

## **Aplicação de Mineração de Dados no Censo da Educação Superior: Estudo de Caso na Base de Docentes de 2014**

**Bruno Boavetura de Oliveira Lacerda<sup>1</sup>, Manuela Amaral de Araújo<sup>2</sup>,  
Saionara da Silva Araújo<sup>3</sup>, Pablo Freire Matos<sup>4</sup>**

<sup>1</sup>Discente de graduação em Sistemas de Informação - IFBA. e-mail: brunoboavetura@gmail.com; <sup>2</sup> Discente de graduação em Bacharelado em Sistemas de Informação - IFBA. e-mail: mn21922192@gmail.com; <sup>3</sup> Discente de graduação em Bacharelado em Sistemas de Informação. e-mail: narabdo@gmail.com; <sup>4</sup> Professor do curso Bacharelado em Sistemas de Informação. e-mail: pablofmatos@ifba.edu.br

1 **RESUMO:** Esse artigo tem o objetivo de apresentar uma aplicação da descoberta de  
2 conhecimento através da mineração de dados, utilizando o algoritmo Apriori, a ferramenta  
3 Weka e soluções/problema envolvendo o tratamento da base de dados. Também são descritos  
4 trabalhos correlatos que abordam o tema em questão. Como base de dados, foi selecionada  
5 uma que passou por avaliações, pesquisas e exames realizados pelo INEP (Instituto Nacional  
6 de Estudos e Pesquisas Educacionais Anísio Teixeira) com relação aos dados do Censo da  
7 Educação Superior. Finalmente, é apresentado o resultado das regras de associação obtidas  
8 por meio da mineração de dados.

9 **Palavras-chave:** análise de dados, descoberta de informação, regras de associação

### **Data Mining on Higher Education Census: Case Study on 2014 Teacher's Database**

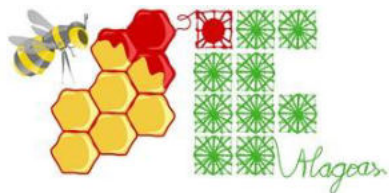
14 **ABSTRACT:** This paper aims at presenting an application of knowledge discovery through  
15 data mining, using the Apriori algorithm, the Weka tool and solutions/problem involving the  
16 database's treatment. It also describes related works on this topic. It was selected a database  
17 that has passed by ratings, research and tests, carried out by INEP (National Institute of  
18 Educational Studies Anísio Teixeira) with respect to the Higher Education Census data.  
19 Finally, it shows the result of the association rules obtained using the data mining process.

20 **KEYWORDS:** data analysis, information discovery, association rules

### **INTRODUÇÃO**

23 Com o avanço e redução de custos da tecnologia, é cada vez mais rápido o aumento da  
24 quantidade de dados gerados. Alguns desses dados poderiam, se bem analisados, agregar  
25 informações importantes para empresas, governos e pessoas. A mineração de dados é uma  
26 tarefa muito utilizada para essa finalidade, pois, como parte de um processo de descoberta de  
27 conhecimento (KDD), busca gerar informações baseadas em dados gerados por determinado  
28 “meio” (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

29 Alguns trabalhos com objetivos semelhantes ao proposto neste artigo podem ser  
30 observados na literatura. Fonseca e Namen (2016) utilizaram a técnica de classificação,



31 através do algoritmo Naïve Bayes, para relacionar características de professores de  
32 matemática com o aprendizado dos alunos na disciplina e identificou fatores positivos e  
33 negativos em relação ao aprendizado dos alunos. Alguns dos pontos positivos foram o alto  
34 percentual de cumprimento do planejamento das aulas, baixo índice de faltas dos professores  
35 e assiduidade dos alunos. Já como pontos negativos pode-se citar a desvalorização salarial do  
36 professor, a grande quantidade de falta dos alunos.

