Universidade Federal de Ouro Preto
Departamento de Computação e Sistemas
Curso Sistemas de Informação

Técnicas de Mineração de Dados aplicadas ao Problema de Cesta de Mercado em Empresa Varejista

Nathália Sales Lage

TRABALHO DE CONCLUSÃO DE CURSO

ORIENTAÇÃO:
PROF. MSc. George Henrique Godim da Fonseca

Julho, 2015
João Monlevade/MG

Nathália Sales Lage

Técnicas de Mineração de Dados aplicadas ao Problema de Cesta de Mercado em Empresa Varejista

Orientador: Prof. MSc. George Henrique Godim da Fonseca

Monografia apresentada ao Curso de Sistemas de Informação do Departamento de Ciências Exatas e Aplicadas, como requisito parcial para aprovação na Disciplina Trabalho de Conclusão de Curso II.

Universidade Federal de Ouro Preto
João Monlevade
Julho de 2015

ATA DE DEFESA

Aos 13 dias do mês de Julho de 2015, às 21 horas, na sala C203, foi realizada a defesa de Monografia pela aluna Nathália Sales Lage, sendo a Comissão Examinadora constituída pelos professores: Prof. MSc. George Henrique Godim da Fonseca, Profa. MSc. Helen de Cássia Souza da Costa Lima e Profa. MSc. Janniele Aparecida Soares. A candidata apresentou a monografia intitulada: *"Técnicas de Mineração de Dados aplicadas ao Problema de Cesta de Mercado em Empresa Varejista"*. A comissão examinadora deliberou, por unanimidade, pela aprovação do candidato, concedendo-lhe o prazo de 15 dias para incorporação no texto final das alterações sugeridas. Na forma regulamentar, foi lavrada a presente ata que é assinada pelos membros da Comissão Examinadora e pelo formando.

João Monlevade, 13 de Julho de 2015.

Prof. MSc. George Henrique Godim da Fonseca
Professor Orientador/Presidente

Profa. MSc. Helen de Cássia Souza da Costa Lima
Professor Convidado

Profa. MSc. Janniele Aparecida Soares
Professor Convidado

Nathália Sales Lage
Formando

FOLHA DE APROVAÇÃO DA BANCA EXAMINADORA

Técnicas de Mineração de dados aplicadas ao Problema de Cesta de Mercado em Empresa Varejista

Nathália Sales Lage

Monografia apresentada ao Departamento de Ciências Exatas e Aplicadas da Universidade Federal de Ouro Preto como requisito parcial da disciplina CEA499 – Trabalho de Conclusão de Curso II, do curso de Bacharelado em Sistemas de Informação, e aprovada pela Banca Examinadora abaixo assinada:

Prof. MSc. George Henrique Godim da Fonseca
Mestre em Ciência da Computação / Universidade Federal de Ouro Preto – MG, Brasil
Orientador
Departamento de Computação e Sistemas - UFOP

Profa. MSc. Helen de Cássia Souza da Costa Lima
Mestre em Ciência da Computação / Universidade Federal de Ouro Preto – MG, Brasil
Examinador
Departamento de Computação e Sistemas – UFOP

Profa. MSc. Janniele Aparecida Soares
Mestre em Ciência da Computação / Universidade Federal de Ouro Preto – MG, Brasil
Examinador
Departamento de Computação e Sistemas - UFOP

João Monlevade, 13 de Julho de 2015

TERMO DE RESPONSABILIDADE

O texto do trabalho de conclusão de curso intitulado "**Técnicas de Mineração de Dados aplicadas ao Problema de Cesta de Mercado em Empresa Varejista**" é de minha inteira responsabilidade. Declaro que não há utilização indevida de texto, material fotográfico, código fonte de programa ou qualquer outro material pertencente a terceiros sem as devidas referências ou consentimento dos referidos autores.

João Monlevade, 13 de Julho de 2014

Nathália Sales Lage

Resumo

O presente trabalho aborda a aplicação do processo de descoberta de conhecimento em base de dados de uma empresa varejista, que tem como objetivo descobrir padrões de cesta de compra através da aplicação de regras de associação nos dados de vendas da empresa. Também faz parte do objetivo do trabalho analisar os produtos vendidos em épocas do ano, a fim de detectar sazonalidades e analisar as vendas em determinadas regiões de clientes que compram na empresa, para detectar as tendências dos clientes em diferentes regiões. Para atingir os objetivos foi utilizado o *WEKA* para gerar as regras e foram realizadas análises através de gráficos que representam o comportamento dos dados. Foi detalhado todo o processo realizado antes da geração das regras e as dificuldades encontradas. Nas análises regionais de vendas destacou-se o conhecimento adquirido sobre o volume das vendas que são realizadas para clientes do bairro Cruzeiro Celeste, mostrando que estes clientes buscam na empresa por Pisos/Porcelanato e Revestimento. Também foi detectado queda nas vendas nos meses de Fevereiro e Abril e crescimento em épocas de pouca chuva. Em relação a sazonalidade, destacou-se a venda de Chuveiros e Resistências no inverno, Impermeabilizante em épocas de chuva e Mangueiras em épocas de pouca chuva.

Agradecimentos

Primeiramente quero agradecer a Deus, por guiar meus passos e me dar forças durante toda esta trajetória. Agradeço ao Ulete Mota pela confiança e por acreditar no trabalho disponibilizando todas as informações necessárias e auxiliando no trabalho sempre que solicitado. Ao meu professor e orientador Msc. George Henrique Godim Fonseca, que sempre esteve disponível quando precisei, me motivando e pela confiança. A minha família que sempre esteve presente, me apoiando e torcendo, me dando forças para seguir em frente e Lourenço pelo apoio e carinho. A todos que de alguma forma contribuíram para este trabalho.

Sumário

Lista de ilustrações

Lista de tabelas

1 Introdução

Atualmente as empresas utilizam sistemas de informação para processar dados de vendas, cadastrar produtos e clientes, manter todo o controle de estoque, controle do financeiro, dentre outras atividades que podem ser realizadas e que são importantes para a gestão do negócio. Grande volume de dados é gerado diariamente, porém uma base de dados, por maior que seja não é informação.

Shaeffer (2003) afirma que muitas vezes, os dados são explorados somente de forma a atender os requisitos do sistema de informação para o qual eles existem. Desta forma, não são extraídas informações valiosas para o conhecimento do negócio que precisam de ferramentas específicas para serem descobertas.

Existem muitas possibilidades para descobrir novos conhecimentos sobre padrões de compra. Uma forma seria o relacionamento de fatores como região do cliente, tipo de produto que ele compra, época do ano que compram determinado produto, a relação de compra de um produto com outro, dentre outras informações. Estas informações podem ser utilizadas para abertura de lojas filiais em regiões potenciais de vendas, realizar promoções de produtos, combinar itens em propagandas ou planejar estratégias de marketing sazonal. Todas estas estratégicas podem ser aplicadas a fim de aumentar a lucratividade do negócio.

Para obter essas informações, as ferramentas de mineração de dados são úteis, pois elas são capazes de gerar estas informações, prever futuras tendências e comportamentos. Dessa forma, a mineração de dados auxilia as empresas na tomada de decisão, reunindo sugestões úteis para capitação de cliente e tornando a empresa mais lucrativa e competitiva no mercado.

Conforme Navathe e Elmasri (2010), a mineração de dados refere-se à mineração ou descoberta de novas informações em termos de padrões ou regras com base em grande quantidade de dados. Para ser útil na prática, a mineração de dados precisa ser executada de modo eficiente em grandes arquivos e banco de dados. Dessa forma, ela ajuda na extração de novos padrões significativos que não podem ser necessariamente encontrados apenas ao consultar ou processar dados e para isso, deve ser precedida por uma preparação significativa nos dados, antes de gerar informações úteis que possam influenciar diretamente as decisões de negócios.

De acordo com Camargo (2002), dentre as várias aplicações da mineração de dados, uma das que tem o maior potencial de interesse do comércio varejista é a que focaliza a descoberta de regras de associação em banco de dados de transações de vendas, que é frequentemente referida como o problema da análise de cestas de compras. O motivo desse interesse refere-se ao fato de que no varejo a maior parte das compras é realizada por impulso. A aplicação da técnica de análise de cesta de compra dá pista acerca do que um cliente poderia ter comprado se alguma sugestão interessante lhe fosse feita.

O estudo de caso apresentado propõe a utilização de uma ferramenta *open source* de mineração de dados denominada *WEKA*. A técnica de mineração de dados a ser utilizada é a Associação, utilizando o algoritmo *Apriori*. Este estudo é realizado na base de dados de uma empresa varejista no ramo de materiais de construção, na qual não há exploração de informações nos registros armazenados. O propósito desse estudo de caso é obter um melhor conhecimento dos dados da empresa, gerando regras para analisar as características sazonais e temporais a fim de conhecer o perfil de clientes e análise de cesta de mercado para conhecer quais produtos estão associadas em transações de venda.

1.1 Objetivos

1.1.1 Objetivos Gerais

O principal objetivo deste trabalho é apresentar um estudo de caso da aplicação da mineração de dados para auxiliar na descoberta de conhecimento em base de dados do varejo. O negócio em questão é a comercialização de materiais de construção. Toda a preparação e transformação dos dados, visa facilitar e tornar eficiente a aplicação de técnicas de associação para a descoberta de conhecimento.

Também é objetivo deste trabalho realizar um estudo de caso que contribua para a comunidade acadêmica e profissional por destacar as dificuldades encontradas nas etapas do processo de descoberta do conhecimento, a fim de minimizar estes problemas em empresas que pretendem aplicar essas técnicas de mineração de dados futuramente e também para apresentar a importância de obter alto nível de conhecimento sobre o negócio.

1.1.2 Objetivos Específicos

Foram estabelecidos alguns objetivos específicos a fim de obter os resultados esperados, sendo eles:

1. Preparação e transformação dos dados, aplicando o processo de descoberta do conhecimento (KDD – *Knowledge Discovery in Databases*);
2. Geração de regras de associação aliada à aplicação do algoritmo *Apriori*;
3. Análises de características sazonais e regionais das vendas e análise de cesta de mercado para conhecer quais produtos estão associados em transações de venda.
4. Apresentação dos resultados obtidos e análise das informações que sejam consideradas úteis para o negócio.

1.2 Trabalhos Relacionados

Shaeffer (2003) apresentou um estudo de caso de mineração de dados no varejo, cujo negocio em questão é a comercialização de móveis e materiais de construção. O objetivo estabelecido como resultados da aplicação da mineração de dados foi conhecer o perfil do cliente que compra na loja, criar uma lista dos produtos mais rentáveis e verificar quais os clientes com maior tendência em realizar compras a estes produtos e conhecer o perfil dos clientes que compram pela primeira vez na loja. O estudo foi realizado com base na metodologia CRISP, que foi adotada para conduzir o processo, é considerada pelo autor um guia fundamental para qualquer projeto nesta área, pois sem metodologia a necessidade de iteração para fazer correções será maior. Foram utilizadas duas ferramentas de mineração de dados, sendo elas: *WEKA* e IBM *Intelligent Miner*. Conclui-se que o *WEKA* se destaca na geração de regras de associação pela facilidade de manipulação dos parâmetros de mineração, já que na outra ferramenta é bastante complexo, limitando a interação. O autor afirma que revisar a classificação dos produtos é trabalhoso, mas de extrema importância e deve-se tomar cuidado com generalizações, pois produtos para diferentes fins que pertencem à mesma categoria geram resultados errados nas minerações. Desta forma, foi possível concluir que a necessidade de padronização na descrição e classificação dos itens de venda foi entendida como um fator determinante de sucesso, tanto para trabalhos deste tipo quanto para a própria informação gerencial e organizacional da empresa. As conclusões deste trabalho foram determinantes para a escolha das metodologias e ferramentas a serem utilizadas no presente trabalho. Os trabalhos se assemelham em relação a geração de regras para análise de cesta de mercado, porém se difere na utilização das ferramentas para atingir esse objetivo. Outro fator que diferencia é que Shaeffer (2003) selecionou todos os atributos que seriam interessantes para a mineração e durante um período coletou esses dados na empresa, adaptando o software utilizado. Já no presente trabalho, a empresa utiliza um software que não pode ser alterado e foram utilizados dados existentes na base de dados.

Camargo (2002) mostrou em seu trabalho que o algoritmo *Apriori* muitas vezes é utilizado em trabalhos com base de dados de supermercados, cujo número de transações costuma ser grande. Entretanto, trouxe o exemplo de base de dados do setor de comércio varejista de confecção e o setor de ferragens, apresentando um tamanho médio de transações bem abaixo do intervalo correspondente aos supermercados. A partir desta constatação, leva a hipótese de que o desempenho da tarefa de descoberta de regras de associação pode ser influenciado decisivamente pelo tamanho médio das transações do banco de dados analisado. Para minimizar esse problema foi desenvolvido um algoritmo que otimiza a tarefa de mineração de regras de associação em banco de dados com baixo tamanho médio de transação a fim de aumentar a eficiência desta tarefa, chamado *MiRABIT* e a ferramenta, chamada MIRA que implementa o algoritmo *MiRABIT*. Com a realização de experimentos constatou-se que a geração de conjuntos de itens candidatos pode ser

otimizada, se executada a cada transação ao invés de ser executada no início da passagem. Esta alteração aumenta o desempenho do processo de mineração de regras de associação em banco de dados com uma baixa média de itens por transação. Com base nesse trabalho, deve-se considerar nos testes realizados o tamanho das transações devido à aplicação do *Apriori*, a fim de obter melhor resultado. Ambos trabalham com análise de regras associativas, entretanto, esse é focado na comparação dos algoritmos para aplicação da regra, propondo adaptações.

Kasahara e Conceição (2008), realizaram a análise de duas ferramentas de mineração de dados, o *WEKA* e o *TANAGRA*. Foi realizado um estudo de caso da situação dos pacientes do Laboratório de Análises Clínicas do Instituto de Previdência e Assistências do Município de Belém (IPAMB), com o intuito de avaliar a qualidade dos padrões minerados. Para esse estudo foi utilizado a regra de associação e escolhido o algoritmo *Apriori*, por ser o mais disseminado e presente nas duas ferramentas analisadas. Como resultado da comparação das ferramentas concluiu-se que o *WEKA* apresenta maior vantagem em relação à interface, também é mais fácil de utilizar. Observou que o *WEKA* mostra os resultados em ordem decrescente de confiança, já o *TANAGRA* mostra em ordem decrescente de Suporte, porém ambas ferramentas obtiveram o mesmo resultado. Semelhantemente a este trabalho, também utilizam regra de associação e utilização do *Apriori*, a diferença está no tipo de informação apresentada no estudo de caso, sendo o setor de saúde e este no setor varejista de matérias de construção. O trabalho foi útil para obter um comparativo em relação ao *WEKA* e outra ferramenta mostrando sua facilidade de uso e vantagens.

Araújo (2009), realizou um estudo de caso onde aplicou a regra de associação em uma base de dados de empresa varejista utilizando a ferramenta de mineração de dados *WEKA*. Os objetivos a serem alcançados com a mineração de dados foram: conhecer o perfil de clientes que frequentam e compram na loja; conhecer clientes que compram na loja pela primeira vez e tentar conhecer produtos que estão associados em transações, para uma análise de cestas de mercado. Porém, foi deparado o problema de que a empresa não havia registro de dados de cliente, com isso o escopo foi reduzido apenas para as associações em transações de venda. Foi concluído que com a utilização da mineração de dados, pode-se realizar buscas eficientes na base de dados, agregando valor ao negócio em questão. Uma das dificuldades encontradas foi com relação à categorias de produtos, sendo que essas foram generalizadas, não sendo possível refletir total precisão dos resultados, sugerindo a empresa a criar categorias e subcategorias para que uma análise mais profunda e detalhada possa ser feita. O trabalho se assemelha com o presente trabalho em relação à utilização de regra de associação para determinar padrões na cesta de mercado, porém este pretende tratar as limitações e dificuldades de forma diferente, para não impactar tanto nos resultados. Também serão realizadas novas análises de características sazonais e temporais das vendas da empresa.

Dias (2014), realizou um estudo da aplicação da mineração de dados em uma base de dados real de uma empresa do ramo alimentício. Foi desenvolvida uma ferramenta

intitulada Mineração Anchieta com o objetivo de gerar as regras de associação, onde o usuário poderá escolher se deseja gerar regras referentes à cidade do cliente ou à data de inclusão do pedido, informando assim o período que deseja analisar. Para criação dessa ferramenta foram realizadas algumas adaptações no algoritmo *Apriori* com o objetivo de fixar os atributos que indicam a região e o período das vendas sem passar pela análise de suporte e dar opção para o usuário escolher o tamanho dos itemsets gerados pelo algoritmo. Foram realizadas análises comparativas da ferramenta desenvolvida com o *WEKA*, confirmando a similaridade entre os resultados encontrados. Foram confirmados padrões de vendas de produtos sazonais, regionais e relação entre alguns produtos. O trabalho se assemelha em relação à utilização de regras de associação para determinar padrões de vendas de produtos sazonais, regionais e análise de cesta de compra, porém o trabalho de Dias (2014) é realizado em uma base de dados de atacado, não obtendo informações sobre os clientes que compraram o produto, mas as empresas que compram, já o presente trabalho é direcionado para as vendas de varejo, analisando as tendências dos clientes finais. Outro fato que diferencia os trabalhos é a criação da ferramenta para geração das regras de associação, que não foi realizada no presente trabalho, pois os testes utilizando o WEKA atenderam as expectativas. Em ambos foram realizadas análises gráficas para demonstrar os padrões encontrados.

