DOI: 10.18605/2175-7275/cereus.v10n2p37-49

ARTIGO ORIGINAL

<<Recebido: 13/06/2018. Aceito: 13/07/2018 >>

REVISTA CEREUS ISSN: 2175-7275

Um Modelo para Predição de Desempenho de Pesquisadores na Grande Área de Conhecimento Ciência da Computação

A Performance Predictive Model for Computer Science Researchers

Hugo Andrei Mendes da Silva[1], Renê Rodrigues Veloso[2], Marcos Flávio Silveira Vasconcelos D'Angelo[3], João Batista Mendes [4]

RESUMO

Avaliar o desempenho de pesquisadores não é uma tarefa fácil, pois, ao ser realizada, essa avaliação deve levar em consideração a área de atuação dos pesquisadores e o contexto em que eles trabalham, uma vez que realidades distintas necessitam de avaliações específicas. Diferentes estudos têm sido feitos para direcionar essas avaliações, não sendo comum encontrar estudos focados em pesquisadores de um mesmo grupo e que trabalham em uma mesma área e região. Tais estudos são importantes para direcionar o pesquisador em relação aos seus pares e indicar o seu potencial dentro de seu grupo de pesquisa ou departamento. Nesse sentido, este trabalho tem o objetivo de apresentar uma análise da produtividade de pesquisadores da grande área da Ciência da Computação, que estão cadastrados na plataforma Lattes, além de desenvolver um modelo preditivo do desempenho de pesquisadores nessa área. O modelo apresentado obteve resultados relevantes, com acurácia média acima de 75%, demonstrando que é possível predizer computacionalmente o desempenho futuro dos pesquisadores. As contribuições deste trabalho podem auxiliar no processo de julgamento de concessão de bolsas pelas instituições de fomento à pesquisa, possibilitando também o direcionamento da carreira acadêmica dos pesquisadores dentro de um mesmo departamento das instituições de ensino.

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

ABSTRACT

Assess the scientific performance of researchers is not always a simple task. 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 evaluations. 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 the Computer Science area that are registered in the Lattes Platform, and develop a predictive model of the performance of researchers in this area. The selected predictive classification model obtained relevant results, with average accuracy above 75%. It can assist in the judging process by the research funding agencies, and to help in academic career of the Computer Science researchers.

Keywords: Researchers, Data Mining, Classification, Model.

[1] Mestre em Modelagem Computacional e Sistemas pela Universidade Estadual de Montes Claros. Atualmente é professor na Faculdade de Ciência e Tecnologia de Montes Claros e na Universidade Estadual de Montes Claros.

E-mail:

hugoandreimendesdasilva@gmail.com

[2] Doutor em Ciência da Computação pela UFMG. Professor efetivo da Universidade Estadual de Montes Claros e na Faculdade de Ciência e Tecnologia de Montes Claros .

[3] Doutor em Engenharia Elétrica pela UFMG. Professor efetivo da Universidade Estadual de Montes Claros.

[4] Doutor em Engenharia Elétrica pela UFMG. Professor efetivo da Universidade Estadual de Montes Claros.

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 (Abbasi et al., 2011).

Uma das formas de avaliar o desempenho de pesquisadores e de programas de pós-graduação, é a análise dos dados cadastrados na plataforma Lattes, que centraliza informações que servem como base para o fomento em pesquisas nacionais (CNPQ, 2015). Mena-Chalco (2013) afirmam que a plataforma Lattes é uma ferramenta de padrão nacional entre os pesquisadores e foi utilizada no tão conhecido projeto ScriptLattes (Mena-Chalco, 2015). Por conseguinte, a plataforma Lattes é uma fonte valiosa para a criação de modelos que auxiliem na tomada de decisão, utilizando informações contextualizadas dentro da realidade da pesquisa nacional.

1.1 Bolsa de Produtividade em Pesquisa

O Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), tem como uma de suas principais atribuições fomentar a pesquisa científica e tecnológica (CNPQ, 2015). Dentre as suas possibilidades de financiamento, a Bolsa de Produtividade em Pesquisa (PQ) é atribuída para doutores e possui um foco voltado para a qualidade das publicações, o que permite entender o processo de avaliação de pesquisadores no Brasil (Wainer & Vieira, 2013).