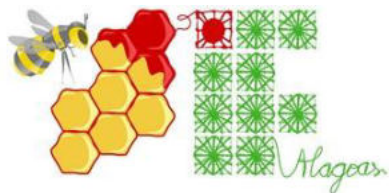
37 Outro trabalho nesse campo buscou apresentar uma análise sobre alunos que  
38 ingressavam ou saíam das Instituições de Ensino Superior, aplicando técnicas de associação,  
39 classificação e agrupamento (PASTA, 2011). Como resultados obtidos, pode-se observar que  
40 a maioria dos ingressantes veio de escolas públicas, e escolheu a instituição por conta de sua  
41 localização, além de pretender abrir seu próprio negócio. Em relação ao que sai da IES, a  
42 escolha da instituição foi influenciada pela matriz curricular dos cursos pretendidos. Um  
43 ponto importante a ser observado pela gestão da instituição é que a maioria dos egressos não  
44 indicaria a instituição.

45 Um estudo sobre o aprendizado de língua portuguesa foi o foco de Namen e Soares  
46 (2011). Nesse trabalho, os autores trouxeram associações obtidas na aplicação do algoritmo  
47 Apriori sobre uma base de dados de uma escola de ensino fundamental do Rio de Janeiro.  
48 Com os resultados da pesquisa observou-se que a falta de incentivo dos pais, ou caso o aluno  
49 exerça alguma atividade de trabalho são fatores que, dentre outros, afetam o aprendizado do  
50 estudante.

51 Nesse sentido, este trabalho busca demonstrar os resultados obtidos a partir da análise  
52 da base de dados do Censo da Educação Superior disponibilizada pelo INEP (Instituto  
53 Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira) com aplicação de algoritmo  
54 de regras de associação e com a utilização da ferramenta Weka (2016).

55 Com a aplicação de tarefas de mineração de dados, padrões são descobertos de forma  
56 mais automática, mais rápida e com maior grau de certeza. A escolha do método de mineração  
57 utilizado é fundamental para a geração de conhecimento. Dentre esses métodos, pode-se citar  
58 alguns: Classificação, Agrupamento, Associação e Regressão (CAMARGO *et al.*, 2016).

59 Para o trabalho proposto, a técnica escolhida foi a de associação, indicada quando se  
60 deseja descobrir regras que indiquem as relações entre os atributos dos dados informados  
61 (SILVA, 2005). As associações descobertas, também chamadas de regras, são no formato X



62  $\Rightarrow Y$ , indicando que quando ocorre X também ocorre Y, sendo o X o determinante da regra e  
63 Y o resultante (SILVA, 2004). Para gerar essas regras, as medidas usadas são, basicamente, o  
64 suporte e a confiança. O primeiro consiste no cálculo de quantas transações onde ocorre X e Y  
65 estão no conjunto completo de dados, enquanto o segundo é dado pelo número de ocorrências  
66 de X e Y sempre que ocorre X (SILVA, 2005). Dentre os algoritmos de mineração de dados,  
67 o utilizado neste trabalho foi o Apriori, que identifica regras de associação entre atributos da  
68 base de dados (SILVA, 2005).

69 Várias são as ferramentas desenvolvidas para tarefas de mineração. A ferramenta Weka  
70 foi a escolhida para a realização desse trabalho. Weka é uma ferramenta para aprendizagem  
71 de máquina e permite que o computador analise grande quantidade de dados e decida, por  
72 meio de algoritmos, quais as informações mais relevantes.

73

## 74 MATERIAL E MÉTODOS

75

### 76 Proposta do trabalho

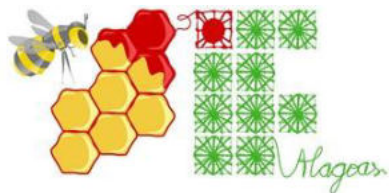
77 Este trabalho tem o objetivo de, através do conceito sobre o processo de descoberta de  
78 conhecimento, aplicar a teoria no desenvolvimento de um projeto prático utilizando a  
79 ferramenta Weka e a Associação como técnica de mineração de dados. Para isso, foi preciso  
80 realizar as seguir etapas:

- 81 1. Realizar uma busca por uma base de dados recente (a partir de 2010) com uma  
82 quantidade razoável de dados;
- 83 2. Realizar seleção e transformação de atributos para obter regras de associação  
84 relevantes;
- 85 3. Utilizar a ferramenta Weka para analisar a base de dados.

