

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Letícia Marques Cardoso

**Análise de clientes de uma distribuidora de
produtos farmacêuticos com Mineração de
dados baseada em Árvore de Decisão**

Uberlândia, Brasil

2017

Letícia Marques Cardoso

Análise de clientes de uma distribuidora de produtos farmacêuticos com Mineração de dados baseada em Árvore de Decisão

Trabalho de conclusão de curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, Minas Gerais, como requisito exigido parcial à obtenção do grau de Bacharel em Ciência da Computação.

Trabalho aprovado. Uberlândia, Brasil, 24 de julho de 2017.

Profa. Maria Adriana Vidigal de Lima
Orientadora

Anilton Joaquim da Silva

William Chaves de Souza Carvalho

Uberlândia, Brasil
2017

Resumo

O processo de descoberta de conhecimento em bases de dados, incluindo a etapa da mineração de dados vem sendo amplamente utilizado como mecanismo para contribuir no processo de tomada de decisão em organizações empresariais. Este trabalho tem como objetivo principal a aplicação de um modelo de mineração de dados baseado em classificação, utilizando como fonte de dados os registros de vendas de uma distribuidora de produtos farmacêuticos. O processo de extração de dados de vendas foi feito para um determinado período e baseou-se na categorização de clientes (Bronze, Prata, Ouro e Diamante) proposta pela equipe responsável da área de marketing.

Palavras-chave: mineração de dados, clientes, classificação, árvore de decisão, estatística descritiva.

Abstract

The process of Knowledge Discovery in Databases, including the data mining stage has been widely used as a mechanism to assist decision-making process in business organizations. The main objective of this study is the application of a data mining model based on classification, to analyze a data source obtained from sales records of a pharmaceutical distributor. The extraction process of sales data was established for a certain period and was also based in the customer categorization (bronze, silver, gold and diamond) held by the business marketing team.

Keywords: data mining, clients, classification, decision tree, descriptive statistics.

Lista de ilustrações

Figura 1 – Exemplo de um processo de descoberta de conhecimento em bancos de dados	13
Figura 2 – Arquitetura de um DW com as fontes de dados e artefatos dos usuários	15
Figura 3 – Estrutura visual da ferramenta <i>RapidMiner</i>	19
Figura 4 – Exemplo de um relatório feito na ferramenta do <i>QlikView</i>	20
Figura 5 – Carregamento das informações dos fornecedores para o primeiro estágio do processo de ETL.	22
Figura 6 – Diagrama ER das tabelas que serão utilizadas.	23
Figura 7 – Regras para a definição dos dados de 2016.	24
Figura 8 – Dados na planilha após a classificação dos clientes de 2016.	25
Figura 9 – Árvore de Decisão resultante do processamento dos clientes de 2016. . .	26
Figura 10 – Regras de associação dos dados dos clientes de 2016.	28
Figura 11 – Pseudocódigo do algoritmo de árvore de decisão (MONARD; BARA-NAUSKAS, 2003).	29
Figura 12 – Arquivo dos dados do 1º trimestre de 2017.	31
Figura 13 – Componentes da ferramenta que permitem carregar arquivos externos.	31
Figura 14 – Exemplo de tabela criada no <i>RapidMiner</i> com dados de 2016.	32
Figura 15 – Configuração da coluna "Cliente" como um indicador.	33
Figura 16 – Etapa de seleção da coluna de identificação dos dados de 2016.	33
Figura 17 – Etapa de seleção da coluna de previsão dos dados de 2016.	33
Figura 18 – Etapa de seleção do algoritmo Árvore de Decisão que será aplicada aos dados de 2016.	34
Figura 19 – Etapa de aplicação do algoritmo selecionado na Figura 18.	34
Figura 20 – Etapa de seleção da coluna de identificação dos dados de 2017.	35
Figura 21 – Aplicação dos dados de 2016 e 2017 no processo completo de classificação.	35
Figura 22 – Saída dos dados dos clientes de 2016.	36
Figura 23 – Parâmetros utilizados no algoritmo de árvore de decisão.	37
Figura 24 – Cálculo de confiança para os clientes categorizados como Diamante de 2017.	38
Figura 25 – Cálculo de confiança para os clientes categorizados como Diamante de 2017.	38
Figura 26 – Outros cálculos produzidos pelo <i>RapidMiner</i>	39
Figura 27 – Gráfico de Dispersão representando a quantidade de dias em relação ao valor bruto por cliente.	40
Figura 28 – Relatório de venda trimestral de 2016.	41
Figura 29 – Relatório de venda trimestral de 2017.	41

Lista de abreviaturas e siglas

BI	Business Intelligence
ETL	Extract, Transform, Load
KDD	Knowledge Discovery in Databases
OLAP	On-line Analytical Processing
ODS	Operational Data Store
AQL	Associative Query Analytical
CD	Centro de Distribuição
SAP	Systems Applications and Products

Sumário

1	INTRODUÇÃO	7
1.1	Visão Geral	7
1.2	Objetivos	8
1.3	Justificativa	9
1.4	Organização do trabalho	10
2	METODOLOGIA	11
3	REFERENCIAL TEÓRICO	13
3.1	<i>KDD - Knowledge Discovery in Databases</i>	13
3.2	<i>Data Warehouse</i>	14
3.3	<i>Data Mining</i>	16
3.4	Técnica de Classificação	17
3.5	Ferramenta para mineração dos dados	18
3.6	Visualização - Ferramenta <i>QlikView</i>	19
4	DESENVOLVIMENTO DO TRABALHO	22
4.1	Processo de ETL	22
4.2	Preparação dos dados de treinamento	23
4.3	Algoritmo de Classificação baseado em <i>Árvore de Decisão</i>	26
5	EXPERIMENTOS E RESULTADOS	30
5.1	Fluxo de processo	30
5.2	Análise dos Resultados Obtidos	40
6	CONCLUSÃO	43
	REFERÊNCIAS	44

1 Introdução

1.1 Visão Geral

A informação, um dos bem mais valiosos atualmente, começou a ser mais valorizada no início da Era da Informação (mais conhecida como Era Digital) com a criação dos microprocessadores, do PC (computador pessoal) e também da evolução das redes e de comunicação digital e suas tecnologias. Peter Druker, considerado o pai da administração moderna, foi o primeiro a nomear este momento como Era da Informação ([WIKI, 2010](#)). Ele defende que esta era se iniciou com a atitude dos soldados americanos, que após voltarem da 2a Guerra Mundial exigiam empregos seguros, como colocações em universidades. Neste cenário de avanços tecnológicos, o trabalho em [FILHO \(2001\)](#) apresenta o questionamento de como pode ser possível tirar proveito dessas tecnologias que colocam à disposição das pessoas um volume cada vez maior de informações.

Em consequência, tanto nas relações entre indivíduos ou empresas, tem-se um grande desafio de como customizar grandes volumes de informação, isto é, como proporcionar de forma eficaz que uma informação precisa seja encontrada de maneira simples e fácil, quando necessário.

Quando um indivíduo tem uma primeira experiência com o termo *Business Intelligence* (BI), tende a pensar que é algo novo. Mas ao contrário do que se pode imaginar, a ideia do *Business Intelligence* já era utilizada em outras civilizações que existiam antes da nossa, por exemplo, quando os povos antigos usavam as informações que a natureza (ou os astros) transmitiam para decidir o que fazer para melhorar a situação da própria civilização.

O termo BI foi marcado pelo *Gartner Group* em meados da década de 1990 (PRIMA, 2008), mas computacionalmente teve suas origens nos sistemas que geravam relatórios mais simples, daqueles que não tinham nenhum poder de análise. De acordo com [Matheus e Parreiras \(2004\)](#), entre o significado de *Business Intelligence* e sua tradução Inteligência Empresarial, existe uma diferença peculiar, notadamente no Brasil: "Se por um lado a inteligência empresarial busca entender os fatores e processos humanos e organizacionais envolvidos na busca de informações, principalmente externas à organização e na posterior tomada de decisões, por outro, o BI, através das tecnologias da informação e da comunicação, tem como objetivo criar sistemas de informação computacionais geralmente a partir de grandes volumes de dados, capazes de prover aos gerentes melhores informações para a tomada de decisão". Com isso, pode-se explicitar qual é o papel do *Business Intelligence*.

O objetivo geral do BI para qualquer organização é, a partir do cruzamento de informações e dados provenientes de diferentes pontos (empresa, cliente, produto, recursos humanos, etc), ter a capacidade de obter outros tipos de dados e informações que podem ser úteis para uma tomada de decisão mais importante, tanto externamente quanto internamente.

A técnica de *data mining* (em português, mineração de dados) faz parte de um processo mais abrangente chamado *Knowledge Discovery in Databases* (em português, Descoberta de Conhecimento em Bancos de Dados ou KDD) é e utilizada para processar uma quantidade de dados maior do que uma exploração de dados tradicional utiliza. Essa técnica pode ser utilizada para vários fins, mas em geral ela procura padrões nessa variedade de dados que são usados por uma empresa na área de vendas, produtos ou clientes. A técnica de *data mining* está sendo amplamente utilizada no mercado empresarial pois proporciona, cada vez mais, a geração de novos dados e informações que podem ser utilizados em seu próprio favor. Esse processo pode ser dificultado nos casos em que as organizações não disponham de seus dados de uma maneira integrada, isto é, quando não existe investimento em um *data warehouse* (ELMASRI; NAVATHE, 2010) ou quando servidores e os dados estão espalhados.

A proposta deste trabalho é aplicar tecnologias de *Business Intelligence* no contexto de uma empresa distribuidora de medicamentos. A ideia é compreender o conjunto total dos dados de vendas, distribuídos pelos diversos setores da empresa, juntá-los e examiná-los em uma ferramenta específica para a mineração e após isso agrupar essas informações em um *dashboard* para que possa ser visualizado por pessoas interessadas.