1.3 Organização do trabalho

O presente trabalho está organizado em 6 capítulos incluindo esta introdução. O processo de descoberta do conhecimento será abordado no Capítulo 2, bem como a regra de associação, o algoritmo *Apriori* e a ferramenta *WEKA*. Na sequência, o Capítulo 3 apresenta a metodologia aplicada para coleta e preparação dos dados. Os resultados obtidos serão apresentados no Capítulo 4, destacando-se as informações que sejam consideradas úteis para o negócio onde está sendo aplicado o processo de mineração de dados. O Capítulo 5 apresenta as conclusões obtidas e sugestões para trabalhos futuros e por fim, o capítulo 6 serão os anexos do trabalho, contendo todos os comandos utilizados para realização do trabalho.

2 Fundamentos Teóricos

2.1 Processo de Descoberta do Conhecimento em Banco de Dados

Segundo Fayyad, Shapiro e Smyth (1996) o processo de descoberta de conhecimento em bases de dados ou KDD (*Knowledge-Discovery in Databases*) trata-se de um processo não trivial, formado por várias etapas, de forma interativa e iterativa, para identificação de padrões válidos, novos, potencialmente úteis e compreensíveis em grandes conjuntos de dados para auxiliar na tomada de decisão.

O processo de KDD pode ser cíclico, pois após uma melhor análise dos dados em cada etapa poderá surgir necessidade de voltar em uma etapa para melhorar ou modificar algo que foi realizado e que poderá influenciar nos dados. Também é um processo interativo entre homem e máquina, pois é necessário que haja uma análise dos dados que estão sendo processados. Com o conhecimento do negócio, pelo especialista, pode-se selecionar as informações úteis para obter um resultado satisfatório. Desta forma, o analista dos dados não precisa ser da área de TI, podendo ser um especialista no negócio em que os dados se referem.

O grande desafio do processo de KDD é preparar a base de dados de forma que não haja dados incoerentes, inconsistentes e com grande quantidade de valores nulos para que seja realizado o processo de mineração de dados e que obtenha resultados satisfatórios.

O KDD envolve as seguintes etapas: seleção, pré-processamento, transformação, mineração dos dados, interpretação dos padrões e assimilação do conhecimento. A Figura 1 mostra como estas etapas estão dispostas, ilustrando todo o processo de KDD.

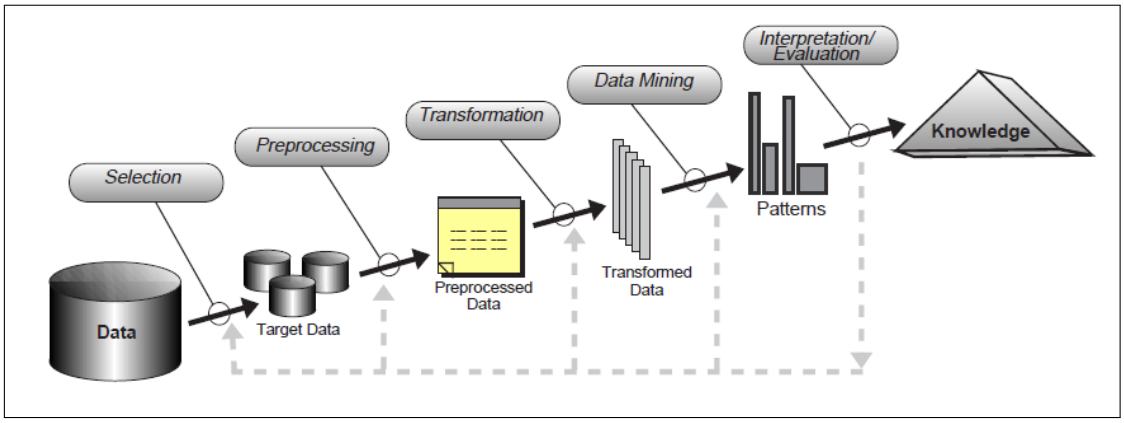


Figura 1 – Processo de descoberta do conhecimento. Fonte: (FAYYAD; SHAPIRO; SMYTH, 1996)

Segundo Fayyad, Shapiro e Smyth (1996) as fases do processo podem ser descritas da seguinte forma:

1. **Entendimento do domínio da aplicação:** buscando associar os objetivos do negócio com os da mineração de dados.

2. **Seleção dos dados:** selecionar da base de dados todos os dados que serão úteis para atingir os objetivos do negócio.

3. **Limpeza dos dados e pré-processamento:** envolve operações básicas, como remoção de ruído de campos com preenchimento incorreto ou sem preenchimento e tratamento dos dados.

4. **Redução e projeção dos dados:** encontrar características úteis para representar os dados de acordo com o objetivo da tarefa. Utilizar métodos de transformação dos dados para reduzir o número efetivo de variáveis ou encontrar representações mais viáveis para os dados.

5. **Transformando os dados para Mineração de dados:** os dados serão transformados de forma que atenda as exigências de execução do método que será utilizado na mineração de dados, por exemplo, sumarização, classificação, associação, regressão, clusterização, e etc.

6. **Mineração de dados:** Execução da tarefa de mineração de dados a fim de obter padrões de interesse para atingir o objetivo do negócio. A execução dos passos anteriores é fundamental para o sucesso desta etapa.

7. **Interpretação dos padrões:** Analisar os padrões encontrados na tarefa de mineração de dados. Caso não seja encontrado padrões interessantes pode-se retornar a qualquer etapa anterior a fim de refinar os dados para conseguir chegar no objetivo do negócio.

8. **Consolidação do conhecimento descoberto:** Documentar o conhecimento descoberto e relatá-lo as partes interessadas a fim de auxiliar na tomada de decisão. Também pode servir para comparação de informações existentes com o conhecimento previamente obtido.

Todas as etapas descritas serão aplicadas na realização do presente trabalho e serão detalhadas no capitulo 3.

2.2 Mineração de Dados

A Mineração de Dados ou *Data Mining* busca extrair novos padrões significativos, tendências nos dados e inferir regras que não podem ser descobertas apenas ao realizar consultas tradicionais, pois estas são limitadas. Fayyad, Shapiro e Smyth (1996) define

a mineração de dados como um passo do processo de KDD que consiste em algoritmos particulares de *Data Mining* que, sobre algumas limitações aceitáveis de eficiência computacional, produzem uma série particular de padrões sobre os dados.

Um exemplo é a realização de consultas em um banco de dados contendo registros de transações de vendas de clientes. Esta consulta poderá responder a uma pergunta como: "Quantos revestimentos foram vendidos para o cliente X na data dd/mm/aaaa?". Esta é considerada uma operação comum da baixa administração da empresa. Porém, as técnicas de mineração de dados buscam atender níveis mais elevados da administração, extraindo informações para auxiliar na tomada de decisão. Estas informações podem ser úteis para o marketing, como mala direta direcionada para potenciais clientes, planejamento de estoque, abertura de novas filiais e outras decisões estratégicas.

Existem diversas tarefas de mineração de dados na literatura, para escolher qual tarefa utilizar é necessário identificar qual objetivo a ser alcançado com a mineração de dados. Neste trabalho será utilizada a tarefa de Associação com destaque ao algoritmo *Apriori*, pois visa encontrar relacionamentos interessantes entre registros do banco de dados. Esta tarefa será descrita de forma detalhada a seguir.

2.2.1 Regra de Associação

Grandes bancos de dados de empresas armazenam milhares de itens e registros de transações de vendas destes itens. Desta forma, este grande volume de informações pode ser analisado a fim de detectar associações importantes entre os itens comercializados, tal que a presença de algum item implica na presença de outros itens na mesma transação. Para isso, pode-se utilizar a tarefa de regra de associação presente na mineração de dados.

Segundo Navathe e Elmasri (2010), a regra de associação tem forma $X \rightarrow Y$, onde X e Y são um conjunto de itens distintos. Essa associação indica que, se um cliente compra X, ele também provavelmente comprará Y. O conjunto de todos os itens de X e Y é chamado de *itemset*, ou seja, conjunto dos itens comprados pelos clientes em uma transação de venda. Uma transação de venda é composta por um identificador da venda, produtos, data, cliente, endereço, quantidade e outros. Um exemplo de transações no contexto de vendas de materiais de construção é demonstrado na Tabela 1.

Tabela 1 – Transações de Vendas de Materiais de Construção

Venda	Produtos	Mês	Ano	Cidade
1	Assento Banheiro, Engate.	Jan	2013	João Monlevade
2	Torneira, Assento Banheiro.	Jan	2013	João Monlevade
3	Anel Vedação, Torneira, Assento Banheiro, Parafuso	Dez	2013	João Monlevade
4	Anel Vedação, Engate, Assento Banheiro, Parafuso	Abr	2013	Bela Vista de Minas
5	Anel Vedação, Engate, Assento Banheiro.	Jun	2013	João Monlevade

Um algoritmo de associação procura por regras de associação que atendam a um suporte e confiança mínimos preestabelecido. Segundo Tan, Stainbach e Kumar (2009), o suporte se refere à frequência com que um *itemset* específico ocorre no banco de dados, ou seja, é o percentual de transações que contêm todos os itens no conjunto de *itemset*. Caso o suporte seja baixo, conclui-se que não existe evidências fortes de que os itens ocorrem juntos, pois o *itemset* ocorrerá em apenas uma pequena fração das transações. Seja $\alpha(X \cup Y)$ a quantia de vezes que um *itemset* aparece nas transações e N o número total de transações, o calculo do suporte é representado por:

$$Suporte = \frac{\alpha(X \cup Y)}{N}$$

Já a outra medida é a confiança, sendo a medida da força das regras, ou seja, a probabilidade de que o cliente compre Y, dado que X foi comprado. Seja $\alpha(Y)$ a ocorrência do *itemset* a direita, e $\alpha(X)$ a ocorrência do *itemset* a esquerda, o cálculo da confiança é representado por:

$$Confiança = \frac{\alpha(Y)}{\alpha(X)}$$

No exemplo de transações da Tabela 1, pode-se observar a ocorrência da seguinte regra:

Regra: $[AnelVedação, AssentoBanheiro] \rightarrow [Parafuso]$.

O suporte desta regra será o número de vezes que a regra aparece, sobre o número total de registros: $Suporte = \frac{2}{5}$

Já a confiança será o número de vezes que a implicação (Y) aparece sobre o número de vezes que o precedente (X) aparece: $Confiança = \frac{2}{3}$

A regra de associação é um processo composto por duas etapas, primeiro gera o conjunto de itens frequentes que atendam ao suporte mínimo e em seguida gera as regras a partir dos itens frequentes que atendam a confiança mínima. O objetivo da mineração de regras de associação, então, é gerar todas as regras possíveis que excedem alguns patamares mínimos de suporte e confiança especificados pelo minerador.

2.2.2 Algoritmo *Apriori*

Segundo os criadores do Algoritmo *Apriori* Agrawal e Srikant (1994), a performance de seu algoritmo é superior a de algoritmos existentes na geração de regras de associação. Segundo Tan, Stainbach e Kumar (2009) o algoritmo considera que cada item é o conjunto candidato de 1 item, verificando o suporte mínimo de cada um e descartando os itens que

não atendem ao suporte. Posteriormente, procura-se os conjuntos candidatos de 2 itens usando apenas os conjuntos de 1 item. Esse processo é repetido até que não restem mais conjuntos de candidatos a serem testados.

O algoritmo 2.1 é detalhado a seguir:

1. Gera o conjunto de *1-itemsets* frequentes;
2. Gera o conjunto de itens candidatos. A função $genCandidatos(F_{k-1})$ irá fundir pares de conjuntos frequentes de tamanho $k-1$, gerando conjuntos frequentes de tamanho k.
3. Separa os candidatos que estão contidos na transação T;
4. Incrementa seu valor de suporte;
5. Adiciona os candidatos que atendem ao suporte mínimo.

Algoritmo 2.1: Apriori

Entrada: *minSup*, *conj. de transações*, T, *conj. de itens* I
Saída: Conjunto de regras encontro F

```
1  k = 1; F = F_k = {i | i ϵ I ∧ σ(i) ≥ minSup × |T|};
2  enquanto F_k ≠ ∅ faça
3      ++k;
4      C_k = genCandidatos(F_{k-1});
5      para cada t ϵ T faça
6          C_k = subset(C_k, t);
7          para cada candidato c ϵ C faça
8              ++σ(c);
9      F = F_k = {c | c ϵ C_k ∧ σ(c) ≥ minSup × |T|};
10     F = F ∪ F_k;
11 retorna s;
```

2.3 *WEKA*

O *WEKA* (*Waikato Environment for Knowledge Analysis*) foi criado pelo Departamento de Ciência da Computação da Universidade de *Waikato*, na Nova Zelândia (WAIKATO, 2015). É uma ferramenta de software livre e desenvolvido em Java.

A ferramenta *WEKA* trabalha com dados no formato especifico *arff*. O arquivo deve seguir um padrão rigoroso de formatação pois o software é muito restritivo quanto à formatação do arquivo de entrada. Deve existir um cabeçalho de atributos para cada valor que existir como um dado e os valores prováveis para cada um deles, no formato a seguir:

@attribute **NOME DO ATRIBUTO** [Valores prováveis que ele poderá assumir]

Para este trabalho, todos os atributos assumem o mesmo valor (?,1). A ausência do item na relação será representada por ? e a presença por 1. Em seguida, vêm os dados e cada linha representa um registro, sendo que cada item do registro esteja na ordem dos atributos informados anteriormente.

Parte de um arquivo gerado para a mineração pode ser visto a seguir:

```
@attribute VARAL {?,1}
@attribute VASELINA {?,1}
@attribute VASO_ASSENTO_MICTORIO {?,1}
@attribute VEDA_PORTA {?,1}
@attribute VEDA_ROSCA {?,1}
@attribute VELA_FILTRO {?,1}
@attribute VENTILADOR {?,1}
@attribute VERNIZ {?,1}
@attribute ZARCAO {?,1}

@data
1,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,1,?,?,?,1,?,?,?,1,?,?,?,?,?,?,?,?,?,?
?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?
?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?
?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?
?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?
?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?
?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?
?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?
?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?
?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,1,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?
?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?
?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?
?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?
```

Com o arquivo criado da forma correta, atendendo a todas exigências da ferramenta, deve-se abri-lo na ferramenta, escolher a tarefa de mineração e o algoritmo a ser utilizado. Neste trabalho, será utilizada a regra de associação que implementa o algoritmo *Apriori*. No Capítulo 3 serão apresentados os parâmetros configurados para geração de resultados da mineração na ferramenta.

3 Desenvolvimento

3.1 Compreensão do domínio da aplicação

A seguir serão destacadas as etapas do Processo de Descoberta de Conhecimento em Base de dados - KDD, com o intuito de compreender o domínio da aplicação, entendendo onde se quer chegar na mineração de dados e seus objetivos.

3.1.1 Entendimento do negócio

O presente estudo de caso foi realizado na empresa Ulete Mota e Cia LTDA, estabelecida em João Monlevade-MG (MOTA, 2015). A empresa oferece produtos e soluções em materiais de construção. Teve inicio de suas atividades no final dos anos 60 e se tornou uma empresa conceituada em toda região do Médio Piracicaba e do Vale do Aço. A empresa dispõe de um software de Gestão Comercial e Frente de Caixa desenvolvido pela EMC Sistemas (COELHO, 2015) para auxiliar na gestão da empresa, controlando estoque e financeiro, além de atender as legislações fiscais, como na emissão de documento fiscal no ato das vendas. Desta forma para a realização do presente trabalho, foram utilizados os registros do banco de dados gerados nas transações de vendas, cadastro de clientes, produtos e outros. A base de dados foi coletada no dia 11 de Junho de 2014. Foi realizada uma entrevista com o gerente da empresa com o intuito de entender o negócio, o funcionamento da empresa atualmente e as informações que os gestores obtêm através do sistema de informações para apoiar a tomada de decisão. Alguns pontos principais da entrevista serão descritos a seguir.

Nathália: Como funciona o processo de vendas da empresa?

Ulete: O processo de venda da empresa pode ser realizado de duas formas: o cliente pega a mercadoria na gôndola e em seguida se direciona para o caixa para efetuar o pagamento ou é atendido por um vendedor, obtendo assim uma venda assistida, pois para este ramo de materiais de construção, muitas vezes o cliente não sabe o que precisa para resolver seu problema e com isso, é necessária a orientação do vendedor. Este vendedor realizará uma pré-venda e este cliente é direcionado ao caixa para realizar o pagamento e receber impresso o seu documento fiscal. É importar ressaltar que, mesmo que a venda seja à vista e direta, é orientado passar por uma pré-venda com o vendedor antes de se direcionar ao caixa, pois mesmo a empresa cadastrando todos os itens, pode acontecer de algum ficar sem cadastro, com código de barras incorreto ou até mesmo sem código de barras, devido ao produto serem pequenos ou serem comprados em embalagens e não haver identificação na gondola, causando assim transtornos no pagamento. Já para vendas no crediário, a pré-venda é necessária para evitar burocracias no caixa e verificar se o

cliente possui o crédito para efetuar a compra.

