Utilizando Regras de Associação para Delinear o Perfil Feminino em Ciência da Computação

Joyce Quintino[1], **Carina T. Oliveira**[1], **Mauro Oliveira**[1]

[1]Laboratório de Redes de Computadores e Sistemas (LAR)
Instituto Federal de Educação, Ciência e Tecnologia do Ceará – Campus Aracati
Aracati – CE – Brasil

joycequintino11@gmail.com,(carina.oliveira,mauro)@ifce.edu.br

Abstract. *In recent years, there has been a significant increase in female participation in undergraduate courses in Brazil. At the same time, the number of evasions has increased, especially in courses such as Computer Science. This situation has alerted to the need for new solutions that contribute to the decision making of professors and educational managers. This work makes use of Association Rules to find correlations between data from female students of the Bachelor of Computer Science at the IFCE Campus Aracati. For instance, the results point out the main characteristics of female students with low performance, such as the study shift, the family income, the mother's educational level and the number of children.*

Resumo. *Nos últimos anos, houve um aumento significativo da participação feminina em cursos de graduação no Brasil. Paralelamente, também aumentou o número de evasões, principalmente em cursos como Ciência da Computação. Tal situação tem alertado para a necessidade de novas soluções que contribuam para a tomada de decisão dos professores e gestores educacionais. Este trabalho faz uso de Regras de Associação para encontrar correlações entre os dados de estudantes mulheres do curso de Bacharelado em Ciência da Computação do IFCE Campus Aracati. Os resultados apontam, por exemplo, as principais características de alunas com baixo desempenho, tais como turno de estudo, renda familiar, grau de instrução da mãe e número de filhos.*

1. Introdução

No Brasil, o número de estudantes mulheres nos cursos superiores tem crescido nos últimos anos. Dados do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) mostram que, nos últimos dez anos, de um total de 6 milhões de matrículas, aproximadamente 3,4 milhões foram de mulheres [INEP 2018]. Ainda de acordo com o INEP, em 2013, o percentual médio de ingressantes mulheres foi de 55% em cursos de graduação presenciais.

De acordo com o Censo da Educação Superior (CES), as mulheres representam 60% dos concludentes nos cursos superiores no Brasil em 2015. No entanto, quando são considerados apenas cursos relacionados às ciências, a participação feminina cai para 41%. Dados do INEP apontam que o ingresso e concluintes do gênero feminino na área de Ciências, Matemática e Computação é inferior ao masculino em todas as regiões do país [INEP 2018].

Segundo Miliszewska e Moore [Moliszewska and Moore 2010], apesar das iniciativas existentes para abordar a escassez feminina na área da Tecnologia da Informação e Comunicação (TIC), o problema persiste. Conforme Moreira *et al.* [Moreira et al. 2014], o maior número de evasões ainda está relacionado ao público feminino, principalmente nos cursos ligados a área de TIC, que apresentam o menor percentual de matrículas de estudantes do sexo feminino.

O cenário do curso superior de Bacharelado em Ciência da Computação do IFCE Campus Aracati, que teve início em 2012, não é diferente. Segundo dados da própria instituição, em 2015, a taxa de desistência feminina foi superior a 40% [Q-ACADÊMICO-IFCE 2018]. As informações do curso podem ser obtidas através do sistema acadêmico da própria instituição. Antes do presente trabalho, a análise dos dados era realizada manualmente, tornando-se um processo demorado. Além disso, estava sujeito à falhas, podendo haver a ocorrência de erros na interpretação e, posteriormente, na tomada de decisão de professores e gestores educacionais do IFCE Aracati.

Assim, dada a importância da análise de dados das estudantes mulheres de cursos de TIC, fundamental no apoio à tomada de decisão de professores e gestores educacionais, destaca-se a necessidade do uso de ferramentas computacionais que possibilitem uma análise mais rápida e organizada dos dados. Ao final, os resultados obtidos da análise permitirão o desenvolvimento de novas metodologias e iniciativas de incentivo para redução dos índices de evasões de mulheres na área de TIC.

Neste trabalho, usa-se Regras de Associação da Mineração de Dados para extração das características das estudantes que cursaram a disciplina de Introdução à Programação do curso de Bacharelado em Ciência da Computação do IFCE Campus Aracati. As regras representam implicações nos dados e são importantes quando necessário visualizar as correlações entre os dados. De acordo com o objetivo do trabalho, usamos a linguagem de programação R e, aplicando o algoritmo Apriori, foi possível obter quais características são frequentes em estudantes de acordo com seu desempenho acadêmico. As características mais encontradas das estudantes foram: escola de origem (pública ou privada), turno de estudo, renda familiar, grau de instrução da mãe e número de filhos. Devido a ocorrência do abandono nos turnos noturno e vespertino, podemos também visualizar o quantitativo de mulheres oriundas da escola pública que cursaram essa disciplina nos diferentes turnos.