O presente trabalho intercorre nos campos da distribuição e venda de medicamentos, e a justificativa para esta escolha vem do fato da autora ter realizado estágio supervisionado numa empresa do ramo, e, por conseguinte, ter adquirido importantes experiências através da realização de atividades na área de *Business Intelligence*. Os dados trabalhados apresentam ótimo potencial para análises e puderam ser utilizados neste trabalho, limitados apenas pela privacidade de algumas informações de clientes dessa distribuidora. Após o término desse estudo pretende-se utilizar o resultado das análises como material adicional para favorecer a tomada de decisão da própria distribuidora, aprovendo os coordenadores de áreas como logística e marketing, por exemplo, de novas informações e conhecimento.

1.2 Objetivos

O objetivo deste estudo é produzir uma análise sobre os clientes de uma determinada distribuidora farmacêutica no intuito de saber quais são os melhores clientes, levando-se em consideração a quantidade de produtos comprados ao longo do mês e a

quantidade de dias em que foram efetuadas compras. Esta análise tem a finalidade específica de estabelecer previsões de como esses clientes podem se comportar futuramente.

Serão utilizados algoritmos para classificar esses clientes em 4 tipos: *Bronze*, *Prata*, *Ouro* e *Diamante*, de acordo com os seguintes atributos:

- valor de venda bruta;
- quantidade de dias que são efetuadas compras; e
- quantidade de produtos que são comprados.

Esta análise será realizada a partir de um estudo sobre os dados disponíveis da distribuidora, partindo do conhecimento obtido sobre os clientes e produtos que ela oferece, filtrando informações importantes sobre cada um.

Além disso, outro objetivo é apresentar um estudo específico sobre o procedimento de coleta e organização e dos dados. Uma das ferramentas utilizadas para o processamento dos dados foi a *Datastage* da IBM ([REDBOOKS, 2008](#)), transformá-los e carregá-los de acordo com o objetivo central.

Após o processamento preliminar dos dados, inicia-se a fase de extração de conhecimento. Nesta fase objetiva-se a sinalização dos clientes em uma lista de compradores potenciais como de acordo com as categorias já pré definidas: *Bronze*, *Prata*, *Ouro* e *Diamante*. A solução para se obter tal conhecimento advém de técnicas simples de mineração de dados, como o uso de algoritmos de classificação baseados em árvores de decisão. Desta forma, propõe-se um mecanismo para determinar com mais facilidade os clientes com potencial.

A ferramenta que será usada para a aplicação dos referidos algoritmos será a *RapidMiner* ([RAPIDMINER, 2014](#)). Os dados são inicialmente obtidos a partir uma planilha eletrônica (arquivo em formato Excel) contendo um histórico sobre como os clientes do ano de 2016 foram categorizados, de outra planilha dos dados de clientes que serão categorizados de acordo com o histórico. As duas planilhas são ligadas para que os algoritmos possam ser aplicados.

1.3 Justificativa

A justificativa principal para a pesquisa é mostrar como o estudo teórico das técnicas de *data mining* é importante para se atingir objetivos específicos no contexto empresarial, apresentando conceitos que são usados na área de *Business Intelligence*, ferramentas usadas nas operações de extração e apresentação dos dados, e a função de cada uma durante a descoberta de conhecimento. Em adição, este trabalho mostra como os algoritmos propostos podem ser aplicados em uma situação real no mercado, a fim de

auxiliar no ponto principal desta investigação: o modelo para análise de comportamento dos clientes de uma distribuidora de medicamentos.

A análise proposta neste trabalho tem o objetivo de atender à uma demanda da área de marketing, que é classificar os clientes de acordo com o valor comprado, os dias em que fazem as compras e a quantidade de produtos adquiridos. Essa classificação define quatro tipos principais de cliente: diamante, ouro, prata, e bronze. Esta divisão por categorias permite que possam ser planejadas campanhas de marketing estratégico, baseadas nos resultados das análises de dados focadas nestes tipos de clientes, de forma a produzir materiais e promoções para alavancar as vendas. Assim, são identificados para os tipos de clientes: produtos mais comprados, dias mais utilizados para as compras, e quantidades adquiridas para que se possa conhecer melhor os clientes e seus comportamentos, melhorar a comunicação e promover novas vendas para clientes que não comprem muito mas que são adeptos à promoções.

1.4 Organização do trabalho

O estudo proposto neste trabalho, com foco no cruzamento de informações e dados provenientes de compras de clientes no intuito de classificá-los e de produzir conhecimento útil para a tomada de decisão, está documentado em seis capítulos. No primeiro capítulo apresenta-se o projeto com uma breve contextualização do tema e da problemática vislumbrada, bem como os objetivos gerais e específicos.

No segundo capítulo é descrita a metodologia de trabalho definida considerando-se as necessidades, restrições, métodos e ferramentas disponíveis para a realização de análises para a descoberta de conhecimento sobre os dados.

O terceiro capítulo trata do referencial teórico necessário para a realização do trabalho, enfocando o tema da descoberta de conhecimento em banco de dados e seus métodos e tecnologias. O quarto capítulo contém o desenvolvimento do trabalho e suas etapas: obtenção e preparação dos dados e aplicação dos algoritmos sobre os mesmos.

O quinto capítulo apresenta e discute os resultados obtidos assim como as possibilidades de análise considerando-se a proposição do trabalho. Por fim, o sexto capítulo contém as considerações finais e os apontamentos para possíveis trabalhos futuros.

2 Metodologia

Como foi dito anteriormente, a sociedade atual tem fácil acesso a várias informações de diferentes fontes e com isso as empresas estão optando por obter a maior quantidade de conhecimento futuro possível para obter vantagens em relação às outras organizações. Para isso, saber o que de fato é a descoberta de conhecimento em banco de dados se tornou um recurso diferencial do analista que trabalha com tecnologia de informação dentro de uma empresa. Além disso, a abertura de caminhos para esse tipo de descoberta em qualquer área que a empresa atue pode auxiliar em futuras inovações e/ou melhorias nos serviços que ela presta termos de complexidade, flexibilidade e criatividade.

A descoberta de conhecimento em banco de dados é uma técnica que precisa de vários pontos a serem ajustados e verificados em relação a necessidade, extraindo apenas dados realmente interessantes pois é um processo mais demorado do que a coleta em si que pode ser feita a todo instante e de diversos dispositivos e fontes.

A preparação dos dados essenciais é feita em uma ferramenta de Extração, Transformação e Carregamento (ETL para *Extract-Transform-Load*), a *Datastage*. A sua função principal é extrair os dados de vários sistemas da empresa para ser capaz de transformar esses dados de acordo com as regras de negócio e por fim armazená-los em um *data warehouse* (REDBOOKS, 2008). Ao final do processo o *Datastage* terá dados consolidados de diversas fontes, inclusive dados provenientes de banco de dados relacionais ou de outros tipos de fonte, com isso a partir dessa ferramenta também teremos acesso aos dados desses bancos de dados (PRASS, 2014).

Foram extraídos do banco de dados, uma base de clientes que efetuaram compras entre janeiro e março de 2016 (a qual será usada como uma base de treinamento) e clientes que compraram no mesmo período de 2017, esse método é chamado de Amostragem Estratificada. No caso deste trabalho a população é identificada a partir de pelo menos dois estratos (1º trimestre de 2016 e 1º trimestre de 2017) e deve ser possível extrair uma amostra de cada um.

Após a extração adotar-se-á uma das técnicas da área de Estatística dentre as várias existentes para descrever e classificar os dados dos clientes, de acordo com condições propostas fazendo uma pré-classificação. Esta técnica é denominada estatística descritiva e o conceito de medida de dispersão será utilizado para que a dispersão possa ser utilizada como critério de categorização, antes da realização da previsão para os clientes do ano de 2017. (DAVILA, 2010).

Com um objetivo bem definido, sendo ele a classificação dos clientes da distribuidora farmacêutica em categorias pré selecionadas, a seleção e preparação dos dados

também é uma etapa que precisa estar bem fundamentada. Para este fim, serão utilizados dados internos relacionados a dados estatísticos no *RapidMiner*, e a partir desta atividade, serão geradas as previsões de dados futuros. Para fazer esta classificação, será utilizada uma técnica baseada em árvore de decisão que gera algumas regras que servirão de base para a predição sobre o comportamento dos clientes.

A partir da consolidação dos dados que são relevantes para a pesquisa, pode-se iniciar um processo de análise desses dados comparando-os com os valores reais existentes na base, que são as compras já efetuadas. Assim, ao final, é possível comparar qual o índice de acertos que esse algoritmo tem em relação aos dados reais da empresa. A próxima seção apresenta os conceitos importantes dentro da área de *Business Intelligence* que foram utilizados, bem como as ferramentas escolhidas para cada etapa.

3 Referencial Teórico

3.1 KDD - Knowledge Discovery in Databases

O KDD consiste em extrair informações importantes de uma massa de dados de um banco de dados qualquer. Esse processo ocorre antes mesmo da mineração dos dados em si, já que prepara toda a estrutura necessária para que a mineração seja feita corretamente e para que isso ocorra temos 5 fases, que vai desde o nível mais baixo que é a seleção dos dados até a sua avaliação. A Figura 1 apresenta uma das várias maneiras que esse processo pode ser visualizado, já que os processos podem ter diversos nomes mas o mesmo objetivo.