Nathália: É proposto para o cliente realizar cadastro na empresa mesmo sendo venda à vista?

Ulete: É sugerido para o cliente realizar o cadastro na empresa, porém atualmente o processo de venda esta facilitado com a utilização do cartão de crédito, pois assim o crédito já é pré-aprovado e o cliente não precisa ter cadastro na empresa para compras, como na venda à vista.

Nathália: A empresa tem conhecimento sobre seus clientes, como os que compram com mais frequência, os que realizam as maiores compras e quais são os inadimplentes?

Ulete: Pelo cadastro do cliente no sistema é analisado apenas os históricos de compra do cliente quando precisa liberar o crédito, pois ocorre do cliente está com débito em aberto na empresa, mas se tiver um histórico de frequência de compras e pagamentos esse crédito não será bloqueado. Já em relação aos clientes que compram na empresa com mais frequência e maior valor de compras é feito um controle interno, sem utilizar o software e o vendedor também consegue identificar esses clientes pelo fato, de estar sempre presente na empresa realizando compras. Além disso, nenhum outro conhecimento é extraído dos clientes.

Nathália: É realizado investimento em marketing e publicidade? Essas publicidades são direcionadas a produtos específicos e publico alvo?

Ulete: Não realizamos muitas campanhas publicitárias pelo fato, da empresa estar no mercado a muito tempo e já estar consolidada na região. Mas, quando realizam algum tipo de publicidade é focado mais no nome da empresa e algumas vezes em algum produto específico.

Nathália: A empresa realizada alguma estratégia para atrair clientes? Quais?

Ulete: A estratégia utilizada para atrair cliente é através de indicação de profissionais como pedreiro, bombeiro, pintor, sendo esses tratados de forma diferenciada.

Nathália: Já foi realizada mineração de dados na empresa?

Ulete: Não. Nunca foi realizado nenhum estudo especifico da base de dados.

Nathália: A empresa utiliza algum tipo de relatório para acompanhamento de vendas por período, região e produtos? Consegue neles extrair todas as informações necessárias para aplicar estratégias da empresa?

Ulete: É utilizado pela empresa apenas as ferramentas que o software de Gestão Empresarial atualmente implantado oferece, como conhecimento do produto mais vendido, acompanhamento de venda por vendedor e outros relatórios de gestão. Com a experiência

dos vendedores é possível identificar alguns produtos sazonais e alguns produtos relacionados na cesta de compra que são mais comuns. Temos uma pesquisa de mercado que identifica as regiões de João Monlevade que mais compram na empresa. Apesar disso, será interessante comparar essa pesquisa com os dados que serão extraídos através deste trabalho.

Nathália: A empresa tem interesse em obter essas informações para aplicar estratégias de marketing?

Ulete: Sim, apesar de ser uma empresa consolidada no mercado é interessante tentar extrair conhecimentos da base de dados para aplicar estratégias e se tornar cada vez mais competitiva no mercado.

3.1.2 Objetivos do negócio

Com base do que foi levantado na entrevista apresentada na seção anterior, foram estabelecidos alguns objetivos para este trabalho de mineração, sendo eles:

1. Análise de cesta de mercado para conhecer quais produtos está associada em transações de venda.
2. Análise de perfil de clientes que compram na empresa.
3. Análise das características de vendas regionais.
4. Análise das características de vendas temporais.

3.1.3 A base de dados da empresa

A base de dados atual de informações da empresa é gerada pelo software da EMC Sistemas, que conta com vários módulos integrados de Gestão Empresarial, Frente de Caixa, Nota Fiscal Eletrônica e SPED (COELHO, 2015). Esta base de dados contém informações sobre os registros de transações de vendas, cadastro de clientes, produtos, realiza o controle de estoque e vários outros dados são coletador diariamente. Foram coletados dados de Janeiro de 2010 à Junho de 2014, totalizando 4 anos e 6 meses.

Realizou-se uma análise de toda a estrutura da base de dados, tabelas, relacionamentos e atributos, para filtrar apenas o que seria necessário para o estudo. Desta forma a modelagem da base de dados reduzida é apresentada na Figura 2 a seguir:

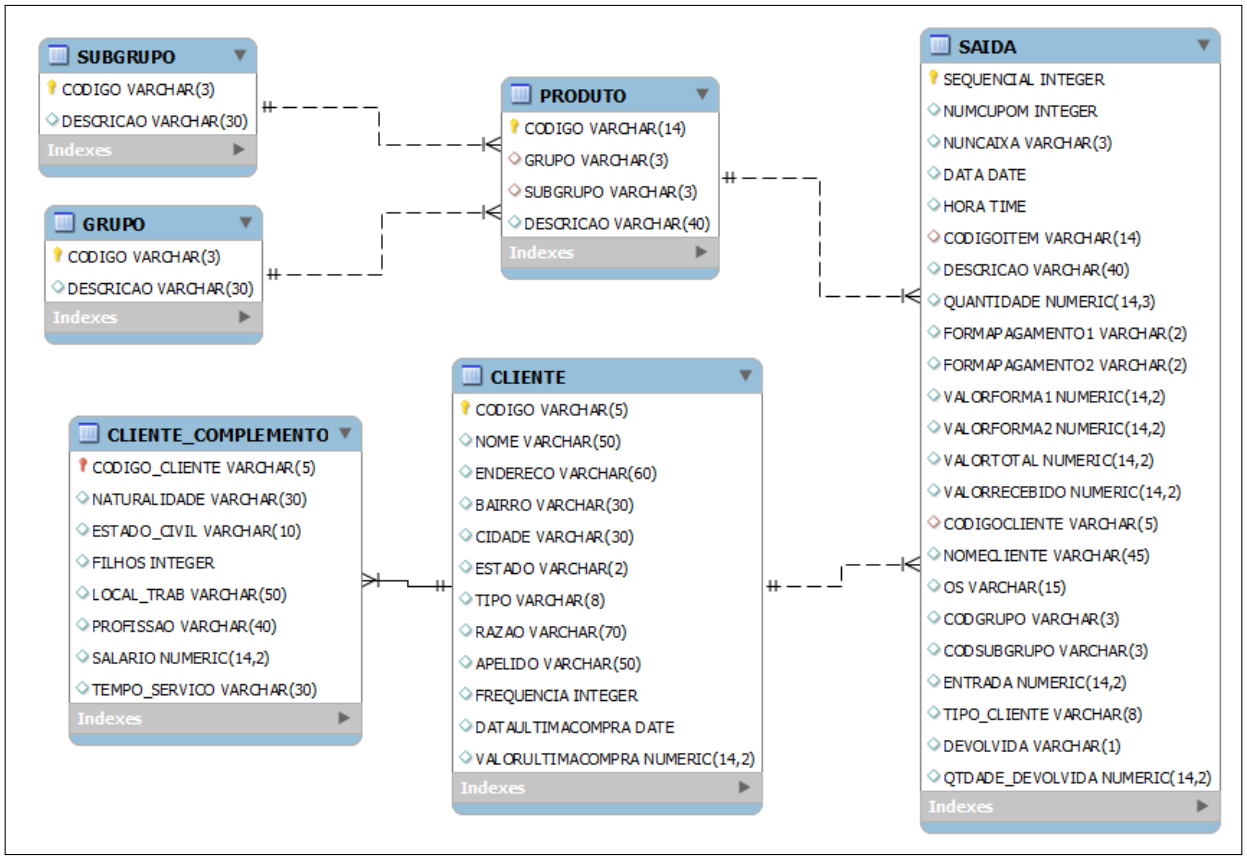


Figura 2 – Modelagem da Base de dados reduzida para realização do trabalho.

3.2 Criação do conjunto de dados a serem minerados

Nesta etapa do projeto, podemos ter o primeiro contato com a base de dados atual da empresa, com o objetivo de identificar problemas na qualidade dos dados, relacionamento entre tabelas, atributos que contém e podem ser úteis e se os atributos que contém são devidamente utilizados. Está é a segunda etapa do KDD de seleção dos dados. Os resultados deste estudo serão descritos a seguir.

3.2.1 Coletando os dados iniciais

O formato do banco de dados utilizado pela empresa é o .GDB, com isso foi instalado na máquina o gerenciador de banco de dados Firebird na versão 2.5. Foi utilizada a ferramenta IBConsole na versão 1.0 para manipulação inicial da base de dados no seu formato original. Foram analisadas as tabelas e os atributos. Via comandos SQL, selecionaram-se os atributos necessários de cada tabela e exportando no formato .TXT. Em seguida, foi realizada a importação dos dados para o banco de dados Microsoft SQL Server 2012 que foi escolhido para que seja realizada a manipulação de dados, por ser considerado um gerenciador de banco de dados robusto. Os dados de transações de vendas selecionados para o presente trabalho estão compreendidos em um intervalo de Janeiro de 2010 à Junho

de 2014, totalizando 4 anos e 6 meses. Também foram selecionados todos os dados de produtos e clientes cadastrados.

3.2.2 Descrevendo os dados

A base de dados atualmente usada pela empresa não pode ser alterada, pois o software não é específico para o negócio da empresa. Trata-se de um produto comercializado que pode atender à mudanças, porém não foi o foco deste trabalho propor alterações no software para complementar a coleta de dados. Outro detalhe que foi analisado é que muitos atributos disponíveis no software não são utilizados pela empresa, ou foram utilizados poucas vezes, como por exemplo, diversos campos do cadastro de cliente. O motivo para o não preenchimento do cadastro completo de clientes deve-se ao fato de haver muitos atributos para ser preenchido e durante a venda de produto pode não haver tempo hábil para coletar todas as informações.

Na tabela cliente, havia campos pouco utilizados e se fossem preenchidos com maior frequência, eles poderiam ser utilizados para determinar algum padrão de venda, como por exemplo, o campo FILHOS, para identificar o número de filhos. De 20.828 clientes cadastrados, apenas 111 possui esta informação. Também havia o campo PROFISSÃO, onde apenas 2.156 cadastros foram preenchidos com este dado. Desta forma não foram utilizados estes dados e o objetivo de conhecer o perfil de clientes foi descartado o escopo do trabalho.

Além dos campos que constam no cadastro de cliente poderia haver campos como Tipo de Moradia, para identificar se o cliente mora em casa ou apartamento, diferenciando assim o perfil de compra. Outro campo seria o tempo de residência, para identificar quanto tempo o cliente mora em sua residência e se for muito tempo pode ser mais provável realizar uma reforma, dentre outros campos que podem ser pensados para determinar um perfil de cliente, porém estes não fazem parte do software utilizado para a empresa, então também não serão utilizados.

Trabalhou-se com dois contextos no que diz respeito a dados: as informações cadastrais da região dos clientes, produtos e as informações das transações de vendas. A Figura 3 e a Figura 4 mostram os atributos extraídos da base de dados da empresa que são considerados úteis para o trabalho de mineração.

Column Name	Data Type
CODIGO	varchar(14)
DESCRICAO	varchar(40)
C1	varchar(40)
C2	varchar(40)
GRUPO	varchar(40)

Figura 3 – Estrutura da tabela Produto.

Column Name	Data Type
NUMCUPOM	int
NUMCAIXA	int
NUMERO_NOTA	int
IDVENDAS	varchar(50)
DATA	date
ANO	numeric(18, 0)
MES	varchar(50)
CODIGOITEM	nvarchar(50)
GRUPO	varchar(40)
VALORITEM	float
QUANTIDADE	int
BAIRRO	varchar(50)
CIDADE	varchar(50)

Figura 4 – Estrutura da tabela Venda.

3.2.2.1 Justificativa para escolha dos atributos das tabelas

Tabela Produto

- *Código e Descrição:* identificar o produto.
- *Grupo:* utilizado para classificar os produtos de forma a desconsiderar características especifica e agrupá-los para que a mineração seja eficiente. Este campo será relacionado com a tabela de vendas de forma que a identificação do item seja considerada o subgrupo.
- *C1:* irá conter o primeiro nome da descrição dos produtos, para facilitar nos comandos SQL de criação dos subgrupos.
- *C2:* irá conter o segundo nome da descrição dos produtos, para facilitar nos comandos SQL de criação dos subgrupos.

Tabela Venda

- *Sequencial:* identificador de registro.
- *Número do Cupom e Número do Caixa:* com a combinação dos dois identificadores é possível saber a cesta de marcado, pois a venda de vários itens é realizada em um cupom e este número do cupom é único para cada caixa, mas o número do cupom pode repetir em caixas diferentes, por isso a necessidade de utilizar o número do caixa.
- *Número Nota:* a empresa pode realizar vendas emitindo o documento fiscal cupom ou nota fiscal, desta forma, quando o campo NUMCUPOM estiver NULL o campo NUMERONOTA estará preenchido e também corresponde a uma cesta de compra.
- *Identificador de venda:* este campo foi criado para combinar os campos NUMCUPOM, NUMCAIXA E NUMERONOTA. Nele foi criado um identificador de venda de forma

que quando a venda foi por cupom fiscal foi preenchido com a combinação dos campos NUMCUPOM + NUMCAIXA e quando a venda for por nota fiscal ficará preenchido com o campo NUMERONOTA, e este campo que será utilizado para identificar a cesta de compra.

- *Data, Ano e Mês:* registra a data que a venda foi realizada que será desmembrada nos campos ANO e MÊS, pois não é interessante para a mineração utilizar o valor completo do campo data considerando o dia. O intuito desse campo é determinar padrões de vendas em determinados mês e ano, por exemplo, com isso pode-se utilizar alguma estratégia de marketing como promoções em alguns meses e determinar produtos sazonais.

- *Código do Item:* o código do produto que será comprado. Este campo será relacionado com a tabela produto e agrupado conforme o grupo determinado.

- *Bairro:* identificar onde o cliente mora e tentar determinar algum padrão de vendas para cada localização, identificar clientes de qual bairro compram mais na empresa e com isso poderá determinar estratégias de marketing nos bairros com menos frequência de compra, com o intuito de aumentar as vendas para potenciais clientes destes bairros.

- *Cidade:* identificar a cidade do cliente e relacionar com o bairro.

3.2.2.2 Quantificando o conjunto de dados

Afim de conhecer o conjunto de dados utilizado na mineração de dados, a tabela 2 apresenta os dados quantificados.

Tabela 2 – Sumarização do conjunto de dados

	Base de dados completa	Selecionados para a Mineração
Produtos	15.872	10.925
Itens de Vendas	2.761.748	1.412.604
Transações de Vendas	892.571	565.614
Período de Registro de Vendas	04/01/1996 até 16/06/2014	01/01/2010 até 16/06/2014

3.2.3 Preparação dos dados e limpeza

Na terceira etapa do KDD será realizada a preparação dos dados e limpeza. Após determinado os dados que serão utilizados na mineração, foi realizada uma análise dos dados e percebeu-se que na tabela de vendas, muitas cestas de mercado não estão vinculadas a clientes cadastrados, com isso muitas transações não possuem informações de clientes como

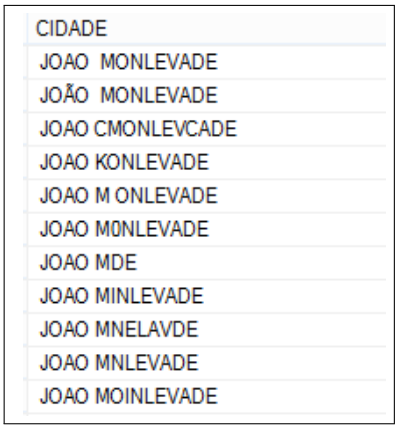


Figura 5 – Campo CIDADE da tabela Vendas. Fonte: Desenvolvido pela autora.

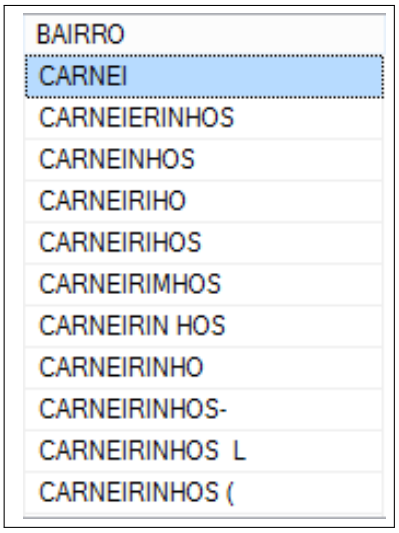


Figura 6 – Campo BAIRRO da tabela Vendas. Fonte: Desenvolvido pela autora.