2. Algoritmo

2.1. Regras de Associação

Para extração de regras precisamos de um conjunto de itens definido aqui como I = $\{i_1, i_2, ..., i_n\}$. Tendo T = $\{i_1, i_2, ..., i_l\}$ o conjunto de instâncias, dizemos que T $\subseteq$ I, tal que $i_l \subset I$ e $l <= n$. Por exemplo, sendo A e B $\subseteq$ de I, temos que uma regra de associação é do tipo A $\rightarrow$ B se A $\cap$ B = $\emptyset$ [Silva et al. 2016].

2.2. Apriori

O Algoritmo Apriori possui duas fases para seu funcionamento. Na primeira, calcula-se a quantidade de vezes que cada item ocorre na base de dados. Esses itens com a frequência maior que o minímo estabelecido na entrada são selecionados e combinados, assim, calculando o suporte dado por $suporte_{regra}(A \cup B)$, visto na Equação 1. O suporte

indica a frequência dos itens que aparecerem juntos em transições individuais no conjunto de dados. Temos $cont(A \cup B)$ representando a contagem dos elementos que aparecem juntos de A e B e $cont(A)$ a quantidade de transições de A.

$$suporte_{regra}(A \cup B) = \frac{cont(A \cup B)}{cont(A)} \quad (1)$$

Este processo se repete até que nenhum outro conjunto frequente seja criado. Na segunda fase, o algoritmo busca as regras de associação combinando os itens entre si e calculando a confiança dada por $confian\text{ç}a_{regra}(A \cup B)$, vista na Equação 2. A confiança pode ser interpretada como uma probabilidade, onde o $suporte_{regra}(A \cup B)$ equivale a A $\rightarrow$ B ocorrer, dado que $suporte_{itemset}(A)$ ocorre, ou seja, a ocorrência de A [Silva et al. 2016].

$$confian\text{ç}a_{regra}(A \cup B) = \frac{suporte_{regra}(A \cup B)}{suporte(A)} \quad (2)$$

3. Resultados

Os resultados deste trabalho utilizam a base de dados das estudantes das turmas 2012.2 até 2017.1 do curso de Bacharelado em Ciência da Computação do IFCE Campus Aracati. O curso tem duração de 4 anos e sua oferta é semestral, com 40 vagas disponíveis, sendo a maioria ocupada por homens [Monteiro et al. 2017] e com turnos vespertino e noturno intercalados a cada semestre.

No intuito de descobrir as principais características correlacionadas das estudantes que cursaram a disciplina de Introdução à Programação, que é ofertada no 1º semestre, usou-se o algoritmo de Regras de Associação Apriori do pacote R arules. Vale destacar que é importante a realização da análise nos dados logo no início do curso, pois, com base nos resultados, iniciativas de incentivo à permanência e êxito das alunas podem ser tomadas.

Para os *scripts* em R, utilizou-se a ferramenta RStudio juntamente com o pacote R sqldf. Este último possibilita que consultas SQL sejam realizadas dentro dos objetos DataFrame do R, evitando a utilização de um SGBD para armazenamento dos dados. A seguir, são apresentadas descrições das principais regras e do pré-processamento de dados.

3.1. Preparação de dados

À princípio, os dados referentes às estudantes foram disponibilizados em forma de planilhas pelo setor de controle acadêmico, porém separadamente, pois algumas informações o sistema acadêmico da instituição não permite que sejam coletadas juntas. Após a coleta, os dados foram submetidos à fase de pré-processamento para seleção, limpeza e criação de atributos.

A base de dados transformada em um objeto R DataFrame para aplicação do algoritmo foi composta pelos seguintes atributos: *renda familiar, escola de origem, grau de instrução da mãe e do pai, estado civil, nº de filhos, descrição da situação da matrícula,*

coeficiente e *situação*. O atributo *situação* foi criado com base na situação das mulheres em relação à disciplina de Introdução à Programação, considerando a primeira vez que a mesma foi cursada, assim evitando a ocorrência de ambiguidade na base. Criou-se também o atributo *coeficiente*, referente ao coeficiente de rendimento das estudantes. Para um processamento mais significativo do algoritmo, esse atributo foi criado seguindo os intervalos: 0-6, 6-7.9 e 8-10 rotulados com ruim, médio e bom, respectivamente.

3.2. Regras Extraídas

As regras encontradas são no formato de implicação, representando a ocorrência das características juntas. O Apriori encontrou as seguintes características com mais frequência: *renda familiar, grau de instrução da mãe, escola de origem, turno* e *coeficiente ruim*. A regra $\{escola = pública, coeficiente = ruim\} \Rightarrow \{turno = noturno\}$ significa que estudantes oriundas de escolas públicas e que possuem um coeficiente ruim, frequentemente estão no turno da noite. Por outro lado, a regra $\{escola = pública, coeficiente = ruim\} \Rightarrow \{turno = vespertino\}$ mostra que esse fato também ocorre no turno vespertino. A Figura 1 apresenta uma representação gráfica do quantitativo de mulheres nos diferentes turnos, com destaque para o turno vespertino com um maior número de estudantes com coeficiente ruim. Isso mostra que mesmo o fato ocorrendo nos dois turnos, as meninas que entram no turno vespertino tendem a ter o coeficiente ruim.