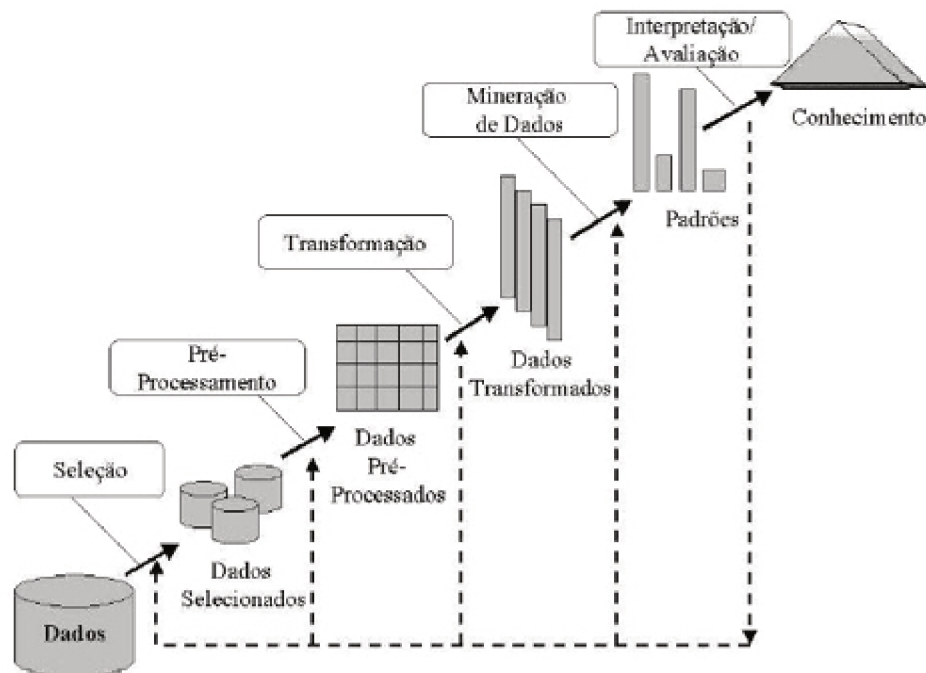


Figura 1 – Exemplo de um processo de descoberta de conhecimento em bancos de dados (Fonte: <http://revistas.ufpr.br/atoz/article/viewFile/41284/25209/154053>)

Alguns trabalhos (FIGUEIREDO, 2004) dividem em 6 as fases do processo de descoberta de conhecimento, que são:

- Definição do problema: está parte é essencial para que possa ser entregue ao usuário final exatamente o que ele precisa, pois com isso pode ser tomadas decisões importantes para a organização como um todo.

- Seleção: no contexto desse trabalho, a seleção correta é obter dados apenas de clientes e produtos, já que o objetivo gera em torno dessas duas áreas.
- Filtragem: é uma consequência da fase de Seleção, sendo a obtenção dos dados em si.
- Enriquecimento: essa etapa por ser vista como a união dos dados de diferentes áreas, ou seja, precisamos analisar clientes porém temos que levar consideração as informações sobre produto, o que seria o enriquecimento sobre o cliente.
- Mineração: consiste na escolha de um algoritmo específico para trabalhar com os dados, será explicado em um tópico especial.
- Visualização: na prática empresarial são usados relatórios onde esses dados são apresentados em forma de gráficos ou tabelas para o usuário final. Essa forma além de ser mais prática e de fácil entendimento, enriquece todo o processo em relação aos diretores da empresa que poderão utilizá-las com uma maior importância.

Essas etapas são pensadas de acordo com a necessidade da empresa e ao final interpretados e avaliados por uma pessoa ou equipe especializada na área de negócio, através de relatórios por exemplo, para que todo o processo tenha maior utilidade para a empresa.

3.2 *Data Warehouse*

O Data Warehouse é um poderoso armazém de dados históricos, utilizado nas empresas com o intuito de registrar as informações que fazem parte das atividades da empresa de uma forma consolidada. É construído a partir de dados provenientes dos bancos de dados relacionais da empresa e tem como finalidade facilitar análises sobre esses dados, gerar relatórios e obter informação e conhecimento relevante para a tomada de decisão estratégica. Assim, um Data Warehouse possui estrutura e características que suportam análise de grande volumes de dados. A Figura 2 ilustra a arquitetura básica de um Data Warehouse com as fontes de dados, o agrupamento do Data Warehouse e as aplicações e usuários finais (ORACLE, 2017).

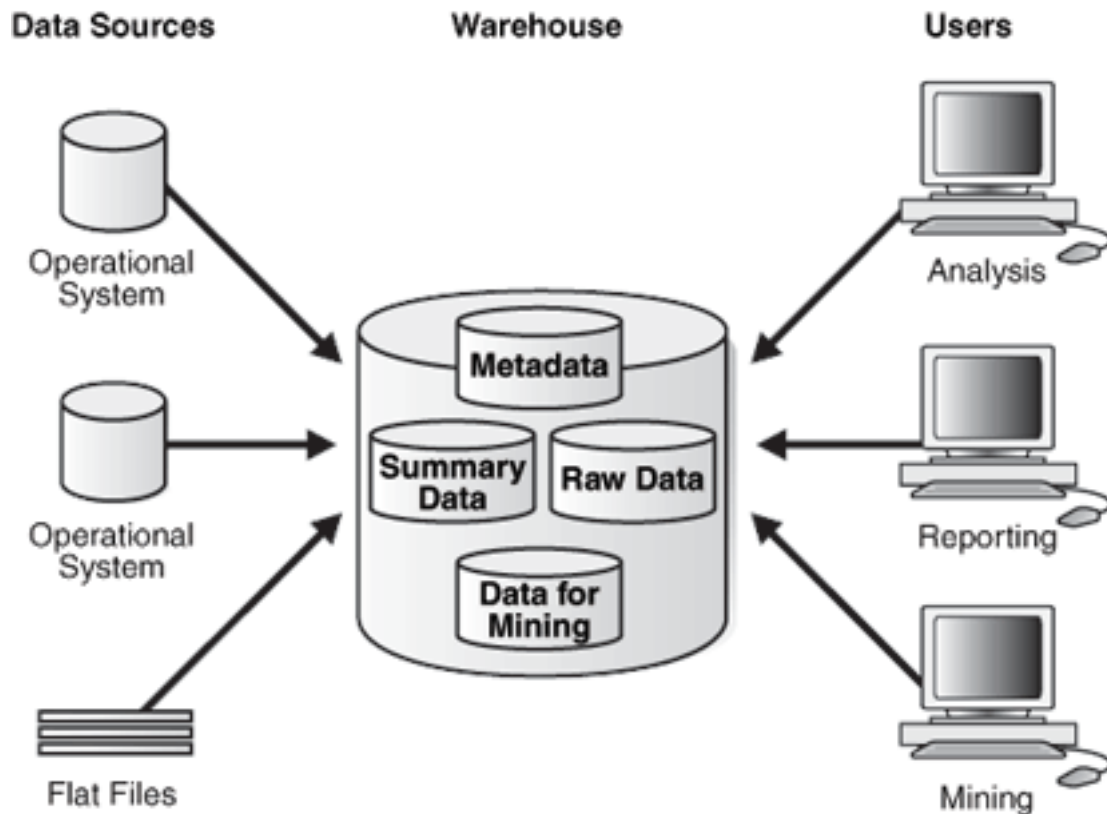


Figura 2 – Arquitetura de um DW com as fontes de dados e artefatos dos usuários. (Fonte: <https://docs.oracle.com/database/121/DWHSG/concept.htm>)

Neste sentido, a primeira atividade é coletar corretamente os dados, percorrendo os diversos bancos de dados da empresa em busca de informações que podem ser importantes. A partir deste ponto, estas informações são tratadas antes de serem armazenadas no *data warehouse*. Desta forma informações desnecessárias são descartadas e aquelas relevantes são ajustadas e padronizadas. Um Data Warehouse bem elaborado deve:

- Permitir que a informação esteja acessível de forma fácil e intuitiva para o usuário.
- Fornecer informações consistentes: todos os dados fornecidos devem ser relevantes, precisos e completos.
- Ser adaptável e flexível à mudanças já que as condições e necessidades do negócio podem se alterar com o passar do tempo.
- Proteger e tornar a informação segura, pois um conjunto de informações críticas para a empresa estão registradas no Data Warehouse.
- Favorecer o processo de tomada de decisão e ser aceito pela equipe de negócios.

Para que apenas informações relevantes e com potencial para serem trabalhadas possam ser coletadas e armazenadas, é necessário um processo para a limpeza desses

dados, selecionando e tratando os que forem mais importantes. Por conseguinte, as organizações adotam o processo de ETL, que significa Extração, Transformação e Carregamento (AFONSO, 2015).

A primeira fase deste processo, uma das obrigatórias e mais importantes, é a extração. Esta fase é importante porque essa será a informação armazenada no *data warehouse* e conduzida a uma área temporária, em que os dados são convertidos em um tipo único já que eles podem vir de fontes diferentes. É a partir deste conjunto de dados que diretores ou donos da empresa tomarão as decisões que influenciarão no negócio da mesma. Após o armazenamento, as informações estão prontas para passar pela próxima fase que é a de transformação e limpeza desses dados. Nesta fase esses dados serão padronizados e transformados de acordo com as regras de negócio da empresa. A última parte desse processo é aquela em que os dados são armazenados no *data warehouse* da forma correta e de acordo com a necessidade da organização.

3.3 *Data Mining*

Devido à grande quantidade de informação que as organizações têm gerado nos últimos anos, iniciou-se um movimento de utilização destes dados para a geração de conhecimento que pudesse complementar e auxiliar tanto nos próprios processos empresariais como na tomada de decisão estratégica. O termo *Data Mining*, em português Mineração de dados, denota um conjunto de técnicas criadas para a realização de análise dos dados armazenados nas bases de dados das organizações, representando o histórico das transações efetuadas. Este processo automatizado e baseado em algoritmos para o processamento de grandes volumes de dados contidos nas bases de dados das organizações empresariais tem como objetivo a descoberta de novos conhecimentos que podem ser aplicados no gerenciamento dos negócios, controle de produção e análise de mercado.

