforme pode ser observado na equação 2.2.

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n}(x_{1i} - x_{2i})^2} \tag{2.2}$$

- ***Decision Tree***
  Os modelos em árvore de decisão, modelos mais utilizados na classificação em aprendizado de máquina, consistem na criação de uma estrutura em árvore que define uma sequência de atributos e uma função que define a classe baseado nos atributos utilizados na árvore [Flach, 2012]. Cada nó da árvore é um teste realizado e as arestas do nó uma possível resposta baseada no atributo do registo.

- ***Naive Bayes***
  Os chamados classificadores baeysianos, que utilizam o teorema de Naive Bayes como princípio para classificação realizada, avaliam a probabilidade de um evento ocorrer baseado na ocorrência de um outro evento. Para sua utilização, esses algoritmos utilizam o princípio da independência entre os atributos da base [Witten & Frank, 2005].

- ***Support Vector Machines***
  O algoritmo de máquina de vetores de suporte consiste em técnicas de classificação e análise de regressão com reconhecimento de padrões nos dados lineares e não lineares. O princípio do algoritmo é criar um espaço com dimensões maiores, e nesse espaço encontrar separações lineares ótimas nos dados [Han et al., 2011].

- ***Random Forest***
  O método *Random Forest* consiste em uma combinação de métodos, conhecido na mineração de dados como *ensemble*. Nesse caso, o algoritmo executa várias vezes, parâmetro de entrada do método, a técnica Decision Tree com atributos selecionados de forma aleatória, motivo pelo nome do algoritmo. O resultado da classificação será a classe com maior indicação pelas árvores da floresta [Flach, 2012].

Todos os algoritmos desenvolvidos nesse trabalho utilizaram a linguagem de programação Python (versão 2.7) [2], que além de ser disponibilizada gratuitamente, está dentre os sistemas mais populares do mundo para desenvolvimento [Harrington, 2012].

Para auxílio na implementação dos algoritmos utilizados, dentre várias bibliotecas públicas e disponíveis da linguagem, foi utilizada a biblioteca gratuita *Scikit Learn*

[2]Disponível em: `https://www.python.org/`

(sklearn [3]), desenvolvida para auxiliar na implementação de algoritmos de mineração de dados com várias ferramentas previamente desenvolvidas [Scikit-Learn-Org, 2016].

2.3 Trabalhos Relacionados

Diferentes trabalhos foram encontrados na literatura especializada a fim de analisar a produtividade de áreas de pesquisas no Brasil. Abbasi et al. [2011], em seu estudo, utilizou uma rede social profissional dos pesquisadores em Sistemas de Informação para prever o desempenho futuro dos mesmos. Propuseram um modelo teórico baseado em medidas de análise de redes sociais (*Social Network Analysis* – SNA) e métodos analíticos para explorar a colaboração (coautoria) em grupos de pesquisa. A investigação correlacionou positivamente com quatro medidas de SNA e chegou a resultados que sugerem o fato de que pesquisadores que trabalham em projetos com outros pesquisadores distintos têm um melhor desempenho, do que os pesquisadores com poucas conexões, baseando em indicadores cienciométricos (*g-index*).

A partir do estudo de Abbasi et al. [2011], Cimenler [2014] realizou um estudo semelhante ao avaliar o impacto de métricas em múltiplas redes sociais diante do conjunto de citações obtidas nas colaborações de pesquisas, mas usando um conjunto de dados diferente para mostrar que o desempenho de um pesquisador deve ser considerado em relação à posição em várias redes. Os autores chegaram a resultados semelhantes, exceto para uma das medidas de SNA, onde diferentemente do estudo anterior, acharam um impacto positivo sobre o desempenho dos pesquisadores que pesquisam com outros pesquisadores com bom desempenho produtivo. Além disso, esse segundo estudo demonstrou que atributos demográficos dos pesquisadores também devem ser considerados quando se investiga o impacto das métricas de redes sociais sobre o desempenho dos pesquisadores.

Alguns trabalhos que analisam a produtividade de grupos de pesquisa no Brasil também se relacionam com essa dissertação, inclusive com a predição da capacidade produtiva. Van Dijk et al. [2014] afirmam que o sucesso acadêmico é previsível e, em seu trabalho, desenvolveram um algoritmo a partir de um modelo computacional para identificar o potencial de pesquisadores em liderar projetos de pesquisa segundo seu histórico acadêmico. A implementação utilizou os atributos número de publicações, o fator de impacto dos periódicos publicados, a relação entre citações recebidas e o número de artigos que receberam mais citações do que a média do periódico onde foi publicado. Tais medições aplicadas em uma base com mais de 25.000 cientistas

[3]Disponível em: `http://scikit-learn.org/`

cadastrados no sistema PubMed of Medicine National Institutes of Health [2015][4], permitiram predizer o potencial de pesquisadores em se tornarem líderes de pesquisas.

Digiampietri et al. [2014] apresentam um estudo que classifica programas de Pós-graduação em Ciência da Computação e gera uma lista ordenada das instituições pelo critério de avaliação utilizado no trabalho, um indicador desenvolvido pelos autores na pesquisa. Nesse trabalho, os autores obtém os dados dos programas através da plataforma Lattes em conjunto com sistemas externos como a classificação Qualis, citações no *Google Scholar*[5] e *Microsoft Academic Search*[6]. Para a análise das características das redes acadêmicas foram usadas métricas específicas de redes computacionais e os resultados foram comparados com indicadores cienciométricos atuais, o que comprovou que os programas com maior presença na topologia da rede de coautoria também apresentam maior produtividade em pesquisa no período de estudo (2004 - 2009). A metodologia utilizada com os dados do sistema Qualis serviu como base para criação de atributos de produtividade ponderada, utilizados nessa dissertação e detalhados no Capítulo 4.

De Lima [2014], em sua dissertação de mestrado, apresenta um estudo com uma análise comparativa entre pesquisadores da área Ciência da Computação no Brasil e suas diferentes subáreas, com intuito de avaliar diferentes perfis de produtividade dentro de uma mesma grande área. Com um direcionamento semelhante ao dessa dissertação, De Lima [2014] questiona índices internacionais utilizados tradicionalmente para medir a produtividade de pesquisa, o que pode apresentar resultados adequados em análises globais de pesquisadores, mas para análise específica do pesquisador ou de uma subárea, pode gerar resultados inadequados. Além disso, em seus resultados e contribuições, De Lima [2014] propõe um método de *ranking* de pesquisadores em diferentes subáreas e compara com a classificação realizada pelo comitê CA-CC do CNPq, onde obteve resultados satisfatórios e aproximados aos método do comitê.

Wanderley [2015], em sua dissertação de mestrado, apresentou uma pesquisa semelhante ao estudo realizado nessa dissertação. Em seu trabalho, o autor desenvolveu 10 modelos de classificação que avalia atributos em análise de redes sociais (ARS) e faz uma predição em relação aos pesquisadores relacionados com Ciência da Computação com intuito de classificar cientistas que recebem financiamento por bolsa de produtividade PQ e que não recebem a bolsa. Para isso, Wanderley [2015] buscou na plataforma Lattes todos os professores vinculados a departamentos de Ciência da Computação no Brasil e relacionou através de contribuições em pesquisa, que gerou uma estrutura em

[4]Pubmed:`http://www.ncbi.nlm.nih.gov/pubmed`
[5]Google Scholar: `http://scholar.google.com`
[6]Microsoft Academic Search: `http://academic.research.microsoft.com`

rede social de contribuição científica. Após desenvolvimento da estrutura de contribuição, foram gerados indicadores avaliativos da rede como Centralidade de Proximidade e Centralidade de Intermediação e, através desses atributos com indicadores de cada pesquisador, os modelos foram desenvolvidos. Ao concluir seu trabalho, o autor conseguiu obter um resultado em acurácia média de 74,53% na validação cruzada, considerado como resultado satisfatório. Pelo fato desse trabalho estar muito relacionado com esta dissertação, foram realizados ajustes que permitiram uma comparação direta entre os resultados obtidos em cada pesquisa, que podem ser analisados detalhadamente no Capítulo 4.

Capítulo 3

Extração dos Atributos

Neste capítulo foram detalhadas as etapas para escolha dos atributos utilizados no modelo preditivo, ou seja, são os critérios adotados para identificar as informações relevantes que caracterizem o potencial de produção científica individual de pesquisadores que trabalham na área Ciência da Computação.

A primeira etapa foi identificar quais pesquisadores seriam analisados na extração de atributos, visto que existe uma grande quantidade de cadastros de pesquisadores na plataforma Lattes e seria inviável a extração manual de todos os registros. Após a definição dos pesquisadores que fariam parte da base de dados em análise, foi realizado um levantamento inicial dos atributos disponíveis na plataforma Lattes e implantadas ações para redução e seleção de quais atributos são relevantes para a predição do potencial produtivo de pesquisadores.

Para uma visão sequencial das ações realizadas na base para a extração dos atributos, as ações realizadas consistem em um agrupamento ponderado, seguido por uma redução por baixa representatividade dos dados, que resultou em uma lista reduzida de atributos. Além disso, outros estudos foram realizados como continuidade do processo, foi feita uma análise nos atributos que restaram para avaliar a possibilidade de criação de outros atributos com novas fontes de dados e uma análise por matriz de correlação para avaliar uma possível redução, que resultou em uma relação final com informações relevantes que foram utilizadas no modelo preditivo.

3.1 Criação do Banco de Dados com Pesquisadores

Com intuito trabalhar com uma base de dados previamente classificada, optou-se por analisar pesquisadores que já receberam financiamento por bolsa de produtividade PQ na área de Ciência da Computação, pois eles já foram classificados pelo comitê

CA-CC na época que receberam a bolsa, são considerados como referência na área nacionalmente e possuem o Lattes devidamente atualizado por exigência do CNPq.

Como a plataforma Lattes não possibilita gerar um relatório com pesquisadores que possuem a bolsa PQ de forma automática, a lista de pesquisadores que foi utilizada nesse trabalho foi criada a partir da base de dados abertos do CNPq [1], informações históricas acerca dos pagamentos realizados pela agência aos pesquisadores, inclusive das bolsas de produtividade, e tem o objetivo de publicar todos os gastos da instituição. Nessa base é possível identificar qual pesquisador recebeu pagamento de bolsa PQ, o nível do pesquisador pelo critério do CA-CC e a data inicial do financiamento.

Após selecionar na base somente os pagamentos de bolsas PQ, foi possível identificar os pesquisadores que já receberam a bolsa de produtividade e qual a respectiva classificação na época, visto que existe na base a data inicial da bolsa. Com isso, foi possível criar uma base com informações dos pesquisadores analisados pelo comitê, pelo menos uma vez, entre 2004 e 2014. Essa base contempla os últimos 11 anos de avaliações realizadas, um total de 1227 registros que podem ser observados de forma detalhada na Tabela 3.1.

Tabela 3.1: Avaliações de Bolsa pelo Comitê CA-CC por Ano

Ano	***PQ 1A***	***PQ 1B***	***PQ 1C***	***PQ 1D***	***PQ 2***	***Total***
2004	5	6	7	13	30	61
2005	1	3	4	21	81	110
2006	2	5	5	11	23	46
2007	7	5	15	31	61	119
2008	4	2	6	18	86	116
2009	3	7	2	12	71	95
2010	10	11	25	27	103	176
2011	2	6	12	34	88	142
2012	8	7	5	15	65	100
2013	1	1	0	8	114	124
2014	5	11	15	28	79	138
Total	**48**	**64**	**96**	**218**	**801**	**1227**

Os arquivos foram extraídos da plataforma em dezembro de 2014 e, consequentemente, não existiam dados de 2015 disponíveis para análise, motivo pelo qual esse trabalho não utilizou dados desse período.

Ao analisar a base de dados, foi possível perceber duplicidades de *IDs* (código de identificação do pesquisador) em diferentes anos, ou seja, um mesmo pesquisador

[1] Disponível em: `http://cnpq.br/dados_abertos`

foi avaliado mais de uma vez dentro do período analisado, o que era esperado devido período de duração das bolsas ser inferior aos 11 anos avaliados.

Como as análises do comitê aconteceram em épocas distintas na carreira do pesquisador, o histórico da produção científica analisado pelo comitê para a classificação era diferente, decidiu-se então manter as duplicidades de *IDs*, pois indicam avaliações diferentes dos pesquisadores e uma das intenções do projeto é entender a análise realizada pelo comitê CA-CC, independente se foi para o mesmo pesquisador em diferentes momentos da carreira.

A próxima etapa foi buscar na plataforma Lattes os dados em XML de cada pesquisador da base, um total de 498 arquivos com informações dos pesquisadores para geração da base de dados utilizada na pesquisa.

As etapas para criação do banco de dados com informações dos pesquisadores pode ser observada na Figura 3.12

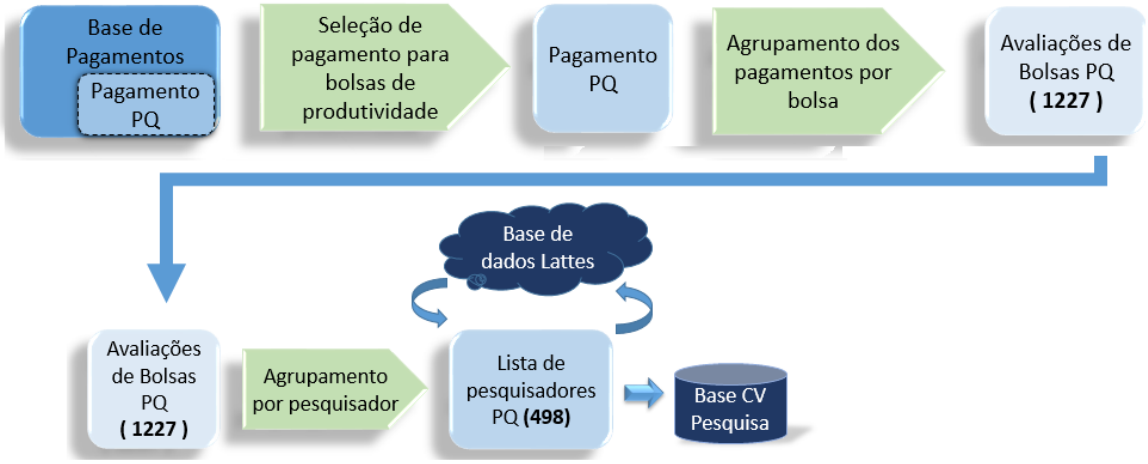

Figura 3.1: Detalhamento da criação da base de dados

3.2 Análise Inicial dos Atributos

Com os currículos extraídos da plataforma Lattes, o ponto de partida para análise dos atributos foi determinar quais informações são disponibilizadas na base Lattes através de um estudo detalhado do *Document Type Definition*[2] (DTD), documento que descreve as informações disponíveis no XML da plataforma que permitiu selecionar inicialmente 106 atributos disponíveis.

[2]Disponível em `http://lmpl.cnpq.br/lmpl/Gramaticas/Curriculo/DTD/Documentacao/DTDCurriculo.pdf`

A partir dessa primeira seleção e, após uma comparação com o relatório de critérios utilizados pelo CA-CC para bolsas de produtividade, foi realizada uma redução do número de atributos por meio de uma análise funcional, visto que alguns atributos não fazem sentido para pesquisadores em Ciência da Computação, como é o caso do atributo de partituras musicais. Dessa forma, retirou-se 47 atributos, o que resultou em uma base de 59 atributos para análise.