86 Para escolher uma base de dados, diversas consultas foram feitas sem sucesso, pois a  
87 maior parte das bases de dados são proprietárias e poucas são disponibilizadas ao público em  
88 geral, devido às questões de segurança e sigilo. Após conseguir bases no site do INEP -  
89 Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira<sup>1</sup>, foram necessários  
90 vários testes, seleção de atributos e transformação em diversos bancos de dados até chegar ao

---

<sup>1</sup> INEP: <http://portal.inep.gov.br>



91 objetivo proposto pelo trabalho que é: selecionar uma base de dados a fim de encontrar  
92 informações relevantes não descobertas outrora. Esse processo foi o mais custoso e  
93 trabalhoso, pois as bases encontradas muitas vezes retornavam associações óbvias e que não  
94 seriam relevantes para nenhuma tomada de decisão.

95

## 96 **Processamento e Análise dos Dados**

97 O processamento e análise dos dados foram realizados utilizando as etapas do processo  
98 de KDD proposto por Fayyad, Piatetsky-Shapiro e Smyth (1996).

99

## 100 **Seleção**

101 A base de dados escolhida foi a do Censo da Educação Superior<sup>2</sup>, realizado pelo INEP,  
102 disponível em 2014. Como o censo se trata de um arquivo composto de várias bases, foi  
103 escolhida a que lista os docentes do ensino superior em todo o Brasil. A mais atual encontrada  
104 foi do ano de 2014 e possui 39.659 instâncias com 50 atributos.

105

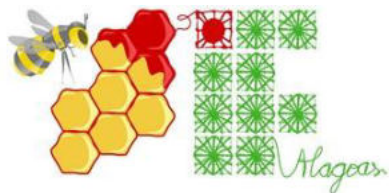
## 106 **Pré-processamento**

107 São necessárias algumas configurações na interface do Weka para que os arquivos  
108 possam ser “lidos”. Um dos erros frequentes ao tentar abrir um arquivo é a codificação de  
109 caracteres. Quando o arquivo possui acentos e caracteres especiais é preciso realizar a conversão  
110 para a codificação UTF-8. No caso deste trabalho, foi utilizada a ferramenta *iconv*, disponível na  
111 grande maioria dos sistemas UNIX, e presente em todas as distribuições do sistema Linux, sendo  
112 o seu uso discriminado a seguir: na linha de comando “`iconv -f ISO_8859-1 -t utf-8`  
113 `DM_DOCENTE.CSV -o docentes-superior-utf8.csv`” (sem aspas), onde o  
114 “DM\_DOCENTE.CSV” foi o arquivo da base de dados antes da formatação e o “`docentes-`  
115 `superior-utf8.csv`” o nome do arquivo após a conversão.

116 Um dos requisitos para executar o algoritmo *Apriori* no Weka é não ter atributos  
117 numéricos. Logo, foi necessária a realização de uma categorização de vários dos atributos que  
118 continham valores 0 e 1 para “não” e “sim”, respectivamente. Para essa tarefa foi usado um

---

<sup>2</sup> Censo da Educação Superior:  
[http://download.inep.gov.br/microdados/microdados\\_censo\\_superior\\_2014.zip](http://download.inep.gov.br/microdados/microdados_censo_superior_2014.zip)



119 *script* na linguagem de programação Python (Figura 1) que também foi aproveitado para  
120 remover atributos que representavam simplesmente uma discretização de outros.

121 O *script* trata os dados seguindo a seguinte sequência: para cada registro, imprime na  
122 tela a instância já tratada, com os atributos não selecionados já removidos e categoriza os  
123 atributos conforme a necessidade. Os valores dos atributos numéricos 0 e 1 são substituídos  
124 por N e S, respectivamente (Linhas 9 a 12, Figura 1). Os atributos categóricos não são  
125 alterados. Através do operador de redirecionamento “>” (sem as aspas), a saída do *script* é  
126 escrita em um novo arquivo que conterá os dados já tratados.