O tópico de mineração de dados é visto como uma parte de um processo maior, o KDD, não sendo menos importante ou relevante. A principal diferença observada entre eles é que nesse ponto podem-se escolher algoritmos que sejam apropriados para efetuar análises dos dados em um intervalo de tempo bem menor do que se fosse feito por um analista. Essa é uma das principais vantagens de se utilizar o processo de mineração de dados.

Neste trabalho serão utilizados dados para identificar perfis de clientes de uma distribuidora de medicamentos para serem aplicados em vendas futuras, em um mesmo intervalo de tempo, e a partir disso filtrar informações relevantes, e fornecer previsões de como esses clientes serão classificados levando em consideração dados anteriores.

O processo de mineração de dados não é um processo independente, ele depende de uma preparação tanto do banco de dados quanto dos dados propriamente ditos. Esse

processo é conhecido como Descoberta de Conhecimento em Banco de Dados (PRASS, 2014), e dentro da ampla utilidade da mineração de dados, são categorizados dois tipos de modelos que são mais comuns, o modelo descritivo e o modelo preditivo. Usando o modelo descritivo, podem-se obter resultados a partir do reconhecimento de semelhanças entre os dados históricos da empresa, e descrever o conhecimento baseado nesses dados de acordo com os interesses e/ou necessidades, em uma linguagem natural.

Já no outro modelo, chamado de preditivo, usam-se as informações que são extraídas dos dados para encontrar padrões que podem proporcionar previsões sobre determinadas situações que envolvem esses dados. Vale considerar que um modelo preditivo pode ser um modelo descritivo, e vice-versa, mas saber a diferença teórica dos conceitos é quase fundamental para entender o objetivo da sua descoberta.

Para que seja possível alcançar o objetivo principal de uma descoberta, seja ela um modelo preditivo ou descritivo, são usados alguns métodos, que são citados em um modelo de formalização (DIAS, 2001) e descritos abaixo:

- **Classificação:** é utilizada para separar os dados em classes definidas de acordo com a necessidade, por exemplo, clientes que comprem o produto A são 5 estrelas e clientes que comprar o produto B são 4 estrelas.
- **Regressão:** é uma função usada para prever o valor de uma variável desconhecida no seu modelo, por exemplo, prever qual será a meta de venda alcançada no próximo mês. É comum ser usado em modelo preditivos devido a sua finalidade.
- **Sumarização:** consiste em encontrar uma descrição mais simples para um conjunto de dados menor do que o seu conjunto de dados original.
- **Clustering** ou **Agrupamento:** seu objetivo é dividir um conjunto de dados em subconjuntos que apresentem alguma característica similar. Seguindo as ideias dos exemplos anteriores, seria agrupar os clientes por regiões geográficas.
- **Associação:** é usada para identificar quais itens estão associados em uma mesma transação. O termo “Análise de cesta de mercado” (em inglês “*Market Basket Analysis*”) deve-se a esse método, em que associam-se os produtos vendidos mais vendidos juntos em uma mesma cesta de compras.

3.4 Técnica de Classificação

As técnicas de classificação são mais usadas para modelos preditivos que são consideradas não supervisionadas por não ser necessário categorizar os dados antes de iniciar

o estudo do modelo. A ideia principal dessa análise é usar medidas de similaridade entre os dados para que possamos classificá-los, já que de uma certa forma a técnica de Classificação e de Agrupamento estão relacionadas.

Especificamente, neste trabalho, será utilizada a técnica de classificação, já mencionada no tópico anterior, para tipificar os clientes em 4 categorias sinalizando se é um bom cliente ou não, sendo Bronze o "pior" cliente e Diamante um cliente com grande potencial para a empresa de acordo com: o valor de venda, a quantidade de produtos e a quantidade de dias em que esses clientes compraram algum produto da distribuidora. Pode-se representar a criação das regras para a classificação através de uma árvore de decisão que será gerada após a execução do processo utilizando-se a ferramenta *RapidMiner*.

Após o processo de classificação desses clientes ter-se-à como resultado, uma previsão de como esses clientes serão classificados do ano seguinte no mesmo período de tempo.

3.5 Ferramenta para mineração dos dados

Atualmente existem várias ferramentas no mercado que são utilizadas para esse tipo de tarefa com os dados, e cada uma tem seus pontos positivos e seus pontos negativos, e o tipo de problema também influencia na hora da escolha. Algumas delas são mais conhecidas, e outras ainda estão se adaptando como foi discutido em (CAMILO; SILVA, 2009), tais como: SAS Enterprise Miner Suite, WEKA, Oracle Data Mining (ODM), IBM Intelligent Miner, RapidMiner, além de várias outras.

Para este trabalho será utilizada a *RapidMiner*. Ela tem algumas vantagens que foram levadas em conta na hora da escolha:

- É uma ferramenta grátis e seu cadastro pode ser feito pelo e-mail institucional.
- A maioria dos algoritmos necessários para mineração já vem incluída na plataforma, ou seja, não é necessária a programação dos mesmos em código fonte.
- É de fácil uso e permite a construção de fluxos de processos para alcançar os resultados.
- Possui muitas opções de tipo de arquivo de entrada, desde conexão com o banco de dados até uma planilha do Excel.
- Interface gráfica moderna e bastante intuitiva, entre outros aspectos de usabilidade.

No *RapidMiner* (RAPIDMINER, 2014) pode-se criar um processo, a gosto do usuário, e ainda alterar os parâmetros de medidas dos componentes, como a taxa de confiança, por exemplo, facilitando a comparação de resultados obtidos. Essa característica

especial das ferramentas de *data mining* permitiu o enriquecimento da análise dos dados e facilitou a compreensão dos resultados já que possui uma vasta possibilidade de algoritmos e operadores facilitando a produção de diferentes processos de mineração, acrescentando o layout prático visto na figura 3.

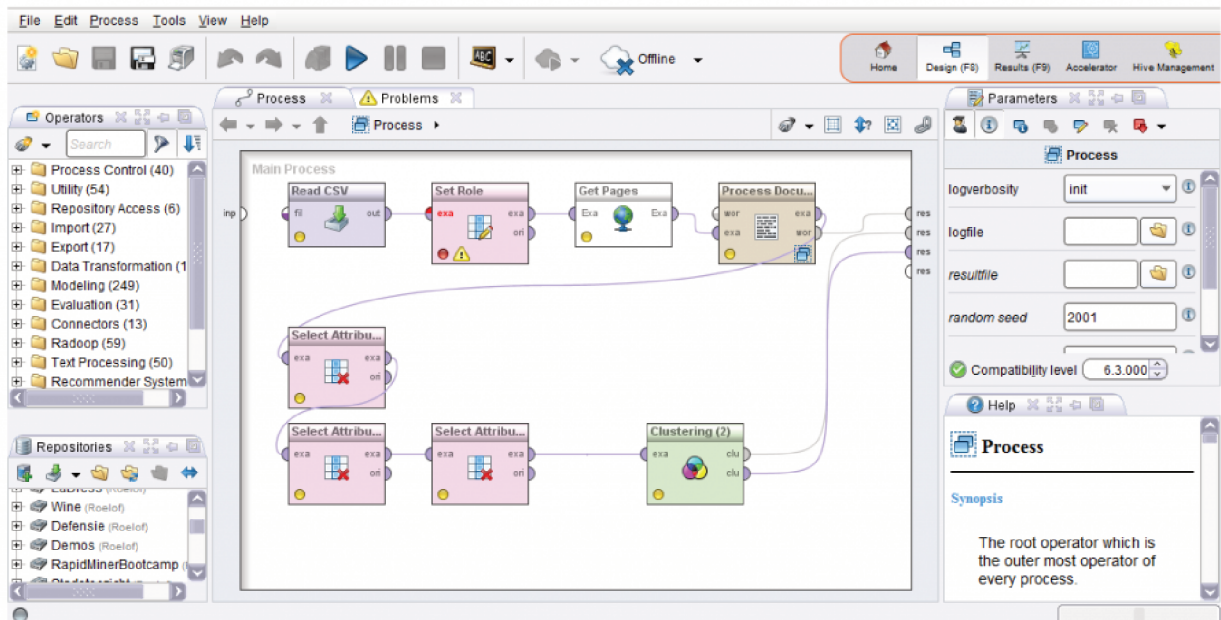


Figura 3 – Estrutura visual da ferramenta *RapidMiner*

3.6 Visualização - Ferramenta *QlikView*

Para que os usuários da área de BI com um todo sejam capazes de visualizar os dados, não basta somente oferecer uma estrutura de *data warehouse*. É necessário que, ao final de todo o processo, o usuário seja capaz de ter ideias sobre os dados, explorando e visualizando os dados de uma forma dinâmica e de vários ângulos, e é nessa parte que o processo OLAP é essencial na comunicação do BI com seus usuários e vem sendo bastante utilizado no contexto gerencial das organizações.

A sigla OLAP significa *On-line Analytical Processing*, o próprio nome já passa a ideia sobre a funcionalidade das ferramentas OLAP, que é de transmitir ao usuário informação mais rápida, simples e flexível e que seja útil para analistas, gerentes ou qualquer outro usuário tomar decisões de acordo com sua área. E é esse tipo de ferramenta que pode ser utilizada para que os usuários possam explorar os dados organizados no *data warehouse* (MARIA, 2007).