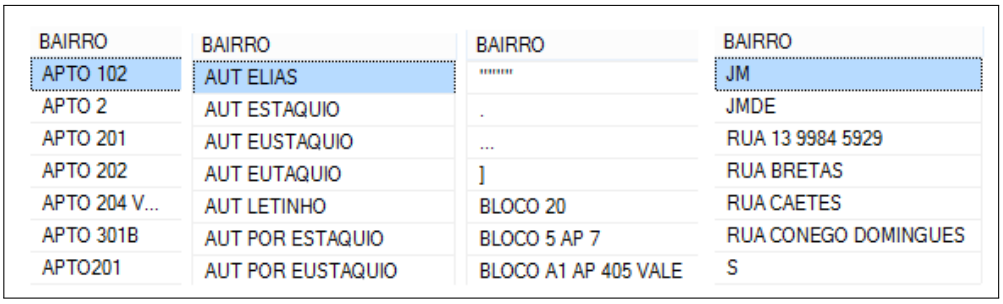
bairro e cidade, que serão úteis para o trabalho. Desta forma, irá influenciar no objetivo da mineração de determinar características regionais de vendas, já que está informação será reduzida e não irá abranger todas as transações selecionadas.

3.2.3.1 Tratamento dos dados de Cidades e Bairros

A qualidade dos dados foi analisada e percebeu-se que há muitos erros de digitação em nome de cidades e bairros. Conforme mostrado na Figura 5 a cidade João Monlevade foi escrita de diversas formas. Também haviam vários bairros escritos de forma incorreta, conforme apresentado na Figura 6, o bairro Carneirinhos foi escrito de diversas formas.

Havia também preenchimento indevido que não foi possível detectar a informação correta para ser corrigida, optando por deixar o campo NULL. Alguns exemplos são apresentados na Figura 7.

A fim de tratar esse problema na base de dados, foram executados comandos SQL para correção, sendo uma análise custosa devido ao volume.



BAIRRO	BAIRRO	BAIRRO	BAIRRO
APTO 102	AUT ELIAS	""""""	JM
APTO 2	AUT ESTAQUIO	.	JMDE
APTO 201	AUT EUSTAQUIO	...	RUA 13 9984 5929
APTO 202	AUT EUTAQUIO	]	RUA BRETAS
APTO 204 V...	AUT LETINHO	BLOCO 20	RUA CAETES
APTO 301B	AUT POR ESTAQUIO	BLOCO 5 AP 7	RUA CONEGO DOMINGUES
APTO201	AUT POR EUSTAQUIO	BLOCO A1 AP 405 VALE	S

Figura 7 – Campo BAIRRO da tabela Venda com valores inválidos. Fonte: Desenvolvido pela autora.

NUMCUPOM	NUMCAIXA	NUMERO_NOTA	IDVENDAS
<null>	<null>	13644	13644
<null>	<null>	13090	13090
105848	7	<null>	105848-7
325952	5	<null>	325952-5
352434	5	<null>	352434-5
154556	7	<null>	154556-7
243644	5	<null>	243644-5

Figura 8 – Tratamento do Identificador de Vendas, campo IDVENDAS da tabela Registro de Vendas. Fonte: Desenvolvido pela autora.

3.2.3.2 Tratamento do campo identificador da cesta de compra

Os campos NUMCUPOM, NUMCAIXA e NUMERO_NOTA são necessários para identificar a cesta de compra. A empresa pode emitir uma venda através do cupom fiscal ou da nota fiscal, desta forma, quando a venda é feita através da nota fiscal os campos NUMCUPOM e NUMCAIXA ficarão nulos, e quando a venda é feita pelo cupom fiscal, não há numeração no campo NUMERO_NOTA, também ficando nulo. Outra questão é o campo NUMCAIXA, cada caixa que emite o cupom fiscal possuem o identificador NUMCUPOM único, porém esse número do cupom pode coincidir e ser o mesmo em caixas diferentes, com isso vendas diferentes podem ter o mesmo NUMCUPOM mas número de caixas distintos.

Para solucionar o problema descrito, foi criado um campo IDVENDA. Neste campo ficará preenchido o número da nota, caso o campo NUMCUPOM esteja NULL e o campo NUMERO_NOTA esteja preenchido e será armazenado o número do cupom junto com o número do caixa caso o campo NUMERO_NOTA esteja NULL e o campo NUMCUPOM e NUMCAIXA esteja preenchido. Na Figura 8 há um exemplo da tabela de como estes campos ficaram na base de dados.

3.2.4 Redução e projeção dos dados

A quarta etapa do KDD, conforme mencionado anteriormente, visa encontrar características úteis para representar os dados, reduzindo assim o número efetivo de variáveis ou encontrando representações mais viáveis para os dados. Desta forma foi realizado essas transformações nos agrupamentos dos produtos e todo essa tarefa será descrita a seguir.

3.2.4.1 Tratamento dos dados de Produto

Para selecionar os produtos cadastrados, utilizou-se o campo da tabela produto DATA_ULTIMA_MOVIMENTACAO para realizar a seleção dos produtos que tiveram movimentação no período determinado. Este campo foi muito útil, pois de 15.872 produtos cadastrados, a seleção resultou em 10.925 produtos cadastrados que foram movimentados neste período, desta forma reduziu o trabalho de preparação dos dados para itens que não foram vendidos no período dos dados coletados.

Em seguida realizou-se uma análise dos produtos cadastrados, pois há características dos produtos que não devem ser levadas em consideração ao realizar a mineração dos dados, como por exemplo, para o produto Escada constam 35 cadastros com diferentes quantidades de degraus e materiais, se a escada é de ferro ou alumínio. Desta forma, estes produtos podem ser classificados em um grupo Escada, pois esta característica depende da necessidade pessoal de cada cliente e não é interessante para a mineração de dados. Com isso será possível obter um melhor resultados na mineração da cesta de mercado.

Percebeu-se que na base de dados da empresa os grupos eram cadastrados de forma genérica, por exemplo, havia um grupo chamado Esquadrias e neste grupo estavam relacionados produtos como escada, basculantes e vários outros produtos. Já os subgrupos estavam sendo utilizados para identificar as marcas dos produtos, como por exemplo, Coral, Tigre e outras. Desta forma houve a necessidade de reformular estes grupos para que possamos classificar os produtos ocultando algumas características. Para que seja possível criar estes subgrupos é necessária uma análise cautelosa e deve ser uma etapa de destaque na execução de qualquer projeto de mineração de dados. Com esta análise bem feita aumenta a credibilidade em relatórios gerenciais, pois uma vez que os produtos não estiverem classificados corretamente, estas informações gerenciais estarão distorcidas.

Para realizar este agrupamento de forma mais rápida, criou-se duas colunas na tabela produto denominadas C1 e C2 através da implementação de uma função no SQL Server 2012. O motivo para isto é que percebeu-se que o primeiro nome dos produtos poderia ser utilizado na maioria dos casos para classificá-los e outros produtos que exigiam uma diferenciação maior poderia utilizar a combinação do primeiro nome cadastrado com o segundo, conforme é mostrado na Figura 9, o produto ESCADA DE ALUMINIO é classificado como ESCADA, desconsiderando a característica de quantidade de degraus. Já na Figura 10 podemos destacar produtos que mesmo com o mesmo nome inicial possuem

CODIGO	DESCRICAO	SUBGRUPO	C1	C2
7896020651055	ESCADA ALUMINIO 7 DEGRAUS MOR	ESCADA	ESCADA	ALUMINIO
7896020651048	ESCADA ALUMINIO 6 DEGRAUS MOR	ESCADA	ESCADA	ALUMINIO
7896020651024	ESCADA ALUMINIO 4 DEGRAUS MOR	ESCADA	ESCADA	ALUMINIO
005101	ESCADA ALUMINIO 3 DEGRAUS MOR	ESCADA	ESCADA	ALUMINIO

Figura 9 – Dados da tabela Produto classificados utilizando C1. Fonte: Desenvolvido pela autora.

CODIGO	DESCRICAO	SUBGRUPO	C1	C2
1569	FIXADOR CAL SACHE 150ML	FIXADOR CAL	FIXADOR	CAL
3680	FIXADOR PORTA ALIANCA RODAPE	FIXADOR PORTA	FIXADOR	PORTA

Figura 10 – Dados da tabela Produto classificados utilizando a combinação de C1 e C2. Fonte: Desenvolvido pela autora.

funções diferentes, como no exemplo, FIXADO DE CAL e FIXADOR DE PORTA, com isso foi utilizado com a combinação dos campos C1 e C2.

Este procedimento foi útil ao executar os comandos para selecionar e alterar os dados, porém mesmo assim, é uma etapa demorada e não pode ser totalmente automática, pois pode haver descrição de produtos com nomes diferentes, mas que representam a mesma coisa. Um exemplo é o grupo Tintas que possui um produto chamado “Rende Muito” e “Super Lavável” que são tintas, porém com características diferentes e que não será interessante para a mineração diferenciar essa categoria. Com isso, além de conhecer os itens é preciso agrupá-los da melhor forma para atingir os objetivos da mineração e tomar cuidado para evitar redundância de grupos.

Haviam 10.925 produtos cadastrados, e estes produtos foram agrupados em 685 grupos. Após esta etapa foram analisados todos os 685 grupos de forma cautelosa a fim de detectar problemas no agrupamento e refinar ainda mais os grupos e ao final foram criados 229 grupos de produtos. Essa redução de agrupamento de produtos será essencial para obter maior eficiência nos resultados da mineração.

3.2.5 Transformação dos dados para Mineração de Dados

Esta seção corresponde à quinta etapa do KDD, onde os dados serão transformados de forma que atenda às exigências do método que será utilizado na mineração de dados. Neste caso, conforme mencionado anteriormente, será utilizado a regra de associação aplicando o algoritmo *Apriori*. A ferramenta utilizada será *WEKA* e será detalhado a seguir as transformações dos dados para a criação do arquivo final de mineração.

Para que possa ser utilizado o algoritmo Apriori do WEKA é necessário que sejam realizadas algumas transformações nos dados e gerado um arquivo com a extensão .ARFF e formatação adequada. Após o tratamento dos dados nas etapas anteriores ao KDD descritas, foi criada uma nova tabela no banco de dados com apenas os campos necessários para a mineração (Figura 11).

	IDVENDA	GRUPO	NUM_MES	BAIRRO	CIDADE
1	291913-6	HIDRAULICO_TUBOS_E_CONEXOES	10	LOANDA	JOÃO MONLEVADE
2	322175-5	COLAS_E_ADESIVOS	11	NOVA ACLIMAÇÃO	JOÃO MONLEVADE
3	291913-6	TOMADA_INTERRUPTOR_PLACA_PINO	10	LOANDA	JOÃO MONLEVADE
4	291913-6	TOMADA_INTERRUPTOR_PLACA_PINO	10	LOANDA	JOÃO MONLEVADE
5	15047	MASSA_CORRIDA_ACRILICA_RAPIDA	04	VILA TANQUE	JOÃO MONLEVADE
6	17671	ESMALTE	09	JK	JOÃO MONLEVADE
7	17671	DILUENTE	09	JK	JOÃO MONLEVADE
8	325956-5	ARGAMASSA	11	LOANDA	JOÃO MONLEVADE
9	14498	HIDRAULICO_TUBOS_E_CONEXOES	03	NOSSA SENHORA DA CONCEIÇÃO	JOÃO MONLEVADE
10	291913-6	TOMADA_INTERRUPTOR_PLACA_PINO	10	LOANDA	JOÃO MONLEVADE
11	17671	SOLVENTE	09	JK	JOÃO MONLEVADE
12	18868	ESMALTE	10	LOANDA	JOÃO MONLEVADE
13	53207-7	TORNEIRAS_MISTURADOR_E_Arejador	09	VALE DO SOL	JOÃO MONLEVADE
14	291913-6	TOMADA_INTERRUPTOR_PLACA_PINO	10	LOANDA	JOÃO MONLEVADE
15	44309-7	DESEMPENO	08	LARANJEIRAS	JOÃO MONLEVADE
16	213217-5	TOMADA_INTERRUPTOR_PLACA_PINO	04	NOSSA SENHORA DO ROSÁRIO	JOÃO MONLEVADE
17	291913-6	Caixa_de_Passagem_Condulete_Sobrepor_Embutir_Tel...	10	LOANDA	JOÃO MONLEVADE
18	318828-5	LAMINA_SERRA	10	CARNEIRINHOS	JOÃO MONLEVADE
19	318828-5	BUJAO	10	CARNEIRINHOS	JOÃO MONLEVADE
20	318828-5	COLAS_E_ADESIVOS	10	CARNEIRINHOS	JOÃO MONLEVADE

Figura 11 – Tabela com atributos selecionados e dados tratados. Fonte: Desenvolvido pela autora.

Porém, os dados ainda não estão no formato necessário para gerar o arquivo de mineração, é necessário que a tabela seja invertida, onde o NUMCUPOM seja as linhas que corresponde a cesta de mercado e as colunas os GRUPOS. Para representar que um produto consta na cesta de compra, será preenchido o atributo desse produto com o número 1 e caso ele não pertença à cesta de compra ficará preenchido com '?'. A tabela foi gerada uma com 565.614 transações (cesta de compra) e 229 colunas (grupos de produtos). A Figura 12 representa parte desta tabela.

Foi criado também esta mesma tabela porém contendo o campo MÊS e outra tabela contendo os campos Bairro e Cidade para cada registro de compra, a fim de gerar as regras de vendas considerando as épocas do ano e determinar comportamento das vendas em diversas regiões.

O passo a passo para criação da tabela será descrito a seguir:

1. Criar a tabela contendo o campo identificador da venda e os campos para cada produto. Também poderá criar com os campos identificando o mês da venda, bairro e cidade do cliente.

2. Executa o comando SQL INSERT que insere os dados da tabela mostrada na Figura 12.

3. Executa o comando SQL UPDATE que altera os dados da tabela de > 1 para 1 e de 0 para ?.

4. Executa o comando SQL ALTER TABLE para excluir o campo identificador da venda, pois não será mais utilizado.

5. Exporta os dados para um arquivo de texto com o delimitador vírgula.

REJUNTE	RELE	REMOVEDOR	REPARO	RESINA	RESISTENCIA	REVESTIMENTO
?	?	?	?	?	?	?
?	?	?	?	?	?	?
?	?	?	?	?	?	?
?	?	?	?	?	?	?
?	?	?	?	?	?	?
?	?	?	?	?	?	?
?	?	?	?	?	?	?
?	?	?	?	?	?	?
?	?	?	?	?	?	?
?	?	?	?	?	?	?
?	?	?	?	?	?	?
?	?	?	?	?	?	?
?	?	?	?	?	?	?
?	?	?	?	?	?	1
?	?	?	?	?	?	?
?	?	?	?	?	?	?
?	?	?	?	?	?	?
?	?	1	?	?	?	?
?	?	?	?	?	?	1
?	?	?	?	?	?	?
?	?	?	?	?	?	?
?	?	?	?	?	?	?

Figura 12 – Parte da tabela Cesta de Compra. Fonte: Desenvolvido pela autora.

6. Formata o arquivo inserindo a @relation e @attributem conforme ilustrado na Figura **??**.

Os comandos SQL utilizados serão inseridos em anexo.

3.2.6 Mineração de dados

A etapa de mineração de dados poderá ser dividida em 4 experimentos, cada um para atingir um objetivo definido anteriormente.

3.2.6.1 Parâmetros utilizados no processo de mineração

Com o arquivo já configurado para atender as exigências do WEKA, ele foi aberto e o sistema fez a leitura. Selecionado a tarefa associação e o algoritmo *Apriori*, foram realizados algumas configurações dos parâmetros conforme Figura 13.

