

# Uso de Regras de Associação para Descoberta de Conhecimento na Produtividade de Açaí no Estado do Amapá

Cassiana Fagundes da Silva<sup>1</sup>, Cleuton T. Rodrigues<sup>2</sup>, Marcus Vinícius B. Monteiro<sup>2</sup>

<sup>1</sup>Faculdade de Balsas  
Km 5 da BR 230, Fazenda Malidere 4 – 65800-000 – Balsas – MA - Brasil

<sup>2</sup>Faculdade Seama  
Av. Nações Unidas, 1201 – Jseus de Nazaré – 68.900-000 – Macapá – AP – Brasil

cassiana@gmail.com, cleuton.torres@gmail.com, marcus.monteiro@politec.com.br

**Abstract.** *This paper presents application in Knowledge Discovery in Databases showing the establish of this process in analysis and discovers associates that influence of the Açaí productivity in State of Amapá. To accomplishing of these experiments were used Instituto de Pesquisas Científicas Tecnológicas do Estado do Amapá (IEPA) informations databases and applied in data mining tool WEKA through the Apriori associates algorythm.*

**Key words:** *Knowledge Discovery in Databases, Açaí, Amapá*

**Resumo.** *Este artigo apresenta estudos da área de Descoberta de Conhecimento em Base de Dados, mostrando a implantação deste processo na análise e descoberta de associações que influenciam a produtividade do açaí no estado do Amapá. Para a realização dos experimentos foram utilizadas as informações contidas na base de dados do Instituto de Pesquisas Científicas e Tecnológicas do Estado do Amapá (IEPA) e aplicadas na ferramenta de mineração de dados WEKA através do algoritmo de associação Apriori.*

**Palavras Chave:** *Descoberta de Conhecimento em Base de Dados, Açaí, Amapá*

## 1. Introdução

Os constantes avanços na área da tecnologia da informação têm viabilizado o armazenamento de grandes e múltiplas bases de dados. Tecnologias como a Internet, sistemas gerenciadores de banco de dados, leitores de códigos de barras, dispositivos de memória secundária de maior capacidade de armazenamento e de menor custo têm contribuído para esse aumento.

A análise das informações contidas nesses volumes de dados é realizada por consultas diretas com acesso a dados específicos para obter as informações necessárias. Porém, essas consultas apresentam-se mais eficazes em pequenas bases de dados, de modo que quando maior a sobrecarga de informação, mais complexo sua exploração.

Surgiu assim, a necessidade de uma nova geração de técnicas e ferramentas com a habilidade de auxiliar os analistas humanos de uma forma “inteligente” e “automática” na procura de informações potencialmente úteis, previamente desconhecidas nos dados.

As ferramentas e técnicas empregadas para análise automática e inteligente destes grandes repositórios são os objetos tratados por uma nova área denominada Descoberta de Conhecimento em Bancos de Dados (DCBD), da expressão em inglês *Knowledge Discovery in Databases* (KDD).

Segundo Carvalho (2002), é cada vez maior o crescimento da demanda por aplicações em KDD nas mais diversas áreas do conhecimento humano. Potencialmente, técnicas de KDD podem ser utilizadas em qualquer domínio de aplicação (saúde, economia, direito, educação etc.) que contenham razoáveis volumes de dados históricos sobre algum assunto.

Para Homma et. al. (2006), o fruto do açaizeiro deve ser o novo produto a ser consumido em todo o país e no mundo como aconteceu com o guaraná e atingindo nichos de mercado de produtos funcionais.

Diante deste cenário, este trabalho tem como objetivo a aplicação da tarefa regras de associação para descobrir quais são os fatores que influenciam na produção do fruto desta riqueza natural abundante no estado do Amapá, e que vem se popularizando tanto no âmbito nacional quanto internacional.

Este artigo está organizado conforme segue. Na seção 2 é descrito o processo de Descoberta de Conhecimento em Base de Dados, tais como suas etapas, tarefas, e descreve também características da tarefa de Associação e do algoritmo *Apriori*, utilizados para descobrir conhecimento na base de dados. A abordagem geral sobre o açaí, tais como características melhorias para o manejo e cultivo e perspectivas para este fruto na mostradas na seção 3. Os experimentos e resultados obtidos são descritos nas seções 4 e 5. Por fim, as conclusões e trabalhos futuros são apresentados na seção 6.