Neste trabalho será utilizada uma ferramenta denominada *QlikView*, fabricada pela QlikTech e mostrada na Figura 4. A *QlikView* possui boa aceitação no mercado

devido a sua facilidade de instalação e de manipulação das informações que podem vir de várias maneiras (GARCÍA, 2012), desde uma planilha excel até um banco de dados relacional.



Figura 4 – Exemplo de um relatório feito na ferramenta do *QlikView*.

Para que se obtenha um ganho de performance, a atratividade é um diferencial do *QlikView*, que associa os dados que podem ser carregados pelo usuário apenas a partir de um conhecimento básico da linguagem de SQL, agregando toda essa informação em um grande arquivo, conhecido como *Data Cloud* que fica disponível em memória (Sá et al., 2011). O *QlikView* utiliza uma tecnologia chamada AQL, *Associative Query Analytical* que permite a associação dos dados enquanto as consultas são feitas na ferramenta (disponível no site da QlikTech), proporcionando uma flexibilidade maior para o usuário já que ele não precisa seguir uma hierarquia nos seus dados.

Após a instalação do *QlikView*, o primeiro passo é gerar o script dos dados que serão carregados. Nessa etapa utilizam-se os comandos na linguagem SQL caso a fonte dos dados seja um banco de dados, para fazer a conexão e a carga dos dados. Uma vantagem do *QlikView* em guardar isso em memória é que a carga dos dados só precisará ser refeita caso ocorra mudança dos dados na fonte, caso contrário pode-se usar esse arquivo que é gerado em toda a aplicação, sendo desnecessário o banco de dados a partir desse ponto. Com os dados carregados, é feita a associação dos dados dentro da ferramenta utilizando o

AQL, que também pode ser visualizado pelo usuário. O último passo é o desenvolvimento da etapa gráfica, onde são gerados gráficos, tabelas e filtros que são fáceis de fazer já que a ferramenta é bem intuitiva, permitindo aos usuários criarem relatórios simples ou mais complexos de acordo com a necessidade.

Como toda ferramenta possui seus pontos positivos e negativos, o *QlikView* tem um ponto negativo no requisito de custo caso a empresa precise de expandir os relatórios para outros usuários devido ao preço da licença que é cobrada separadamente para cada um que queira acesso a esses relatórios. Neste trabalho, em específico, a *QlikView* será utilizada para que auxilie a visualização dinâmica dos resultados encontrados ao final de todo o processo.

4 Desenvolvimento do trabalho

4.1 Processo de ETL

A fase de Extração Transformação e Carga é bastante crítica num Data Warehouse, por envolver a movimentação dos dados de origem nos sistemas transacionais obedecendo as regras de negócio. O primeiro passo a ser dado é a definição das fontes de dados e a captura dos mesmos. Usando a ferramenta *Datastage*, foi possível executar a extração dos dados do sistema de gestão empresarial SAP e carregá-los para a área de *Business Intelligence*.

Os dados passam por um fluxo de carga que vai desde o sistema de origem, o SAP, até a tabela mais completa do BI que é chamada de tabela de Fatos. Nesta tabela estão as relações entre as tabelas de Dimensões e os próprios atributos das Fatos, garantindo que as informações estejam relacionadas. Essa forma de estruturação de tabelas em Fatos e Dimensões é definida em [Kimball \(2013\)](#).

A Figura 5 mostra o processo de Extração dos dados no SAP, sendo que no primeiro passo os dados são apenas trazidos do SAP e armazenados em tabelas chamadas ODS, *Operational Data Store*. Uma ODS é responsável por integrar os dados operacionais da empresa em um módulo para que depois disso comece a ser processada. O componente azul, é o responsável por capturar os dados do SAP, e após a coleta inicial ela é transferida para os *transformers* (representados pelos desenhos de seta azul com vermelho), onde alguns campos são modificados com base na necessidade final do usuário, e a partir daí são armazenados em tabelas no banco de dados transacional.

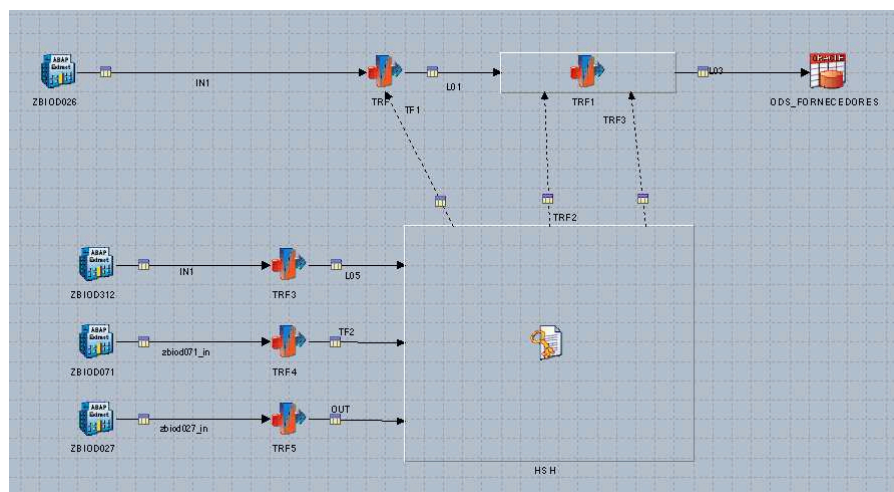


Figura 5 – Carregamento das informações dos fornecedores para o primeiro estágio do processo de ETL.

Depois disso os dados operacionais são carregados para tabelas especiais que são chamadas de "Dimensões" e a última parte são as tabelas de histórico que são chamadas de "Fatos". A extração dos dados será feita por uma tabela que dentro do BI da distribuidora engloba dados de clientes, produtos, fornecedores, valor de venda, valor de devolução, cálculos de ICMS e impostos, entre outros cálculos contábeis.

4.2 Preparação dos dados de treinamento

Os dados que serão utilizados têm como origem no sistema denominado SAP ERP. A sigla ERP vem de *Enterprise Resource Planning*, e pode ser traduzida como Sistema Integrado de Gestão Empresarial. Essa ferramenta visa integrar todos os sistemas da empresa em um único sistema, o que torna a extração de informações para a análise proposta neste trabalho muito mais rápida. A Figura 6 ilustra um diagrama de ER das tabelas que serão usadas para filtrar os dados que serão analisados.

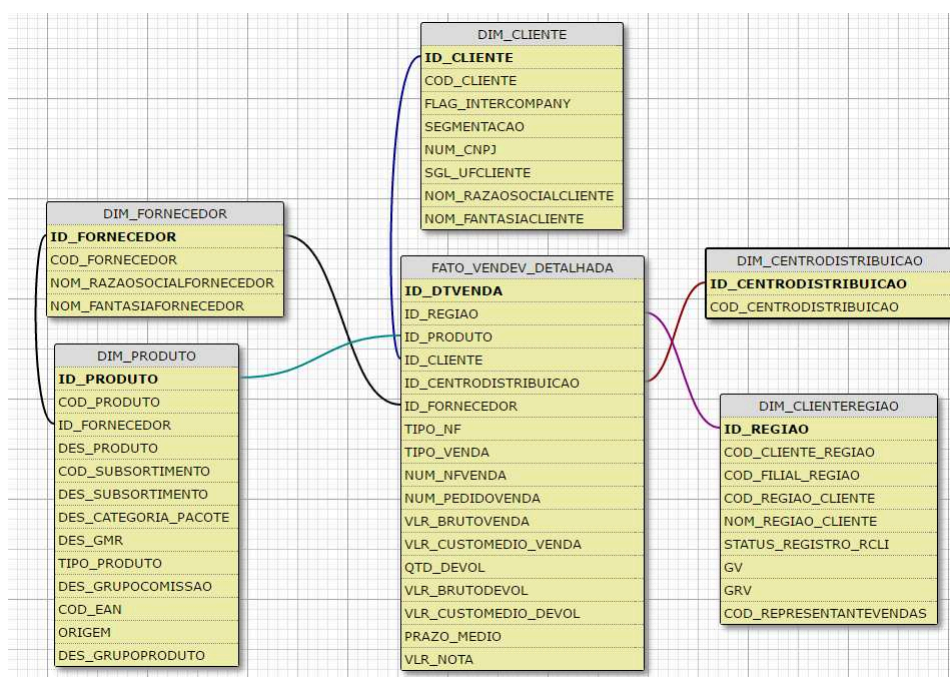


Figura 6 – Diagrama ER das tabelas que serão utilizadas.

Como apresentado no diagrama da Figura 6, tem-se um modelo estrela na relação dos dados, onde a tabela Fato é composta pelas chaves primárias das dimensões e alguns atributos quantitativos (valor de venda e devolução brutos, quantidade de devolução, valor de nota fiscal e entre outros), e as dimensões são compostas pelos atributos descritivos (código de produto, descrição do produto, nome do fornecedor, etc.), conforme definição proposta em Kimball (2013).

Uma análise mais rápida e eficiente dos dados pode ser feita com a preparação de um subconjunto chamado de dados de treinamento. No presente trabalho, o conjunto completo reúne informações dos clientes, relacionados aos produtos e às vendas de uma distribuidora de medicamentos, e os dados de treinamento contemplam essas informações para o primeiro trimestre de 2016.

Definido o conjunto de dados de treinamento, a próxima etapa é fazer a classificação prévia dos clientes deste grupo pois esta será a métrica para previsão dos clientes desejados. Para essa classificação, será utilizada a técnica de estatística descritiva que é responsável por resumir e descrever os dados quando estão em uma grande quantidade, para facilitar a análise dos mesmos.

Os dados históricos foram extraídos e estão relacionados pelo código do cliente, que será a identificação usada neste estudo. Para cada cliente será fornecida a quantidade de dias em que ele efetuou compra, a quantidade total de produtos distintos que ele comprou e qual o valor bruto da venda. Partindo-se desses atributos foram aplicados pesos em seus valores, foi calculada a média ponderada, e em seguida, foram definidas faixas onde cada um dos clientes listados se encaixaria.