As bolsas PQs são requisitadas pelos pesquisadores de alta produtividade e, quando existe o recurso disponível, as bolsas são destinadas aos que atendem os critérios exigidos pelo Comitê Assessor (CA). O CA da Ciência da Computação (CA-CC), possui um critério de julgamento próprio[1]. Nesse critério existem requisitos necessários que são considerados para classificação do candidato à bolsa. É possível observar no critério de julgamento do CA-CC que, o único atributo com exigência mínima é o tempo de doutorado, sendo 3 anos para iniciar a bolsa como PQ 2 e, posteriormente, 8 anos para PQ 1. Os demais atributos avaliados são comparados e analisados para direcionar a classificação do pesquisador

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

dentro de cada nível.

Diante disso, esta pesquisa propõe a utilização das informações de pesquisadores cadastrados na plataforma Lattes que, com o auxílio de conhecimentos da Ciência da Informação, Ciência da Computação e da Modelagem Computacional, apresenta o desenvolvimento de um modelo preditivo para avaliação do potencial de pesquisadores. Devido a necessidade de um modelo específico para cada área, pois existem diferenças entre os critérios dos comitês de julgamento, optou-se por fechar o escopo do trabalho na grande área Ciência da Computação, focando em uma área de interesse dos autores.

A redação do presente artigo é dividida da seguinte forma: a metodologia é apresentada na Seção 2. Posteriormente, na Seção 3 são apresentados os resultados com os modelos preliminares e em sequência, ainda na Seção 3, é apresentado o modelo final e suas características. A Seção 4 discute os experimentos e o modelo de predição obtido. Finalmente, a conclusão do artigo é realizada na Seção 5.

2. MATERIAIS E MÉTODOS

De forma geral, a pesquisa apresentada neste trabalho foi dividida em duas etapas. Na primeira, foi feita a obtenção dos dados de currículos cadastrados na plataforma Lattes, em seguida foi realizado um estudo da estrutura dos currículos, resultando na seleção dos atributos e extração dos dados avaliados como relevantes para análises. Em seguida, na segunda etapa, foram realizados testes em diferentes tipos de modelos, sendo que um deles foi selecionado como resultado do trabalho.

Todos os algoritmos desenvolvidos utilizaram a linguagem de programação Python (versão 2.7), que está dentre os sistemas mais populares do mundo para desenvolvimento (Harrington, 2012). Para auxílio na implementação dos algoritmos, foi utilizada a biblioteca *Scikit Learn* (Scikit-Learn-Org, 2015).

A apresentação de resultados foi dividida em duas etapas, a primeira etapa consiste em identificar quais pesquisadores e os atributos que seriam analisados, e a segunda etapa consiste na criação, validação e ajustes do modelo preditivo.

2.1 Extração dos Atributos

Optou-se por analisar os pesquisadores da área Ciência da Computação que já receberam bolsa PQ, pois eles já foram classificados pelo comitê CA-CC na época que

receberam a bolsa, são considerados como referência nacionalmente e possuem o Lattes devidamente atualizado por exigência do CNPq. A lista de pesquisadores utilizada foi criada com os dados abertos do CNPq[2], base que disponibiliza qual pesquisador recebeu pagamento de bolsa PQ, o nível do pesquisador e a data inicial do financiamento. Em sequência, foram coletados os dados acadêmicos de cada pesquisador na plataforma Lattes e verificadas quais informações estão disponibilizadas na base através do *Document Type Definition*[3] (DTD). Após uma comparação com os critérios do CA-CC, foi realizada uma seleção funcional de atributos que resultou em uma base inicial com 1227 análises, referentes a 498 pesquisadores, realizadas entre 2004 e 2014 e uma lista de 59 atributos selecionados.

Para possibilitar os testes computacionais, foram selecionados somente os pesquisadores com os critérios mínimos, tanto para bolsa PQ 1 quanto para PQ 2. Dessa forma, foram mantidas somente as análises dos pesquisadores com tempo de doutorado igual ou superior a 8 anos na época da avaliação, o que resultou em 915 avaliações na base. As etapas realizadas inicialmente podem ser observadas na Figura 1.

