**DATA MINING ATRAVÉS DA REGRA DE ASSOCIAÇÃO APRIORI**

Arivaldo Pereira Silva Filho[[1]](#footnote-1)

Samuel Bueno da Silva[[2]](#footnote-2)

João Mauricio Hypólito[[3]](#footnote-3)

**RESUMO**

Este artigo tem como objetivo demonstrar por meio de um caso de uso, a utilização da técnica de data mining utilizando a regra de associação *Apriori*. Para a aplicação de tal técnica de mineração foi utilizado o software WEKA, uma ferramenta gratuita de *Data Mining* e o método de mineração de dados testagem de hipóteses. O banco de dados utilizado está disponibilizado no site do IBGE, um arquivo de extensão DBF com os dados de natalidade do estado de São Paulo no ano de 2000. Foi feito toda a preparação do banco e aplicação da regra de associação *Apriori*, por fim serão mostrados os resultados obtidos por meio desta técnica.

Palavras-chave: Data Mining, Associação, *Apriori*.

**ABSTRACT**

This article aims to demonstrate through a small use case, the use of the technique of data mining through association rule apriori. For the application of this technique was used mining software Weka, a free tool DM. The database used was obtained from the IBGE website, a DBF file with the data of birth rate of the state of São Paulo in 2000. It was done all the preparation and implementation of the bank's association rule Apriori finally show the results obtained by this technique.

Keywords: Data Mining, Association, Apriori

# 

# INTRODUÇÃO

Por meio da mineração de dados é possível analisar e estabelecer relações entre dados que a “olho nu” não seria possível perceber.

O *Data Mining* tem como função a mineração dessas informações das bases de dados trazendo relações importantes entre os dados. Por intermédio de softwares é possível utilizar técnicas e métodos de *Data Mining* com o intuito de obter informações importantes a serem analisadas.

**Objetivo**

Este artigo tem como objetivo mostrar na prática, especificamente uma técnica de mineração de dados utilizando a regra de associação *Apriori* por meio de um software para a execução do mesmo.

**Justificativa**

O presente artigo toma como justificativa o fato de que a mineração de dados é uma importante ferramenta de obtenção de conhecimento para as organizações e suporte à decisão, portanto, saber como utilizar suas técnicas é fundamental para obter informações consistentes e confiáveis.

**Revisão Bibliográfica**



“A mineração de dados pode ser realizada de três diferentes formas, em função do nível de conhecimento que se tenha do problema estudado.” (CARVALHO, 2005, p.15), a técnica a ser utilizada neste artigo será a de testagem de hipóteses.

Quando se possui algum conhecimento sobre o campo de atuação da empresa ou alguma ideia sobre que relação nova se está buscando, pode-se definir uma hipótese e verificar sua confirmação ou refutação através da metodologia do *data mining* denominada testagem de hipóteses. (CARVALHO, 2005, p.15).

Utiliza-se a testagem de hipóteses quando se deseja saber se uma informação ou dado tem relação ou não com a outra (o), ou seja, quando se busca atingir algum objetivo.

# Regra de Associação Apriori

O Algoritmo *Apriori* foi escolhido como foco do artigo após pesquisas constatarem, tanto na web quanto em artigos que este é um dos mais utilizados algoritmos em mineração de dados para regras de associação.

Este algoritmo trabalha por meio de uma busca aprofundada nos dados, gerando assim conjuntos de padrões. Os padrões mais frequentes são mantidos e os menos, são eliminados.



# METODOLOGIA

O artigo dividiu-se em duas partes, a primeira é caracterizada como uma pesquisa exploratória, pois de acordo com Gil (2002), ao se desenvolver um projeto de pesquisa esta deve ser a primeira etapa a ser cumprida, posteriormente à pesquisa bibliográfica aplicou-se os conceitos através de um estudo de caso, onde foi possível obter os resultados. O WEKA foi configurado para conectar diretamente no banco de dados onde estavam as informações, o que eliminou a necessidade da geração do arquivo de extensão ARFF.

O banco de dados foi obtido por meio site do IBGE, onde se encontrava em formato DBF e para uma manipulação mais fácil foi executada a migração para o MySQL. Para o processo de migração da base foi desenvolvido uma pequena aplicação em Delphi, que conecta facilmente com DBF, gerando um arquivo SQL com os inserts da tabela original.