A distribuição dos clientes entre as faixas de classificação considerou as quatro categorias citadas e o maior valor que o cálculo da média ponderada poderia atingir. A média foi calculada colocando-se:

- Peso 1 no valor bruto de venda;
- Peso 2 na quantidades de dias em que houve alguma compra para o cliente; e
- Peso 3 na quantidade de produtos.

Desta forma, esse cálculo propiciou uma forma eficiente de nivelamento dos clientes. A Figura 7 apresenta a definição das regras para que os dados sejam separados em quatro faixas distintas.

```
= IF((([VALOR] + 2 * [DIAS] + 3 * [PRODUTO]) < '100000,00'), 'BRONZE',
    IF((([VALOR] + 2 * [DIAS] + 3 * [PRODUTO]) > '100000,00')
    AND (([VALOR] + 2 * [DIAS] + 3 * [PRODUTO]) <= '500000,00'), 'PRATA',

    IF((([VALOR] + 2 * [DIAS] + 3 * [PRODUTO]) > '500000,00')
    AND (([VALOR] + 2 * [DIAS] + 3 * [PRODUTO]) <= '1000000,00'), 'OURO',

    IF((([VALOR] + 2 * [DIAS] + 3 * [PRODUTO]) > '1000000,00')
    AND (([VALOR] + 2 * [DIAS] + 3 * [PRODUTO]) < '10000000,00'), 'DIAMANTE'))))
```

Figura 7 – Regras para a definição dos dados de 2016.

Após aplicadas as regras sobre os dados de vendas realizadas no ano de 2016, obteve-se como resultado uma tabela com a identificação dos clientes e seus atributos a serem considerados e sua classificação prévia encontrada, conforme Figura 8.

Cliente	VALOR	DIAS	PRODUTO	CLASSE T
309936	7.198.943,83	47	1004	DIAMANTE
42084	4.534.133,16	53	503	DIAMANTE
317668	3.675.606,09	19	82	DIAMANTE
47926	3.394.041,74	39	611	DIAMANTE
346742	3.060.999,09	40	539	DIAMANTE
314374	2.380.767,92	35	489	DIAMANTE
344246	2.240.834,31	62	2057	DIAMANTE
327333	1.715.067,11	60	1752	DIAMANTE
345518	1.163.340,14	13	66	DIAMANTE
98369	1.022.011,05	12	32	DIAMANTE
305361	903.926,28	37	299	OURO
304058	877.696,87	41	1546	OURO
41136	876.323,83	62	2205	OURO
40134	873.592,15	62	2337	OURO
318697	849.048,50	47	375	OURO
304767	745.941,20	68	1650	OURO
27819	688.699,76	65	1662	OURO
140232	656.204,59	55	816	OURO
40465	645.146,39	67	775	OURO
82743	617.588,50	68	1791	OURO
329184	599.668,98	62	1820	OURO
43810	593.116,39	70	1234	OURO

Figura 8 – Dados na planilha após a classificação dos clientes de 2016.

A partir da utilização dos dados de clientes e das definições de categorias para o primeiro trimestre de 2016, foi aplicado o algoritmo de classificação *Decision Tree* contido no *RapidMiner* e o resultado encontrado foi a árvore de decisão ilustrada na Figura 9.

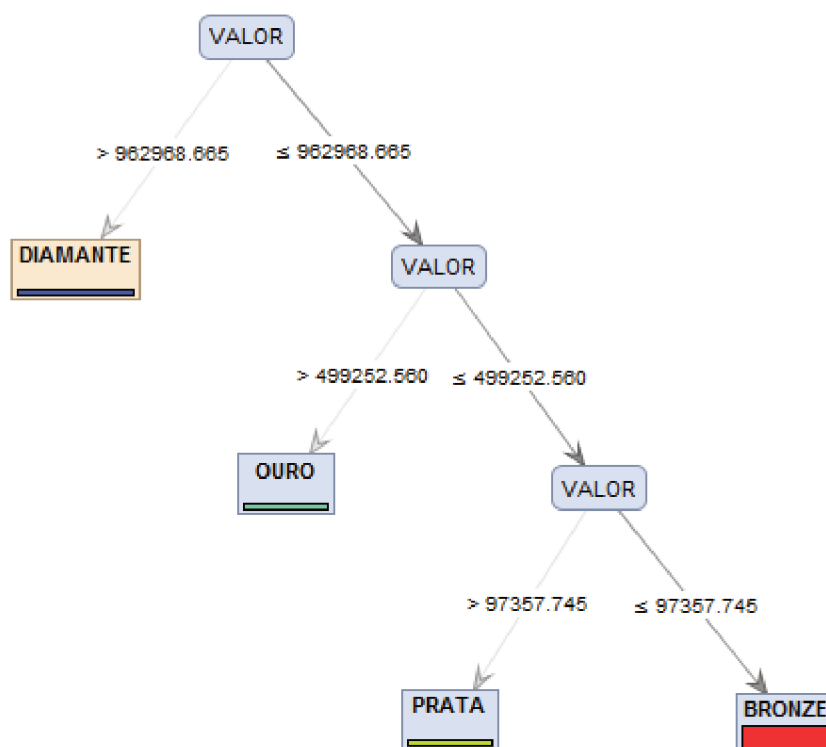


Figura 9 – Árvore de Decisão resultante do processamento dos clientes de 2016.

4.3 Algoritmo de Classificação baseado em Árvore de Decisão

Um algoritmo de mineração de dados é um conjunto de heurísticas e cálculos onde ele primeiro analisa os dados fornecidos, procurando padrões específicos ou tendências, e após iterações ele encontra parâmetros para a criação do modelo final. Esses modelos permitem identificar a probabilidade de resultados futuros com base em dados históricos disponíveis, isto é, eles aproveitam os resultados obtidos de testes passados para treinar e desenvolver novos modelos de previsão, que podem ser através de clusters, modelos matemáticos, regras de associação ou árvores de decisão. No caso deste trabalho, serão usados os resultados obtidos em 2016 para estabelecer previsões sobre as possibilidades de resultados em 2017 pelas árvores de decisão.

A tarefa de escolher qual o algoritmo que deve produzir um resultado satisfatório para o problema, pode ser bastante complicada, já que um algoritmo pode ser usado para mais de um tipo de solução. Conhecer os tipos dos dados também é fundamental para a escolha do método mais adequado, sendo estes categorizados basicamente em quantitativos e qualitativos. Os dados quantitativos são representados por valores numéricos e podem ser discretos e contínuos. Já os dados qualitativos contêm os valores nominais. Assim,

antes de se aplicar algoritmos de mineração é necessário explorar, conhecer e preparar os dados. Para a escolha do algoritmo usando árvore de decisão, foram considerados:

- Tipo de algoritmo: necessário que o algoritmo pudesse prever uma variável discreta, com base em outros valores do conjunto de dados, isto é, a sua classificação entre as categorias teria que ser decidida considerando valor da venda, quantidade de dias que tiveram vendas e a quantidade de produtos vendidos.
- Tipo de tarefa: essa decisão envolve totalmente o objetivo final pois são os passos que serão percorridos até ele ser atingido. Para esse estudo o objetivo é, resumidamente, prever a classificação de um cliente colocando-os em "categorias".

Tendo esses dois pontos definidos, foi concluído que o melhor algoritmo a ser usado seria a Árvore de Decisão, além de ser um modelo de fácil entendimento pois é baseada em regras que podem ser expressas em linguagem natural.

Árvores de Decisão são alguns dos vários métodos que podem ser utilizados na mineração de dados, entretanto é mais usada quando o problema é de classificação, especificamente quando se trata de categorização (LE MOS; STEINERAND; NIEVOLA, 2005). Esse método é amplamente utilizado pois os resultados obtidos são mais fáceis de serem compreendidos, ainda mais quando é relacionado com as regras de associação que representam esses resultados, isso porque elas se baseiam na escolha do atributo que possui mais relevância dentro do modelo.

No caso deste trabalho, como foram escolhidas apenas 3 variáveis e elas são de tipo discretas (ou seja, são definidas entre um intervalo de valores (GRUPOALVO, 2009)), estabeleceu-se que apenas o valor de venda bruto seria definido como relevante, conforme mostra a Figura 10. Como se trata de um exemplo real, não é possível modificar os dados encontrados, porém esse resultado poderia ser diferente caso a regra de cálculo dos clientes de 2016 fosse alterada, já que essa classificação de 2016 serve de base para o algoritmo classificar os clientes de 2017.

As Árvores de decisão são ditas estruturas de dados recursivas que utilizam a técnica de dividir-para-conquistar para representar a solução dos problemas. Essa técnica consiste nos seguintes passos (TOFFOLO; CARVALHO, 2017):

- Divisão: o problema inicial é dividido em subproblemas menores.
- Conquista: esses subproblemas gerados são resolvidos recursivamente e independentes.
- Combinação: os resultados dos subproblemas são combinados com a finalidade de encontrar uma solução para o problema inicial.

Além disso, as árvores são constituídas por nós, os quais são: os nós folha e os nós de decisão. Os nós folha são os últimos nós da árvore e são os que, neste trabalho, representam as classes em que os clientes são classificados. Já os nós de decisão são os que representam o atributo escolhido possui uma relevância na atual etapa do algoritmo (MO-NARD; BARANAUSKAS, 2003), e com base nestes nós, cada caminho adquirido neste processo, a partir do nó raiz até um nó folha, forma uma regra de associação.