127

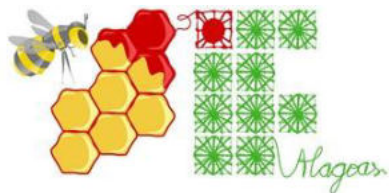
```
1 # uso do script:
2 # python3 tratar-dados.py > NOME_DO_NOVO_ARQUIVO
3
4 import csv
5
6 def categorizar (valor):
7     try:
8         teste = int (valor)
9         if (teste == 0):
10            texto = "N"
11        else:
12            texto = "S"
13    except:
14        texto = valor
15
16    return texto
17
18 if __name__ == '__main__':
19     # abre o arquivo CSV
20     with open('docentes-superior-utf8.csv', 'r') as f:
21         # cria o objeto que vai ler os registros com separador "|"
22         reader = csv.reader(f, delimiter='|')
23         # para cada linha, gera somente novos registros com os
24         # atributos selecionados, com separador ";"
25         for linha in reader:
26
27             # Campos usados (base zero - lista iniciada em 0)
28             for i in [1,3,5,6,10,12,14,16,22,25,37,38,39,40,41,42,43,44,45,46,47,48]:
29                 valor = categorizar(linha[i])
30                 # Usa ? em valores nulos
31                 print ('?' if (valor == '') else valor, end=';')
32         print ('')
```

128  
129 **Figura 1.** Script Python (comentado) utilizado para tratamento dos dados. IFBA, 2016.

### 130 131 **Formatação**

132 Ao tentar abrir a base de dados no Weka, podem ocorrer algumas mensagens de erro,  
133 devido às configurações conflitantes com o padrão de geração do arquivo, já que a ferramenta  
134 prevê a vírgula (,) como separador de atributo e o script gerou arquivos com ponto e vírgula  
135 (;) como separador. Para isso, antes de clicar em abrir, é preciso visualizar a opção “*invoke*  
136 *options dialog*”, opção que vem, por padrão, desmarcada.

137 Caso não seja selecionada e o arquivo possuir erros de sintaxe, o programa não  
138 prossegue. Após clicar em abrir, será disponibilizada uma tela onde a opção



139 “enclosureCharacters” deve ser preenchida somente com “;” (sem as aspas) e o campo  
140 “fieldSeparator” deve ser preenchido com “;” (sem as aspas).

141

## 142 **Mineração de Dados**

143 Para conseguir obter regras de associação válidas, é necessário remover alguns atributos  
144 e testar os resultados com valores diferentes para o suporte e a confiança. O algoritmo não  
145 fica ativo para alguns atributos numéricos, pois ele só trabalha com dados discretos e  
146 nominais (SANTOS, 2005). Portanto, é preciso removê-los. A própria interface do Weka, ao  
147 abrir o banco de dados, disponibiliza informações sobre os atributos na página inicial. Dos 50  
148 atributos, foram selecionados 12, conforme mostrados na Tabela 1.

149

150 **Tabela 1.** Atributos selecionados no pré-processamento. IFBA, 2016.

Atributo	Significado
NO_IES	Nome da IES
DS_CATEGORIA_ADMINISTRATIVA	Nome da Categoria Administrativa
DS_ORGANIZACAO_ACADEMICA	Nome da Organização Acadêmica
IN_CAPITAL_IES	Informa se a IES (reitoria/sede administrativa) está localizada na capital
DS_SITUACAO_DOCENTE	Nome da situação do docente na IES
DS_ESCOLARIDADE_DOCENTE	Informa o nome do grau de escolaridade do docente
DS_REGIME_TRABALHO	Nome do regime de trabalho do docente
DS_SEXO_DOCENTE	Nome do sexo do docente
DS_COR_RACA_DOCENTE	Nome da cor/raça do docente
IN_ATU_EXTENSAO	Informa se o docente atua em atividades de extensão
IN_ATU_PESQUISA	Informa se o docente atua em pesquisa no âmbito de projetos e programas da IES
IN_BOLSA_PESQUISA	Informa se o docente possui bolsa de pesquisa