Segue abaixo uma amostra dos registros do banco de dados utilizado para aplicar a mineração de dados, onde constam os campos utilizados para a testagem de hipóteses, são eles idade\_mae e gravidez:



Figura 1: Tabela de registros

Fonte: Elaborada pelos autores

Após a migração houve a necessidade de uma análise dos dados para verificar inconsistências. Como resultado muitos dados foram tratados e houve a necessidade da remoção de dados não interessantes para o processo. Por exemplo, percebeu-se a necessidade da criação de faixas de valores para a idade da mãe, pois era uma abrangência muito grande de valores, tornando a aplicação do método *Apriori* inviável. Abaixo segue a tabela com os códigos das faixas de idade modificadas e suas descrições:

Tabela 1: Tabela dos códigos e faixa de idade

|  |  |
| --- | --- |
| Código de Idade | Faixa de Idade |
| 1 | Entre 11 e 20 anos |
| 2 | Entre 21 e 30 anos |
| 3 | Entre 31 e 40 anos |
| 4 | Entre 41 e 50 anos |
| 5 | Acima de 51 |

Fonte: Elaborada pelos autores

Houve também a necessidade de alterar os campos definidos com inteiro para texto, pois o *Weka* não permitiu a utilização do algoritmo *Apriori* com os campos definidos com *Integer*.

Por meio da testagem de hipóteses e do algoritmo *Apriori*, selecionado no *menu* de opções do *Weka*, procurou-se saber se há relações entre a idade da mãe e o tipo da gravidez ocorrida, por exemplo, se foi única, gêmeos, trigêmeos, etc.

Pelo fato também de se trabalhar com o tipo de gravidez, segue abaixo uma tabela contendo os códigos e a descrição de cada tipo de gravidez:

Tabela 2: Tabela dos códigos e descrição da gravidez

|  |  |
| --- | --- |
| Código da Gravidez | Descrição |
| 9 | Ignorado |
| 1 | Única |
| 2 | Dupla |
| 3 | Tripla ou Mais |

Fonte: Elaborada pelos autores

# RESULTADOS E DISCUSSÕES

Com a execução do algoritmo através do *Weka* obteve-se os seguintes resultados, ilustrados na figura abaixo e explicados posteriormente:

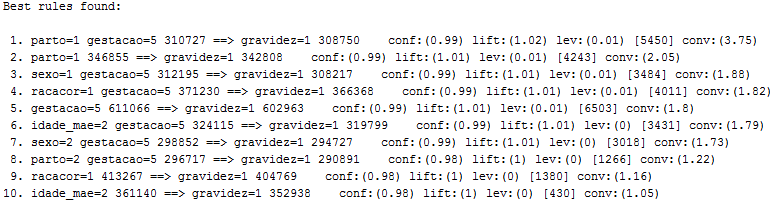


Figura 2: Resultado WEKA

Fonte: Elaborada pelos autores

Ao executar o *data mining*, como retorno obteve-se 10 possíveis correlações. Pelo fato de o foco do artigo ser somente verificar se há relações entre a idade da mãe e o tipo de gravidez da mesma, as relações que não atendem a esse requisito foram descartadas, sendo utilizada somente duas, que diz respeito às relações 6 e 10.

Ao analisar as correlações percebeu-se, de acordo com a tabela dos tipos de gravidez e de idade que, quando a mãe está na faixa de idade 2, que é correspondente a faixa de idade entre 21 e 30 anos, o tipo de gravidez ocorrida foi de código 1, que corresponde à uma gravidez única, ou seja, nasceu somente um filho no parto.

.

**CONCLUSÕES**

Com base nos resultados obtidos após a mineração ser executada, o software foi capaz de encontrar duas associações entre os dados das mães no que diz respeito à idade e o tipo da gravidez da mesma como mostrado na figura 2.

As relações encontradas apresentam um nível de confidência (*conf)* de 0.99, que significa o percentual de ocorrência da mesma e quão confiável ela é, ou seja, as duas relações apresentadas possuem um nível alto de confidência de 99%.

# REFERÊNCIAS

BRASIL. Ministério da Saúde (2012). Nascidos. Disponível em <http://tabnet.datasus.gov.br/cgi/sinasc/dados/nov\_indice.htm#tabela>. Acesso em 20 ago de 2013.

CARVALHO, Luis Alfredo Vidal: Datamining: A mineração de Dados no Marketing, Medicina, Economia, Engenharia e Administração. Rio de Janeiro: Ciência Moderna, 2005.

GIL, Antônio Carlos. Como elaborar projetos de pesquisa. 4ª ed. São Paulo: Atlas, 2002.

1. Aluno do curso de Análise de Sistemas e Tecnologia da Informação – FATEC Campus Ourinhos/ SP. E-mail: arivaldops@gmail.com [↑](#footnote-ref-1)
2. Aluno do curso de Análise de Sistemas e Tecnologia da Informação – FATEC Campus Ourinhos/ SP. E-mail: ss.infosilva@gmail.com [↑](#footnote-ref-2)
3. Professor Orientador: João Mauricio Hypólito – FATEC Campus Ourinhos/ SP. E-mail: joao.hypolito@fatec.sp.gov.br [↑](#footnote-ref-3)