## **2. Descoberta de Conhecimento em Base de Dados**

Segundo Fayyad (1996), o termo KDD, foi criado em 1989 como referência ao amplo processo para encontrar conhecimento nos dados e enfatizar uma aplicação em especial o método MD.

Para Macedo e Matos (2009), KDD será empregado para todo o processo de extração de conhecimento dos dados. Neste contexto, conhecimento significa relacionamento e padrões entre elementos de dados. O termo Mineração de Dados deveria ser utilizado para os estágios de descoberta do processo de KDD.

O processo de KDD é um conjunto de atividades contínuas para o compartilhamento de conhecimento descoberto a partir de bases de dados, segundo Fayyad (1996), esse conjunto é dividido em seis etapas, como mostra a Figura 1.

Podem-se distribuir os passos do processo de KDD, vistos na Figura 1, em três etapas primordiais: o Pré-processamento, a Mineração de Dados e o Pós-processamento [Feldens 1997].

A etapa de pré-processamento é uma das mais demoradas podendo tomar até 80% de todo tempo necessário para o processo completo, devido às muito conhecidas

dificuldades de integração de bases de dados heterogêneas. Nesta são realizados os seguintes passos: definição dos objetivos, coleta de dados, limpeza e pré-processamento dos dados e transformação de dados [BARROSO E NETO, 2006].

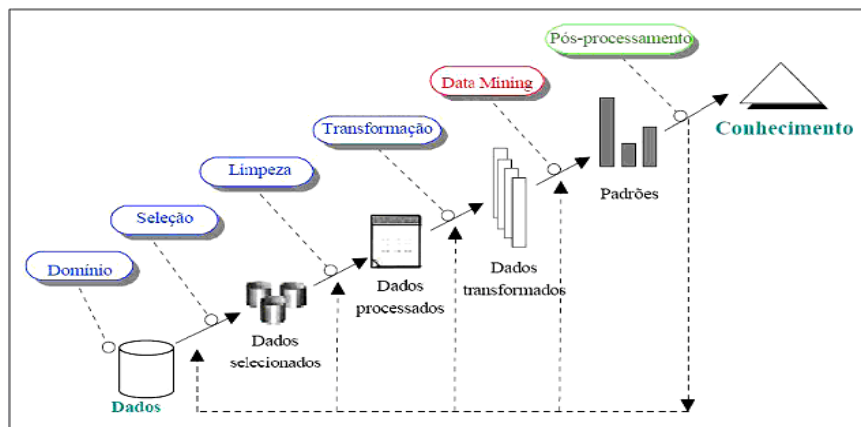


Figura 1 – Etapas do processo de KDD.

A Extração de Padrões é direcionada ao cumprimento dos objetivos definidos na identificação do problema. Esta etapa é um processo iterativo, podendo ser necessária sua execução por diversas vezes até que se ajuste ao conjunto de parâmetros visando à obtenção de resultados mais adequados aos objetivos preestabelecidos. Cabe ressaltar também que, esta compreende a escolha da tarefa de MD a ser empregada, a escolha do algoritmo e a extração dos padrões propriamente dita [Goldschmidt e Passos 2005].

A fase de pós-processamento envolve a visualização, análise e a interpretação do modelo de conhecimento gerado pela etapa de MD. Em geral, é nesta etapa que o especialista em KDD e o especialista no domínio da aplicação avaliam os resultados obtidos e definem novas alternativas de investigação dos dados [Goldschmidt e Passos 2005].

Conforme descrito na seção anterior, a tarefa de MD utilizada para este trabalho foi Regras de Associação (RA) e o algoritmo *Apriori*. A técnica RA consiste na identificação de padrões intrínsecos ao conjunto de dados, ou seja, encontrar conjuntos de itens que ocorram simultaneamente e de forma freqüente em um banco de dados [Goldschmidt e Passos 2005]. Já o algoritmo Apriori é considerado um clássico na extração de Regras de Associação. Este algoritmo faz buscas recursivas no Banco de Dados à procura dos conjuntos freqüentes (conjuntos que satisfazem um suporte mínimo estabelecido).