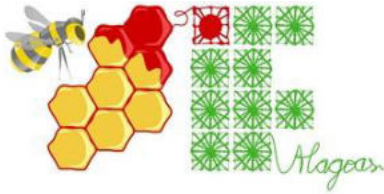
151  
152

153 O algoritmo Apriori é baseado em suporte e confiança, sendo de suma importância a  
154 escolha desses parâmetros para a descoberta de regras de associação relevantes. No Weka, a  
155 configuração do Apriori: foi realizada da seguinte maneira: lowerBoundMinSupport = 0.40,  
156 metricType = Confidence, minMetric = 0.99 e numRules = 30.

157

## 158 **RESULTADOS E DISCUSSÃO**

159 Através das regras obtidas (Figura 2), é possível demonstrar o quanto os professores do  
160 ensino superior em exercício pouco exercem atividades de pesquisa e extensão. Essa  
161 informação pode ser utilizada para que alguma medida seja realizada com o objetivo de  
162 incentivar e alterar essa realidade presente no ensino superior.



163

```
Associator output
Instances: 396595
Attributes: 12
NO_IES
DS_CATEGORIA_ADMINISTRATIVA
DS_ORGANIZACAO_ACADEMICA
IN_CAPITAL_IES
DS_SITUACAO_DOCENTE
DS_ESCOLARIDADE_DOCENTE
DS_REGIME_TRABALHO
DS_SEXO_DOCENTE
DS_COR_RACA_DOCENTE
IN_ATU_EXTENSAO
IN_ATU_PESQUISA
IN_BOLSA_PESQUISA

=== Associator model (full training set) ===

Apriori
=====
Minimum support: 0.4 (158638 instances)
Minimum metric <confidence>: 0.99
Number of cycles performed: 12

Generated sets of large itemsets:

Size of set of large itemsets L(1): 9
Size of set of large itemsets L(2): 10
Size of set of large itemsets L(3): 2

Best rules found:
1. IN_ATU_EXTENSAO=N 298463 ==> DS_SITUACAO_DOCENTE=Em exercicio 298463 <conf:(1)> lift:(1.03) lev:(0.03) [9946
2. IN_ATU_PESQUISA=N 277300 ==> DS_SITUACAO_DOCENTE=Em exercicio 277300 <conf:(1)> lift:(1.03) lev:(0.02) [9235
3. IN_ATU_EXTENSAO=N IN_ATU_PESQUISA=N 239989 ==> DS_SITUACAO_DOCENTE=Em exercicio 239989 <conf:(1)> lift:(1.03)
4. DS_SEXO_DOCENTE=Masculino IN_ATU_EXTENSAO=N 166992 ==> DS_SITUACAO_DOCENTE=Em exercicio 166992 <conf:(1)> lift:(1.03)
```

**Figura 2.** Tela do Weka com as regras encontradas. IFBA, 2016.

164  
165  
166  
167

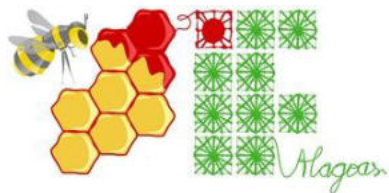
A regra  $IN\_ATU\_EXTENSAO=N \Rightarrow DS\_SITUACAO\_DOCENTE=Em\ exercicio$  mostra que 298.463 (75,25% do total) docentes em exercício não atuam em projetos de extensão. E de acordo com a regra  $IN\_ATU\_PESQUISA=N \Rightarrow DS\_SITUACAO\_DOCENTE=Em\ exercicio$ , a atuação dos docentes em pesquisa é de apenas 30,07%. A terceira regra,  $IN\_ATU\_EXTENSAO=N \wedge IN\_ATU\_PESQUISA=N \Rightarrow DS\_SITUACAO\_DOCENTE=Em\ exercicio$  deixa evidente que 60,50% dos docentes não atuam nem em pesquisa e nem em extensão.