Segundo Waikato (2015) será explicado a seguir cada parâmetro:

1. *Car:* configurado como *True* para que as regras de associação de classe habilitadas sejam extraídas, em vez da regra geral da associação se for configurado como *False*. Esse parâmetro foi útil na mineração de dados utilizando dados como Mês e Região, pois estes atributos foram classificados para que fossem geradas regras considerando-

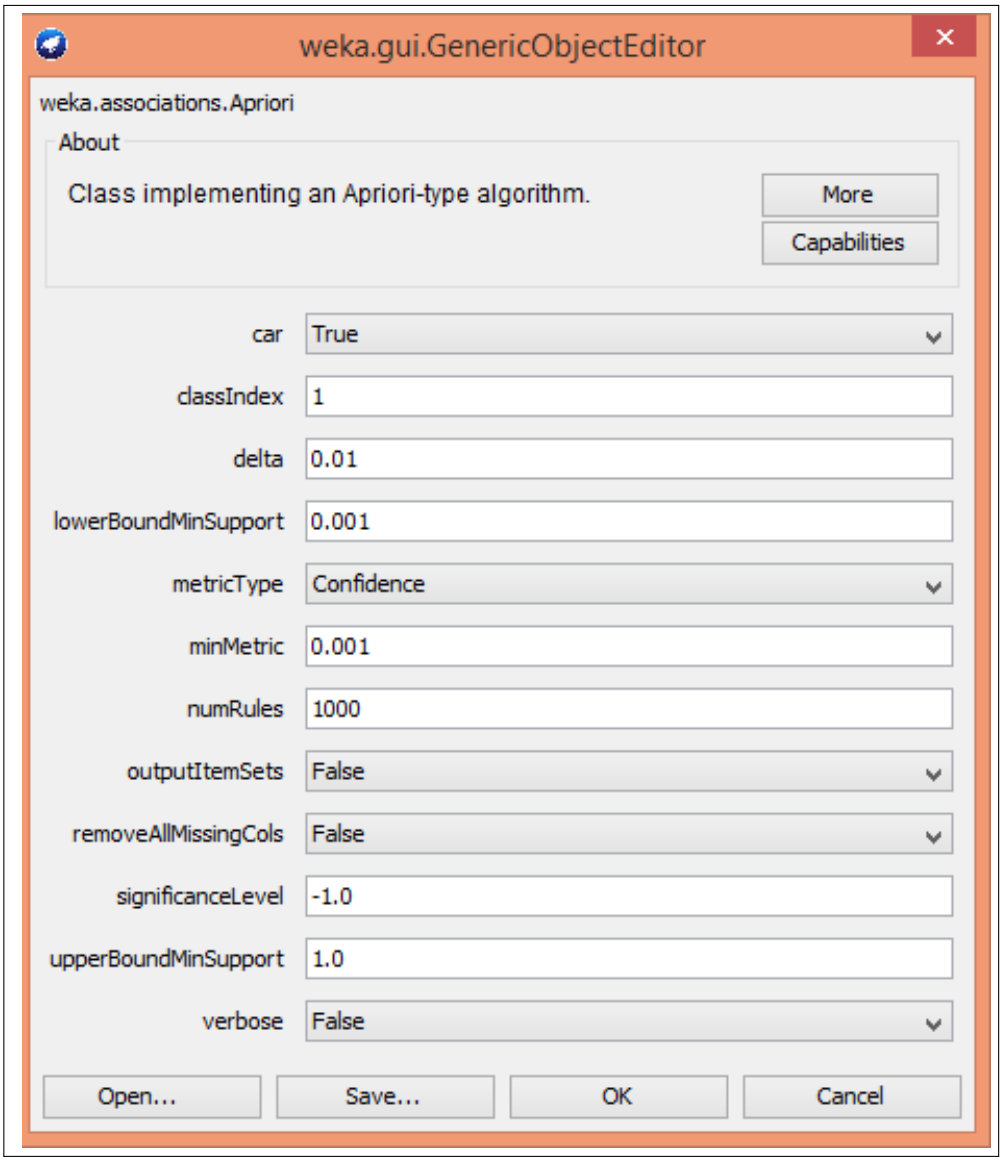


Figura 13 – Parâmetros de Configuração do *Apriori* no Weka Fonte: WEKA.

os. Para a análise de cesta de compra esse atributo foi considerado como *false*, pois não é necessário classificar os atributos da cesta.

2. *classIndex:* é o índice do atributo classe. Se for definido como -1, o ultimo atributo é tomado como classe. Nesse caso foi configurado como 1 para que o primeiro atributo seja considerado como classe, o atributo mês e região em cada etapa do processo.

3. *Delta:* a cada iteração, o *Apriori* diminui o suporte pelo valor especificado em delta.

4. *lowerBoundMinSupport:* define o limite inferior para o suporte mínimo.

5. *metricType:* define a métrica que será utilizada para classificação das regras. Neste trabalho foi utilizado a métrica confiança.

6. *minMetric:* é o valor mínimo para a métrica escolhida.

7. *numRules:* números de regras para encontrar.

8. *outputItemSets:* se configurado como *true*, na saída, além de exibir as regras mineradas, exibirá também os itemsets frequentes.

9. *removeAllMissingCols:* remova colunas com todos os valores em falta. Foi configurado como *false* pois os dados foram tratados antes da mineração não havendo nenhum campo com valores em falta.

10. *significanceLevel:* o nível de significância. Não foi alterado, mantendo o valor -1 padrão.

11. *upperBoundMinSupport:* limite superior para apoio mínimo. Não foi alterado, mantendo o valor 1.0 padrão.

12. *Verbose:* se ativado o algoritmo será executado no modo detalhado.

No próximo Capítulo serão apresentados os resultados obtidos.

4 Resultados e Discussão

Neste Capítulo foram apresentados os resultados e análises que podem ser úteis ao negócio em questão. Serão apresentados os resultados extraídos da mineração de dados e os possíveis padrões de compras nas cestas de mercado detectados. Também serão realizadas análises de vendas regionais, como o comportamento das vendas em determinadas regiões ao longo dos anos e o ramo de produtos que se destacam nesses comportamentos. E por fim será mostrado o comportamento das vendas de produtos nas épocas dos anos e análises que foram obtidas após este estudo.

4.1 Análise de Cesta de Compras

O suporte utilizado para gerar as regras foi baixo, devido à grande variedade de produtos vendidos na empresa e ao grande volume de registros encontrados. Uma regra que demonstre uma confiança alta ou de 100%, independente do suporte mínimo, é muito provável que será uma informação útil, se não for óbvia. Algumas regras, com confiança alta, podem ser vistas na tabela 3.

Tabela 3 – Regras de Associação da cesta de compra.

Confiança	Regra: X →	Y
0.99	[ACAB BANHEIRO] + [CAIXA SIFON] +[VEDA ROSCA] →	[TUBOS E CONEXÕES]
0.99	[ACAB BANHEIRO] + [COLAS E ADESIVOS] + [LIXAS] +[VEDA ROSCA] →	[TUBOS E CONEXÕES]
0.99	[LIXAS] + [TINTA] +[TRINCHA OU PINCEL] →	[ROLO DE PINTURA]
0.98	[BUCHA FIXAÇÃO] + [PRATELEIRA] →	[PARAFUSO]
0.97	[ABRAÇADEIRA] + [BUCHA FIXAÇÃO] + [TUBOS E CONEXÕES] →	[PARAFUSO]
0.97	[MASSA CORRIDA] + [TINTAS] →	[ROLO DE PINTURA]
0.96	[ANEL VEDAÇÃO] + [CUBA OU LAVATÓRIO] + [VASO SANITÁRIO] →	[PARAFUSO]
0.95	[ABRAÇADEIRA] + [BUCHA FIXAÇÃO] →	[PARAFUSO]
0.94	[BUCHA FIXAÇÃO] + [CABO ELÉTRICO] + [COLAS E ADESIVOS] →	[PARAFUSO]
0.94	[ESPAÇADOR] + [REJUNTE] + [REVESTIMENTO] →	[ARGAMASSA]
0.92	[ANEL VEDAÇÃO] + [REVESTIMENTO] →	[PARAFUSO]
0.91	[ENGATE] + [PARAFUSO] + [TORNEIRA] + [VÁLVULA LAVATÓRIO] →	[REVESTIMENTO]
0.91	[COLAS E ADESIVOS] + [ESMALTE] + [MASSA CORRIDA] + [SOLVENTE] →	[TINTAS]
0.90	[BUCHA FIXA] + [TUBOS E CONEX] + [TOMADAS] + [PARAFUSO] →	[REVESTIMENTO]
0.90	[ACAB BANHEIRO] + [CAIXA DAGUA] →	[TUBOS E CONEXÕES]
0.90	[ANEL VEDAÇÃO] + [CUBA OU LAVATÓRIO] + [PARAFUSO] →	[VASO SANITÁRIO]
0.90	[PASTA LUBRIFICANTE] →	[TUBOS E CONEXÕES]
0.90	[BUCHA FIXAÇÃO] + [COLAS E ADESIVOS] + [LAMPADAS E LUMIN] →	[PARAFUSO]
0.90	[ANEL VEDAÇÃO] + [ENGATE] + [TUBOS E CONEXÃO] →	[PARAFUSO]
0.89	[COLAS E ADESIVOS] + [MASSA CORRIDA] + [SOLVENTE] + TINTAS →	[LIXAS]
0.87	[ESPAÇADOR] + [REVESTIMENTO] →	[ARGAMASSA]
0.86	[ANEL VEDAÇÃO] + [ARGAMASSA] →	[PARAFUSO]
0.86	[ABRAÇADEIRA] + [COLAS E ADESIVOS] + [PARAFUSO] →	[BUCHA FIXAÇÃO]
0.78	[ESPAÇADOR] + [REJUNTE] →	[ARGAMASSA]
0.75	[CABO ELÉTRICO] + [CANALETA] →	[TOMADAS]
0.74	[ABRAÇADEIRA] + [REGULADOR DE GÁS] →	[MANGUEIRA DE GÁS]
0.74	[CABO ELÉTRICO] + [TERMINAL E CONECTORES] →	[TUBOS E CONEXÕES]
0.73	[CABO ELÉTRICO] + [DISJUNTOR] + [LÂMPADAS E LUMINÁRIAS] →	[TOMADAS]
0.69	[LIXAS] + [LONA] →	[TINTA]
0.69	[CHUVEIRO] + [ENGATE] →	[TUBOS E CONEXÕES]
0.67	[LONA] + [TINTA] →	[LIXAS]
0.66	[FUNDO E SELADOR] + [ROLO DE PINTURA] + [TINTA] →	[LIXAS]
0.63	[DILUENTE] + [ESMALTE] + [SOLVENTE] →	[LIXAS]
0.63	[LIXAS] + [LONA] →	[COLAS E ADESIVOS]
0.62	[GESSO] + [TINTA] →	[LIXAS]
0.62	[ARGAMASSA] + [MASSA CORRIDA] →	[LIXAS]
0.56	[BUJÃO] + [COLAS E ADESIVOS] + [TUBOS E CONEXÕES] →	[ACAB BANHEIRO]
0.55	[COLAS E ADESIVOS] + [TEXTURA] →	[LIXAS]
0.54	[ESTOPA] + [LIXAS] →	[COLAS E ADESIVOS]
0.52	[VÁLVULA] + [VEDA ROSCA] →	[ACAB BANHEIRO]
0.51	[SOLVENTE] + [TINTAS] →	[LIXAS]

As listas de regras geradas é muito extensa. A ferramenta de mineração tem capacidade de gerar muitas informações deste tipo. A conclusão que pode-se chegar destas regras é que como a empresa busca realizar sempre a venda assistida, com acompanhamento de um vendedor especializado, as regras são próximas do óbvio. Neste tipo de ramo de comercialização, o cliente muitas vezes não sabe o que vai precisar e estas vendas assistidas

serve para orientá-lo a levar algum produto relacionado com o que ele irá precisar.

A empresa poderá verificar a estratégias de organização dos itens atualmente, a fim de detectar o melhor posicionamento dos produtos e orientar os vendedores menos experientes sugerir aos clientes estas associações de produtos geradas.

4.2 Análises das Vendas

Analisando as vendas da empresa, sem considerar dimensões de ano, mês e regiões, detectamos alguns comportamentos gerais da empresa.

Na análise regional das vendas da empresa serão mostradas as regiões em destaque no faturamento e os principais produtos adquiridos pelos clientes destas regiões. Este estudo tem como objetivo detectar o comportamento das vendas da empresa nas cidades e bairros que seus clientes residem. Foi constatado que 96% das vendas da empresa é efetuada na cidade de João Monlevade onde a empresa esta fixada, 3% correspondem a cidades vizinhas como Bela Vista de Minas, Rio Piracicaba, Alvinópolis, São Gonçalo do Rio Abaixo, São Domingos do Prata, Nova Era, Itabira, Belo Horizonte e Santa Maria de Itabira. O restante representam as demais cidades que não serão detalhadas, pois a frequência de compras é muito reduzida se analisadas individualmente. Para a cidade de João Monlevade a análise será mais detalhada por representar um grande número de vendas, com isso haverá uma divisão por bairros. As cidades vizinhas serão analisadas sem divisões de bairros.

Em relação à venda geral de produtos da empresa analisou-se o volume de vendas (em reais) dos produtos mais vendidos no período de Janeiro de 2010 a Dezembro de 2013. O segmento líder de vendas na empresa são as Tintas com faturamento aproximado de 5 milhões, em seguida os Revestimentos com faturamento aproximado de 4 milhões e em terceiro lugar os Tubos e Conexões Hidráulicas com aproximadamente 2,5 milhões. Na Figura 14 pode-se verificar os produtos mais vendidos.

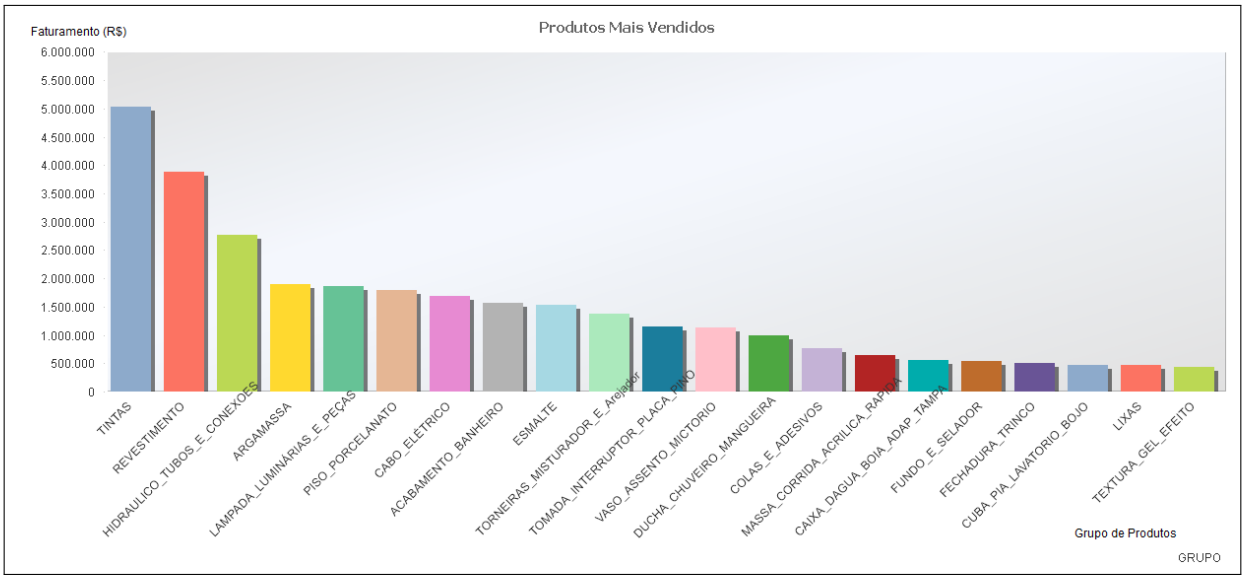


Figura 14 – Gráfico de ranking de vendas.

Analisando o comportamento geral das vendas mensais no período de 2010 a 2013, pode-se perceber queda nos primeiros meses, porém mais acentuada em Fevereiro e Abril, devido aos feriados de Carnaval e Semana Santa e também devido ao mês de Fevereiro ser menor. Já nos meses de Maio, Junho, Julho e Agosto, há um aumento nas vendas, devido ao fato de ser época de pouca chuva, sendo propício para realização de reformas e construções. De Setembro a Dezembro há uma queda, mas não é muito acentuada como no inicio do ano, pois nesta época os clientes fazem reformas para o novo ano que se inicia. A Figura 15 comprova esses fatos.

4.2.1 Análise Regional das Vendas

A análise regional na cidade de João Monlevade foi realizada com base nos dados de bairro dos clientes preenchidos na venda. Porém, de todas as vendas em João Monlevade apenas 53% estão vinculadas ao cadastro de clientes contendo assim, as informações de bairros. Desta forma, o resultado poderá ficar comprometido por representar um número grande de transações sem a informação necessária, mas será analisada a tendência das vendas nas regiões de João Monlevade.

O bairro que representa maior faturamento de vendas na empresa é o bairro Carneirinhos. O motivo principal se deve ao fato da empresa estar situada nele, com isso os clientes que residem nele recorrem à empresa por estarem mais próximos. Outro motivo é por ser um bairro tradicional na cidade e muito conhecido, os bairros vizinhos acabam sendo registrados como Carneirinhos. Um exemplo são os moradores do bairro Alvorada, que antigamente era considerado como Carneirinhos, e ainda utiliza o nome antigo para informar seu endereço ao invés de considerar o nome atual. Este fator também poderá causar alterações no resultado. A Figura 16 mostra o faturamento de vendas (em reais),

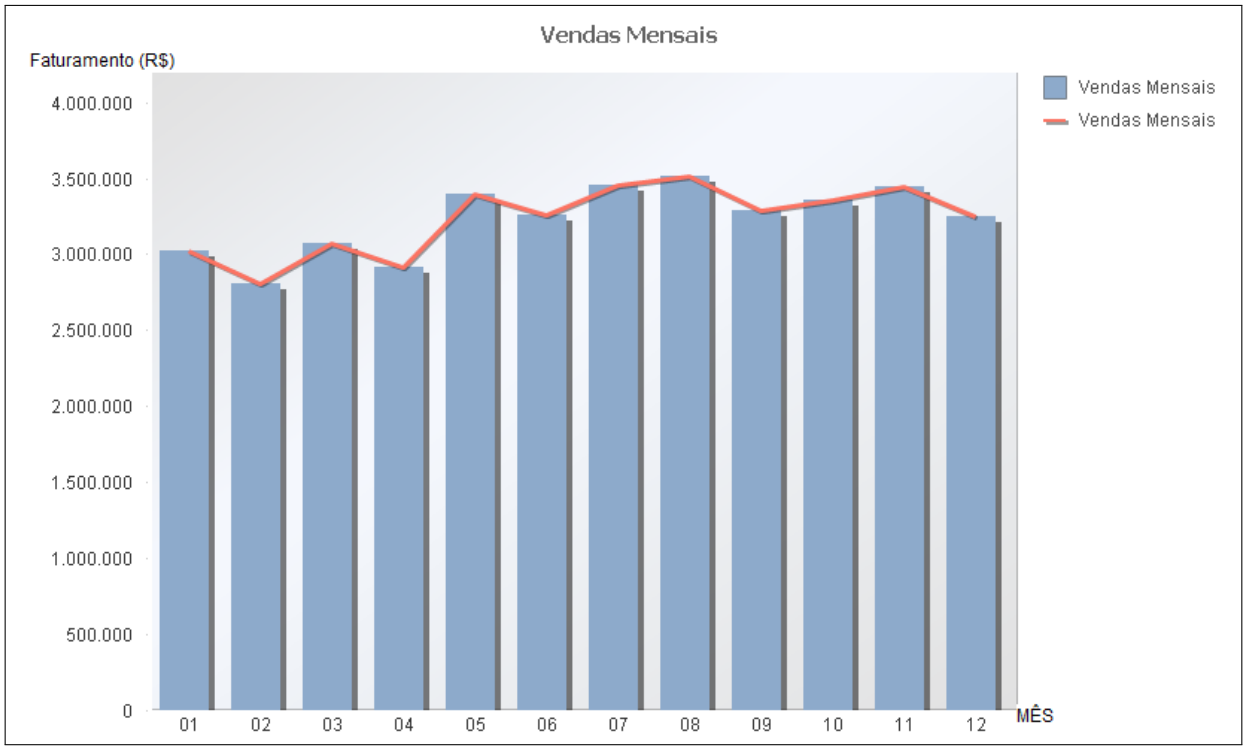


Figura 15 – Comportamento mensal das vendas.