A ferramenta escolhida para o estudo de caso deste trabalho foi o WEKA<sup>1</sup> (*Waikato Environment for Knowledge Analysis*). Essa ferramenta contempla uma série de tarefas e métodos do processo de KDD. Segundo Goldschmidt e Passos (2005), o WEKA permite a abertura direta de arquivos ARFF, CSV e C45, porém apenas consegue manipular os ARFF. Este é um arquivo ASCII usado para definir atributos e seus valores. O WEKA possui um formato próprio, o ARFF, o qual temos que descrever o domínio do atributo, pois o mesmo não pode ser obtido automaticamente pelo seu valor.

<sup>1</sup> Disponível em <http://www.cs.waikato.ac.nz/ml/weka/>

Na próxima seção, são descritos as características do Açaí bem como o seu manejo para melhor entender os experimentos realizados.

### 3. Açaí

O açaí é um dos alimentos mais consumidos no Amapá e também uma importante fonte de renda para milhares de famílias de ribeirinhos. São eles que abastecem as amassadeiras e fornecem esta valorosa matéria-prima para exportadores de polpa e suco para outros estados e diversos países.

Segundo Dra. Roberta Ogata<sup>2</sup>, o açaí é um fruto com grandes propriedades: é rico em ferro; contém elevada quantidade de vitamina E, sendo, portanto um antioxidante natural; tem grande quantidade de fibras; tem um teor considerável de proteínas; os teores de potássio e cálcio são elevados, o que faz do açaí um alimento bastante completo.

As peculiaridades desta riqueza natural abundante no estado do Amapá, tem promovido o aumento de sua procura fora da região Norte, de modo particular no Rio de Janeiro, onde seu “vinho” já é servido nas praias, e academias diretamente ao consumidor, onde a demanda por esse produto, até há pouco considerado exótico, é crescente e começa a ganhar popularidade entre os cariocas e turistas. Em nível internacional, o produto ainda é pouco conhecido, porém, sendo o açaí um produto de características originais, um bom trabalho de *marketing* contribuirá muito para a conquista desse mercado.

Com o aumento da demanda, é preciso garantir a coleta o ano inteiro, inclusive na entressafra. Segundo Homma et al (2006) um dos caminhos é investir em pesquisas a médio e longo prazos.

A expansão do consumo da polpa de açaí esta promovendo escassez do produto, ou seja, a oferta não esta suprindo a procura, fato que contribui para a exclusão social dos consumidores de menor poder aquisitivo, os modos de produção em geral primitivos, a pouca disseminação da cultura, falta de políticas de incentivo contribuem para este quadro. No estado do Amapá este fato vem sendo contornado.

Há sete anos a Embrapa Amapá (Macapá-AP) iniciou, em parceria com o Instituto de Pesquisas Científicas e Tecnológicas do Estado do Amapá (IEPA), um projeto chamado “Desenvolvimento de tecnologias para o manejo e cultivo de açaizais para produção de frutos”, para agregar valores a este produto e garantir sua produção em larga escala, possibilitando ao produto novo nexo de mercado.

### 4. Experimentos

Para a realização dos experimentos o primeiro passo no processo de KDD foi a definição dos objetivos. Neste trabalho o principal objetivo foi analisar quais são os fatores que influenciam na produtividade do açaí no estado do Amapá, descobrindo

---

<sup>2</sup> Disponível em <http://www.sposito.com.br/artigosroberta.asp>

Em paralelo com a definição dos objetivos e seleção dos dados foi realizada a fase de limpeza e pré-processamento dos dados. Isto porque no momento da elaboração da tabela com os dados selecionados foram tomados os cuidados necessários e pertinentes a esta fase como eliminação dos registros em branco, redundantes e inconsistentes.

Conforme os dados ilustrado na Figura 2, foi necessário a transformação dos dados selecionados para um formato que a ferramenta de MD *Weka* compreendesse, a fim de suprir quaisquer limitações existentes no algoritmo utilizado, no caso deste trabalho, foi escolhido o algoritmo *Apriori*. Este algoritmo aceita somente campos nominais, ou seja, não trabalha com números, por este motivo houve a necessidade de transformar alguns campos da tabela com os dados selecionados, conforme descrito a seguir.