174

## CONCLUSÕES

176  
177  
178  
179  
180  
181

A mineração de dados é um valioso recurso para obtenção de conhecimento. Principalmente na atualidade, onde a quantidade de dados aumenta de forma exponencial e a utilização somente de recursos humanos para análise desses dados se torna insuficiente. As ferramentas desenvolvidas, como o Weka, disponibilizam um recurso importante para análise e obtenção de informações valiosas para a tomada de decisão em diversos ramos empresariais e acadêmicos.



182 Futuramente, se poderia aplicar essa pesquisa em outros dados disponibilizados no  
183 próprio site do INEP, para que os dados sejam utilizados não só para consulta e estatística e  
184 sim como fonte de conhecimento para identificar comportamentos comuns que implicam em  
185 determinados resultados que, se identificados, poderiam ser um diferencial na tomada de  
186 decisões na educação básica e superior. É importante ressaltar que outros atributos poderiam  
187 ser utilizados, e outras regras relevantes poderiam ser descobertas, o que pode ser feito em  
188 trabalhos futuros.

189

## 190 REFERÊNCIAS

191 CAMARGO, A. *et al.* Mineração de dados eleitorais: descoberta de padrões de candidatos a  
192 vereador na região da campanha do Rio Grande do Sul. **Revista Brasileira de Computação**  
193 **Aplicada**, Passo Fundo, v. 8, n. 1, p.64-73, 2016.

194

195 FAYYAD, U. M.; PIATETSKY-SHAPIRO, G; SMYTH, P. **From Data Mining to**  
196 **Knowledge Discovery in Databases**. AAAIIntelligence, p. 37-54, 1996.

197

198 FONSECA, S. O; NAMEN, A. A. Mineração em Bases de Dados do INEP: Uma Análise  
199 Exploratória para Nortear Melhorias no Sistema Educacional Brasileiro. **Educação em**  
200 **Revista**, BH, v. 32, n. 1, p.133-157, 2016.

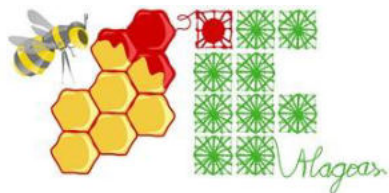
201

202 NAMEN, A. A; SOARES, A.C. S. Mineração de dados relacionados ao aprendizado de  
203 língua portuguesa: um estudo exploratório. In: ENCONTRO DE MODELAGEM  
204 COMPUTACIONAL, 14., 2011, Nova Friburgo. **Anais...** Nova Friburgo: Rede Sirius, 2011.  
205 p. 295 - 304.

206

207 PASTA, A. **Aplicação da técnica de data mining na base de dados do ambiente de gestão**  
208 **educacional**: um estudo de caso de uma instituição de ensino superior de Blumenau-SC. 2011.  
209 153 f. Dissertação (Mestrado) - Curso de Computação Aplicada, Universidade do Vale do  
210 Itajaí, São José, 2011.





- 211 SANTOS, R. **Weka na munheca.** 2005. Disponível em:  
212 <[www.ambientelivre.com.br/downloads/doc\\_download/81-weka-na-munheca.html](http://www.ambientelivre.com.br/downloads/doc_download/81-weka-na-munheca.html)>. Acesso  
213 em 15 jul. 2016.  
214
- 215 SILVA, G. C. **Mineração de regras de associação aplicada a dados da secretaria**  
216 **municipal de saúde de Londrina - PR.** 94 f. Dissertação de Mestrado - UFRGS, Porto  
217 Alegre, 2005.  
218
- 219 SILVA, M. P. S. **Mineração de Dados - Conceitos, Aplicações e Experimentos com Weka.**  
220 2004. Disponível em: <<http://www.lbd.dcc.ufmg.br/colecoes/erirjes/2004/004.pdf>>. Acesso  
221 em 15 jul. 2016.  
222
- 223 WEKA. **Machine Learning Group at the University of Waikato: Downloading and**  
224 **installing Weka.** 2016. Disponível em:  
225 <<http://www.cs.waikato.ac.nz/ml/weka/downloading.html>>. Acesso em 15 jul. 2016.