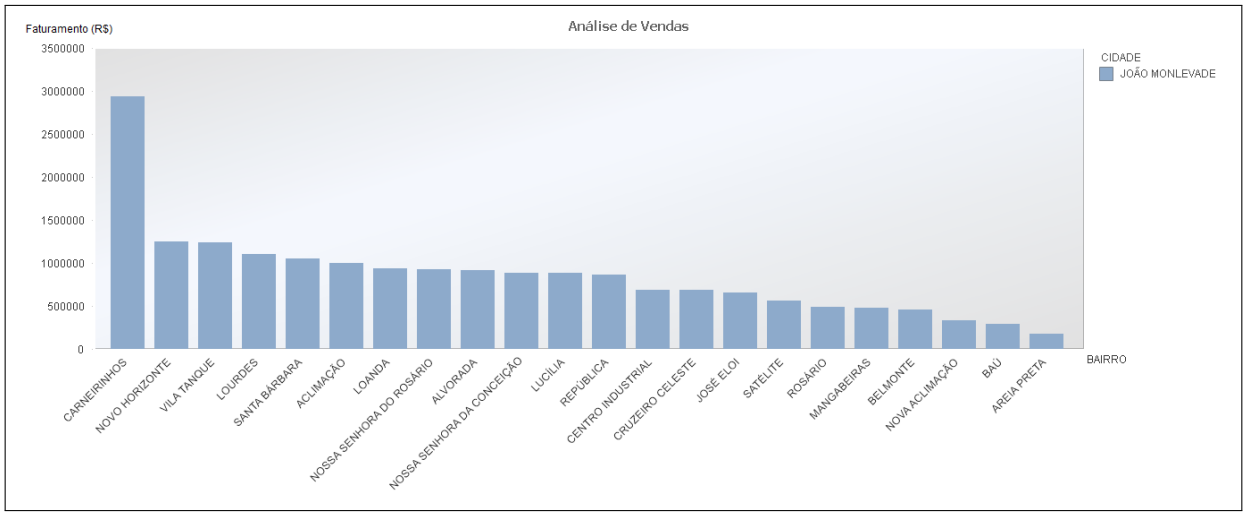


Figura 16 – Vendas por bairros de João Monlevade.

de Janeiro de 2010 a Dezembro de 2013, dos principais bairros de clientes que compram na empresa.

Analisando as vendas anuais, há uma queda nas vendas no do bairro Carneirinhos (Figura 17), porém este fato pode estar relacionado com a nomenclatura dos bairros explicada anteriormente, pois um dos bairros que era considerado Carneirinhos, o bairro Alvorada, teve um crescimento considerável das vendas (Figura 18), provavelmente os moradores estão considerando o nome atual do bairro.

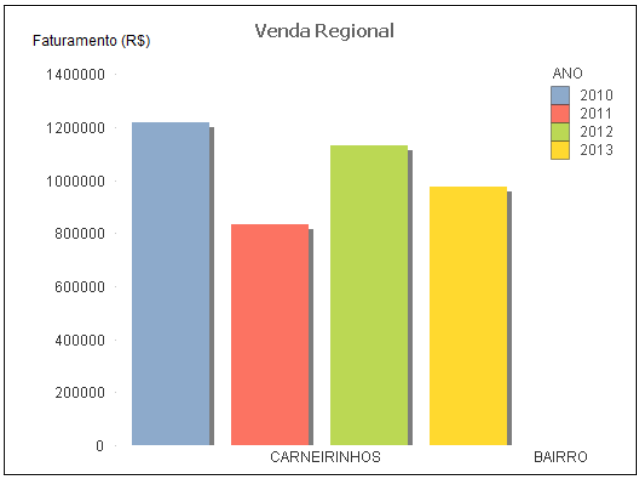


Figura 17 – Venda anual do bairro Carneirinhos.

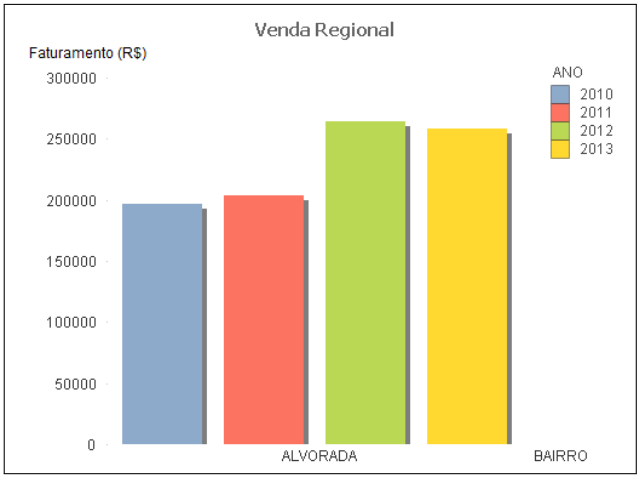


Figura 18 – Venda anual do bairro Alvorada.

4.2.1.1 Venda de Produto na Região de João Monlevade

Foi realizada uma análise dos produtos mais vendidos em alguns bairros. Na maioria dos bairros o comportamento da venda de produtos é semelhante ao comportamento geral dos produtos mais vendidos (Figura 14), porém há alguns bairros com comportamento diferenciado que será comentado a seguir.

Os clientes do bairro Cruzeiro Celeste compram muito Revestimento e em segundo lugar no ranking das compras por faturamento (R$) (Figura 19), vêm as Tintas que no geral é o que mais se vende na empresa. Clientes deste bairro também procuram Pisos/Porcelanatos. O bairro Cruzeiro Celeste é sede de outros depósitos de materiais de construção, concorrentes do Ulete Mota. Com isso, a análise se torna interessante pelo fato de que alguns clientes preferem comprar no Ulete Mota, ao invés de comprar nos depósitos mais próximos, estes segmentos. Já outros produtos como tubos e conexões hidráulicas, cabos elétricos, não há tanta procura como nos outros bairros da cidade, provavelmente pois os cliente optam por comprar nos depósitos mais próximos.

Foi possível observar também que os clientes do bairro Nossa Senhora do Rosário

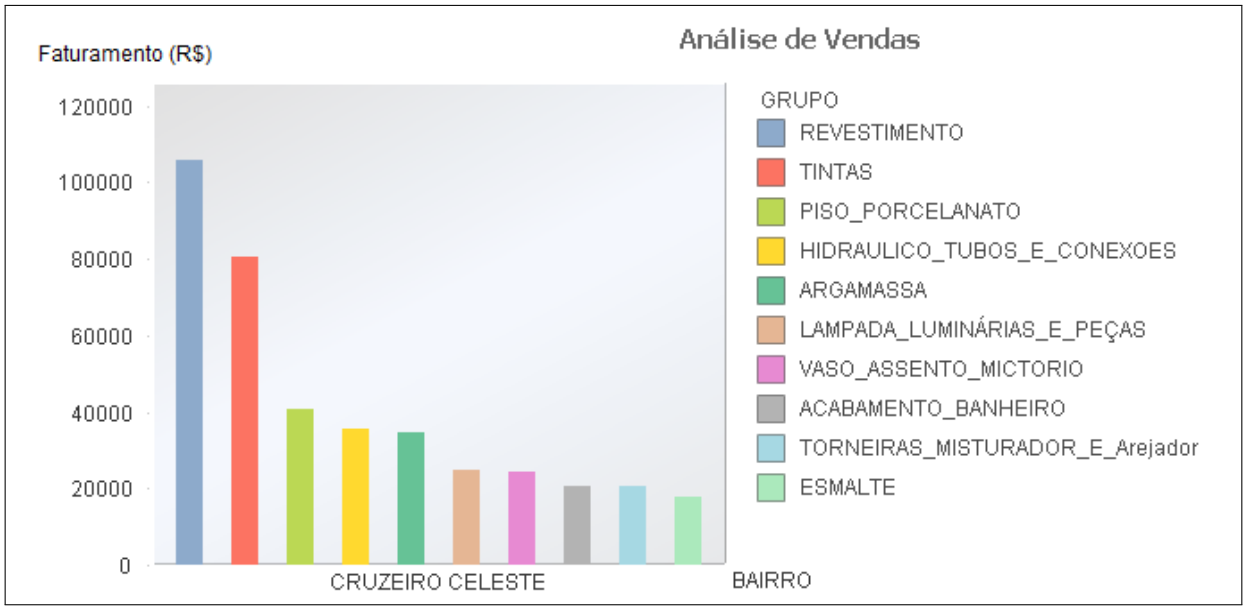


Figura 19 – Vendas de produtos no bairro Cruzeiro Celeste.

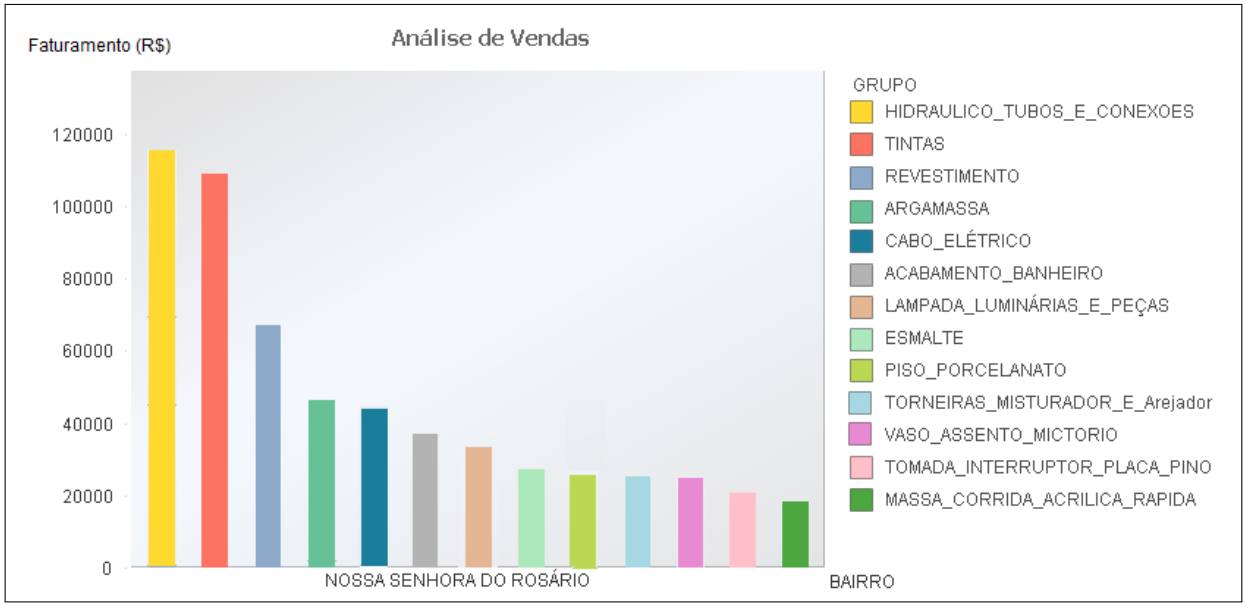


Figura 20 – Vendas de produtos no bairro Nossa Senhora do Rosário.

(Figura 20) compram mais Tubos e Conexões Hidráulicas. Em segundo lugar vêm as Tintas e em terceiro os Revestimentos, porém com um volume bem abaixo que os outros bairros analisados.

4.2.1.1.1 Regras geradas na venda de produtos regionais por bairros de João Monlevade

A fim de detectar informações sobre as vendas nos bairros de João Monlevade, foram geradas as regras de associação. Porém a confiança da regra é muito baixa, devido à grande quantidade de bairros, produtos e transações da base de dados. Para minimizar esse problema, foram geradas regras de associação dos bairros com maior faturamento (R$), para tentar detectar se existe um perfil de compras de clientes nos bairros selecionados.

Na seleção dos bairros, o bairro carneirinhos impactou muito nos resultados, pois a grande maioria das regras envolviam este bairro, por ser o mais frequente. Sendo assim, foram realizadas análises com este bairro e também excluindo-o da listagem para que fossem geradas regras dos outros bairros. Mesmo com esta divisão as confianças não foram altas, sendo que a maior confiança gerada nas regras foi de 0.35, devido ao grande número de variedade de produtos.

As regras do bairros envolvendo o bairro carneirinhos que tiveram maior confiança envolviam produtos como: Gesso Rápido, Tintas, Lixas, Colas e Adesivos e Massa Corrida. A confiança variando de 0.30 a 0.35, conforme mostrado na Tabela 4. Estas regras envolvendo Gesso Rápido só foram observadas nesse bairro.

Tabela 4 – Regras de vendas de produtos por bairro

Confiança	Regra: X →	Y
0.35	[GESSO] + [TINTAS] →	[BAIRRO=CARNEIRINHOS]
0.34	[GESSO] + [LIXAS] + [TINTAS] →	[BAIRRO=CARNEIRINHOS]
0.33	[COLAS E ADESIVOS] + [GESSO] + [TINTAS] →	[BAIRRO=CARNEIRINHOS]
0.30	[GESSO] + [MASSA CORRIDA] + [TINTAS] →	[BAIRRO=CARNEIRINHOS]

As regras geradas envolvendo os principais bairros, considerando os de maior faturamento (R$) para a empresa, pode-se destacar as seguintes:

- O bairro Loanda se destacou na compra de Verniz, sendo o de maior frequência com confiança 0.29. E o outro bairro que gerou regra para esse produto produto foi o Santa Bárbara, com confiança de 0.13. Também no bairro Loanda outro produto que se destacou pela frequência de compras neste bairro foram as Ferramentas de Polir com confiança de 0.47 e esse produto juntamente com o Esmalte Sintético a confiança é ainda maior, sendo de 0.63.

Tabela 5 – Regras de vendas de produtos por bairro

Confiança	Regra: X →	Y
0.29	[VERNIZ] →	[BAIRRO=LOANDA]
0.47	[ESMALTE SINTÉTICO] + [VERNIZ] →	[BAIRRO=LOANDA]
0.13	[VERNIZ] →	[BAIRRO=SANTA BÁRBARA]
0.63	[FERRAMENTA DE POLIR] + [ESMALTE SINTÉTICO] →	[BAIRRO=LOANDA]
0.47	[FERRAMENTA DE POLIR] →	[BAIRRO=LOANDA]

- O bairro Nossa Senhora do Rosário se destacou pela frequência de compras de Baterias e Pilhas, com confiança de 0.32. Também foram geradas regras para o bairro Nossa Senhora do Rosário envolvendo Tubos e Conexões. Conforme foi apresentado na Figura 20 é o produto que representa maior faturamento (R$) neste bairro. As regras envolvem Abraçadeiras e Colas e Adesivos com confiança de 0.28.

Tabela 6 – Regras de vendas de produtos por bairro

Confiança	Regra: X →	Y
0.32	[BATERIA E PILHAS] →	[BAIRRO=N. SRA. DO ROSÁRIO]
0.28	[ABRAÇADEIRA] + [COLAS E ADESI] + [TUBOS E CONEX] →	[BAIRRO=N. SRA. DO ROSÁRIO]
0.25	[CAIXA DAGUA] + [VEDA ROSCA] →	[BAIRRO=N. SRA. DO ROSÁRIO]

- O bairro Vila Tanque, é o terceiro com maior faturamento (R$). As regras que destacam neste bairro envolvem produtos como Esmalte Sintético, Massa Corrida, Solvente e Tintas, com confianças de 0.23.