Para uma comparação adequada com os critérios do comitê CA-CC e possibilitar a identificação de quais atributos são relevantes, foram selecionados somente os pesquisadores que poderiam ser classificados em qualquer nível de bolsa PQ, ou seja, que se enquadram nos critérios mínimos necessários tanto para bolsa PQ 1 quanto para bolsa PQ 2. Um exemplo de critério que poderia interferir na análise dos atributos é o tempo de doutorado, pois um pesquisador com tempo de doutorado inferior a 8 anos estaria sempre classificado como PQ 2, mesmo com uma aulta produtividade científica.

Dessa forma, dos 1227 pesquisadores previamente listados (Tabela 3.1), foram mantidos somente os pesquisadores com tempo de doutorado igual ou superior a 8 anos em relação à época da avaliação do comitê, pois esses poderiam se enquadrar tanto em PQ1 quanto em PQ2. Este novo levantamento resultou em uma base com 915 pesquisadores avaliados pelo CNPq, conforme exposto na Tabela 3.2.

Tabela 3.2: Amostra da Base Completa para Análise dos Atributos

Ano	***PQ 1A***	***PQ 1B***	***PQ 1C***	***PQ 1D***	***PQ 2***	***Total***
2004	5	6	7	13	15	46
2005	1	3	4	20	44	72
2006	2	5	5	10	7	29
2007	7	5	15	27	34	88
2008	4	2	6	18	62	92
2009	3	7	2	12	30	54
2010	10	11	25	27	63	136
2011	2	6	12	34	63	117
2012	8	7	5	15	45	80
2013	1	1		8	75	85
2014	5	11	15	28	57	116
Total	**48**	**64**	**96**	**212**	**495**	**915**

As etapas para criação do banco de dados e o processo de seleção dos pesquisadores com o mesmo potencial de classificação pode ser observado na Figura 3.2

Com os registros selecionados, foi contabilizada a produtividade científica nos 59 atributos para os últimos 10 anos que precedem aos anos de análise do comitê, período

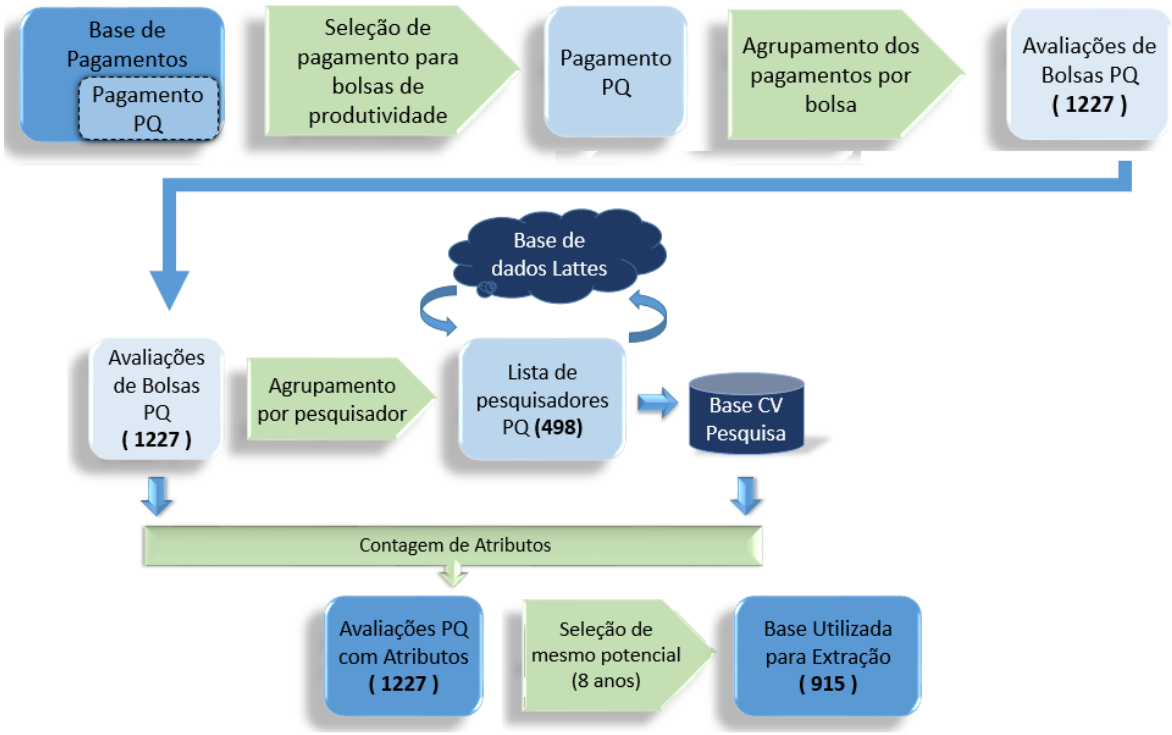

Figura 3.2: Detalhamento da criação da base de dados com mesmo potencial

avaliado para o nível mais elevado da bolsa (PQ 1A). Um exemplo de como os atributos foram contabilizados pode ser observado na Tabela 3.3, ou seja, um pesquisador que foi avaliado em 2014 (na Tabela 3.3 denominado como Pesquisador 1) teve o período produtivo avaliado entre 2005 e 2014.

Tabela 3.3: Exemplo de Contagem dos Atributos

Pesquisador	***Ano de Avaliação do CNPq***	***Período Contabilizado***
Pesquisador 1	2014	2005 - 2014
Pesquisador 2	2013	2004 - 2013
⋮	⋮	⋮
Pesquisador 4	2006	1997 - 2006
Pesquisador 5	2005	1996 - 2005

Foram realizados alguns testes e comparações com algoritmos de mineração de dados nos 915 registros com os 59 atributos para criação de um parâmetro inicial de análise e para possibilitar o acompanhamento da evolução no processo de redução quantitativa dos atributos. O resultado dessa primeira análise foi registrado em um gráfico em linhas paralelas na Figura 3.3, gráfico com a quantidade contabilizada de cada atri-

buto para os pesquisadores com bolsa PQ 1A e 1B x PQ 2, ou seja, pesquisadores em níveis extremos da classificação feita pelo CNPq.

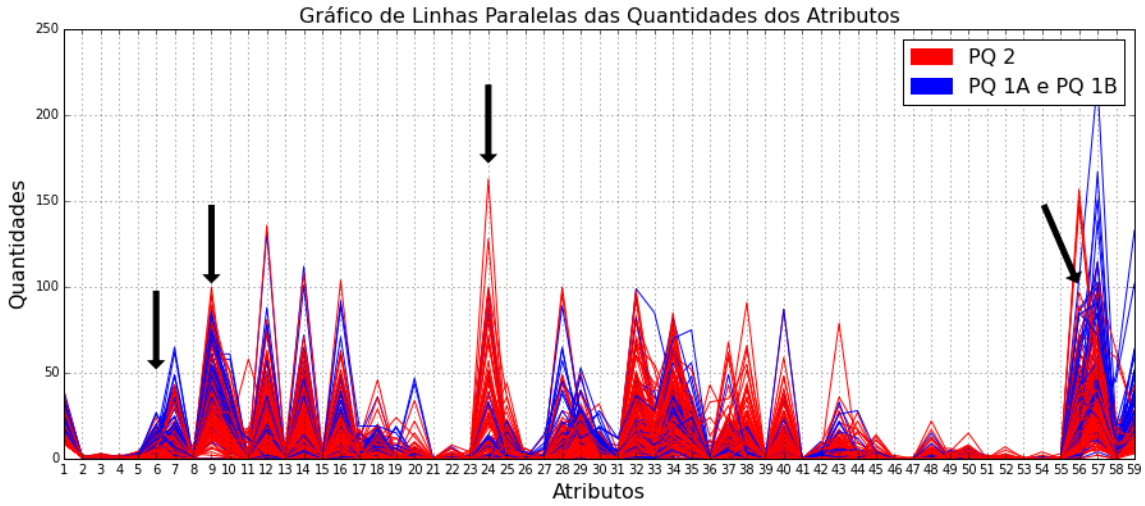

Figura 3.3: Avaliação dos atributos segundo perfil dos pesquisadores

No gráfico de linhas paralelas são criadas linhas que indicam a quantidade em cada atributo para cada registo, ou seja, é possível visualizar toda contabilização da base de dados em um único gráfico. Foram selecionados somente os níveis extremos da classificação feita pelo CNPq para uma melhor visualização das separações entre as linhas no gráfico, o que já indicaria possíveis separações entre os diferentes níveis de classificação.

É possível observar na Figura 3.3 uma tendência de separação em alguns atributos como 6, 9, 24 e 57, ou seja, são informações que possivelmente são relevantes e que podem auxiliar o modelo preditivo na classificação do potencial produtivo dos pesquisadores, uma vez que são os atributos que mais diferem entre os grupos analisados.

Para melhor visualização da diferença entre os atributos entre os grupos, foi gerado um novo gráfico em linhas paralelas com a média das quantidades para cada atributo, que pode ser visualizado na Figura 3.4. Com esse gráfico, foi possível perceber de uma forma mais clara a separação entre os grupos em alguns atributos.

Com o intuito de minimizar o impacto da discrepância quantitativa entre os diferentes atributos, como acontece entre os atributos 24 e 27, foi gerado um novo gráfico com a média normalizada dos atributos, que pode ser visualizado na Figura 3.5. Tal normalização, chamada de min-max, trata-se de um pré-tratamento que redistribui os dados entre 0 e 1, mas preserva a relação entre os dados originais e os dados normalizados [Han et al., 2011].

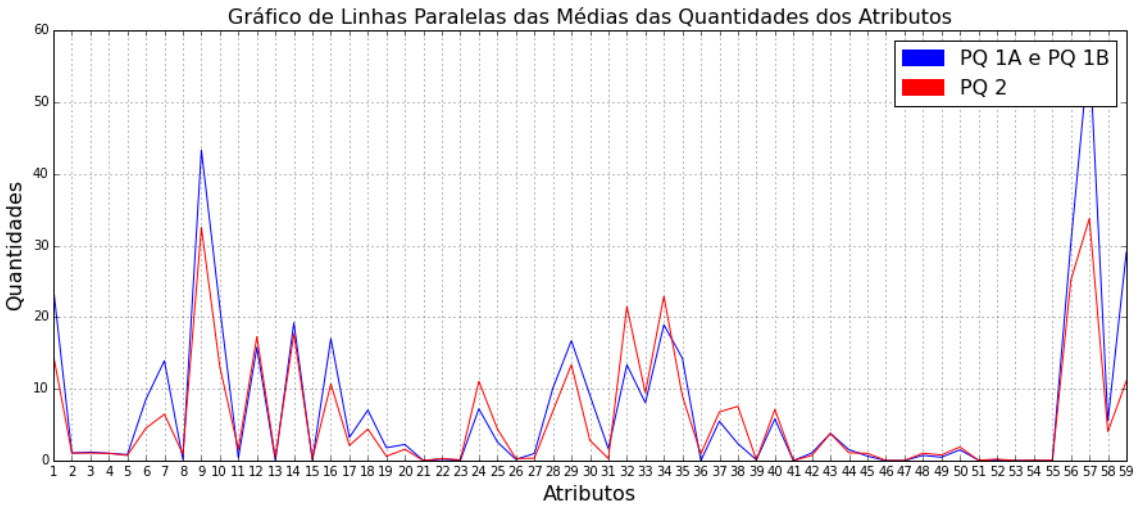

Figura 3.4: Avaliação da média dos atributos segundo perfil dos pesquisadores

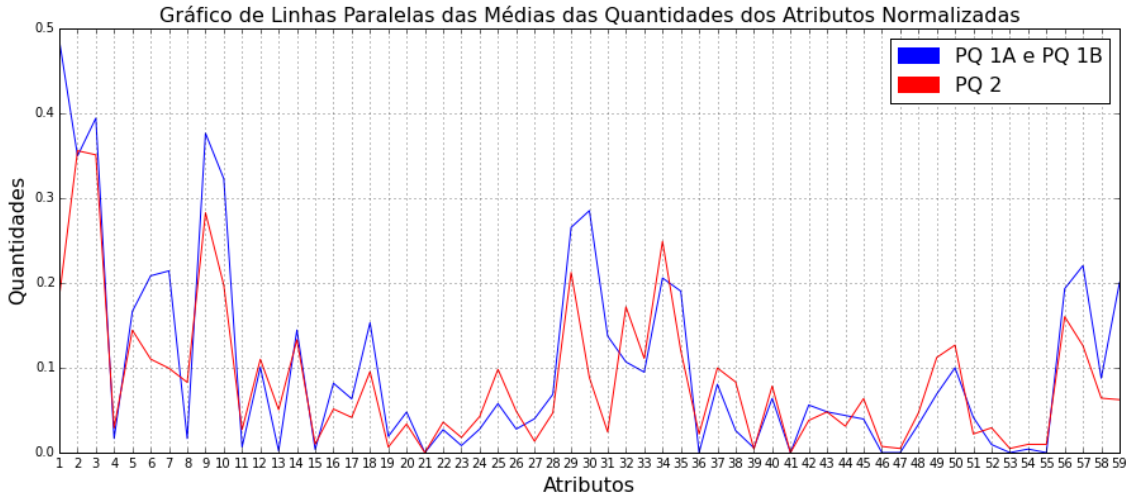

Figura 3.5: Avaliação da média normalizada dos atributos segundo perfil dos pesquisadores

Foi possível observar nos três gráficos que, somente com uma contagem inicial dos atributos selecionados na primeira análise, já existe uma tendência para identificar quais informações são relevantes, visto que os grupos analisados se diferem em alguns atributos, fato que indica a possível existência de grupos bem delimitados entre os pesquisadores.

Comparando os resultados apresentados nas Figuras 3.4 e 3.5, é possível observar que, após a normalização, a relação das quantidades em cada grupo de pesquisadores nos atributos se manteve. Entretanto, foi possível identificar novos atributos relevantes

que estavam mascarados pela discrepância quantitativa em relação a outros atributos que já se destacava mesmo antes da normalização, como é o caso do atributo 49, que na Figura 3.4 não indica uma tendência de separação e na Figura 3.5 indica.

Gonçalves [2000], em seu trabalho de mestrado, aplicou normalização nos dados para classificação de grupos de pesquisadores em diferentes áreas de concentração, o que auxiliou no processo de mineração desenvolvido em seu trabalho, e trouxe melhores resultados. Sendo assim, diante dos achados na literatura e corroborados pelo presente estudo, todos os dados dos próximos resultados gerados pelos mecanismos de mineração desenvolvidos nesse trabalho serão avaliados após a normalização dos dados.

Assim, dando continuidade ao processo de análise dos atributos extraídos até o momento e com o intuito de ter um parâmetro inicial da qualidade dos primeiros atributos selecionados, foi proposto um agrupamento dos pesquisadores através do algoritmo *K-means*. Optou-se por cinco grupos, semelhante a classificação feita pelo CNPq, sendo os resultados expostos na Tabela 3.4.

Tabela 3.4: Resultado do agrupamento com 5 grupos - 59 atributos

		Agrupamento				
		Grupo 1	**Grupo 2**	**Grupo 3**	**Grupo 4**	**Grupo 5**
CNPq	**PQ 1A**	8	5	1	**34**	0
	PQ 1B	24	7	4	26	3
	PQ 1C	32	6	18	31	9
	PQ 1D	48	17	70	42	35
	PQ 2	48	8	**288**	48	103

É possível observar uma tendência de polarização, ou seja, de agrupamentos dos pesquisadores em níveis extremos quando comparados à classificação do CNPq. Os grupos criados pelo algoritmo ao analisar os 59 atributos possuem grupos de pesquisadores com mesma classificação realizada pelo CNPq, o que significa que mesmo sem uma análise detalhada dos atributos é possível encontrar uma tendência de separação correta, mostrando que os testes e análises realizados estão em um direcionamento correto e satisfatório.