No caso do campo tamanho da área, estes períodos foram representados pelas letras P, M e G que significam respectivamente: pequenas propriedades, abrangendo áreas com até 50 hectares; médias propriedades, áreas com 51 a 99 hectares e grandes propriedades para áreas com 100 ou mais hectares. Este procedimento será ilustrado na Tabela 1.

**Figura 2 – Base de dados no Excel.**

As mesmas transformações foram realizadas nos campos tempo de atividade, menor e maior preço, quantidade produzida por dia e quantidade produzida por semana,

pois os mesmos se apresentavam de forma semelhante ao campo tamanho da área, conforme Tabelas 2, 3, 4, 5 e 6.

**TABELA 1 – Transformação do campo tamanho da área.**

<b>Tamanho da área (hectares)</b>	<b>Intervalos (período)</b>
De 1 a 50	P
De 51 a 99	M
100 ou mais	G

**TABELA 2 - Transformação do campo tempo de atividade**

<b>Tempo de atividade (anos)</b>	<b>Representação dos Intervalos (período)</b>	<b>Significado</b>
De 1 a 10	C	Curto tempo
De 11 a 30	M	Médio tempo
31 ou mais	L	Longo tempo

**TABELA 3 – Transformação do campo menor preço**

<b>Menor preço (R\$)</b>	<b>Representação dos Intervalos</b>	<b>Significado</b>
De 5 a 25	MPA	Menor preço no intervalo <b>A</b> (5 a 25)
De 26 a 40	MPB	Menor preço no intervalo <b>B</b> (26 a 40)
De 41 a 60	MPC	Menor preço no intervalo <b>C</b> (5 a 25)

**TABELA 4 – Transformação do campo preço**

<b>Maior preço (R\$)</b>	<b>Representação dos Intervalos</b>	<b>Significado</b>
De 20 a 60	MAPA	Menor preço no intervalo <b>A</b> (20 a 60)
De 61 a 90	MAPB	Menor preço no intervalo <b>B</b> (61 a 90)
De 91 a 120	MAPC	Menor preço no intervalo <b>C</b> (91 a 120)

Através do Excel, foi possível transformar os dados para o arquivo ARFF, onde os campos títulos se tornaram os atributos e os valores dos campos os dados sobre esses atributos, conforme Figura 3.

**TABELA 5 – Transformação do campo quantidade produzida por dia**



## 5. Resultados

Na análise dos resultados procurou-se selecionar as regras de associação de maior interesse para o estudo de caso. Cabe ressaltar que as regras geradas são resultantes de um conjunto de atributos relevantes, bem como a variações desses atributos juntamente com alternâncias do intervalo de confiança e suporte do algoritmo *Apriori*.

A primeira iteração envolveu os atributos relevantes para a análise do escoamento da produção, tais como: local de venda, paga frete, possui transporte e despesas, bem como os parâmetros padrões da ferramenta WEKA (confiança de 90% e número de regras (NR) igual a 10), conforme ilustra a Figura 4.

1.	Local_de_venda=porto	272 ==>	Vende_p/_atravessa-dores=\$	272	conf: (1)
2.	Paga_frete=N	Local_de_venda=porto	223 ==>	Vende_p/_atravessa-dores=\$	223 conf: (1)
3.	Possui_transporte=S	Local_de_venda=porto	194 ==>	Vende_p/_atravessa-dores=\$	194 conf: (1)
4.	Possui_transporte=S	Paga_frete=N	Local_de_venda=porto	155 ==>	Vende_p/_atravessa-dores=\$ 155 conf: (1)
5.	Despesas=N	Local_de_venda=porto	134 ==>	Vende_p/_atravessa-dores=\$	134 conf: (1)
6.	Despesas=S	Local_de_venda=porto	134 ==>	Vende_p/_atravessa-dores=\$	134 conf: (1)
7.	Despesas=N	Paga_frete=N	Local_de_venda=porto	119 ==>	Vende_p/_atravessa-dores=\$ 119 conf: (1)
8.	Despesas=S	Paga_frete=N	137 ==>	Vende_p/_atravessa-dores=\$	134 conf: (0.98)
9.	Despesas=S	171 ==>	Vende_p/_atravessa-dores=\$	167	conf: (0.98)
10.	Despesas=S	Possui_transporte=S	128 ==>	Vende_p/_atravessa-dores=\$	125 conf: (0.98)

Figura 4 – Regras para escoamento da produção de açaí

A Regra 1 mostra que 272 (51 % do total da produção) é vendida para o porto e destes 51% de produção utilizam atravessadores com uma confiança de 100%. Cabe ressaltar que atravessadores são fornecedores que compram o açaí e vendem a um preço maior do que o vendido pelo produtor no porto. Já a Regra 4 define que 155 (60 %) do total de produtores possuem transporte, não pagam frete e vendem para o porto. E destes 60% vendem para atravessadores com uma confiança de 100%.