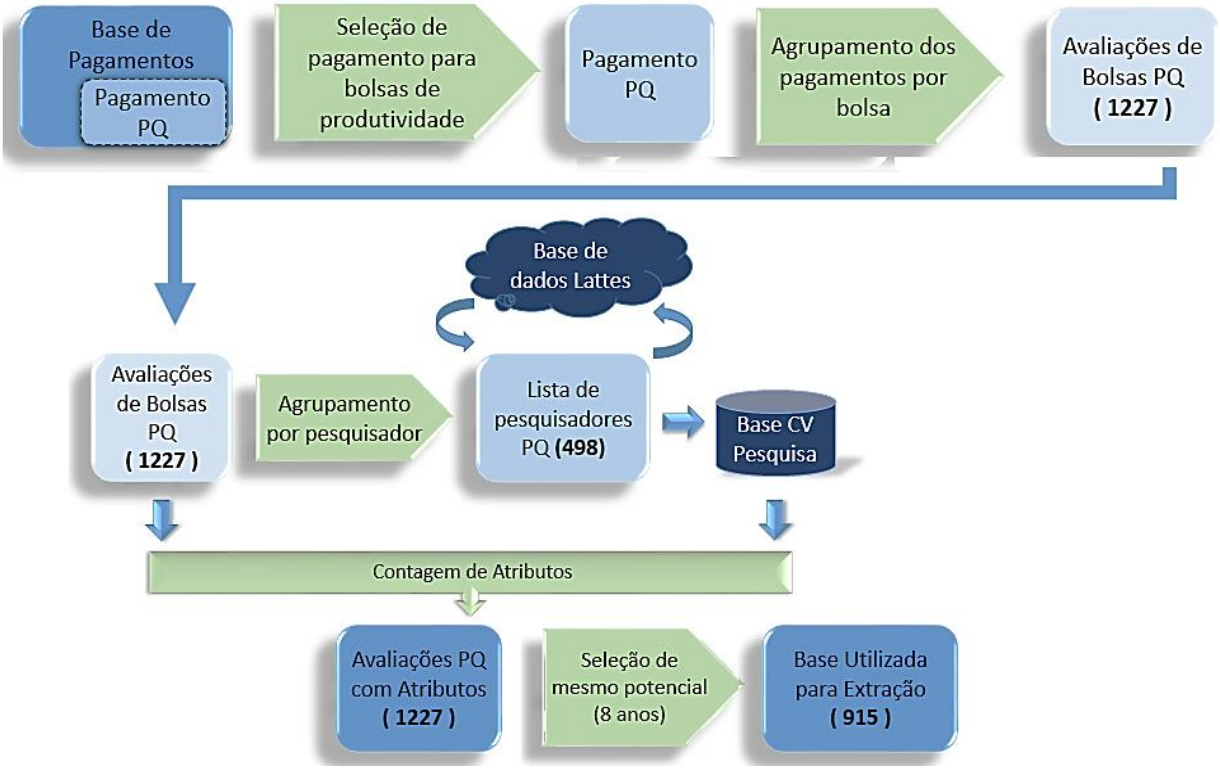

Figura 1 - Detalhamento da criação da base de dados com mesmo potencial.

[2] Disponível em: http://cnpq.br/dados_abertos

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

Com os 59 atributos selecionados inicialmente, foi contabilizada a produtividade científica dos pesquisadores nos 10 anos que precedem a análise do comitê, período avaliado para o nível mais elevado da bolsa (PQ 1A). Em sequência, foram realizados os seguintes testes para redução dos atributos: agrupamento ponderado, redução por baixa representatividade, ponderação de atributos e redução por correlação de variáveis.

Esses métodos, que consistem em avaliar a possibilidade de unificar atributos de forma ponderada ou eliminar algum, caso esse não seja determinante na análise, resultaram em uma lista reduzida com 15 atributos. Após a redução de atributos, foi realizado um agrupamento dos pesquisadores em cinco grupos através do algoritmo *K-means* para comparação dos grupos criados pelo algoritmo e a classificação realizada pelo CNPq. Essa comparação é apresentada na Tabela 1. Como é possível observar, os grupos criados pelo algoritmo têm concentrações em cada classe do CNPq. Essa concentração aconteceu de forma mais clara nas classes das extremidades, evidenciadas na Tabela (PQ 1A no Grupo 2 com 32 registros e PQ 2 no Grupo 5 com 323 registros). Como os grupos foram criados com os 15 atributos, indica que a escolha dos atributos foi adequada nessa etapa da pesquisa.