Para complementar a análise do agrupamento, foi gerado o gráfico em linhas paralelas dos grupos 3 e 4 (Figura 3.6) que se assemelham aos grupos PQ 1A e PQ 2 (Figura 3.5). Essa semelhança acontece devido predominância dessas classificações nos respectivos grupos, o grupo 3 possui uma predominância de pesquisadores PQ 2 e o grupo 4 uma predominância de pesquisadores PQ 1A.

Ao comparar os gráficos obtidos com a classificação feita pelo CNPq (Figura 3.5) e o gráfico dos grupos criados pelo algoritmo de agrupamento (Figura 3.6), é possível

observar uma semelhança das linhas paralelas e consequentemente no comportamento de alguns atributos, o que reforça o correto direcionamento da seleção de atributos para a construção do modelo.

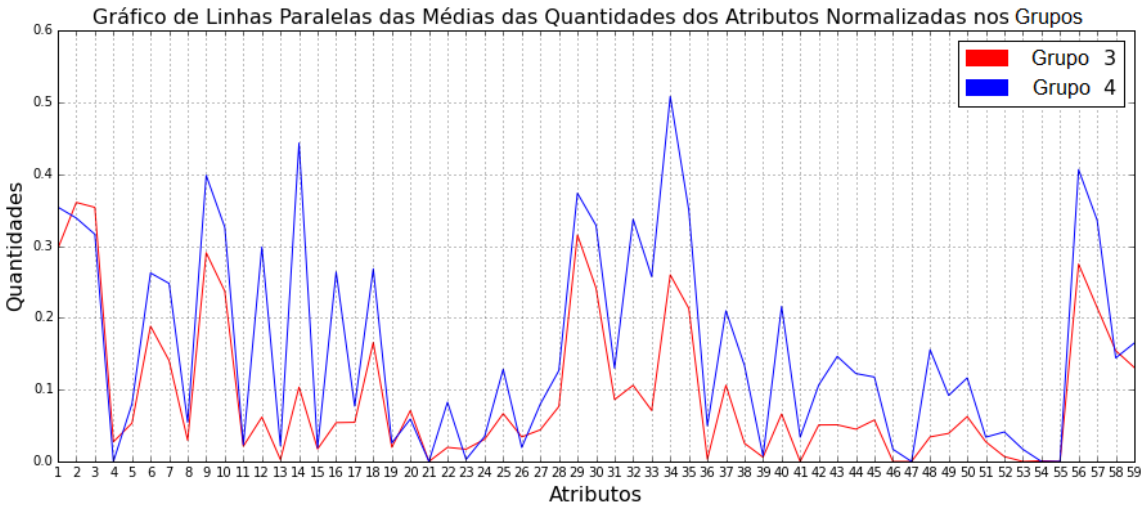

Figura 3.6: Avaliação da média normalizada dos atributos segundo grupos pesquisadores criados pelo algoritmo

Após os resultados iniciais, encontrados na análise inicial dos atributos, foram realizados testes mais específicos com intuito de analisar cada atributo e avaliar a necessidade de permanecer na base para criação do modelo preditivo. O fluxo com os testes realizados pode ser observado na Figura 3.7

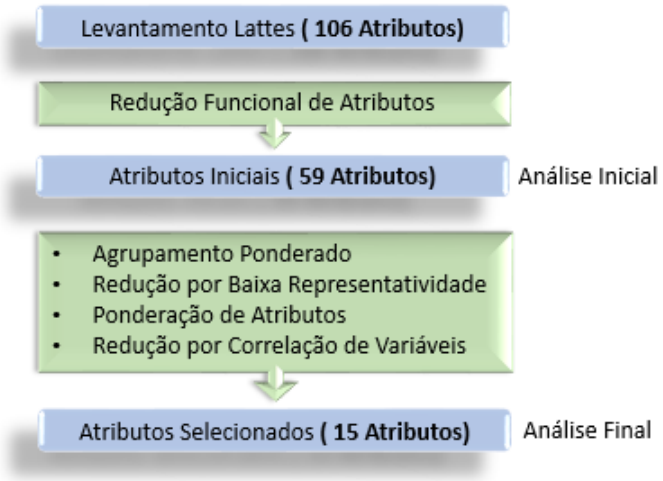

Figura 3.7: Fluxo de atividades para seleção dos atributos

3.2.1 Agrupamentos Ponderados

Alguns dos 59 atributos selecionados correspondem ao desdobramento ou detalhamento de outros atributos, como é o caso de Orientações Concluídas, que tem variações como Orientação de Iniciação Científica e Orientação de Mestrado. Sendo assim, foram criados alguns agrupamentos ponderados dos atributos. Esse tipo de agrupamento evita a utilização de vários atributos contemplando a mesma atividade científica. A ponderação atribui pesos aos atributos considerando a importância e complexidade de cada função. Por exemplo, orientação de uma iniciação científica é mais simples que a orientação de um pós-doutorado. Nesse caso atribui-se um maior fator ao segundo atributo.

Foram realizados alguns testes com diferentes ponderações para os grupos de atributos e foram analisados os resultados para cada grupo, semelhante ao teste realizado e apresentado na Tabela 3.1. Os agrupamentos finais dos atributos, que obtiveram os melhores resultados, podem ser observados de forma detalhada com suas respectivas ponderações na tabela do apêndice A.4, sendo: Participações em Projeto Ponderado (do atributo 12 ao 16), Orientações Concluídas (do atributo 29 ao 33), Participações em Bancas (do atributo 34 ao 39), Participações em Eventos (do atributo 40 ao 48) e Orientações em Andamento (do atributo 49 ao 55).

3.3 Redução por Baixa Representatividade

Na mineração de dados, técnicas para redução do conjunto de atributos podem ser aplicadas para obter uma representação reduzida da base de dados analisada com resultados possivelmente semelhantes, ou melhores, que os obtidos com a base completa. Uma das estratégias utilizadas é a exclusão de atributos irrelevantes ou de relevância fraca, visto que não interferem fortemente no processo de mineração, e podem ser avaliados por medidas estatísticas como média, mediana e variância [Han et al., 2011].

Com isso, a redução de atributos por baixa variância e representatividade de valores foi realizada a partir de uma análise de medidas estatísticas para cada atributo, listada na tabela do apêndice A.3. Para uma melhor visualização e escolha de quais atributos não tem representatividade, foi criado um gráfico *boxplot* que apresenta a distribuição de cada atributo normalizado, conforme apresentado na Figura 3.8.

Alguns atributos da base de dados possuem medianas e médias baixas, próximas ou iguais a zero, o que indica uma baixa utilização pelos pesquisadores em geral dessas informações e, consequentemente, permite a exclusão do atributo na base de dados.

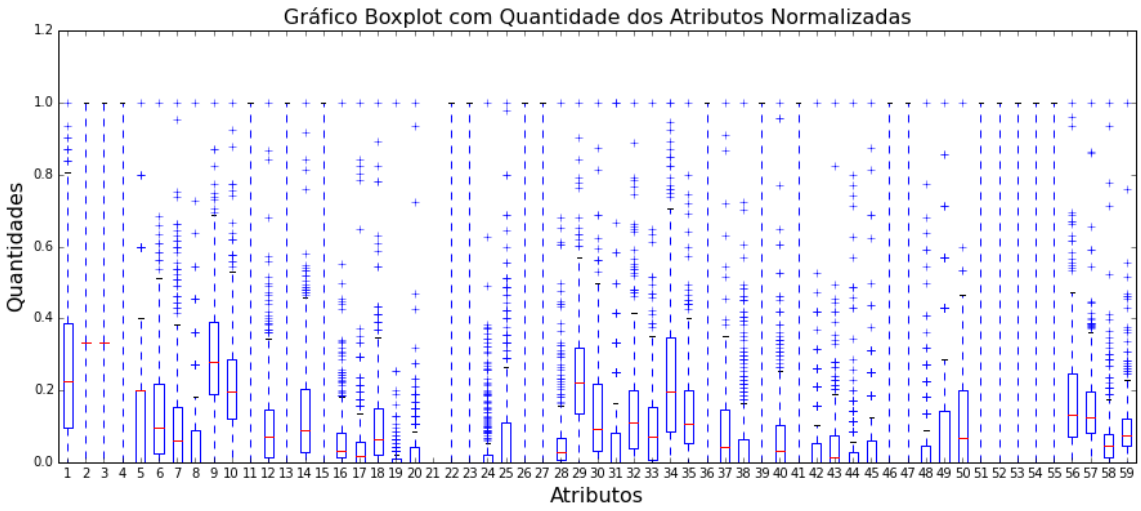

Figura 3.8: Análise estatística da base por atributo

3.4 Ponderação de Atributos

Atributos relacionados com a produção de artigos científicos e apresentação de trabalhos em eventos devem ser avaliados com cautela quando o objetivo do trabalho é avaliar a produtividade do pesquisador. Analisar somente a quantidade de publicações não é suficiente, sendo também necessário avaliar a qualidade das publicações [De Lima, 2014].

Assim, os atributos que contém informações com a quantidade de artigos publicados nacionais e internacionais foram substituídos por atributos ponderados que levaram em consideração a qualidade da publicação, para isso foram criados novos atributos que levaram em consideração a qualidade da publicação do pesquisador.

Para tanto, foi considerada a metodologia utilizada por Digiampietri et al. [2014], que utilizaram informações de outros sistemas para criar critérios de ponderação. Foi criada uma ponderação proporcional à qualificação Qualis do periódico publicado, sendo: A1=100, A2=85, B1=70, B2=50, B3=20, B4=10, B5=5 e C=1, mesma ponderação proposta por Digiampietri et al. [2014], diferindo apenas quanto ao fato de considerar peso 1 em vez de zero para publicações em revistas com Qualis C e em revistas que não estão na base Qualis.

Após a redução funcional, agrupamentos de atributos e redução por baixa representatividade, os 59 atributos foram reduzidos para 16, como apresentado na tabela do apêndice A.2.

3.5 Redução por Correlação de Variáveis

Os 16 atributos resultantes foram utilizados em uma análise de seleção de atributos por correlação de variáveis, método que avalia todos os atributos através de uma matriz de correlação. Esse método visa identificar atributos que transmitem a mesma informação ao modelo e podem ser representados por apenas um atributo. O resultado da matriz de correlação pode ser observado na Figura 3.9.

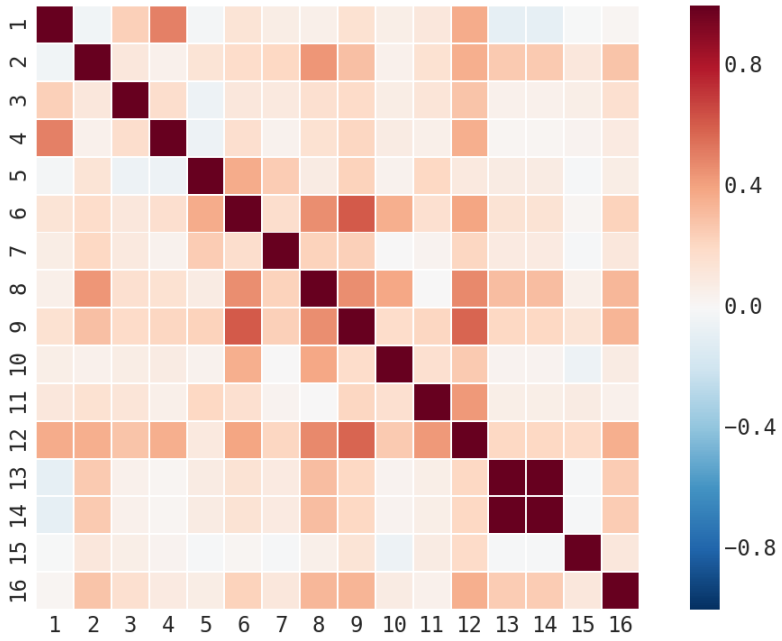

Figura 3.9: Matriz de correlação dos 16 atributos normalizados

Pelo gráfico da Figura 3.9, é possível observar que existe uma alta correlação positiva entre o atributo de **participações em bancas** (atributo 13) e **participações em eventos** (atributo 14) que podem ser representados por um dos dois atributos.

Dessa forma, somente o atributo de participação em bancas foi mantido, devido a relevância de participações em bancas ser superior em detrimentos a participação em eventos, o que resulta em uma lista de 15 atributos conforme consta na tabela do apêndice A.1.

3.6 Análises dos 15 Atributos Finais

Após os ajustes realizados nos atributos, que resultou em uma lista com 15 atributos, foi realizado um novamente o agrupamento dos pesquisadores em cinco grupos através

do algoritmo *K-means*, que está representado na Tabela 3.5 em uma comparação dos grupos criados pelo algoritmo e a classificação realizada pelo CNPq.

Como é possível observar, após a redução de atributos, houve uma melhora no agrupamento em relação ao agrupamento inicial com 59 atributos (Tabela 3.4), uma vez que aumentou a tendência de polarização nos grupos.

Tabela 3.5: Resultado de agrupamento com 5 Grupos - 15 atributos

		Agrupamento				
		Grupo 1	**Grupo 2**	**Grupo 3**	**Grupo 4**	**Grupo 5**
CNPq	**PQ 1A**	8	**32**	5	1	2
	PQ 1B	12	25	6	13	8
	PQ 1C	9	25	10	30	22
	PQ 1D	11	44	12	67	78
	PQ 2	15	58	6	93	**323**

Outra forma de analisar o agrupamento proposto com os 15 atributos é através do gráfico de linhas paralelas para os principais grupos, no caso desse último agrupamento os grupos 2 e 5, que pode ser observado na Figura 3.10. Nesse gráfico, é possível observar que, com os 15 atributos, a separação entre os valores médios normalizados no gráfico é maior que com 59 atributos, o que caracteriza grupos mais distintos, possibilitando resultados melhores para o modelo que será criado.

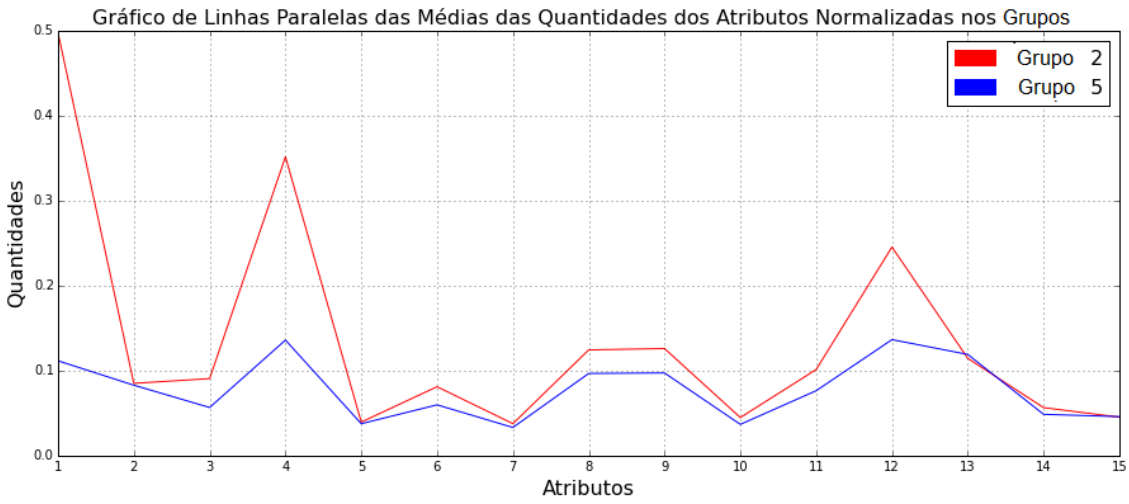

Figura 3.10: Avaliação da média normalizada dos atributos segundo grupos com 15 atributos

Outras informações importantes extraídas dessa análise é que alguns atributos são mais separáveis do que outros, como o **tempo de conclusão do doutorado**, e

que alguns atributos apresentam maior representatividade no grupo 2 que no grupo 5, ou seja, possivelmente são atributos que os pesquisadores de alta produtividade não utilizam mais em sua carreira científica.