Tabela 7 – Regras de vendas de produtos por bairro

Confiança	Regra: X →	Y
0.23	[ESMALTE] + [MASSA CORRIDA] + [SOLVENTE] + [TINTAS] →	[BAIRRO=VILA TANQUE]
0.23	[FUNDO E SELADOR] + [LIXAS] + [MASSA CORRIDA] →	[BAIRRO=VILA TANQUE]
0.20	[ARGAMASSA] + [COLAS E ADESIVOS] →	[BAIRRO=VILA TANQUE]
0.16	[TINTA PISO] →	[BAIRRO=VILA TANQUE]

Outras regras foram geradas para os bairros, mas a confiança diminui ainda mais, e não se pode gerar muitas conclusões destas informações.

4.2.2 Vendas em Cidades Vizinhas

A primeira análise realizada foi em relação ao comportamento das vendas nos últimos anos, de 2010 a 2013 para detectar se houve aumento ou queda as vendas nas cidades. Na Figura 21, pode-se verificar que de todas as cidades vizinhas analisadas, a cidade em que os clientes mais compram na empresa é Bela Vista de Minas, provavelmente por ser a cidade mais próxima de João Monlevade. Podemos observar também que sua frequência de compra não oscilou muito durante os anos. Já a cidade de Rio Piracicaba não houve variações de 2010 para 2011, mas houve um crescimento significativo até 2013. Este crescimento se deu pelo aumento das vendas significativo dos produtos Revestimento e Pisos/Porcelanato (Figura 22).

Através da análise individual de vendas das cidades vizinhas que houveram crescimento ao longo dos anos São Domingos do Prata, São Gonçalo do Rio Abaixo e Nova era, foi possível concluir que clientes destas cidades buscam na empresa por variedade e qualidade de revestimento e pisos, que provavelmente não encontram a melhor relação de custo/benefício nos depósitos de sua região e essa necessidade aumentou consideravelmente de 2012 em diante.

4.2.2.1 Regras geradas na venda de produtos regionais em cidades vizinhas

Analisando as regras de associação geradas pela ferramenta, constatamos que de todas as cidades vizinhas, a cidade de Rio Piracicaba é a que mais compra revestimentos na

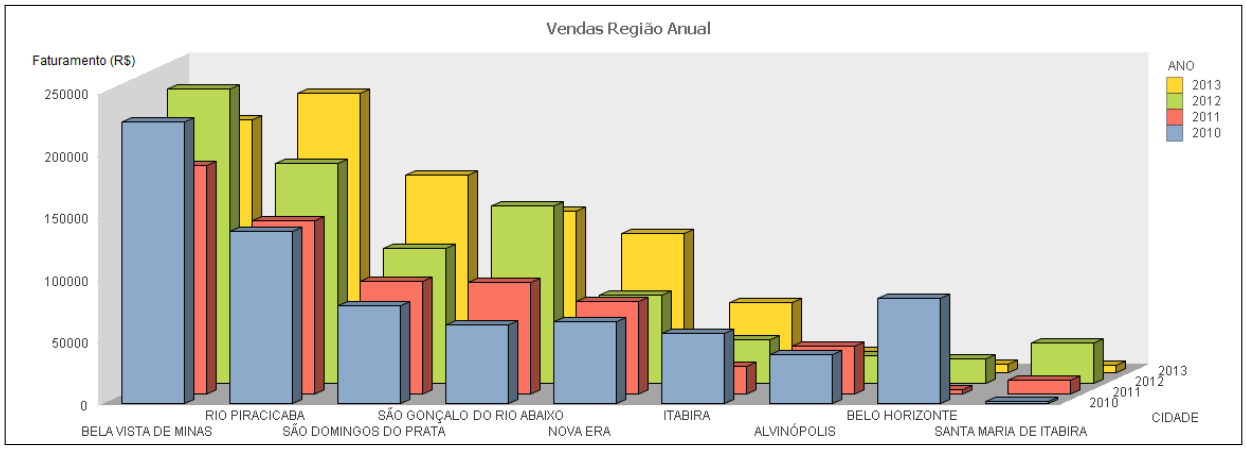


Figura 21 – Venda anual em cidades vizinhas.

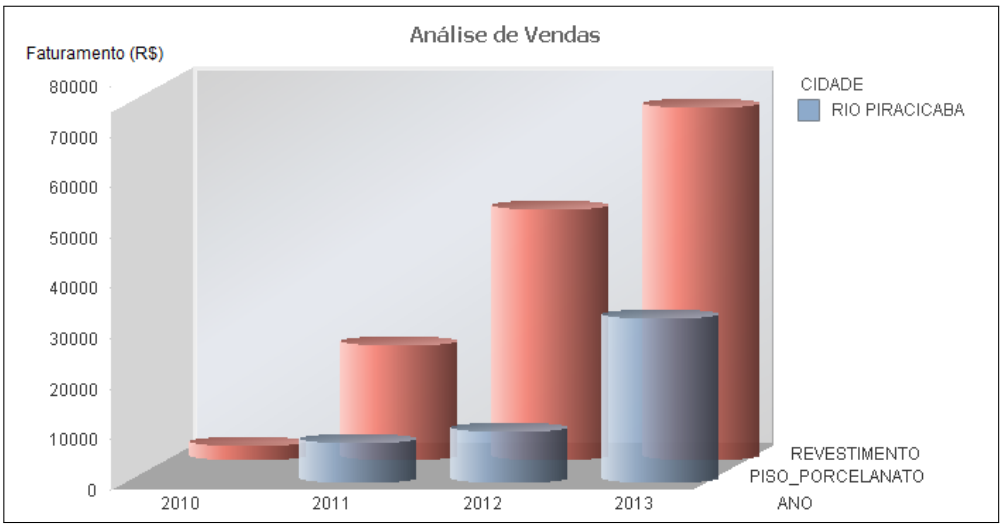


Figura 22 – Produto causadores da evolução das vendas da cidade de Rio Piracicaba.

empresa, isso pode ser percebido através da confiança que é mais alta, mesmo não sendo a cidade de clientes que esta em primeiro lugar no ranking das que mais compram na empresa, esta está em segundo. Pode-se comprovar esta informação nas regras geradas conforme apresentado na Tabela 8.

Tabela 8 – Regras de venda de Revestimento por bairros

Confiança	Regra: X →	Y
0.28	[REVESTIMENTO] →	[CIDADE=RIO PIRACICABA]
0.27	[REVESTIMENTO] →	[CIDADE=BELA VISTA DE MINAS]
0.13	[REVESTIMENTO] →	[CIDADE=SÃO DOMINGOS DO PRATA]
0.13	[REVESTIMENTO] →	[CIDADE=NOVA ERA]
0.12	[REVESTIMENTO] →	[CIDADE=SÃO GONÇALO DO RIO ABAIXO]

Em relação a venda de tintas, a cidade que mais compra é Bela Vista de Minas, seguida

de Rio Piracicaba e depois as outras cidades com confianças bem menores (Tabela 9).

Tabela 9 – Regras de venda de Tintas por bairros

Confiança	Regra: X →	Y
0.35	[TINTAS] →	[CIDADE=BELA VISTA DE MINAS]
0.29	[TINTAS] →	[CIDADE=RIO PIRACICABA]
0.10	[TINTAS] →	[CIDADE=SÃO GONÇALO DO RIO ABAIXO]
0.10	[TINTAS] →	[CIDADE=SÃO DOMINGOS DO PRATA]

A venda de Pisos/Porcelanatos têm a mesma tendência de vendas nas cidades próximas, podemos verificar nas regras geradas com confiança de $0,20$ a $0,22$.

Tabela 10 – Regras de venda de Piso/Porcelanato por bairros

Confiança	Regra: X →	Y
0.22	[PISO E PORCELANATO] →	[CIDADE=SÃO GONÇALO DO RIO ABAIXO]
0.21	[PISO E PORCELANATO] →	[CIDADE=SÃO DOMINGOS DO PRATA]
0.21	[PISO E PORCELANATO] →	[CIDADE=BELA VISTA DE MINAS]
0.20	[PISO E PORCELANATO] →	[CIDADE=RIO PIRACICABA]

Outras regras para a cidade de Rio Piracicaba são descritas abaixo. Regras envolvendo argamassa, rejunte e revestimento com confiança de $0,41$. De todas as cidades esta foi a única que gerou regras contendo fechadura/trinco, provavelmente com o intuito de aumentar a segurança das casas ou apenas por motivos de reformas ou trocas (Tabela 11).

Tabela 11 – Regras de venda de Rio Piracicaba

Confiança	Regra: X →	Y
0.36	[ARGAMASSA] + [REJUNTE] + [REVESTIMENTO] →	[CIDADE=RIO PIRACICABA]
0.36	[FECHADURA OU TRINCO] →	[CIDADE=RIO PIRACICABA]
0.32	[TUBOS E CONEXÕES] + [TORNEIRAS] →	[CIDADE=RIO PIRACICABA]
0.30	[FUNDO E SELADOR] →	[CIDADE=RIO PIRACICABA]

De todas as cidades vizinhas, conforme mencionado, a cidade de Bela Vista de Minas possui os clientes que mais compram na empresa. Suas compras estão muito relacionadas com Tintas, Revestimentos, Pisos/Porcelanato, porém há várias regras geradas com a relação de outros produtos e todas com confiança alta em relação ás outras cidades. Alguns exemplos de regras podem ser vistos na Tabela 12.

Tabela 12 – Regras de venda em Bela Vista de Minas

Confiança	Regra: X →	Y
0.52	[LAMPADA LUMINÁRIAS E PEÇAS] + [TOMADAS] →	[CIDADE=BELA VISTA DE MINAS]
0.51	[COLAS E ADESIVOS] + [TUBOS E CONEXÕES] →	[CIDADE=BELA VISTA DE MINAS]
0.50	[CABO ELÉTRICO]] + [COLAS E ADESIVOS] →	[CIDADE=BELA VISTA DE MINAS]
0.50	[TOMADA E INTERRUPTOR] →	[CIDADE=BELA VISTA DE MINAS]
0.48	[TUBOS E CONEXÕES] →	[CIDADE=BELA VISTA DE MINAS]
0.47	[VEDA ROSCA] →	[CIDADE=BELA VISTA DE MINAS]

Para a cidade São Domingos do Prata, só foram geradas regras dos produtos Pisos/-Porcelanatos, Tintas, Revestimento, Argamassa e Lampadas/Luminárias. Provavelmente a regra de vendas de outros produtos foram geradas com suporte muito baixo e pelo número de regras geradas foram podadas. Conclui-se também que os clientes dessa cidade procuram mais por Pisos/Porcelanatos, devido a confiança da regra ser alta, em relação aos outros produtos (Tabela 13).

Tabela 13 – Regras de venda em São Domingos do Prata

Confiança	Regra: X →	Y
0.21	[PISO E PORCELANATO] →	[CIDADE=SÃO DOMINGOS DO PRATA]
0.13	[REVESTIMENTO] →	[CIDADE=SÃO DOMINGOS DO PRATA]
0.10	[LAMPADA LUMINÁRIA E PEÇAS] →	[CIDADE=SÃO DOMINGOS DO PRATA]
0.10	[TINTAS] →	[CIDADE=SÃO DOMINGOS DO PRATA]
0.10	[ARGAMASSA] →	[CIDADE=SÃO DOMINGOS DO PRATA]

4.2.3 Análise de vendas de produtos nas épocas do ano

Nesta seção serão apresentados alguns produtos que têm um comportamento diferenciado em determinadas épocas do ano, seja mensal ou anual. Estas análises foram realizadas através de gráficos

Um produto que pode ser considerado sazonal vendido na empresa são as Duchas ou Chuveiros e também as Resistências de Chuveiros. Pode-se constatar (Figura 23) que as vendas destes produtos aumentam consideravelmente no inverno, pois nesta época as pessoas utilizam o chuveiro na maior potência para que fique mais quente e aumenta as chances da resistência queimar e as pessoas comprarem resistências para trocar ou mesmo um novo chuveiro. O gráfico a seguir mostra como estes dois tipos de produtos têm o mesmo comportamento de aumento de vendas em Maio até Agosto.

Produto que se destaca nas épocas do ano, são as mangueiras de jardim, pois em épocas de pouca chuva, a venda das mangueiras aumentam consideravelmente (Figura 24).

O principal produto de venda da empresa é a Tinta, analisando o comportamento de vendas mensal pode-se observar que acompanha a evolução geral das vendas da empresa

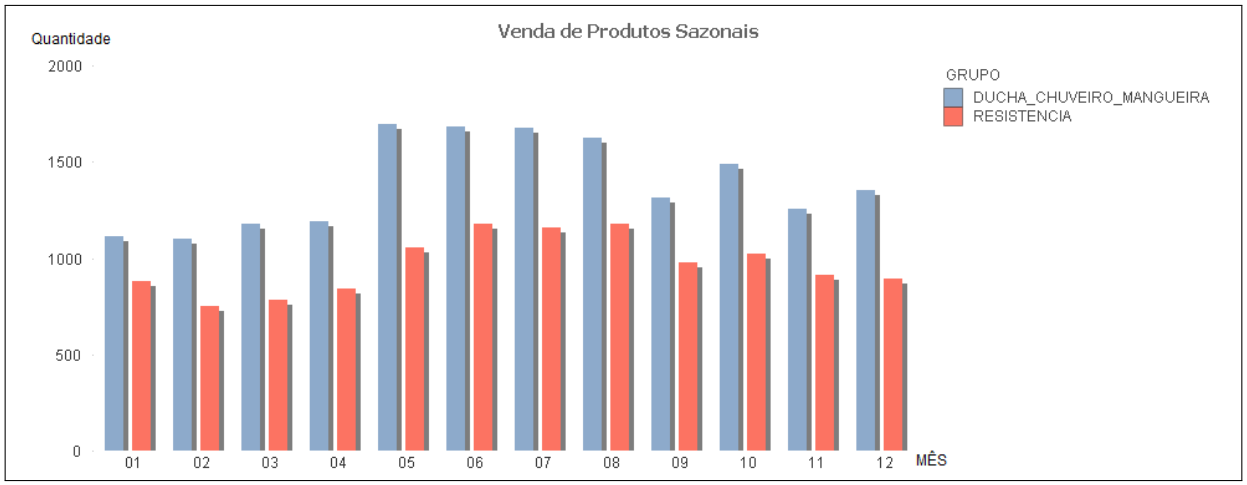


Figura 23 – Vendas de Duchas ou Chuveiros e Resistências de Chuveiro.

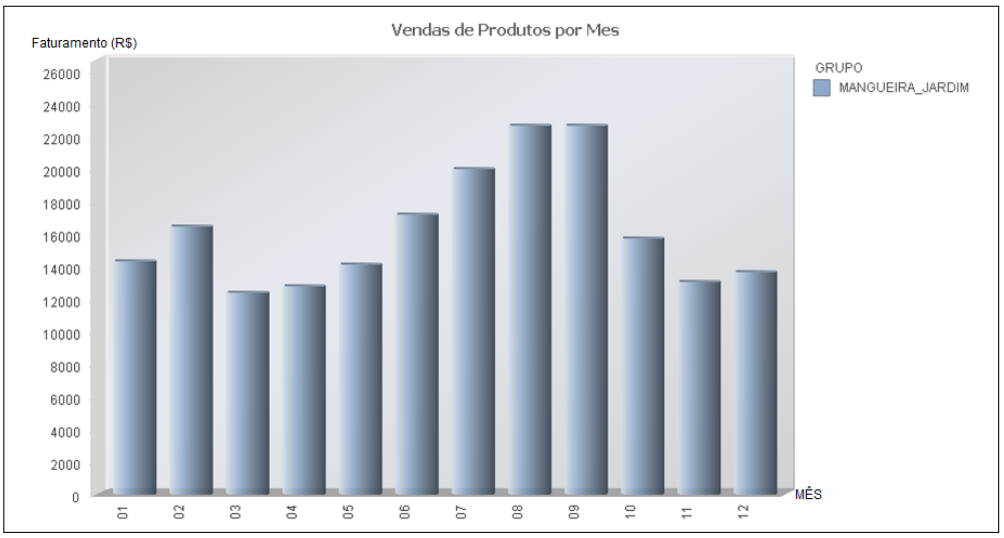


Figura 24 – Vendas de Mangueiras de Jardim.

(Figura 25, com isso observamos os mesmos motivos da queda em Fevereiro e Abril, também o crescimento das vendas em época de pouca chuva.

Em relação a venda do Verniz, foi detectado um comportamento ao longo dos anos impactante, pois em 2010 vendia-se um volume muito maior que nos anos seguintes (Figura 26). Este fator se deve às construções atuais, pois houve uma diminuição na utilização de madeiras, como por exemplo as janelas de madeira que antes eram muito utilizadas e atualmente usa-se alumínio ou vidro temperado.

A venda de Lâmpadas, Luminárias e Peças aumenta consideravelmente em Dezembro (Figura 27), porém este fato é desconhecido pela empresa, visto que não comercializam iluminações voltadas para festas natalinas. Mas se torna um fato curioso, pois esse comportamento se dá em todos os anos, levando a possibilidade de ser devido á iluminação para festas de fim de ano.