Ainda com os atributos referentes ao escoamento da produção novas iterações foram realizadas modificando o intervalo de confiança para 70% e aumentando o número de regras para 20. Pode-se observar com essa modificação que o escoamento da produção continua sendo pelo vendido para atravessadores (Regra 1), assim como a Regra 2 que mostra que os produtores não pagam frete e vendem a produção no porto para atravessadores.

Para certificar que as regras de associação eram consistentes, independente dos parâmetros adotados, outra iteração foi realizada com esses atributos, reduzindo então a confiança para 50% e aumentando o número de regras para 30. Novamente pode-se observar os mesmos resultados das iterações anteriores.

A segunda iteração envolve os atributos referentes ao período da safra, menor preço da safra, período entre-safra, menor preço entre-safra e o local de venda com uma confiança de 90 % e o número de regras igual a 10. Onde a regra número 1 mostra que 272 (51 %) do total de vendas são vendidos no porto. E destas vendas 268 (98 %) é vendido pelo menor preço no período de safra, com uma confiança de 99 %. Resgatando as informações descritas na Tabela 5 o menor preço no período de safra varia entre R\$ 5,00 e R\$ 25,00.

A regra número 2 mostra que 63 (12 %) do total é vendido pelo menor preço na entre-safra variando entre R\$ 5,00 e R\$ 25,00. Assim como na safra o menor preço também varia entre R\$ 5,00 e R\$ 25,00 com uma confiança de 95 %.



Sobre o quesito produtividade na safra, ilustrado na Figura 5, destacam-se os atributos mão-de-obra, quantidade produzida por semana, financiamento e assistência técnica.

1. Mão-de-obra=familiar 419 ==> Quantidade_produzida/semana_Safra=QPSSA 415	conf:(0.99)
2. Mão-de-obra=familiar Financiamento=N 398 ==> Quantidade_produzida/semana_Safra=QPSSA 394	conf:(0.99)
3. Assistência_tecnica=N Quantidade_produzida/semana_Safra=QPSSA 421 ==> Financiamento=N 408	conf:(0.97)
4. Quantidade_produzida/semana_Safra=QPSSA Paga_frete=N 398 ==> Financiamento=N 382	conf:(0.96)
5. Paga_frete=N 427 ==> Financiamento=N 408	conf:(0.96)
6. Assistência_tecnica=N 461 ==> Financiamento=N 440	conf:(0.95)
7. Mão-de-obra=familiar 419 ==> Financiamento=N 398	conf:(0.95)
8. Mão-de-obra=familiar Quantidade_produzida/semana_Safra=QPSSA 415 ==> Financiamento=N 394	conf:(0.95)
9. Mão-de-obra=familiar 419 ==> Quantidade_produzida/semana_Safra=QPSSA Financiamento=N 394	conf:(0.94)
10. Financiamento=N Paga_frete=N 408 ==> Quantidade_produzida/semana_Safra=QPSSA 382	conf:(0.94)

**Figura 5 – Produtividade na safra (confiança 90 % e NR =10)**

A Regra 1 mostra que 419 (78%) do total da mão-de-obra é familiar. E para esta mão-de-obra familiar 76% produzem até 40 sacas por semana com uma confiança de 90%.

A quarta iteração envolve os atributos manejo, tempo de atividade, tamanho da área e mão-de-obra, quantidade produzida por semana na safra e entre-safra com a confiança de 90 % e o número de 10 regras. Analisando os resultados observa-se que: Regra 1 mostra que 277 (78%) do total do tamanho da área de produção é Pequena, ou seja até 50 hectares, e a produção é através de mão-de-obra familiar. E desta produção 51% (277) da quantidade produzida varia de 1 a 40 sacas/semana com confiança de 100%. Já a Regra 2 mostra que 237 (51%) do total das áreas tem manejo e mão-de-obra familiar. E destas 43% da quantidade de sacas produzidas por semana varia de 1 a 40 com confiança de 100%.