Antes de concluir os trabalhos relativos à extração de atributos, é importante analisar os casos de discrepância que aconteceram nesse último agrupamento. De acordo com os dados da Tabela 3.5, existem casos de PQ 1A que foram inseridos no grupo 5, grupo com predominância de pesquisadores PQ 2, e existem casos de PQ 2 no grupo 2, grupo com predominância de pesquisadores PQ 1A, ou seja, pesquisadores classificados no maior nível pelo CNPq que foram atribuídos pelo algoritmo aos grupos com menor produtividade científica, o que indica uma possível discrepância do algoritmo.

O resultado com uma análise gráfica sobre os registros que apresentaram divergências entre o agrupamento e a classificação do CNPq é exposto na Figura 3.11.

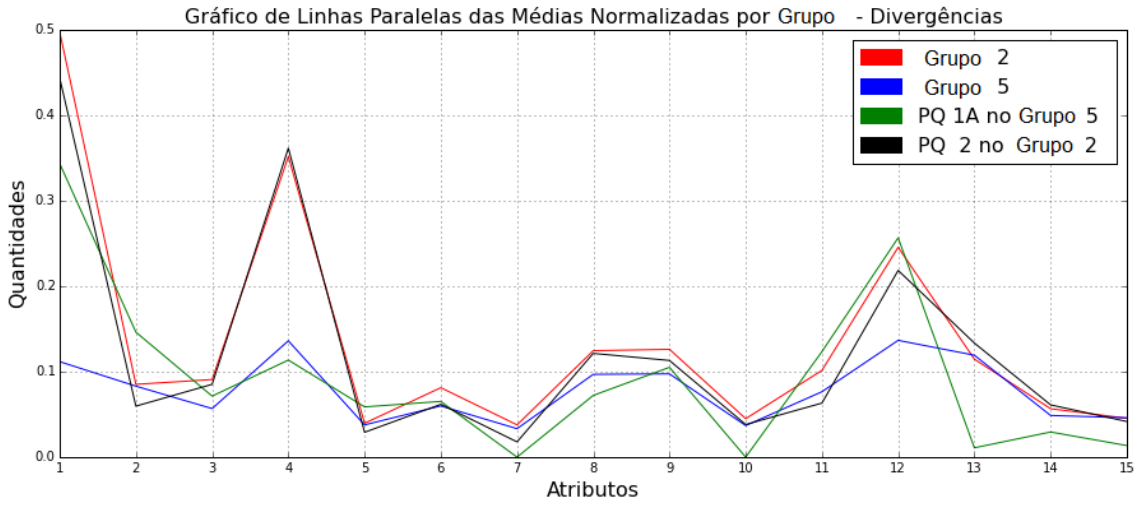

Figura 3.11: Avaliação das divergências entre os grupos e a classificação do CNPq

É possível observar que esses casos interferem como ruído para o agrupamento e, consequentemente, na construção do modelo final, uma vez que os atributos desses pesquisadores estão discrepantes com os demais pesquisadores de mesma classificação no CNPq. Entretanto, ambas as situações são esperadas como resultado do algoritmo, como pode ser avaliado no comportamento das curvas no gráfico da Figura 3.11.

Outro ponto que pode justificar alguns dos casos de divergência é a indisponibilidade da vaga no nível que o pesquisador pertence. Somente o fato do pesquisador possuir os atributos para o nível, não implica necessariamente na classificação pelo CNPq, é preciso ter vaga e orçamento disponível para o financiamento da bolsa [CNPq, 2016].

Como resultado final dessa etapa, foram gerados os gráficos com o comportamento

dos atributos selecionados para cada grupo e cada nível de classificação real do CNPq. Os resultados podem ser observados nas Figuras 3.12 e 3.13.

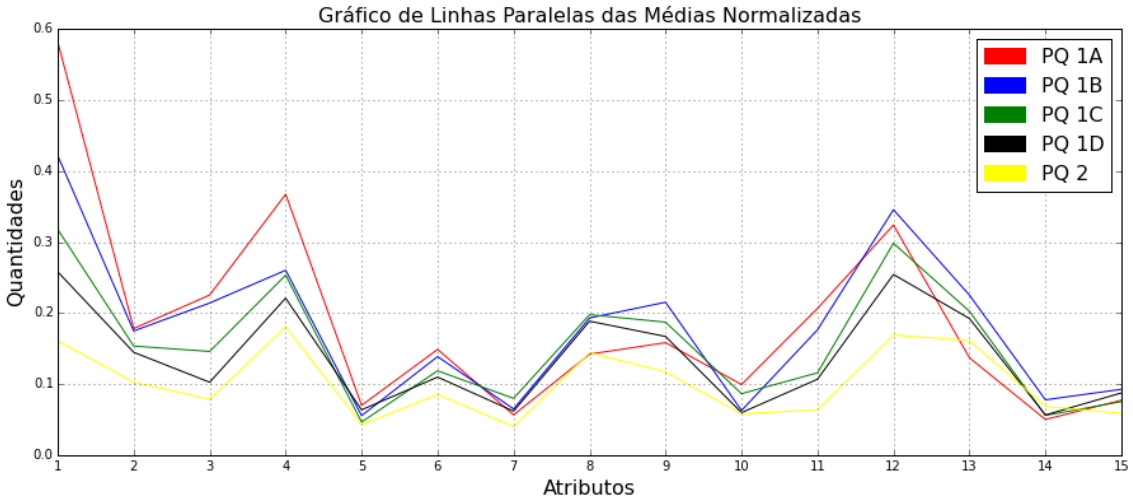

Figura 3.12: Avaliação da classificação do CNPq com os atributos finais

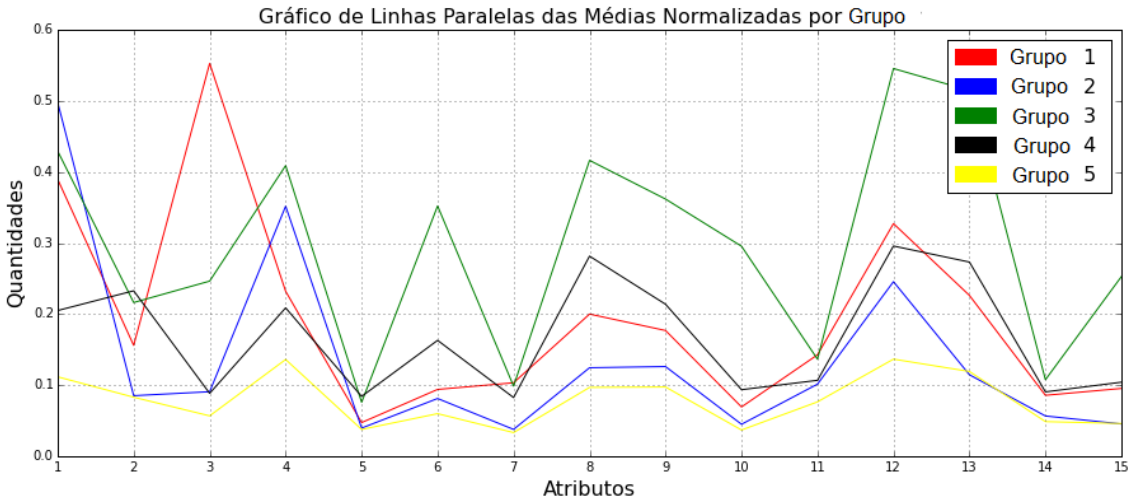

Figura 3.13: Avaliação dos grupos criados pelo agrupamento com os atributos finais

Visto que os resultados foram satisfatórios, pois foi possível selecionar atributos que criaram grupos distintos de pesquisadores através do agrupamento com as quantidades produzidas, fixou-se esses 15 atributos, que estão detalhados na tabela do apêndice A.1, como fonte de informação que será utilizada no modelo preditivo apresentado no próximo capítulo.

Capítulo 4

Construção e Ajuste do Modelo

Neste capítulo é descrito o processo de ajuste nos parâmetros dos classificadores analisados, escolha do algoritmo com melhores resultados para desenvolvimento do modelo e ajustes para aperfeiçoar a capacidade preditiva do potencial científico.

Para isso, após a seleção dos atributos descrita no capítulo 3, o primeiro passo para desenvolvimento do modelo foi identificar diferentes algoritmos de classificação, com diferentes princípios de funcionamento, que foram propostos como base para modelos distintos de classificação para o problema. Com a definição dos algoritmos, cada classificador teve seus parâmetros ajustados para melhor performance e, com os modelos ajustados, foi realizada uma comparação entre os resultados para identificar o melhor modelo de classificação, dentre os algoritmos analisados.

Com a definição do melhor modelo de classificação para o problema, foram realizados testes para comparação com outros trabalhos. Posteriormente foi realizada uma avaliação da capacidade preditiva, objetivo principal dessa dissertação, e, por último, foram realizados ajustes para possibilitar a utilização do modelo em diferentes grupos de pesquisa.

4.1 Seleção da Amostra

A amostra utilizada nessa pesquisa consiste em dados de pesquisadores relacionados com a área Ciência da Computação cadastrados na plataforma Lattes. Diante da base de dados já utilizada na seleção de atributos, descrita no capítulo 3, foram acrescentados registros de pesquisadores doutores que não receberam bolsa de produtividade PQ no período avaliado (2004 a 2014). Com isso, foi possível montar uma base de análise com pesquisadores classificados pelo CNPq em todos os níveis de bolsa e pesquisadores que não tiveram bolsa de produtividade.

Da mesma forma que na etapa de seleção de atributos, porém, agora com intuito de auxiliar no desenvolvimento do modelo, foram selecionados da base de dados, somente os registros de pesquisadores com 8 anos ou mais de tempo de doutorado em relação ao ano de avaliação do comitê. O resultado, que pode ser observado na Tabela 4.1, foi uma base com 1143 registros de pesquisadores avaliados entre 2004 e 2014, além de pesquisadores que não receberam a bolsa PQ nesse período.

Tabela 4.1: Distribuição Anual da Amostra de Dados Completa

Ano	***PQ 1A***	***PQ 1B***	***PQ 1C***	***PQ 1D***	***PQ 2***	***Sem Bolsa***	***Total***
2004	5	6	7	13	15	0	16
2005	1	3	4	20	44	0	72
2006	2	5	5	10	7	0	29
2007	7	5	15	27	34	0	88
2008	4	2	6	18	62	0	92
2009	3	7	2	12	30	0	54
2010	10	11	25	27	63	0	136
2011	2	6	12	34	63	0	117
2012	8	7	5	15	45	0	80
2013	1	1		8	75	0	85
2014	5	11	15	28	57	228	344
Total	**48**	**64**	**96**	**212**	**495**	**228**	**1143**

A partir desses registros, foi contabilizada a produtividade científica de acordo com os 15 atributos identificados no capítulo 3, sendo para cada pesquisador os últimos 10 anos que precedem a avaliação do comitê CA-CC. Assim, se o pesquisador foi avaliado em 2004, foi considerada a produção nos anos de 1995 a 2004, para a quantificação da sua produtividade científica (conforme exemplificado anteriormente na Tabela 3.3).

Para os pesquisadores que não recebem bolsa PQ, como os currículos foram selecionados em 2014, para a quantificação da sua produtividade científica foi considerada a produção nos anos de 2005 a 2014. Após a quantificação da produtividade de cada pesquisador, foi criada uma base de dados que reúnem essas informações para definição e ajustes dos modelos classificadores.

4.2 Definição dos Modelos e Ajustes dos Parâmetros

Como não foi possível encontrar estudos que apontam algoritmos de classificação adequados para identificação do potencial de pesquisadores em Ciência da Computação,

foi proposto uma exploração entre classificadores disponíveis na biblioteca *Scikit Learn* e que são comuns na aplicação de mineração de dados. Os algoritmos utilizados nesse trabalho para construção dos modelos iniciais foram: *k-Nearest Neighbors*, *Random Forest*, *Decision Tree*, *Naive Bayes* e *Support Vector Machines*.

O algoritmo *Decision Tree* foi utilizado no modelo com os parâmetros padrões da biblioteca *Scikit Learn*. Já os algoritmos *k-Nearest Neighbors*, *Random Forest*, *Naive Bayes* e *Support Vector Machines* tiveram alguns de seus parâmetros ajustados previamente para aplicação no problema em questão, permitindo assim uma comparação justa entre os modelos gerados pelos classificadores.

Para avaliar os classificadores e testar os ajustes dos parâmetros, foi utilizado no conjunto de dados descritos na Tabela 4.1, o método de validação cruzada *K-Fold* estratificada com k=10 (*10-fold*), exemplificada na Figura 4.1. Esse método separa a base de dados em 10 sub-bases, sendo nove para treinamento e uma para teste, de forma que cada sub-base tenha a mesma proporção de classes que as demais, permitindo comparar estatisticamente os resultados obtidos em cada execução de teste [Japkowicz & Shah, 2011].

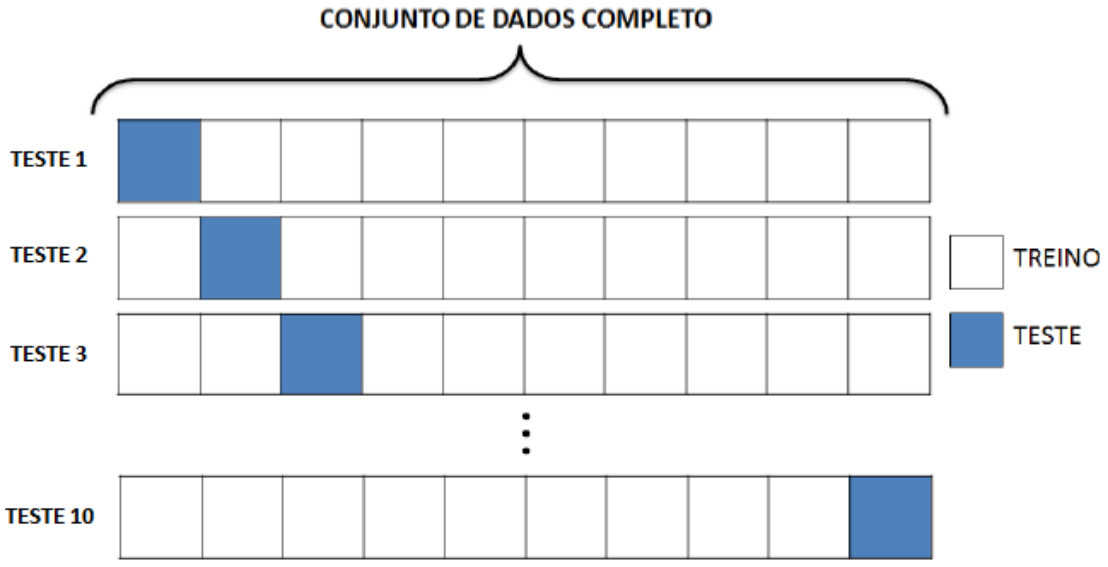

Figura 4.1: Processo de validação cruzada *10-fold* estratificada

Foram realizados vários testes em validação cruzada para cada algoritmo com diferentes configurações dos parâmetros, sendo armazenada a acurácia média em cada execução do modelo, ou seja, o total de registros com classificação correta em relação ao total de registros existentes na base de teste. Ao término desse procedimento, as médias das acurácias foram analisadas para identificar o melhor resultado e, por consequência,

determinar a melhor configuração dos parâmetros.