Tree

```

VALOR > 962968.665: DIAMANTE {DIAMANTE=10, OURO=0, PRATA=0, BRONZE=0}
VALOR ≤ 962968.665
|   VALOR > 499252.560: OURO {DIAMANTE=0, OURO=20, PRATA=0, BRONZE=0}
|   VALOR ≤ 499252.560
|   |   VALOR > 97357.745: PRATA {DIAMANTE=0, OURO=0, PRATA=373, BRONZE=1}
|   |   VALOR ≤ 97357.745: BRONZE {DIAMANTE=0, OURO=0, PRATA=0, BRONZE=6885}

```

Figura 10 – Regras de associação dos dados dos clientes de 2016.

Sabendo de todos estes pontos sobre árvores de decisão, então é importante saber sobre o seu processo de construção para que a conferência do resultado seja bem sucedida. Alguns algoritmos são bastante utilizados nas bibliografias, e em Zuben e Attux (2010) foram citados por serem os principais: ID3, C4.5 e CART.

Na Figura 11 temos os passos do pseudocódigo base dos algoritmos de árvores de decisão, onde considera-se T como o conjunto de treinamento e C as classes para classificação.

Algoritmo Base: Árvores de Decisão

Passos:

1. T contém um ou mais exemplos, todos pertencentes à mesma classe C_j . Nesse caso, a árvore de decisão para T é um nó folha identificando a classe C_j ;
2. T não contém exemplos. Novamente, nessa situação, a árvore é uma folha mas a classe associada à folha deve ser determinada a partir de informação além de T . Por exemplo, a classe mais freqüente para o nó pai desse nó pode ser utilizada;
3. T contém exemplos que pertencem a várias classes. Nesse caso a idéia é refinar T em subconjuntos de exemplos que são (ou aparentam ser) conjuntos de exemplos pertencentes a uma única classe.

Normalmente, um teste é escolhido, baseado em um único atributo que possui resultados mutuamente exclusivos (na realidade, cada indutor tem sua própria forma de escolher o atributo que será utilizado no teste). Sejam os possíveis resultados do teste denotados por $\{O_1, O_2, \dots, O_r\}$. T é então particionado em subconjuntos $T_1, T_2, \dots, T_r$, nos quais cada T_i contém todos os exemplos em T que possuem como resultado daquele teste o valor O_i . A AD para T consiste em um nó interno identificado pelo teste escolhido e uma aresta para cada um dos resultados possíveis;

4. Os passos 1, 2 e 3 são aplicados recursivamente para cada subconjunto de exemplos de treinamento de forma que, em cada nó, as arestas levam para as subárvores construídas a partir do subconjunto de exemplos T_i ;

Figura 11 – Pseudocódigo do algoritmo de árvore de decisão (MONARD; BARANAUSKAS, 2003).

O indutor citado no passo 3 do algoritmo é o que vai ser escolhido dependendo do critério de seleção na escolha do nó de predição, isto é, qual será o atributo de relevância. Dentro do *RapidMiner* temos 4 tipos de critérios: ganho de informação, razão de ganho, índice gini e precisão (RAPIDMINER, 2014).

- Ganho de Informação: a entropia de todos os atributos é calculada. O atributo com entropia mínima é selecionado para dividir. Este método tem uma tendência para selecionar atributos com um grande número de valores.
- Razão de Ganho: é uma variante do ganho de informação. Ele ajusta o ganho de informações para cada atributo para permitir a amplitude e uniformidade dos valores dos atributos.
- Índice Gini: a divisão em um atributo escolhido dá uma redução no índice médio de gini dos subconjuntos resultantes.
- Precisão: esse atributo é selecionado para dividir que maximiza a precisão de toda a árvore.

5 Experimentos e resultados

5.1 Fluxo de processo

O *RapidMiner* é a ferramenta onde será criado um processo para que sejam carregados os dados que foram organizados em arquivos de planilhas eletrônicas (arquivos em Excel) e depois obter os resultados desejados. Os arquivos de entrada contém as seguintes colunas:

- CLIENTE: identificador ou código do cliente.
- VALOR: valor monetário de venda em reais.
- DIAS: quantidades de dias em que foi efetuada alguma venda.
- PRODUTO: quantidade de produtos vendidos ao longo desses dias.
- CLASSE T: classificação do cliente entre Diamante, Ouro, Prata e Bronze.

A última coluna é preenchida apenas no arquivo onde estão resgistrados os dados do primeiro trimestre de 2016, pois o arquivo com os dados do mesmo período de 2017, será a coluna definida como coluna previsível. A Figura 12 mostra como o arquivo com os dados de 2017 foi preparado, se diferenciando da Figura 8 pela ausência de dados na última coluna.

Ciente	VALOR	DIAS	PRODUTO	CLASSE T
309936	9.209.264,38	56	1150	
336715	8.661.252,65	30	168	
348070	8.171.241,99	37	182	
346742	5.442.209,09	34	581	
47926	3.716.810,44	39	625	
336868	3.169.187,20	20	120	
31182	3.145.802,61	20	112	
305361	2.131.742,08	55	502	
345956	1.881.680,45	21	51	
140232	1.843.077,94	62	1231	
317668	1.835.306,76	43	301	
354293	1.796.984,06	8	348	
344246	1.684.514,17	64	1537	
42084	1.660.302,94	43	231	
327333	1.235.435,50	65	1416	
345518	1.224.151,42	12	80	
304058	891.729,48	32	1535	
314374	843.327,61	24	191	
352285	742.997,37	24	79	
40134	726.769,32	49	2109	
48758	687.474,22	63	1542	
347093	669.428,97	26	125	

Figura 12 – Arquivo dos dados do 1º trimestre de 2017.

Após selecionado o caminho desses arquivos, eles são carregados (pelos componentes apresentados na Figura 13) para a ferramenta que sugere algumas configurações para os dados que estão sendo carregados, dentre elas mudança de tipo de dados (de real para texto, ou inteiro para real).

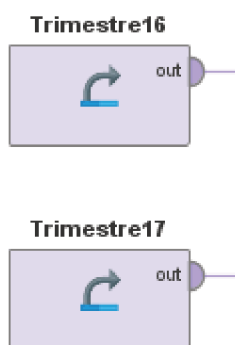


Figura 13 – Componentes da ferramenta que permitem carregar arquivos externos.

Concluído o carregamento do arquivo, ele é armazenado em um repositório local e é exibido como na Figura 14.

Row No.	Cliente	VALOR	DIAS	PRODUTO	CLASSE T
1	309936	7198943.830	47	1004	DIAMANTE
2	42084	4534133.160	53	503	DIAMANTE
3	317668	3675606.090	19	82	DIAMANTE
4	47926	3394041.740	39	611	DIAMANTE
5	346742	3060999.090	40	539	DIAMANTE
6	314374	2380767.920	35	489	DIAMANTE
7	344246	2240834.310	62	2057	DIAMANTE
8	327333	1715067.110	60	1752	DIAMANTE
9	345518	1163340.140	13	66	DIAMANTE
10	98369	1022011.050	12	32	DIAMANTE
11	305361	903926.280	37	299	OURO
12	304058	877696.870	41	1546	OURO
13	41136	876323.830	62	2205	OURO
14	40134	873592.150	62	2337	OURO
15	318697	849048.500	47	375	OURO
16	304767	745941.200	68	1650	OURO
17	27819	688699.760	65	1662	OURO

Figura 14 – Exemplo de tabela criada no *RapidMiner* com dados de 2016.

Com os arquivos de entrada devidamente inseridos no fluxo, é preciso ajustá-los para que o algoritmo de Árvore de decisão funcione corretamente ao decorrer do processo.

Primeiramente foi construído um fluxo para os dados de treinamento: os clientes do primeiro trimestre de 2016. Então pode-se aplicar o componente *"Set Role"* (o qual definirá quais serão os papéis de cada coluna dentro do processo), na saída do arquivo para que se possa identificar o papel da coluna "Cliente" como apenas identificador, conforme Figura 15.

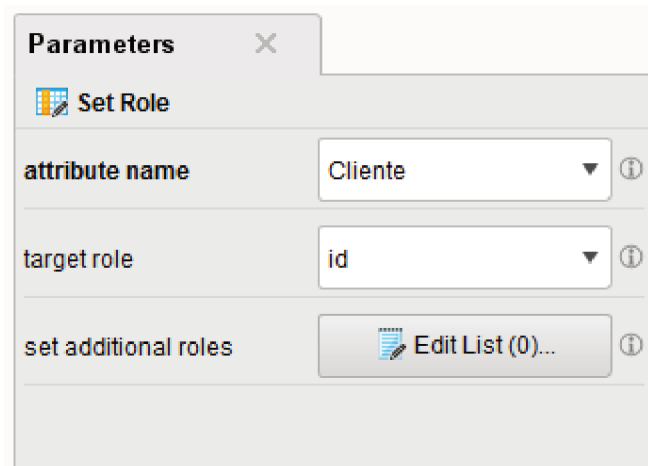


Figura 15 – Configuração da coluna "Cliente" como um indicador.

Desta forma, a coluna "Cliente" não influenciará no resultado da classificação. O fluxo do processo para esta etapa está ilustrado na Figura 16.



Figura 16 – Etapa de seleção da coluna de identificação dos dados de 2016.

Com a coluna de identificação selecionada e feita a separação do que serão os atributos influenciadores no cálculo e os identificadores, é necessário selecionar também a coluna de previsão, isto é, a coluna que auxiliará na classificação dos dados de 2017. Outro componente do tipo "Set Role" será necessário para tornar a coluna "Classe T" como rótulo. Esta etapa pode ser visualizada na Figura 17.



Figura 17 – Etapa de seleção da coluna de previsão dos dados de 2016.

Para que fosse possível extrair a árvore de decisão dos clientes de 2016, foi necessário aplicar os componentes *Decision Tree* e *Apply Model*, representados nas Figuras 18 e 19 respectivamente.