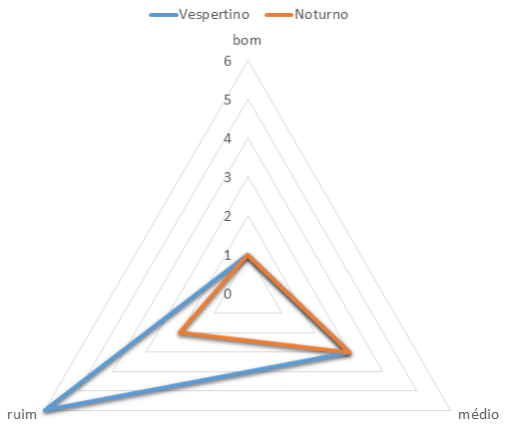

Figura 1. Quantitativo de estudantes mulheres do Bacharelado em Ciência da Computação do IFCE Aracati oriundas de escola pública nos turnos vespertino e noturno.

Outras regras mostram que a *situação financeira* e *grau de instrução da mãe* também encontram-se relacionados ao coeficiente ruim. Juntas, essas regras influenciam para o abandono. A regra $\{renda = até1salário\} \Rightarrow \{coeficiente = ruim\}$ mostra que estudantes com renda familiar até um salário mínimo, frequentemente, possuem o coeficiente ruim. A regra $\{graudeinstruçãodamãe = fundamentalincompleto, nºdefilhos = 0, renda = até1salário\} \Rightarrow \{coeficiente = ruim\}$ também mostra a relação entre a renda familiar até um salário mínimo, grau de instrução da mãe correspondente ao ensino fundamental incompleto e número de filhos igual a zero, todas influenciando para um coeficiente ruim.

Apesar da relação com o número de filhos igual a zero, fator importante a ser observado, as estudantes sentem dificuldades nas disciplinas durante o curso. De acordo com

a regra $\{escola = pública\} \Rightarrow \{coeficiente = ruim\}$, as estudantes de escolas públicas tendem a ter um coeficiente ruim. Isso mostra que outras questões estão relacionadas, por exemplo, o ensino básico que não é de qualidade e, por isso, as estudantes não têm uma base adequada de matemática, português e outras disciplinas importantes para acompanhar o curso.

4. Conclusões

Neste artigo é apresentada uma pesquisa referente aos dados das estudantes que cursaram a disciplina de Introdução à Programação do curso de Bacharelado em Ciência da Computação do IFCE Campus Aracati. A mesma foi realizada por meio do uso de Regras de Associação, cujo objetivo foi extrair as principais características que influenciam na evasão. Com a aplicação do algoritmo Apriori, as características de turno, renda familiar, coeficiente ruim, grau de instrução da mãe, quantidade de filhos e escola de origem foram as mais encontradas como fatores implicantes na desistência das estudantes. Além disso, mostrou-se que o turno da tarde apresenta o maior número de meninas oriundas da escola pública e que possuem um coeficiente ruim.

Como trabalhos futuros, pretende-se usar outras tarefas de Mineração de Dados na análise exploratória para melhorar os indicadores. Além disso, pode ocorrer o enriquecimento da base utilizada para facilitar a possível predição de evasão feminina.

Agradecimentos

Os autores agradecem ao IFCE (PROEXT e PRPI) e FUNCAP (BPI/2016-2018) pelo financiamento do trabalho.

Referências

INEP (2018). Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). Disponível em `http://www.inep.gov.br/`.

Moliszewska, I. and Moore, A. (2010). Encouraging Girls to Consider a Career in ICT: A Review of Strategies. *Journal of Information Technology Education Innovations in Practice*, 9.

Monteiro, R. d. S., Marinho, J. M., Braga, R. B., Viana, M. d. N., and de Oliveira, C. T. (2017). Delineando o Perfil Feminino Discente do Bacharelado em Ciência da Computação do IFCE Campus Aracati. In *XI Women in Information Technology (WIT)/CSBC*.

Moreira, J., Mattos, G., and Reis, L. (2014). Um Panorama da Presença Feminina na Ciência da Computação. *Encontro Internacional da Rede Feminista Norte e Nordeste de Estudos e Pesquisa sobre a Mulher e Relações de Gênero*.

Q-ACADÊMICO-IFCE (2018). Sistema Q-ACADÊMICO do IFCE. Disponível em `https://qacademico.ifce.edu.br/`.

Silva, A. L. D., Peres, M. S., and C., B. (2016). *Introdução à Mineração de Dados Com Aplicações em R*. Elsevier.