Outra informação utilizada no ajuste dos parâmetros foram as probabilidades de classificação em cada execução dos testes. Nos algoritmos utilizados, é possível obter em cada classificação a probabilidade do registro pertencer à classe prevista, ou seja, qual a confiança que o modelo tem em afirmar que o registro pertence a determinada classe. Isso implica que, além de acertar qual é a classe que o registro pertence, medida pela acurácia, é importante também o classificador ter confiança na classe que ele escolheu como prevista [Scikit-Learn-Org, 2016].

Em cada teste realizado na validação cruzada para ajuste dos parâmetros, além da classe prevista pelo modelo e acurácia, foram armazenadas as probabilidades de cada previsão. Isso permite avaliar qual configuração do modelo tem maior confiabilidade na classificação realizada, informação que pode interferir na decisão dos parâmetros escolhidos para cada modelo.

4.2.1 Ajuste do *k-Nearest Neighbors*

Dentre os parâmetros do algoritmo *k-Nearest Neighbors*, o números de vizinhos utilizados na classificação tem impacto direto no resultado do modelo. Com isso, foi proposta uma avaliação do resultado para diferentes valores de K, mantendo os demais parâmetros do classificador com os valores padrões da biblioteca *Scikit Learn*.

Na Figura 4.2 é possível observar a distribuição da média das probabilidades e das acurácias para diferentes valores de K no modelo criado com o algoritmo e permite identificar qual a melhor configuração desse parâmetro.

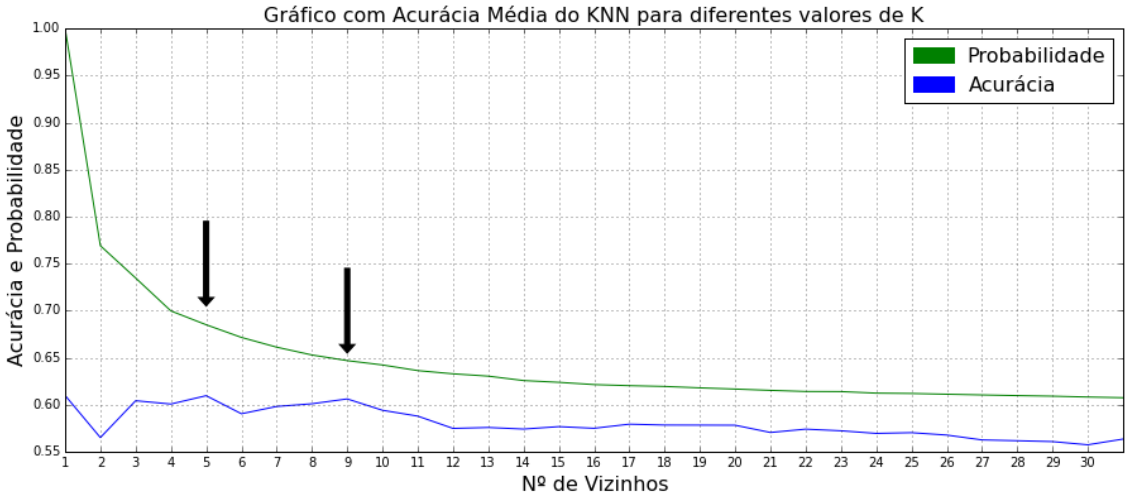

Figura 4.2: Acurácia média do *k-Nearest Neighbors* com variação dos valores de K

Após execução do teste para vários valores de K e análise gráfica da Figura 4.2, observou-se que as melhores configurações identificadas para o modelo foram K=5 e K=9, cujas as acurácias foram aproximadamente 61%. Porém, apesar das duas opções terem resultados semelhantes em termos de acurácia média, a probabilidade média de classificação com K=5 (aproximadamente 68%) foi superior a do modelo com k=9 (aproximadamente 65%). Dessa forma, K=5 foi eleito como a melhor configuração desse classificador.

4.2.2 Ajuste do *Random Forest*

No caso do algoritmo *Random Forest*, o principal parâmetro de configuração que foi avaliado é a quantidade de estimadores existentes na floresta. Novamente, foi proposto vários testes do modelo com pequenas variações do número de estimadores para avaliar o impacto no resultado.

A Figura 4.3 apresenta os valores de acurácia e probabilidade para quantidades variadas de estimadores para o modelo.

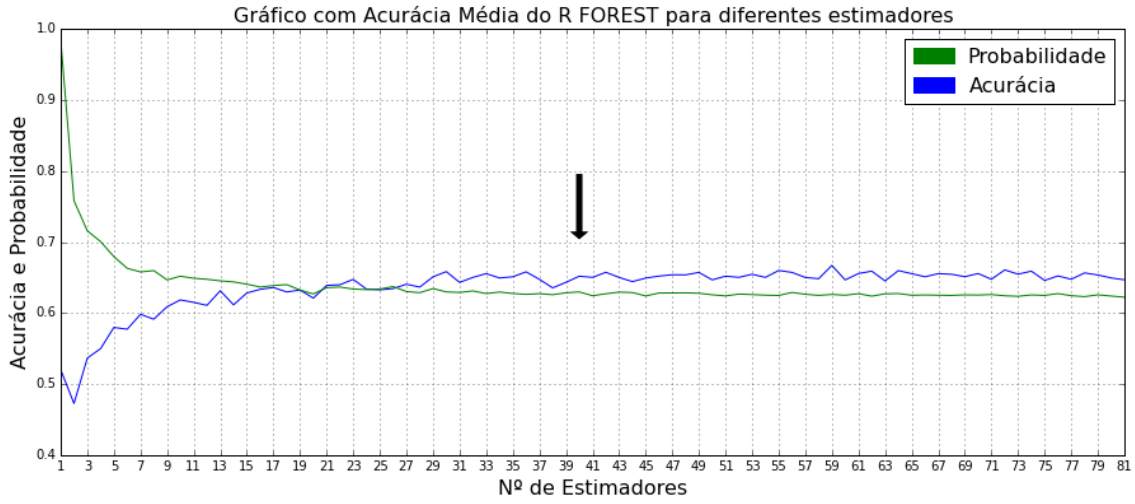

Figura 4.3: Acurácia Média do *Random Forest* com Diferentes Estimadores

Ao avaliar os testes com variação da quantidade de estimadores por análise gráfica, disponível na Figura 4.3, observa-se que a acurácia do classificador oscila em torno de 65% a partir de 40 estimadores, configuração que também estabiliza a probabilidade de classificação do algoritmo. Com isso, essa configuração foi definida como apropriada para o modelo em questão.

4.2.3 Ajuste do *Naive Bayes*

Os algoritmos baeysianos que são disponíveis na biblioteca *Scikit Learn* são: *Gaussian Naive Bayes*, *Multinomial Naive Bayes* e *Bernoulli Naive Bayes*. No caso desse modelo, o ajuste no parâmetro de avaliação foi escolher o algoritmo que seria utilizado para classificação.

Foram realizados testes de validação cruzada com os três algoritmos, cujas acurácias e probabilidades de classificação podem ser observadas na Tabela 4.2. Após comparação entre os resultados, observou-se que o modelo *Bernoulli* apresentou uma acurácia ligeiramente superior ao *Gaussian*, entretanto o modelo Gaussian apresentou uma probabilidade de classificação muito superior aos demais modelos testados, sendo considerado como modelo escolhido.

Tabela 4.2: Avaliação com Diferentes Configurações do Algoritmo *Naive Bayes*

Parâmetro	***Acurácia (%)***	***Probabilidade (%)***
Gaussian	**50,49**	**81,00**
Multinomial	43,64	44,12
Bernoulli	52,66	60,48

4.2.4 Ajuste do *Support Vector Machines*

Para o algoritmo *Support Vector Machines*, quando usado para classificação, o parâmetro que foi avaliado nessa análise foi o *kernel*, que pode ser ajustado como: *linear*, *poly*, *rbf* e *sigmoid*. Esse parâmetro diferencia a forma de separação das classes pelo algoritmo com diferentes tipos de função.

Foram realizados testes para comparação das configurações possíveis e os melhores resultados encontrados, tanto para acurácia, quanto para probabilidade, foram para a configuração com *kernel* = *linear*, conforme resultados mostrados na Tabela 4.3. Decidiu-se então, manter essa configuração para esse modelo.

Tabela 4.3: Avaliação do Modelo com Algoritmos *Support Vector Machines*

Parâmetro	***Acurácia (%)***	***Probabilidade (%)***
linear	**55,90**	**56,23**
poly	43,31	48,43
rbf	46,89	54,67
sigmoid	43,31	43,30

4.2.5 Parâmetros Ajustados para os Modelos

Sendo assim, a Tabela 4.4 apresenta os cinco modelos criados com suas respctivas configurações de escolhas para serem avaliados nas próximas etapas da pesquisa.

Tabela 4.4: Modelos de Classificação com os Respectivos Parâmetros Ajustados

Modelo	***Classificador***	***Abreviação***	***Parâmetro Ajustados***
Modelo 1	*k-Nearest Neighbors*	KNN	n_neighbors = 5
Modelo 2	*Random Forest*	R FOREST	n_estimators = 40
Modelo 3	*Decision Tree*	D TREE	Deafautl da biblioteca
Modelo 4	*Naive Bayes*	GNB	Classificador GaussianNB
Modelo 5	*Support Vector Machines*	SVM	kernel='linear'

4.3 Comparação dos Modelos

Após os ajustes nos parâmetros dos modelos, foram analisados os resultados produzidos pelos cinco modelos para avaliação do desempenho de cada algoritmo de classificação. Foi utilizada a validação cruzada *10-Fold* estratificada (Figura 4.1) para execução dos cinco modelos otimizados.

Os resultados dos testes para cada algoritmo com a média das acurácias e qual a classificação dentre os cinco modelos, apresentado entre parênteses, foram relacionados na Tabela 4.5 e a distribuição nas 10 execuções da validação na Figura 4.4.

Tabela 4.5: Resultado da Validação Cruzada para os Modelos - Acurácia (%)

Teste	***Modelo 1***	***Modelo 2***	***Modelo 3***	***Modelo 4***	***Modelo 5***
Teste 1	59,82 (1º)	52,99 (3º)	52,36 (4º)	46,15 (5º)	56,41 (2º)
Teste 2	54,70 (2º)	60,68 (1º)	52,99 (3º)	47,86 (5º)	50,42 (4º)
Teste 3	68,96 (1º)	66,37 (2º)	56,89 (3º)	42,24 (5º)	56,89 (3º)
Teste 4	66,37 (2º)	68,96 (1º)	61,20 (4º)	57,75 (5º)	62,06 (3º)
Teste 5	66,95 (2º)	75,65 (1º)	55,65 (4º)	53,04 (5º)	57,39 (3º)
Teste 6	60,52 (2º)	61,40 (1º)	50,00 (5º)	50,87 (4º)	52,63 (3º)
Teste 7	57,52 (2º)	67,25 (1º)	52,21 (5º)	57,52 (2º)	55,75 (4º)
Teste 8	52,21 (3º)	61,94 (1º)	45,13 (5º)	52,21 (3º)	56,37 (2º)
Teste 9	57,65 (2º)	63,06 (1º)	54,05 (3º)	46,84 (5º)	53,15 (4º)
Teste 10	64,86 (2º)	71,17 (1º)	54,95 (4º)	50,45 (5º)	57,65 (3º)
Acurácia Média	**60,96**	**64,95**	**53,52**	**50,49**	**55,90**
Desvio Padrão	5,3	6,0	4,0	4,6	3,0
1º-2º-3º-4º-5º Lugar	2-7-1-0-0	8-1-1-0-0	0-0-3-4-3	0-1-1-1-7	0-2-5-3-0

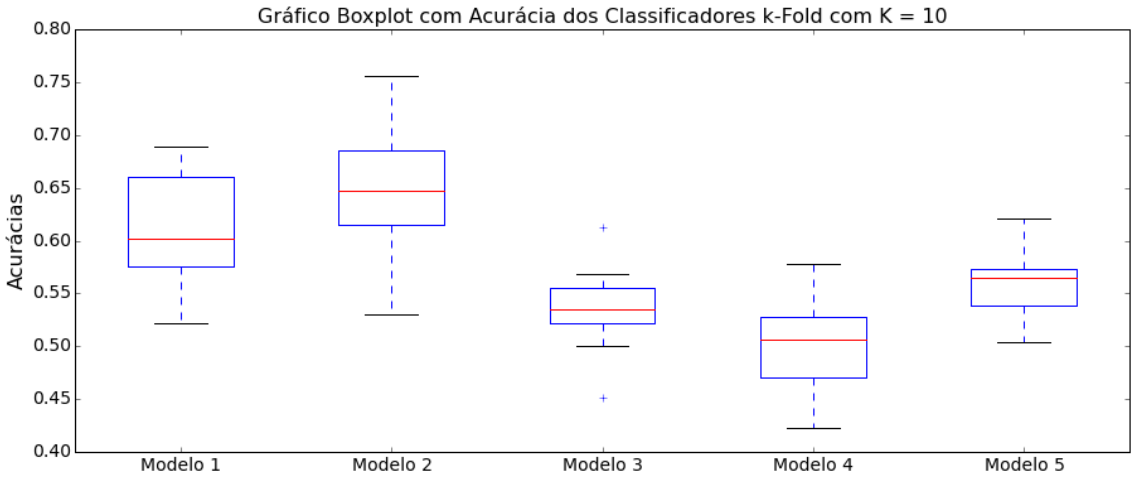

Figura 4.4: Distribuição da Acurácia dos Classificadores - Validação Cruzada

A princípio, os Modelos 1 e 2 foram os que resultaram em melhores acurácias nos testes, como é possível observar na Tabela 4.5. Em duas vezes, o Modelo 1 alcançou o melhor resultado, enquanto nas outras 8 vezes o Modelo 2 obteve o melhor resultado. O que também é possível de observar no Figura 4.4, com o gráfico *boxplot* dos resultados, onde que os dois primeiros modelos se destacam em relação aos demais.

Para escolher qual o melhor modelo dentre os avaliados para o problema, foi realizada uma comparação estatística entre os resultados dos testes. Como os resultados dos testes são cinco amostras independentes e não paramétricas, sem informações acerca da distribuição dos dados, foi utilizado o teste de hipótese de amostras independentes *Wilcoxon-Mann-Whitney*, que avalia se as amostras pertencem a uma mesma população, o que inviabiliza comparações estatísticas em caso positivo. Nesse caso, a hipótese nula ($P\text{-}Value > 0{,}05$) é o resultado do teste para amostras de uma mesma população [Gold, 2007]. Os resultados obtidos com o teste de hipótese entre as amostras com resultados do modelo de melhor performance (Modelo 2) e os demais modelos pode ser observado na Tabela 4.6.

Tabela 4.6: Teste de Hipótese entre Modelo 2 e Demais Modelos

Comparação	***P-Value***
Modelo 2 com Modelo 1	0,0652
Modelo 2 com Modelo 3	0,0003
Modelo 2 com Modelo 4	0,0001
Modelo 2 com Modelo 5	0,0011

Observa-se pelo resultados obtidos com o teste de hipótese apresentado na Tabela 4.6 que a hipótese nula é descartada na comparação do Modelo 2 com os Modelos 3, 4, e 5, dessa forma, pode-se afirmar que a acurácia do Modelo 2 é estatisticamente superior à acurácia dos Modelos 3, 4 e 5.

Ao comparar o Modelo 2 com o Modelo 1, a hipótese nula não é descartada, o que não permite uma afirmação estatística a respeito da comparação. Entretanto, apesar de não existir diferença estatisticamente significativa entre os modelos, optou-se pelo Modelo 2 (*Random Forest*) como o modelo de classificação do trabalho, uma vez que sua acurácia foi superior em 80% dos testes apresentados na Tabela 4.5 (8 dos 10 testes no *10-fold* estratificado).