Após várias iterações na base de dados da produtividade do açaí, utilizando diferentes valores de confiança e número de regras, chega-se às seguintes considerações: a grande maioria dos produtores de açaí no estado do Amapá (concentração da produção nas Ilhas e Mazagão) utilizam mão-de-obra familiar, possuem transporte, não contam com despesas diversas e assistência técnica. O produto é geralmente cultivado em áreas pequenas e a grande maioria da venda do produto é no porto.

## 5. Conclusões

A aplicação de técnicas de KDD mostrou-se útil para descoberta de conhecimento que estava oculto na base de dados que foi utilizada. A eficácia e consistência das regras de associação geradas pela ferramenta, foram comprovadas pelo analista de KDD, juntamente com o especialista no domínio da aplicação.

Uma análise geral feita com as regras geradas, mostram que a produção do açaí no Amapá ainda é praticada de maneira “amadora”, mais como meio de subsistência, pois a maior parte dos produtores possuem outra atividade, utilizam pequenas e médias áreas para produção e empregam mão-de-obra familiar. Os produtores por alguma razão não tem acesso a financiamento e somente uma parcela insignificante dos produtores

possuem assistência técnica. Esses fatores fazem com que a produção seja pouco expressiva, apesar de todas as condições de clima e solo para a produção em potencial.

Como trabalho futuro objetiva-se refazer o experimento adotando uma base de dados do estado do Pará, outro grande produtor de Açaí, e então realizar um cruzamento de dados com o intuito de descobrir se os problemas e associações encontrados no Estado do Amapá também se repetem no Pará. A partir disso, uma análise mais aprofundada será feita para melhor atender o desenvolvimento sustentável dos produtores de Açaí no Norte do país.

## **6. Referências**

AGRAWAL, Rakesh et al. Fast Algorithms for Mining Association Rules. In: FAYYAD, Usama M. et al. *Advances in Knowledge Discovery and Data Mining*. Menlo Park: AAAI Press, 1996. 611p.

BARROSO, Bruno da Costa; NETO, Pedro Nolasco Ferreira. “Descoberta de Conhecimento na Base de Dados de uma Locadora de Filmes”. 2006, 72 pg. Monografia de Trabalho de Conclusão de Curso. Universidade Federal do Pará.

CARVALHO, Luís Alfredo Vidal de. *Datamining: A mineração de dados no Marketing, Medicina, Economia, Engenharia e Administração*. 2.ed. São Paulo: Editora Érica, 2002. 238p.

FAYYAD, U; Piatetsky-Shapiro, G; Smyth, P. “Advances in Knowledge Discovery and Data Mining”. *Califórnia América Association for Artificial Intelligence*. 1996. 611p.

FELDENS, M. A. *Descoberta de conhecimento aplicada à detecção de anomalias em base de dados*. Porto Alegre: PPGCC da UFRGS, 1996.

GOLDSCHMIDT, Ronaldo; PASSO, Emmanuel. *Data Mining. Um guia Prático: Conceitos, Técnicas, Ferramentas, Orientações e Aplicações*. São Paulo: Editora Campus, 2005. 261pg.

HOMMA, A. K. O. ; NOGUEIRA, O. L. ; MENEZES, A.J.E.A. ; Carvalho, J.E.U. ; NICOLI, C. M. L. ; Matos, G.B . Açaí: novos desafios e tendências. *Amazônia JCR*, v. 1, p. 7-23, 2006.

MACEDO, D. C. ; MATOS, Simone Nasser . *Data Mining: Uma ferramenta de auxílio para a Gestão do Conhecimento em cenário de mudanças*. In: ADM, 2009, Ponta Grossa. ADM, 2009.

PAIXÃO, Rodinei Silva da. “Mineração de Dados para Descoberta de Conhecimento em um Sistema de Inquéritos Policiais”. 2005, 68 pg. Monografia de Trabalho de Conclusão de Curso. Faculdade SEAMA. Macapá-AP.

REZENDE, Solange Oliveira. *Sistemas Inteligentes - Fundamentos e Aplicações*. 1ª ed. São Carlos: Manole, 2003. 525p.