Tabela 1 - 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 a seleção dos atributos é através do gráfico de linhas paralelas que pode ser observado na Figura 2. Nesse gráfico, é possível observar que, com os 15 atributos, existe uma separação entre os valores médios normalizados entre os tipos de classificação, o que caracteriza grupos mais distintos, possibilitando resultados melhores para o modelo que será criado. Visto que os resultados foram satisfatórios, fixou-se esses 15 atributos que serão utilizados no modelo preditivo.

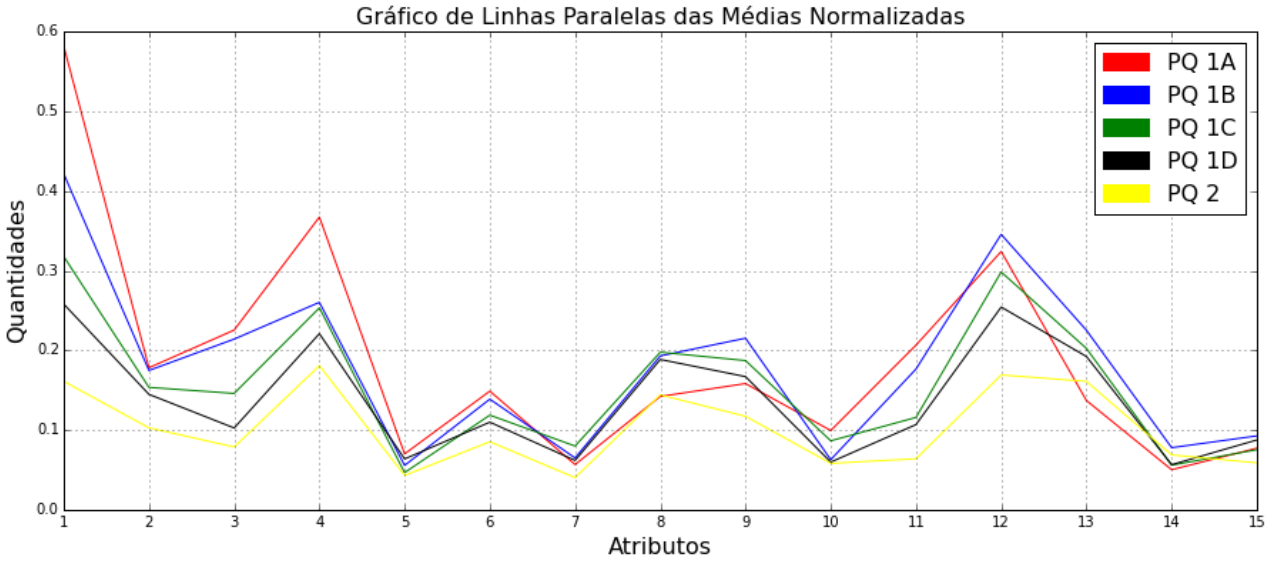

Figura 2 - Avaliação da classificação do CNPq com os atributos finais.

3. RESULTADOS

A amostra de dados utilizada nessa etapa foi a mesma da etapa anterior, acrescentados os registros de pesquisadores doutores que não receberam bolsa de produtividade PQ no período avaliado. Isso resultou em uma base com 1143 registros de pesquisadores. Após criação da amostra e seleção dos atributos, foram selecionados diferentes algoritmos de classificação, que foram propostos como base para modelos de classificação. Cada classificador teve seus parâmetros ajustados e, com os modelos ajustados, foi realizada uma comparação entre os resultados para identificar o melhor modelo, dentre os algoritmos analisados.

Para avaliar os classificadores, testar os ajustes dos parâmetros e comparação dos resultados nas próximas etapas, foi utilizado o método de validação cruzada *K-Fold* estratificada com k=10 (*10-fold*), permitindo comparar estatisticamente os resultados (Japkowicz & Shah, 2011).

Os cinco modelos criados com suas respectivas configurações de escolhas, que serão utilizados nas próximas etapas da pesquisa, podem ser observados na Tabela 2. Após os ajustes nos parâmetros dos modelos, foram analisados os resultados produzidos pelos cinco modelos para avaliação do desempenho de cada algoritmo de classificação. A Tabela 3 relaciona os resultados dos testes com a média das acurácias e qual a classificação dentre os cinco modelos, apresentado entre parênteses.