4.3.1 Discussão

Wanderley [2015], em seu trabalho de mestrado, desenvolveu 10 modelos de classificação por métricas de redes sociais (ARS) em uma abordagem semelhante a desenvolvida nessa pesquisa com validação cruzada. Entretanto, em sua dissertação, a autora teve como objetivo classificar o pesquisador entre duas possibilidades, Com Bolsa PQ e Sem Bolsa PQ, ou seja, uma classificação binária.

A pesquisa obteve como seu principal resultado 74,65% na acurácia média do modelo escolhido, levando em consideração os 10 testes da validação cruzada, que foi considerada pela autora como um bom resultado de classificação.

O classificador desta dissertação obteve como resultado 64,95% (Tabela 4.5) na acurácia média do modelo *Random Forest*. Em contrapartida, deve-se considerar que o presente estudo propõe um classificador com seis possibilidades de classes e, visto que a probabilidade de falha é maior devido número de classes possíveis ser superior, o resultado obtido também pode ser considerado como satisfatório.

Ainda assim, com intuito de uma comparação direta entre os resultados dos modelos de classificação dos trabalhos, foram realizados os mesmos testes por validação cruzada *10-Fold* estratificada para execução do modelo proposto por este estudo, porém considerando uma classificação binária, semelhante à realizada em Wanderley [2015], pesquisadores Com Bolsa PQ e Sem Bolsa PQ, conforme detalhado na Figura 4.5.

Dentre as possíveis métricas para mensurar a performance de seu modelo, Wanderley [2015] utilizou a acurácia, sensibilidade e especificidade para avaliar seus resultados. Sendo a sensibilidade a proporção de predições positivas em relação aos valores positivos reais, ou seja, valores previstos como bolsista PQ em relação ao total que realmente são bolsistas PQ, e a especificidade a proporção entre as predições negati-

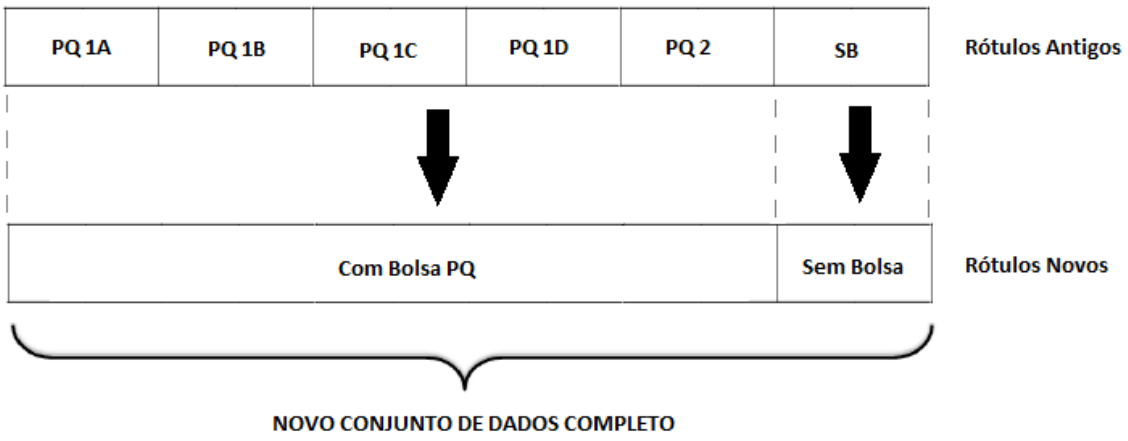

Figura 4.5: Novos rótulos para classificação binária

vas em relação aos valores negativos reais, ou seja, quantidade de previstos como não bolsista em relação ao total que realmente não são bolsistas [Zhu et al., 2010].

Os resultados com a acurácia, sensibilidade e especificidade dos testes para o Modelo 2 (*Random Forest*) ajustado para a classificação binária podem ser observados na Tabela 4.7.

Tabela 4.7: Resultado da Validação Cruzada para Classificação Binária (%)

Teste	***Acurácia***	***Sensibilidade***	***Especificidade***
Teste 1	86,32	86,17	86,95
Teste 2	88,03	91,48	73,91
Teste 3	92,24	95,68	78,26
Teste 4	96,55	100	82,26
Teste 5	93,04	94,56	86,95
Teste 6	93,85	100	69,56
Teste 7	93,80	100	69,56
Teste 8	91,15	98,88	60,86
Teste 9	95,49	97,75	86,36
Teste 10	90,09	96,62	63,63
Média	**92,05**	**96,11**	**75,86**

Com a alteração dos rótulos para uma classificação binária, foi possível fazer uma comparação com o modelo proposto por Wanderley [2015]. É possível observar que, no resultado médio dos 10 testes, o resultado obtido pelo trabalho foi superior nas três medidas, acurácia, sensibilidade e especificidade, o que demonstra a qualidade do modelo preditivo deste trabalho.

Apesar desta dissertação avaliar os pesquisadores que receberam bolsa nos último anos, semelhante ao que foi feito por Wanderley [2015], é importante ressaltar que o ideal para uma comparação direta seria avaliar os resultados do Modelo 2 (*Random Forest*) com exatamente a mesma base de dados utilizada em Wanderley [2015].

4.4 Análise Detalhada do Modelo

Após a definição do modelo preditivo, foi proposta uma análise detalhada do classificador para o teste de validação cruzada, com intuito de identificar as discrepâncias na classificação e entender quais motivos impactaram na acurácia do teste. Para isso, foram selecionados os testes com os piores resultados para o Modelo 2 (*Random Forest*), Teste 1 e Teste 2 conforme é possível observar na Tabela 4.5, e feita uma análise dos casos que o classificador errou em sua predição.

Nas Tabelas 4.8 e 4.9, computou-se a matriz de confusão do Teste 1 e Teste 2 realizados pela validação cruzada, sendo a informação em S.B. referente aos pesquisadores sem bolsa de produtividade. Os dados previstos pelo classificador estão nas colunas, enquanto os dados da classificação real do CNPq estão nas linhas da tabela.

Tabela 4.8: Matriz de Confusão da Classificação no Teste 1 - Acurácia 52,99%

		Resultado do Modelo - Teste 1					
		PQ 1A	**PQ 1B**	**PQ 1C**	**PQ 1D**	**PQ 2**	**S.B.**
CNPq	**PQ 1A**	**2**	0	0	1	2	0
	PQ 1B	2	**0**	0	1	2	2
	PQ 1C	0	0	**0**	4	4	2
	PQ 1D	0	0	1	**5**	39	5
	PQ 2	0	0	1	5	**39**	5
	SB	0	0	0	0	3	**20**

Tabela 4.9: Matriz de Confusão da Classificação no Teste 2 - Acurácia 60,68%

		Resultado do Modelo - Teste 2					
		PQ 1A	**PQ 1B**	**PQ 1C**	**PQ 1D**	**PQ 2**	**S.B.**
CNPq	**PQ 1A**	**3**	1	0	0	1	0
	PQ 1B	1	**1**	1	2	2	0
	PQ 1C	0	1	**1**	2	4	2
	PQ 1D	0	0	2	**8**	11	1
	PQ 2	0	0	0	4	**41**	5
	SB	0	0	0	0	6	**17**

Nas duas situações, Tabelas 4.8 e 4.9, é possível observar casos de erros na classificação do modelo em relação à avaliação feita pelo comitê CA-CC. Existem casos de pesquisadores PQ 1A sendo classificados pelo modelo como PQ 2, o que gera a necessidade de uma avaliação dos atributos desses casos, visto divergência elevada entre a classificação real e a prevista pelo modelo, ou seja, a classificação real é o nível mais elevado de bolsa, enquanto o classificador previu o menor nível de bolsa.

Para avaliação dos atributos nos dois casos de discrepância, foram gerados os gráficos de linhas paralelas com as médias dos atributos normalizados para os pesquisadores PQ 1A, PQ 2 e, no mesmo gráfico, a média dos atributos dos casos discrepantes, conforme pode ser observado nas Figuras 4.6 e 4.7, com a representação do Teste 1 e Teste 2 respectivamente.

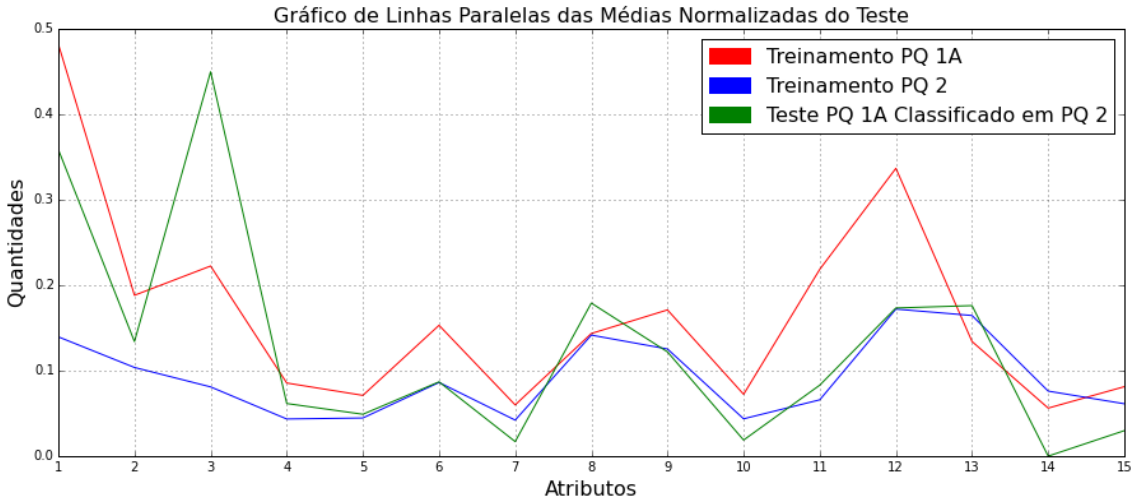

Figura 4.6: Avaliação das divergências encontradas no Teste 1

É possível observar, por análise gráfica nas Figuras 4.6 e 4.7, que os casos com divergências entre a previsão do classificador e a avaliação real do CA-CC eram esperadas. Ao avaliar o comportamento médio dos atributos dos casos divergentes nos dois testes, fica claro que a contagem de alguns atributos não se comporta como deveria.

Nos resultados do primeiro teste, apresentado no gráfico da Figura 4.6, nos atributos iniciais (do 1 ao 4) a linha das divergências está mais próximo ao do PQ 1A, porém nos atributos restantes (do 5 ao 15) o comportamento é muito próximo ao do PQ 2, ou seja, na maioria dos casos a linha do gráfico da divergência se aproxima de um pesquisador PQ 2, motivo da previsão do modelo.

Nos resultados do segundo teste, apresentado no gráfico da Figura 4.7, a linha das divergências está mais semelhante aos dos PQ 1A, porém, também tem um com-

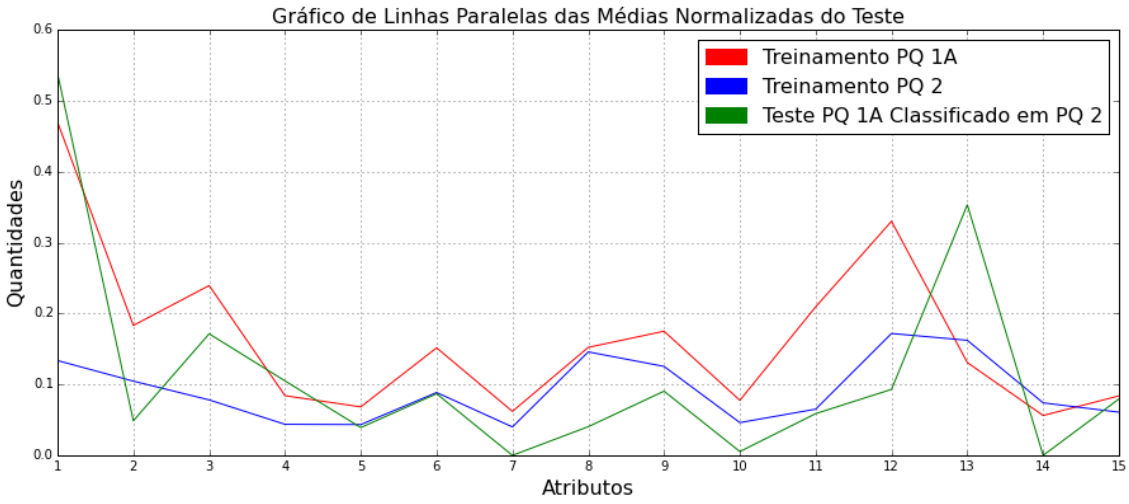

Figura 4.7: Avaliação das divergências encontradas no Teste 2

portamento próximo aos pesquisadores que são PQ 2, o que explica o resultado do classificador nesse teste.

Isso conclui que, levando em consideração os 15 atributos selecionados, o classificador está correto e que é possível classificar computacionalmente o desempenho dos pesquisadores. Porém, também indica que possivelmente outros atributos são analisados pelo comitê em suas avaliações anuais de bolsa, ou que existem características subjetivas nessas avaliações que não podem ser contabilizadas via algoritmos computacionais.

4.5 O Modelo de Predição

Após validação do modelo com dados dos pesquisadores avaliados entre 2004 e 2014, foi realizado um teste para analisar a capacidade do modelo em predizer o desempenho futuro do pesquisador. Para isso realizou-se um estudo retrospectivo, utilizando informações do currículo Lattes.

A base de dados para esse teste de predição foi construída a partir da base de dados com 1143 pesquisadores (Tabela 4.1), dos quais foram selecionados os pesquisadores avaliados entre 2011 e 2014, para permitir uma avaliação da capacidade de predição do modelo com uma antecedência superior a 5 anos.

Além disso, foram selecionados somente os pesquisadores com tempo de doutorado superior a 15 anos. Dessa forma, todos os pesquisadores a serem avaliados terão no mínimo 5 anos de doutorado em 2005, ano que será utilizado como referência nesse

teste. O resultado da base criada para o teste de predição, apresentado na Tabela 4.10, compôs uma amostra com 350 pesquisadores.

Tabela 4.10: Avaliações de Bolsa CA-CC por Ano - Teste de Predição

Ano	***PQ 1A***	***PQ 1B***	***PQ 1C***	***PQ 1D***	***PQ 2***	***Sem Bolsa***	***Total***
2011	2	5	9	18	24	0	58
2012	8	7	4	8	13	0	40
2013	1	1	0	3	30	0	35
2014	5	11	14	19	27	141	217
Total	**16**	**24**	**27**	**48**	**94**	**141**	**350**

Diferentemente das etapas anteriores, que consideraram a produtividade científica dos pesquisadores nos 10 anos que antecediam a avaliação pelo comitê CA-CC (conforme exemplificado anteriormente na Tabela 3.3), nesta etapa foi proposta uma avaliação da predição do modelo através da análise da produtividade científica dos pesquisadores nos 10 anos anteriores à 2005. Em resumo, o modelo irá fazer um prognóstico do potencial de produtividade científica do pesquisador com uma antecedência de 6 a 9 anos.

Na Tabela 4.11, que apresenta um exêmplo dos dados para esse teste, o Pesquisador 1 por exemplo, foi avaliado em 2014. Nesse teste de predição, o objetivo é classificar os pesquisadores com os dados de 2005 qual foi a classificação em 2014, ou seja, uma avaliação de qual a posição futura do pesquisador.

Ainda na Tabela 4.11, também é possível observar o motivo de selecionar somente os pesquisadores com tempo de doutorado igual ou superior a 15 anos, que é o tempo de doutorado em 2005 acima de 4 anos. Por exemplo, um pesquisador que foi avaliado em 2014, se avaliarmos a menor situação da seleção, 15 anos de doutorado em 2014, ao avaliar a situação em 2005 o tempo de doutorado seria 5 anos, tempo de doutorado suficiente para uma boa produtividade na época, o que permite avaliar o modelo com predição futura.