Conforme mostra a Figura 28, as vendas de Argamassa e Revestimento têm praticamente o mesmo comportamento, pois no mês que há aumento das vendas de Revestimento,

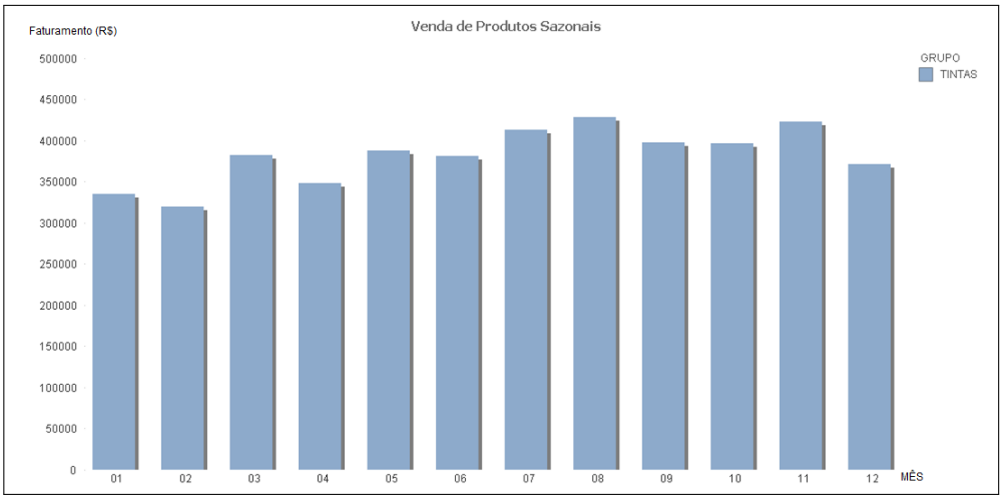


Figura 25 – Vendas de Tintas.

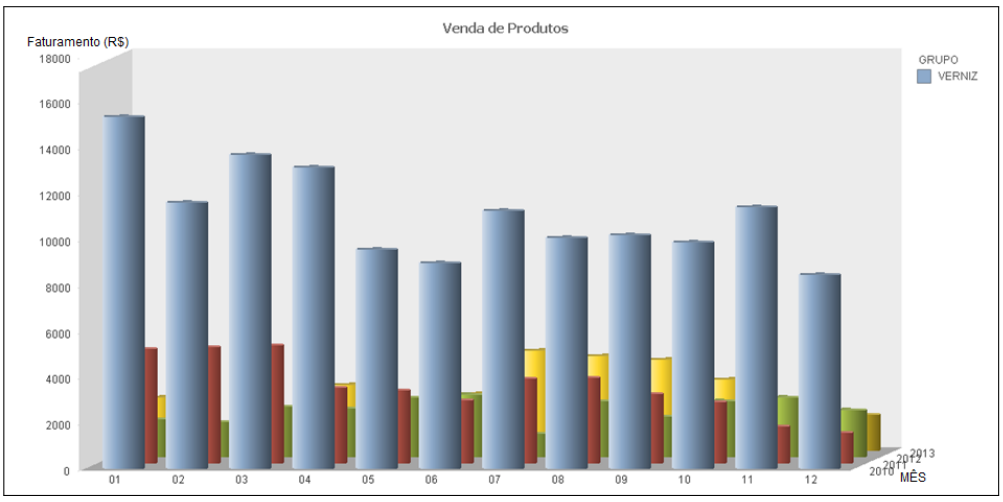


Figura 26 – Vendas de Verniz durante os anos.

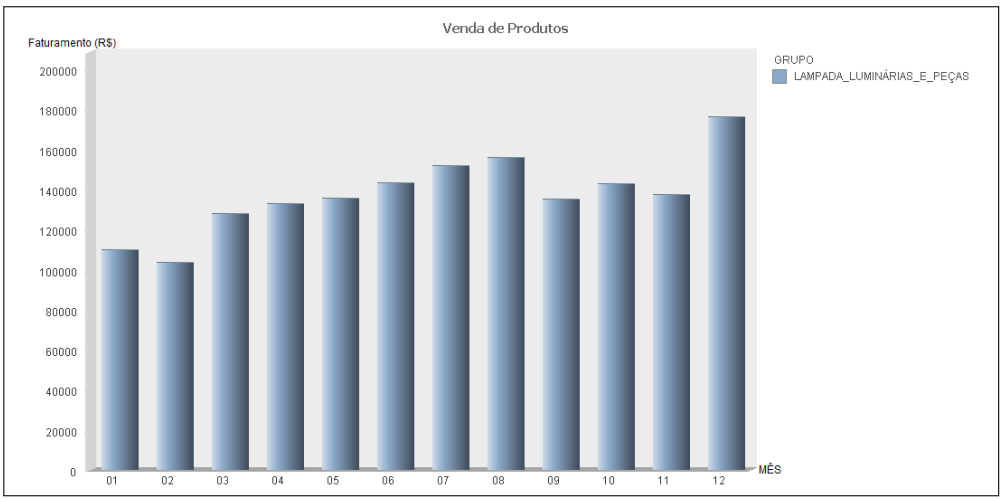


Figura 27 – Vendas de Lâmpadas, Luminárias e Peças.

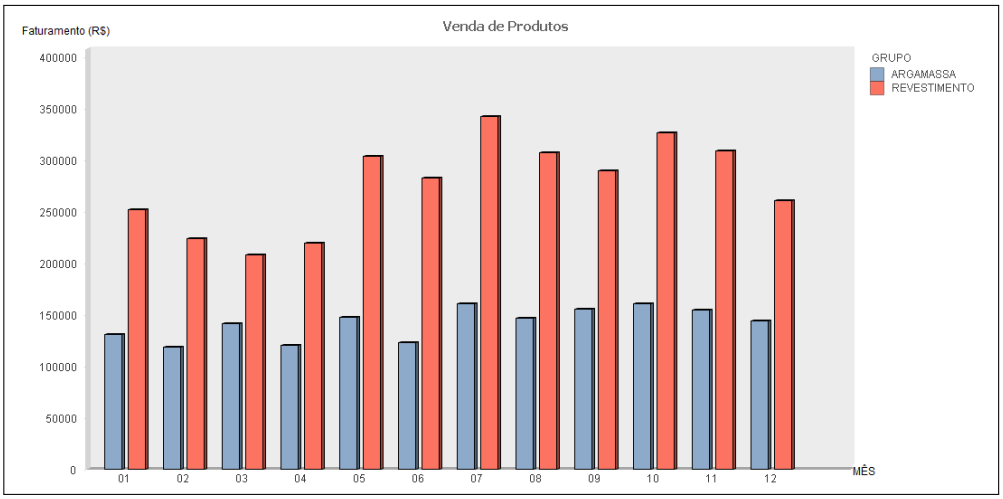


Figura 28 – Vendas de Argamassas e Revestimentos.

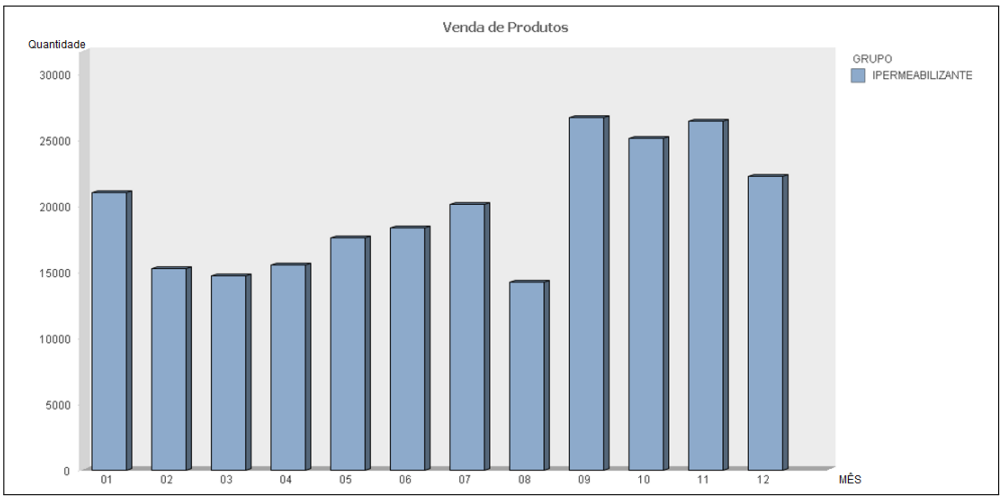


Figura 29 – Vendas de Impermeabilizante.

também há aumento de vendas de Argamassa. Com isso verificamos como estão relacionados estes produtos e comprovou-se as regras de associação geradas. Também é possível verificar como as vendas de revestimento caem nos meses de Fevereiro, Março e Abril.

A venda de Impermeabilizante aumenta muito no mês de Janeiro e ao final do ano por ser época de chuva. Esta informação é de conhecimento da empresa e foi confirmada na Figura 29.

5 Considerações Finais

O trabalho apresentou a aplicação do processo de descoberta do conhecimento KDD em uma empresa varejista, aplicando a mineração de regras de associação nas cestas de compras. Também foram realizados estudos dos dados da empresa das vendas regionais e venda de produtos em determinadas épocas do ano.

A primeira percepção diz respeito a qualidade dos dados, pois há na base de dados muita informação incorreta, como o cadastro dos bairros de clientes, que por ser um campo livre, os operadores do sistema cadastram os bairros sem padrão, com abreviações e erros de digitação. Também foi detectado que este campo é utilizado para gravar outras informações, sem ser as destinadas para o objetivo dele.

Outro fato que deve ser considerado são os agrupamentos dos produtos. O software possui no cadastro de produtos um campo grupo e subgrupo, porém possuem produtos agrupados de forma incorreta e muitos grupos criados não são ideais para o trabalho de mineração realizado. Com isso foi realizado um reagrupamento dos produtos para atender os objetivos do trabalho. Este processo e trabalhoso e demanda tempo, desta forma é importante as empresas que pretendem realizar este tipo de trabalho futuramente, mantenha seus dados e cadastrados corretos e da melhor forma possível.

É interessante para as empresas obterem o maior número de informações em suas bases de dados, porém isso só será possível se manter os seus dados atualizados e cadastrados corretamente. Nesta estudo foi necessário a redução do escopo do trabalho que tinha como objetivo analisar o perfil de clientes que compram na empresa, mas que não foi possível por não haver dados suficientes para esta análise. Das vendas registradas na empresa, cerca da metade havia vínculo com o cadastro de cliente e este cadastro estava apenas com os dados básicos do cliente, impossibilitando a realização do estudo.

Foi possível identificar os produtos mais vendidos pela empresa, e estes se confirmam com a percepção da empresa. O bairro que mais compra também se confirma. Um conhecimento encontrado que a empresa não conhecia foi o volume de vendas no bairro Cruzeiro Celeste, pois de acordo com a empresa há comércios do mesmo ramo na região e isso poderia impactar na procura desses clientes na empresa. Porém foi comprovado que ele está entre os 15 bairros que mais compram, e o que estes clientes mais procuram são os Revestimentos, provavelmente devido à variedade e qualidade deste segmento que a empresa oferece.

Foram confirmadas as sazonalidades dos produtos chuveiro e resistência para chuveiro, pois as vendas aumentam consideravelmente no inverno, que já era conhecido pela empresa. E também a sazonalidade de impermeabilizantes que foi citada pela empresa, pois é um produto muito procurado em épocas de chuva. Já a venda de mangueiras de jardim, que foi comprovado o aumento das vendas deste produto em épocas de pouca chuva, a empresa

não conhecia.

Analisados os comportamentos das vendas em épocas do ano, foram detectados queda nas vendas em Fevereiro e Abril e crescimento nas épocas do ano em que chove pouco. Este comportamento é conhecido pela empresa, pois nos meses de queda há feriados longos e o fato de chover pouco propicia para realização de obras externas. A venda em cidades vizinhas não era de conhecimento da empresa, e foi constatado que estes clientes procuram mais pelos pisos/porcelanatos e revestimento, levando a acreditar que estes estão em busca de maior variedade e qualidade neste segmento. Desta forma, conclui-se que o trabalho foi bem aplicado, com resultados úteis e verídicos.

5.1 Trabalhos Futuros

Foram encontradas informações para tomada de decisão na empresa, porém não fez parte do escopo desta trabalho aplicar e analisar os resultados obtidos pela empresa após a descoberta dos conhecimentos e aplicação dos mesmos. Também não houve um acompanhamento de um especialista do negócio para analisar as informações geradas.

Desta forma como continuidade do trabalho, propõe-se que sejam analisadas as informações com um especialista e aplicar as estratégias na empresa como organização de itens nas prateleiras, treinamento de vendedores orientando á sugerir itens relacionados na cesta de mercado para clientes no ato da compra. Também podem ser aplicadas estratégias de marketing em bairros de clientes, direcionadas à produtos específicos e realizar promoções de itens que são pouco vendidos em algumas épocas do ano, a fim de aumentar o seu faturamento.

Outra proposta seria a empresa começar a cadastrar os dados dos clientes e vincular esses cadastros nas vendas, pois muito informação é perdida quando não se faz este vínculo. Em seguida analisar estar informações para tentar detectar perfis de clientes que compram na empresa. Aplicada as estratégias, analisar se a empresa obteve um retorno satisfatório no aumento das vendas e se conseguiu atrair mais clientes.

6 ANEXOS

6.1 Principais comandos SQL usados

Este anexo exibe os principais comandos SQL usados no desenvolvimento e manipulação de dados deste trabalho.

- Criar a tabela contendo o campo identificador da venda e os atributos para cada produto.

  CREATE TABLE CESTA_DE_COMPRA
  (IDVENDA VARCHAR(30),
  ABRACADEIRA VARCHAR(10),
  ACABAMENTO_BANHEIRO VARCHAR(10),
  ARGAMASSA VARCHAR(10),
  . . .
  VERNIZ VARCHAR(10));

- Comando de inserção dos dados na tabela. O comando conta todas as ocorrências do atributo do identificado da venda e preenche a tabela com 0 caso não haja o procuto na cesta de compra e a quantidade do produto na cesta de compra caso haja.

  INSERT INTO CESTA_DE_COMPRA (IDVENDA, ABRACADEIRA, ACABAMENTO_BANHEIRO, ARGAMASSA, . . . VERNIZ)
  SELECT IDVENDA, [ABRACADEIRA], [ACABAMENTO_BANHEIRO], [ARGAMASSA],. . . [VERNIZ]
  FROM
  (
  SELECT IDVENDA, GRUPO FROM REGISTRO_DE_VENDAS WHERE
) PT
  PIVOT
  (COUNT(GRUPO) for GRUPO in ([ABRACADEIRA], [ACABAMENTO_BANHEIRO], [ARGAMASSA], . . . [VERNIZ]) pvt

- Comando que altera os valores da tabela de 0 para ? e > 1 para 1. Este comando é executado para todos os grupos de produtos.

  UPDATE CESTA_DE_COMPRA SET [ABRACADEIRA] = 1 WHERE [ABRACADEIRA] > 1;
  . . .

UPDATE CESTA_DE_COMPRA SET [VERNIZ] = 1 WHERE [VERNIZ] > 1;

UPDATE CESTA_DE_COMPRA SET [ABRACADEIRA] = '?' WHERE [ABRACADEIRA] = 0;
. . .
UPDATE CESTA_DE_COMPRA SET [VERNIZ] = '?' WHERE [VERNIZ] = 0;

- Comando que altera a tabela excluindo o campo identificador de venda, pois no arquivo de mineração de dados ele não é necessário, mas para a criação da tabela ele é útil.

  ALTER TABLE CESTA_DE_COMPRA DROP COLUMN IDVENDA

Referências

AGRAWAL, R.; SRIKANT, R. Fast algorithms for mining association rules. The International Conference on Very Large Databases, 1994.

ARAÚJO, A. L. V. Aplicação de regra de associação para auxílio na gestão de vendas de uma empresa varejista utilizando a ferramenta weka. 2009.

CAMARGO, S. d. S. *Mineração de Regras de Associação no Problema da Cesta de Compras Aplicada ao Comércio Varejista de Confecção*. Dissertação (Mestrado) — Universidade Federal do Rio Grande do Sul, 2002.

COELHO, E. M. *EMC SISTEMAS*. 2015. Disponível em: <http://www.emcsistemas.com.br/>.

DIAS, A. P. Aplicação de regra de associação para mineração de padões em vendas de produtos. 2014.

FAYYAD, U.; SHAPIRO, G. P.; SMYTH, P. Knowledge discovery and data mining: Towards a unifying framework. *Second International Conference on Knowledge Discovery and Data Mining KDD-96*, p. 4,5, 1996.

KASAHARA, C. N.; CONCEIçãO, F. W. S. Análises de ferramentas de mineração de dados. 2008.

MOTA, U. *Ulete Mota - O supermercado da Construção*. 2015. Disponível em: <http://www.uletemota.com.br/>.

NAVATHE, S. B.; ELMASRI, R. *Sistemas de Banco de Dados*. [S.l.]: Pearson Education, 2010. 698 – 705 p.

SHAEFFER, A. G. *Data Mining no Varejo: Estudo de caso para loja de meteriais de construção*. Dissertação (Mestrado) — Universidade Federal do Rio Grande do Sul, 2003.

TAN, P.-N.; STAINBACH, M.; KUMAR, V. *Introdução ao Data Mining*. [S.l.]: Pearson Education, 2009. Capítulo 6 p.

WAIKATO, T. U. of. *Weka 3: Data Mining Software in Java*. 2015. Disponível em: <http://www.cs.waikato.ac.nz/ml/weka/>.