Tabela 2 - Modelos de Classificação com os Respectivos Parâmetros Ajustados

Modelo	Classificador	Abreviação	Parâmetros
Modelo 1	K-Nearest Neighbors	KNN	n neighbors = 5
Modelo 2	Random Forest	R FOREST	n estimators = 40
Modelo 3	Decision Tree	D TREE	default da biblioteca
Modelo 4	Naive Bayes	GNB	Classificador GaussianNB
Modelo 5	Support Vector Machines	SVM	kernel='linear'

Tabela 3 - 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**	**55,9**
Desvio Padrão	5,3	6	4	4,6	3
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

A princípio, os Modelos 1 e 2 foram os que resultaram em melhores acurácias nos testes. Porém, para escolher qual o melhor modelo dentre os avaliados para o problema, foi realizada uma comparação estatística entre os resultados. Como os resultados dos testes são cinco amostras independentes e não paramétricas, foi utilizado o teste de *Wilcoxon-Mann-Whitney* (Gold, 2007). Os resultados obtidos com o teste de hipótese entre o modelo de melhor desempenho (Modelo 2) e os demais modelos, pode ser observado na Tabela 4.

Tabela 4 - 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 pelos resultados obtidos com o teste de hipótese 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 a esses modelos. 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, 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 3 (8 dos 10 testes no *10-fold* estratificado).

4. DISCUSSÃO

Wanderley (2015), em seu trabalho de mestrado, desenvolveu 10 modelos de classificação por métricas de redes sociais (ARS) em uma abordagem semelhante ao desenvolvido neste trabalho. Entretanto, em sua dissertação, a autora teve como objetivo classificar o pesquisador entre duas possibilidades, que são *Com Bolsa PQ* e *Sem Bolsa PQ*, ou seja, uma classificação binária que obteve como seu principal resultado 74,65% na acurácia média, acompanhada por 78,24% na sensibilidade média e 72,83% na especificidade média do modelo escolhido.

O classificador apresentado neste trabalho obteve como resultado 64,95% na acurácia média do modelo *Random Forest* (Tabela 3). 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 ao 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 considerando uma classificação binária, semelhante à realizada em (Wanderley, 2015).

Dentre as possíveis métricas para mensurar o desempenho do 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 negativas em relação aos valores negativos reais, ou seja, quantidade de previstos como não bolsista em relação ao total que realmente não são bolsistas (Zhu, 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 5.

Tabela 5 - 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	93,8	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
Teste 11	92,05	96,11	75,86
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). Observa-se 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 pesquisa avaliar os pesquisadores que receberam bolsa nos últimos 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).

4.1 O Modelo de Predição

Ao final, com a definição do melhor modelo de classificação para o problema, foi realizada uma avaliação da capacidade preditiva, objetivo principal desse trabalho e, por último, foram realizados ajustes para possibilitar a utilização do modelo em diferentes grupos de pesquisa. Para isso, foi construída uma base iniciada nos 1143 pesquisadores, 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 que foram avaliados tinham no mínimo 5 anos de doutorado em 2005, ano que será utilizado como referência nesse teste. O resultado foi uma base para teste de predição com uma amostra de 350 pesquisadores.

Diferentemente das etapas anteriores, que consideraram a produtividade científica dos pesquisadores nos 10 anos que antecederam a avaliação pelo comitê CA-CC, nesta etapa foi proposta uma avaliação da predição do modelo através da análise da produtividade científica dos pesquisadores nos 10 anos anteriores à 2005 (1996 - 2005). Em resumo, o modelo fez um prognóstico do potencial de produtividade científica do pesquisador com uma antecedência de 6 a 9 anos.

Semelhante ao que foi realizado anteriormente, foram executados novos testes por validação cruzada com os dados da predição pelo Modelo 2 (*Random Forest*). As médias das acurácias nos testes de predição podem ser observados na Tabela 6. 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 6 - 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
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. Os pesquisadores avaliados em 2011, por exemplo, o CNPq utilizou dados de 2002 a 2011, sendo que no teste de predição foram utilizados os dados de 1996 a 2005. Isso demonstra que é possível predizer computacionalmente qual o desempenho futuro dos

pesquisadores em Ciência da Computação com o modelo preditivo desta pesquisa.