Tabela 4.11: Exemplo de Contagem dos Atributos para Predição

Pesquisador	*Ano de Avaliação no CNPq*	*Período Contabilizado*	*Tempo de Doutorado em 2005*
Pesquisador 1	2014	1996 - 2005	5
Pesquisador 2	2013	1996 - 2005	6
Pesquisador 3	2012	1996 - 2005	7
Pesquisador 4	2011	1996 - 2005	8

A partir daí, foi contabilizada a produtividade científica desses pesquisadores nos 10 anos que antecedem o ano de 2005 (de 1996 a 2005), utilizado como referência, para predizer qual seria a classificação desses pesquisadores nos anos de 2011 a 2014, classificação real do CA-CC, e então comparar os resultados.

Semelhante ao que foi realizado anteriormente, foram executados novos testes por validação cruzada com os dados para esse teste da predição pelo Modelo 2 (*Random Forest*). As médias das acurácias nos testes de predição podem ser observados na Tabela 4.12. Pelos resultados obtidos, a acurácia média do teste de predição foi de 62,11%, o que pode ser considerado um bom resultado, visto que são 6 possibilidades de classificação pelo modelo com no mínimo 6 anos de antecedência.

Tabela 4.12: Acurácia Média da Validação Cruzada para Predição - Acurácia (%)

Teste	***Modelo 2 - R FOREST***
Teste 1	55,26
Teste 2	64,86
Teste 3	54,05
Teste 4	51,35
Teste 5	62,85
Teste 6	59,99
Teste 7	50,00
Teste 8	72,72
Teste 9	81,25
Teste 10	68,75
Acurácia Média	**62,11**
Desvio Padrão	9,5

Por se tratar de um modelo preditivo futuro, o teste fez uma previsão de qual seria o nível do pesquisador em avaliações realizadas no período de 2011 até 2014, com dados entre 1996 e 2005, ou seja, com informações de um período anterior ao que foi utilizado pelo CNPq. Conforme exemplificado na Tabela 3.3, os pesquisadores avaliados em 2011, por exêmplo, utilizaram dados de 2002 a 2011, sendo que no teste de predição foram utilizados os dados de 1996 - 2005. Isso demonstra que é possível predizer computacionalmente qual o desempenho futuro dos pesquisadores em Ciências da Computação com o modelo preditivo desta dissertação.

4.6 Ajuste do Modelo Preditivo

Todo desenvolvimento do modelo preditivo foi baseado nas classificações realizadas pelo comitê CA-CC de pesquisadores brasileiros. Com os resultados obtidos, é possível

demonstrar a possibilidade de avaliação da classificação de bolsas de pesquisa e investimento dos órgãos de fomento através de algoritmos de classificação, como realizado até o momento no trabalho.

Apesar disso, é comum as instituições de ensino avaliar cientistas em potencial para inclusão em grupos de pesquisa, que não necessariamente recebem financiamento por bolsa PQ, ou seja, a base de treinamento seria composta em sua maioria por pesquisadores sem bolsa. Nesse caso, se o teste de predição for realizado da forma apresentada, possivelmente todos os registros serão direcionados para a classificação de pesquisador sem bolsa, maioria na base de treinamento.

Para contornar essa situação, seria interessante que a base de dados de pesquisadores que, não necessariamente são proprietários de bolsas de produtividade, receba um rótulo antes da predição. Nesse caso, um rótulo em comparação com um grupo de pesquisa previamente conhecido na instituição avaliada. Para isso, foi proposto os mesmos testes preditivos realizados anteriormente, porém considerando como rótulo da base o resultado de um agrupamento prévio com os atributos utilizados no modelo, conforme observado na Figura 4.8.

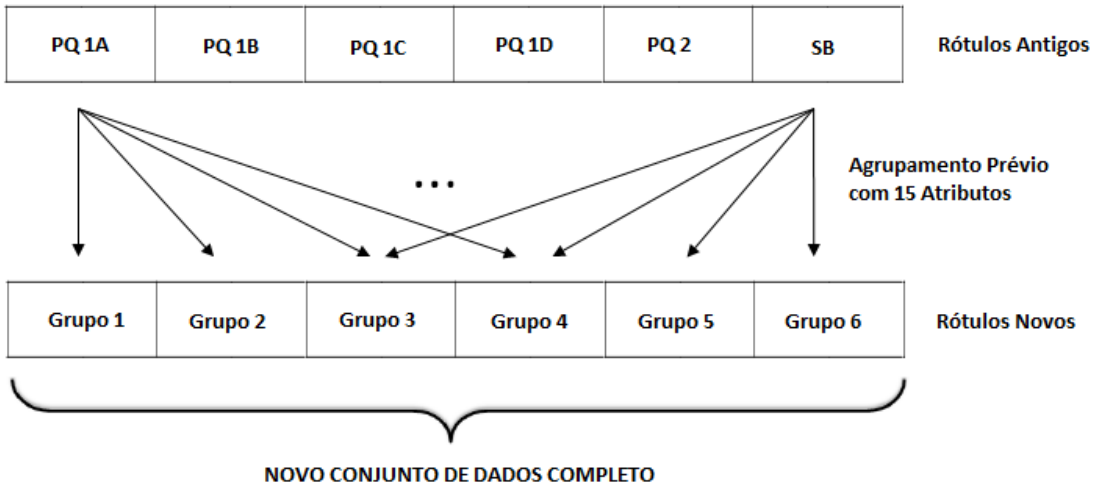

Figura 4.8: Novo arupamento prévio

É possível observar pela Figura 4.8 que os registros da base foram agrupados e gerados novos rótulos. Esse tipo de teste possibilita o modelo preditivo classificar baseado em rótulos criados pelo agrupamento prévio na base de treinamento, ou seja, é possível trabalhar com o modelo preditivo baseado em pesquisadores que não necessariamente recebem as bolsas PQ.

Para realização do teste ajustado no classificador, foi proposta uma validação cru-

zada com *10-fold* estratificada, sendo criados seis grupos de pesquisadores previamente na base através do algoritmo *K-means*.

O agrupamento foi realizado na mesma base de dados utilizada para predição futura (Tabela 4.10), considerando os 15 atributos selecionados e os dados históricos de produtividade dos 10 anos que antecedem a avaliação do CNPq. Em resumo, os pesquisadores avaliados em 2014 foram agrupados com os dados de 2005 até 2014, os pesquisadores avaliados em 2013, foram agrupados com os dados de 2004 até 2013 e assim sucessivamente para os pesquisadores avaliados em 2012 e 2011. O resultado do agrupamento prévio para criação de novos rótulos é apresentado na Tabela 4.13.

Tabela 4.13: Distribuição dos Pesquisadores em Novos Rótulos

		Agrupamento					
		Grupo 1	**Grupo 2**	**Grupo 3**	**Grupo 4**	**Grupo 5**	**Grupo 6**
CNPq	**PQ 1A**	2	2	3	2	3	4
	PQ 1B	3	3	7	4	3	4
	PQ 1C	3	3	3	2	8	8
	PQ 1D	5	4	6	1	15	17
	PQ 2	21	7	2	6	16	42
	S.B.	72	35	0	7	9	18

Nota-se pelo dados apresentados na Tabela 4.13 que, com o agrupamento realizado, é esperada uma melhora na acurácia do modelo preditivo, visto que o próprio agrupamento já cria rótulos baseados em grupos de produtividade semelhante. Porém, isso não reduz a importância do teste, visto que a classificação preditiva será realizada com dados históricos diferentes aos realizados pelo agrupamento. Ou seja, ainda assim é feita uma avaliação do desempenho do pesquisador no futuro em relação aos dados avaliados.

Com o resultado do agrupamento, a validação cruzada com *10-fold* estratificada foi executada, levando em consideração como rótulo o grupo criado pelo agrupamento prévio. Já os dados utilizados para classificação, em todos os casos, foi a produtividade dos 10 anos que antecedem o ano de 2005, ou seja, produção de 1996 até 2005, conforme foi apresentado anteriormente na Tabela 4.11.

O resultado com as médias das acurácias nos testes para o algoritmo *Random Forest* ajustado com agrupamento prévio, é apresentado na Tabela 4.14.

Com o modelo ajustado, foi possível obter um resultado superior em relação ao método de classificação com as classes reais do CNPq, o que era esperado devido ao agrupamento prévio. Mas ao comparar os dois métodos de predição futura (Tabelas 4.12 e 4.14), houve uma variação de 62,11% para 77,67%, uma diferença alta de

Tabela 4.14: Resultado da Validação Cruzada para Predição com Agrupamento - Acurácia (%)

Teste	***Modelo 2 - R FOREST***
Teste 1	81,57
Teste 2	78,37
Teste 3	86,48
Teste 4	83,78
Teste 5	94,28
Teste 6	71,42
Teste 7	73,82
Teste 8	66,66
Teste 9	71,87
Teste 10	68,75
Acurácia Média	**77,67**
Desvio Padrão	8,3

15,56%. Essa variação é explicada pelo fato de que a classificação real do CNPq possivelmente leva em consideração outros atributos ou informações subjetivas, que não foram avaliados pelo modelo preditivo, como já destacado anteriormente.

4.7 Considerações Finais

Após os testes realizados para ajustes e comprovação de viabilidade do modelo preditivo, conclui-se que o modelo, dentre os diversos avaliados nesse trabalho, que apresentou melhores resultados para classificação foi o algoritmo *Random Forest* com o parâmetro de estimadores igual a 40.

Além da identificação do modelo adequado ao problema, é possível observar que apesar da acurácia do classificador ser relevante, existem casos divergentes da classificação real, que são considerados como erro do modelo. Porém, após uma análise gráfica, verificou-se que esses casos eram esperados, o que sugere a possibilidade de o comitê utilizar outras informações no critérios para avaliação do CA-CC.

Outra possibilidade é que um pesquisador não desce o nível de classificação do CNPq, ou seja, classificações anteriores podem interferir em novas avaliações de uma forma subjetiva, mesmo com uma possível queda de produtividade recente pelo pesquisador.

Apesar de conseguir alcançar 77,67% de acurácia média em um modelo preditivo de potencial produtivo de pesquisa, o que demonstra a relevância dos resultados desse algoritmo para classificação de pesquisadores que trabalham na área de Ciência da

Computação, existem várias possibilidades de pesquisa que podem ser desenvolvidas em continuidade nesse projeto, possibilitando aprimorar ainda mais os resultados obtidos neste trabalho.

Capítulo 5

Conclusões e Trabalhos Futuros

Nesta dissertação de mestrado foi realizada a classificação dos pesquisadores brasileiros com relevante produtividade científica, financiados pelo CNPq, que atuam na grande área Ciência da Computação.

A análise e comparação dos pesquisadores foi feita através de alguns atributos selecionados e disponíveis na plataforma Lattes em conjunto com informações do sistema Qualis. Além disso, foi desenvolvida uma metodologia de análise preditiva do potencial de pesquisa do cientista, ao avaliar o desempenho do classificador com base em seus dados históricos disponíveis na plataforma Lattes, para predizer qual seria sua classificação 10 anos após os dados analisados. Diante de tais resultados, foi possível obter um modelo de predição com o algoritmo *Random Forest* para os pesquisadores de Ciência da Computação, sendo a classificação realizada em comparação com pesquisadores financiados pelo CNPq e com pesquisadores sem financiamento, permitindo aplicação do modelo em diferentes grupos de pesquisadores.

Também foi desenvolvido um modelo computacional para predição do potencial produtivo dos pesquisadores em Ciência da Computação, baseado em informações do sistema Lattes de pesquisadores considerados como referência nacional na área, o que permitiu alcançar o objetivo geral do trabalho. O modelo de melhor resultado foi ajustado para qualquer grupo de pesquisadores em diferentes realidades produtivas no Brasil, o que gerou um resultado de 77,67% em acurácia média, considerado satisfatório devido condições do problema.

5.1 Trabalhos Futuros

Apesar da extensa análise realizada sobre a produção científica dos pesquisadores de alta produtividade em Ciência da Computação no Brasil, ainda existem oportunidades

de trabalhos futuros que podem abordar aspectos importantes como:

- **Evolução do classificador com novos atributos:** Apesar dos resultados do modelo de classificação serem considerados satisfatórios, é possível aperfeiçoar os atributos com novas informações do próprio Lattes que não foram analisadas, como coautoria, ou informações de outros sistemas. Objetiva-se aqui melhorar os resultados do classificador, principalmente para avaliação de bolsa de produtividade com análise de anos recentes. Uma pesquisa nessa linha poderia auxiliar ou até mesmo automatizar todo processo de avaliação e julgamento do CA-CC, o que tornaria o processo de seleção autônomo, permanecendo manual somente a tarefa de comprovação de documentos na plataforma Lattes.

- **Complementação do modelo preditivo com análise de colaboração e orientação:** Essa possibilidade consiste em avaliar o impacto de redes de colaboração entre pesquisadores no modelo preditivo e inserir redes de orientador/pesquisador, onde que o resultado da classificação do orientador interfere no potencial do orientando, visto que existe a possibilidade de uma seleção entre os possíveis orientadores de maiores resultados científicos.

- **Desenvolver um direcionamento preciso de evolução:** Desenvolver em paralelo ao modelo preditivo de potencial, uma estrutura de evolução para cada atributo, o que mostra ao pesquisador o que ele precisa evoluir claramente em cada atributo, para alcançar a predição diante da produtividade atual e até mesmo outras classificações, caso pesquisador consiga evoluir sua capacidade produtiva.

Referências Bibliográficas

Abbasi, A.; Altmann, J. & Hossain, L. (2011). 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*, 5(4):594--607. ISSN 1751-1577.

Borko, H. (1968). Information science: What is it? *American Documentation, Wiley Online Library*, 19(1):3--5.

Cimenler, O. (2014). *Social Network Analysis of Researchers Communication and Collaborative Networks Using Self-reported Data*. Doctor of philosophy, University of South Florida.

CNPq (2016). Conselho nacional de desenvolvimento cientifico e tecnologico. Disponível em: http://www.cnpq.br. Acesso em: 16 de ago. 2015.

Da Costa Côrtes, S.; Porcaro, R. M. & Lifschitz, S. (2002). *Mineração de dados-Funcionalidades, técnicas e abordagens*. PUC. Disponível em: ftp://139.82.16.194/pub/docs/techreports/02_10_cortes.pdf.

De Lima, H. A. (2014). *Análise comparativa de pesquisadores considerando características de múltiplas áreas de pesquisa*. Mestrado em ciência da computação, Departamento de Ciência da Computação, Universidade Federal de Minas Gerais.

Digiampietri, L. A.; Mena-Chalco, J. P.; de Melo, P. O. V.; Malheiro, A. P.; Meira, D. N.; Franco, L. F. & Oliveira, L. B. (2014). Brax-ray: an x-ray of the brazilian computer science graduate programs. *PloS one*, 9(4):e94541.

Flach, P. (2012). *Machine learning: the art and science of algorithms that make sense of data*. Cambridge University Press. ISBN 978-1-107-09639-4 and 978-1-107-42222-3.

Garcia, J. F. D. (2015). *Uma análise da produção científica em Ciência da Computação na América Latina*. Mestrado em ciência da computação, Departamento de Ciência da Computação, Universidade Federal de Minas Gerais.

Gold, A. (2007). Understanding the mann-whitney test. *Journal of Property Tax Assessment and Administration*, 4(3):55.

Gonçalves, A. L. (2000). *Utilização de técnicas de mineração de dados em bases de C & T: uma análise dos grupos de pesquisa no Brasil*. Mestrado em engenharia de produção, Universidade Federal de Santa Catarina.

Han, J.; Pei, J. & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier, second edition edição. ISBN 13: 978-1-55860-901-3 and 10: 1-55860-901-6.

Harrington, P. (2012). *Machine learning in action*, volume 5. Manning Greenwich, CT. ISBN 9781617290183.

Japkowicz, N. & Shah, M. (2011). *Evaluating learning algorithms: a classification perspective*. Cambridge University Press. ISBN 978-0-521-19600-0.

Macias-Chapula, C. A. (1998). O papel da informetria e da cienciometria e sua perspectiva nacional e internacional. *Ciência da informação*, 27(2):134--140. ISSN 0100-1965.

Mazloumian, A.; Eom, Y.-H.; Helbing, D.; Lozano, S. & Fortunato, S. (2011). How citation boosts promote scientific paradigm shifts and nobel prizes. *PloS one*, 6(5):e18975.

Mena-Chalco, J. P. & Cesar-Jr, R. M. (2016). scriptlattes. Disponível em: http://scriptlattes.sourceforge.net/. Acesso em: 16 de ago. 2015.

Mena-Chalco, J. P. & Júnior, C. (2013). Prospecção de dados acadêmicos de currículos lattes através de scriptlattes. *Bibliometria e Cientometria: reflexões teóricas e interfaces. São Carlos: Pedro & João*, pp. 109--128.

Mugnaini, R. (2006). *Caminhos para adequação da avaliação da produção científica brasileira: impacto nacional versus internacional*. Doutorado em ciência da informação, Escola de Comunicações e Artes, Universidade de São Paulo.

Mugnaini, R.; Jannuzzi, P. & Quoniam, L. (2004). Indicadores bibliométricos da produção científica brasileira: uma análise a partir da base pascal. *Ciência da Informação*, 33(2):123--131.

of Medicine National Institutes of Health, U. N. L. (2015). Pubmed - ncbi. Disponível em: http://www.ncbi.nlm.nih.gov/pubmed. Acesso em: 16 de abr. 2015.

Scikit-Learn-Org (2016). scikit-learn, machine learning in python. Disponível em: http://scikit-learn.org/. Acesso em: 16 de ago. 2015.

Thomaz, P. G.; Assad, R. S. & Moreira, L. F. P. (2011). Uso do fator de impacto e do índice h para avaliar pesquisadores e publicações. *Arq. bras. cardiol*, 96(2):90--93. ISSN 0066-782X.

Van Dijk, D.; Manor, O. & Carey, L. B. (2014). Publication metrics and success on the academic job market. *Current Biology*, 24(11):R516--R517. ISSN 0960-9822.

Wainer, J. & Vieira, P. (2013). Avaliação de bolsas de produtividade do cnpq e medidas bibliométricas: correlações para todas as grandes áreas. *Perspectivas em Ciência da Informação*, 18(2):60--78.

Wanderley, A. J. (2015). *Um Modelo para Avaliação de Relevância Científica Baseado em Métricas de Análise de Redes Sociais*. Mestrado em informática, Universidade Federal da Paraíba.

Witten, I. H. & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, second edição. ISBN 0-12-088407-0.

Zaki, M. J. & Meira Jr, W. (2014). *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press. ISBN 978-0-521-76633-3.

Zhu, W.; Zeng, N.; Wang, N. et al. (2010). Sensitivity, specificity, accuracy, associated confidence interval and roc analysis with practical sas® implementations. *NESUG proceedings: health care and life sciences, Baltimore, Maryland*, pp. 1--9.

Apêndice A

Tabelas com Atributos

A.1 Lista de 15 atributos Finais Utilizados no Modelo

Tabela A.1: Tabela com 15 Atributos Selecionados

ID	*Atributo*
1	FORMACAO-ACADEMICA-TITULACAO-DOUTORADO-TEMPO
2	PREMIACAO
3	DIRECAO-E-ADMINISTRACAO
4	ENSINO-POS-GRADUACAO
5	LIVRO-PUBLICADO-OU-ORGANIZADO
6	CAPITULO-DE-LIVRO-PUBLICADO
7	ORGANIZACAO-DE-EVENTO
8	TRABALHO-EM-EVENTOS-NACIONAIS
9	TRABALHO-EM-EVENTOS-INTERNACIONAIS
10	ARTIGO-PUBLICADO-PONDERADO-NACIONAIS
11	ARTIGO-PUBLICADO-PONDERADO-INTERNACIONAIS
12	ORIENTACOES-CONCLUIDAS
13	PARTICIPACOES-EM-BANCA
14	ORIENTACAO-EM-ANDAMENTO
15	PARTICIPACAO-EM-PROJETO-PONDERADO

A.2 Lista de 16 Atributos

Tabela A.2: Tabela com 16 Atributos

ID	*Atributo*
1	FORMACAO-ACADEMICA-TITULACAO-DOUTORADO-TEMPO
2	PREMIACAO
3	DIRECAO-E-ADMINISTRACAO
4	ENSINO-POS-GRADUACAO
5	LIVRO-PUBLICADO-OU-ORGANIZADO
6	CAPITULO-DE-LIVRO-PUBLICADO
7	ORGANIZACAO-DE-EVENTO
8	TRABALHO-EM-EVENTOS-NACIONAIS
9	TRABALHO-EM-EVENTOS-INTERNACIONAIS
10	ARTIGO-PUBLICADO-PONDERADO-NACIONAIS
11	ARTIGO-PUBLICADO-PONDERADO-INTERNACIONAIS
12	ORIENTACOES-CONCLUIDAS
13	PARTICIPACOES-EM-BANCA
14	PARTICIPACAO-EM-EVENTOS
15	ORIENTACAO-EM-ANDAMENTO
16	PARTICIPACAO-EM-PROJETO-PONDERADO

A.3 Lista com os 59 Atributos Analisados Inicialmente

Tabela A.3: 59 Atributos Iniciais e Informações Complementares

ID	*Atributo*	*Média*	*Variância*	*Mediana*	*Ação*	*Ponderação*
1	FORMACAO-ACADEMICA-TITULACAO-DOUTORADO-TEMPO	44,05	17,1	16		
2	FORMACAO-ACADEMICA-TITULACAO-GRADUACAO	0,07	1,06	1	Retirar	
3	FORMACAO-ACADEMICA-TITULACAO-MESTRADO	0,17	1,04	1	Retirar	
4	FORMACAO-ACADEMICA-TITULACAO-DOUTORADO	0,02	1,02	1	Retirar	
5	FORMACAO-ACADEMICA-TITULACAO-POS-DOUTORADO	0,65	0,79	1	Retirar	
6	PREMIACAO	36,74	6,06	4		
7	DIRECAO-E-ADMINISTRACAO	112,01	7,79	4		
8	PESQUISA-E-DESENVOLVIMENTO	1,62	0,73	0	Retirar	
9	ENSINO	373,73	34,07	32	Retirar	
10	ENSINO-POS-GRADUACAO	109,45	14,72	13		
11	ENSINO-ESPECIALIZACAO	18,89	1,2	0	Retirar	
12	PARTICIPACAO-EM-PROJETO-GRADUACAO	448,39	17,02	11	Agrupar	1
13	PARTICIPACAO-EM-PROJETO-ESPECIALIZACAO	1,45	0,31	0	Agrupar	1
14	PARTICIPACAO-EM-PROJETO-MESTRADO-ACADEMICO	425,21	19,15	12	Agrupar	5
15	PARTICIPACAO-EM-PROJETO-MESTRADO-PROF	2,08	0,29	0	Agrupar	5
16	PARTICIPACAO-EM-PROJETO-DOUTORADO	400,38	13,98	7	Agrupar	10
17	LIVRO-PUBLICADO-OU-ORGANIZADO	33,26	2,93	1		
18	CAPITULO-DE-LIVRO-PUBLICADO	40,67	5,22	3		
19	TEXTO-EM-JORNAL-OU-REVISTA	28,72	1,14	0	Retirar	
20	SOFTWARE	19,68	1,78	0	Retirar	
21	PATENTE	0	0	0	Retirar	
22	PRODUTO-TECNOLOGICO	0,92	0,24	0	Retirar	
23	PROCESSOS-OU-TECNICAS	0,37	0,13	0	Retirar	
24	TRABALHO-TECNICO	649,66	10,51	0	Retirar	
25	APRESENTACAO-DE-TRABALHO	52,64	4,16	0	Retirar	
26	DESENVOLVIMENTO-DE-MATERIAL-DIDATICO-OU-INSTRUCIONAL	0,55	0,22	0	Retirar	
27	EDITORACAO	4,26	0,61	0	Retirar	
28	ORGANIZACAO-DE-EVENTO	282,3	9,3	4		
29	ORIENTACOES-CONCLUIDAS-PARA-MESTRADO	85,64	15,07	14	Agrupar	5
30	ORIENTACOES-CONCLUIDAS-PARA-DOUTORADO	22,13	4,68	3	Agrupar	10

Tabela A.3: 59 Atributos Iniciais e Informações Complementares

ID	*Atributo*	*Média*	*Variância*	*Mediana*	*Ação*	*Ponderação*
31	ORIENTACOES-CONCLUIDAS-PARA-POS-DOUTORADO	2,52	0,7	0	Agrupar	20
32	OUTRAS-ORIENTACOES-CONCLUIDAS	397,2	19,21	14	Agrupar	0
33	OUTRAS-ORIENTACOES-CONCLUIDAS-IC	109,29	8,97	6	Agrupar	1
34	PARTICIPACAO-EM-BANCA-DE-MESTRADO	362,83	22,24	18	Agrupar	5
35	PARTICIPACAO-EM-BANCA-DE-DOUTORADO	108,67	10,71	8	Agrupar	15
36	PARTICIPACAO-EM-BANCA-DE-EXAME-QUALIFICACAO	10,98	0,54	0	Agrupar	0
37	PARTICIPACAO-EM-BANCA-DE-APERFEICOAMENTO-ESPECIALIZACAO	76,25	6,5	3	Agrupar	1
38	PARTICIPACAO-EM-BANCA-DE-GRADUACAO	141,8	6,03	0	Agrupar	0
39	OUTRAS-PARTICIPACOES-EM-BANCA	1,35	0,14	0	Agrupar	0
40	PARTICIPACAO-EM-CONGRESSO	157,9	7,8	3	Agrupar	1
41	PARTICIPACAO-EM-FEIRA	0	0	0	Agrupar	1
42	PARTICIPACAO-EM-SEMINARIO	3,44	0,86	0	Agrupar	1
43	PARTICIPACAO-EM-SIMPOSIO	63,58	4,25	1	Agrupar	1
44	PARTICIPACAO-EM-OFICINA	17,08	1,37	0	Agrupar	1
45	PARTICIPACAO-EM-ENCONTRO	3,86	0,89	0	Agrupar	1
46	PARTICIPACAO-EM-EXPOSICAO	0,01	0,01	0	Agrupar	1
47	PARTICIPACAO-EM-OLIMPIADA	0	0	0	Agrupar	1
48	OUTRAS-PARTICIPACOES-EM-EVENTOS-CONGRESSOS	6,23	0,94	0	Agrupar	1
49	ORIENTACAO-EM-ANDAMENTO-DE-MESTRADO	1,36	0,65	0	Agrupar	5
50	ORIENTACAO-EM-ANDAMENTO-DE-DOUTORADO	3,76	1,7	1	Agrupar	10
51	ORIENTACAO-EM-ANDAMENTO-DE-POS-DOUTORADO	0,07	0,05	0	Agrupar	20
52	ORIENTACAO-EM-ANDAMENTO-DE-APERFEICOAMENTO-ESPECIALIZACAO	0,41	0,17	0	Agrupar	3
53	ORIENTACAO-EM-ANDAMENTO-DE-GRADUACAO	0	0	0	Agrupar	0
54	ORIENTACAO-EM-ANDAMENTO-DE-INICIACAO-CIENTIFICA	0,06	0,03	0	Agrupar	1
55	OUTRAS-ORIENTACOES-EM-ANDAMENTO	0,05	0,02	0	Agrupar	0
56	TRABALHO-EM-EVENTOS-NACIONAIS	599,1	28,12	21		
57	TRABALHO-EM-EVENTOS-INTERNACIONAIS	948	42,01	34		
58	ARTIGO-PUBLICADO-NACIONAIS	46,2	4,57	3	Retirar	
59	ARTIGO-PUBLICADO-INTERNACIONAIS	166,42	16,52	13	Retirar	

A.4 Agrupamento Ponderado

Tabela A.4: Tabela com Agrupamentos Ponderados

ID	*Atributo*	*Ponderação*	*Atributo Agrupado*
12	PARTICIPACAO-EM-PROJETO-GRADUACAO	1	PARTICIPACAO-EM-PROJETO-PONDERADO
13	PARTICIPACAO-EM-PROJETO-ESPECIALIZACAO	1	
14	PARTICIPACAO-EM-PROJETO-MESTRADO-ACADEMICO	5	
15	PARTICIPACAO-EM-PROJETO-MESTRADO-PROF	5	
16	PARTICIPACAO-EM-PROJETO-DOUTORADO	10	
29	ORIENTACOES-CONCLUIDAS-PARA-MESTRADO	5	ORIENTACOES-CONCLUIDAS
30	ORIENTACOES-CONCLUIDAS-PARA-DOUTORADO	10	
31	ORIENTACOES-CONCLUIDAS-PARA-POS-DOUTORADO	20	
32	OUTRAS-ORIENTACOES-CONCLUIDAS	0	
33	OUTRAS-ORIENTACOES-CONCLUIDAS-IC	1	
34	PARTICIPACAO-EM-BANCA-DE-MESTRADO	5	PARTICIPACOES-EM-BANCA
35	PARTICIPACAO-EM-BANCA-DE-DOUTORADO	15	
36	PARTICIPACAO-EM-BANCA-DE-EXAME-QUALIFICACAO	0	
37	PARTICIPACAO-EM-BANCA-DE-APERFEICOAMENTO-ESPECIALIZACAO	1	
38	PARTICIPACAO-EM-BANCA-DE-GRADUACAO	0	
39	OUTRAS-PARTICIPACOES-EM-BANCA	0	
40	PARTICIPACAO-EM-CONGRESSO	1	PARTICIPACAO-EM-EVENTOS
41	PARTICIPACAO-EM-FEIRA	1	
42	PARTICIPACAO-EM-SEMINARIO	1	
43	PARTICIPACAO-EM-SIMPOSIO	1	
44	PARTICIPACAO-EM-OFICINA	1	
45	PARTICIPACAO-EM-ENCONTRO	1	
46	PARTICIPACAO-EM-EXPOSICAO	1	
47	PARTICIPACAO-EM-OLIMPIADA	1	
48	OUTRAS-PARTICIPACOES-EM-EVENTOS-CONGRESSOS	1	
49	ORIENTACAO-EM-ANDAMENTO-DE-MESTRADO	5	ORIENTACAO-EM-ANDAMENTO
50	ORIENTACAO-EM-ANDAMENTO-DE-DOUTORADO	10	
51	ORIENTACAO-EM-ANDAMENTO-DE-POS-DOUTORADO	20	
52	ORIENTACAO-EM-ANDAMENTO-DE-APERFEICOAMENTO-ESPECIALIZACAO	3	

Tabela A.4: Tabela com Agrupamentos Ponderados

ID	*Atributo*	*Ponderação*	*Atributo Agrupado*
53	ORIENTACAO-EM-ANDAMENTO-DE-GRADUACAO	0	
54	ORIENTACAO-EM-ANDAMENTO-DE-INICIACAO-CIENTIFICA	1	
55	OUTRAS-ORIENTACOES-EM-ANDAMENTO	0	