4.2 Ajuste do Modelo de Predição

Todo desenvolvimento do modelo preditivo foi baseado nas classificações realizadas pelo comitê CA-CC. Apesar disso, é comum as instituições de ensino avaliarem cientistas em potencial para inclusão em seus grupos de pesquisa, que não necessariamente recebem financiamento por bolsa PQ. Para contornar essa situação e tornar o modelo para qualquer realidade, 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.

Os registros da base foram agrupados e gerados novos rótulos para treinamento e classificação. 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 foram criados previamente seis grupos de pesquisadores na base através do algoritmo *K-means*. O agrupamento foi realizado na mesma base de dados utilizada anteriormente para predição futura, considerando os 15 atributos selecionados e os dados históricos de produtividade dos 10 anos que antecedem a avaliação do CNPq.

Com o resultado do agrupamento, foram realizados os mesmos testes usando 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. O resultado com as médias das acurácias nos testes para o algoritmo *Random Forest* ajustado com agrupamento prévio, é apresentado na Tabela 7.

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. Entretanto, ao comparar os dois métodos de predição futura (Tabelas 6 e 7), houve uma variação de 62,11% para 77,67%, uma diferença alta de 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.

Tabela 7 - 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

5. CONSIDERAÇÕES FINAIS

O presente trabalho apresentou uma análise da produtividade de pesquisadores da grande área da Ciência da Computação, que estão cadastrados na plataforma Lattes, além da extração dos atributos mais relevantes para a construção de um modelo preditivo do desempenho futuro desses pesquisadores. Tendo como base o algoritmo *Random Forest*, o modelo apresentado obteve resultados relevantes em relação ao estado-da-arte, com acurácia média acima de 75%.

Como trabalhos futuros, pretende-se desenvolver novos modelos para outras áreas, criar um direcionamento preciso ao pesquisador, mostrando os atributos que o mesmo precisa evoluir para crescimento na carreira, e possíveis melhorias no classificador com mais informações e análises, como identificar quem foram os orientadores do pesquisador durante sua vida acadêmica.

REFERÊNCIAS

ABBASI, A.; ALTMANN; J. & HOSSAIN, L. **Identifying the effects of co-authorship networks on theperformance of scholars: A correlation and regression analysis of performance measures and social network analysis measures**, Journal of Informetrics, vol 5, 594-607, 2011.

CNPq, **Conselho Nacional de Desenvolvimento Cientifico e Tecnológico**, Disponível em: <http://www.cnpq.br>. Acesso em: 16 de ago. 2015.

GOLD, A. **Understanding the mann-whitney test**, Journal of Property Tax Assessment

and Administration, UNIVERSITY OF ULSTER, vol 4, n3, 55, 2007.

HARRINGTON, P. **Machine learning in action**, Manning Greenwich, CT, vol 5, ISBN 9781617290183, 2012.

JAPKOWICZ, N.; SHAH, M. **Evaluating learning algorithms: a classification perspective**, Cambridge University Press, ISBN 978-0-521-19600-0, 2011.

MENA-CHALCO, J. P.; JÚNIOR, C. **ScriptLattes**, Disponível em: <http://scriptlattes.sourceforge.net>. Acesso em: 16 de ago. 2015.

MENA-CHALCO, J. P.; JÚNIOR, C. **Prospecção de dados acadêmicos de currículos Lattes através de Scriptlattes, Bibliometria e Cientometria: reflexões teóricas e interfaces**. São Carlos: Pedro & João, 109-128, 2013.

SCIKIT-LEARN-ORG, **scikit-learn, Machine Learning in Python**, Disponível em: <http://scikit-learn.org>. Acesso em: 16 de ago. 2015.

WAINER, J; VIEIRA, P. **Avaliação de bolsas de produtividade do CNPq e medidas bibliométricas: correlações para todas as grandes áreas**, Perspectivas em Ciência da Informação, vol 18, n2, 60-70, 2013.

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

ZHU, W. el al **Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS R ? implementations**, NESUG proceedings: health care and life sciences, Baltimore, Maryland, 1-9, 2010.