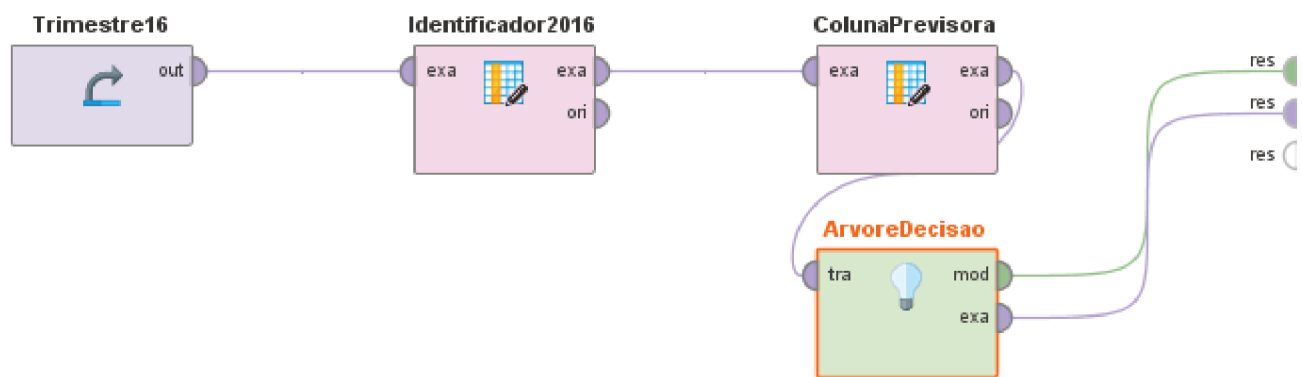


Figura 18 – Etapa de seleção do algoritmo Árvore de Decisão que será aplicada aos dados de 2016.

Esses dois componentes são aplicados um após o outro pois são dependentes, o primeiro especifica qual é o algoritmo que será aplicado no processo e o segundo faz a aplicação propriamente dita como o nome já diz.

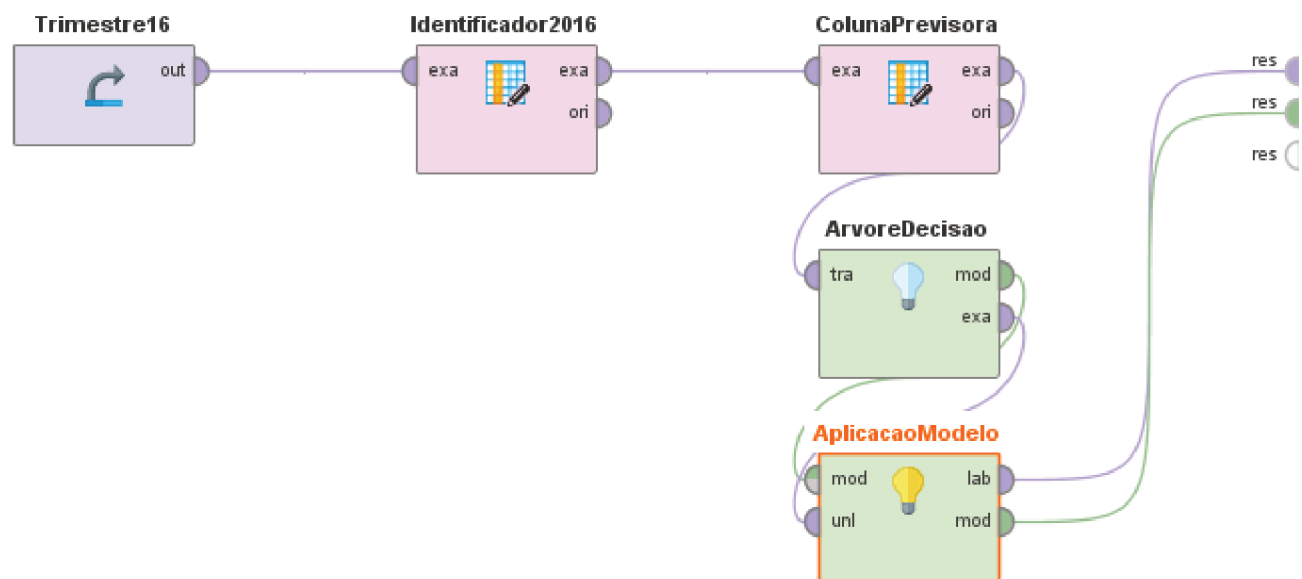


Figura 19 – Etapa de aplicação do algoritmo selecionado na Figura 18.

Todo esse processo e passos que foram feitos para os dados dos clientes de 2016 terão que ser replicados para os clientes de 2017. A Figura 20, mostra que o primeiro passo da escolha da coluna de identificação foi feita da mesma maneira.

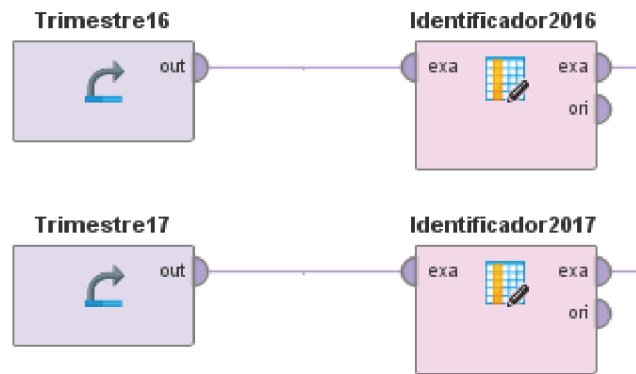


Figura 20 – Etapa de seleção da coluna de identificação dos dados de 2017.

Na Figura 12 a coluna "Classe T" no arquivo dos dados de 2017 está vazia, justamente para que ela seja preenchida com a previsão que será feita pelo algoritmo. Então para que isso seja possível, serão aplicadas a saída do componente nomeado ArvoreDecisao e a saída do componente nomeado Identificador2017 na entrada do componente AplicacaoModelo. A Figura 21 resume o processo completo.

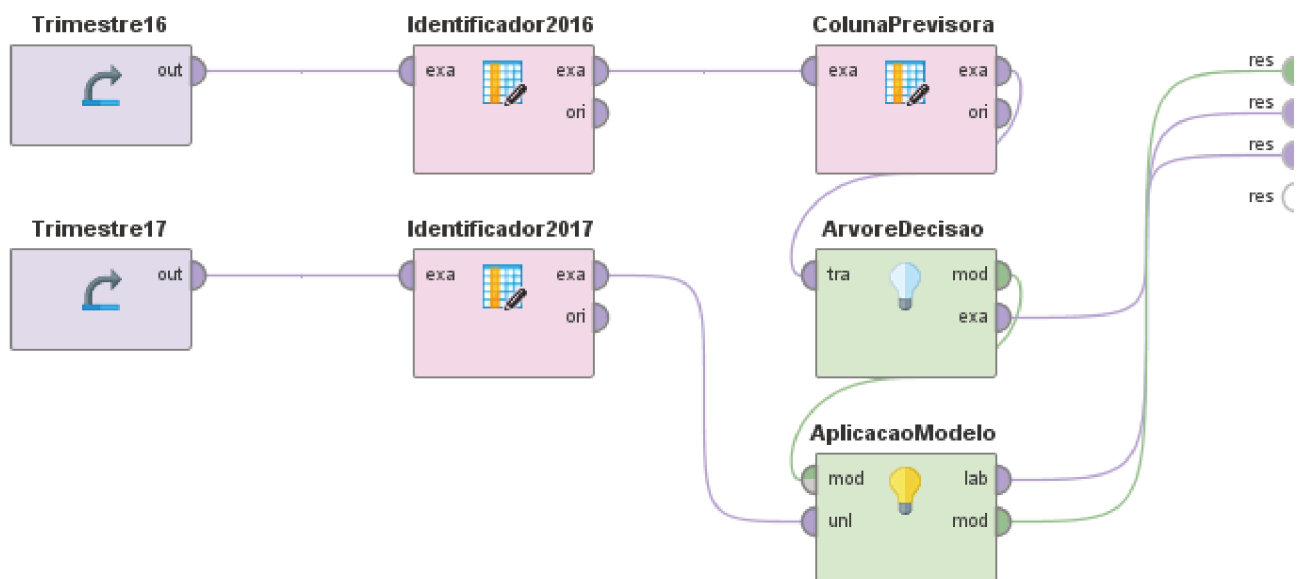


Figura 21 – Aplicação dos dados de 2016 e 2017 no processo completo de classificação.

Pode-se perceber que foram aplicadas três saídas ao final do fluxo, que foram os resultados mostrados na Figura 22.

Row No.	Cliente	CLASSE T	VALOR	DIAS	PRODUTO
1	309936	DIAMANTE	7198943.830	47	1004
2	42084	DIAMANTE	4534133.160	53	503
3	317668	DIAMANTE	3675606.090	19	82
4	47926	DIAMANTE	3394041.740	39	611
5	346742	DIAMANTE	3060999.090	40	539
6	314374	DIAMANTE	2380767.920	35	489
7	344246	DIAMANTE	2240834.310	62	2057
8	327333	DIAMANTE	1715067.110	60	1752
9	345518	DIAMANTE	1163340.140	13	66
10	98369	DIAMANTE	1022011.050	12	32
11	305361	OURO	903926.280	37	299
12	304058	OURO	877696.870	41	1546
13	41136	OURO	876323.830	62	2205
14	40134	OURO	873592.150	62	2337
15	318897	OURO	849048.500	47	375
16	304767	OURO	745941.200	68	1650
17	27819	OURO	688699.760	65	1662
18	140232	OURO	656204.590	55	816

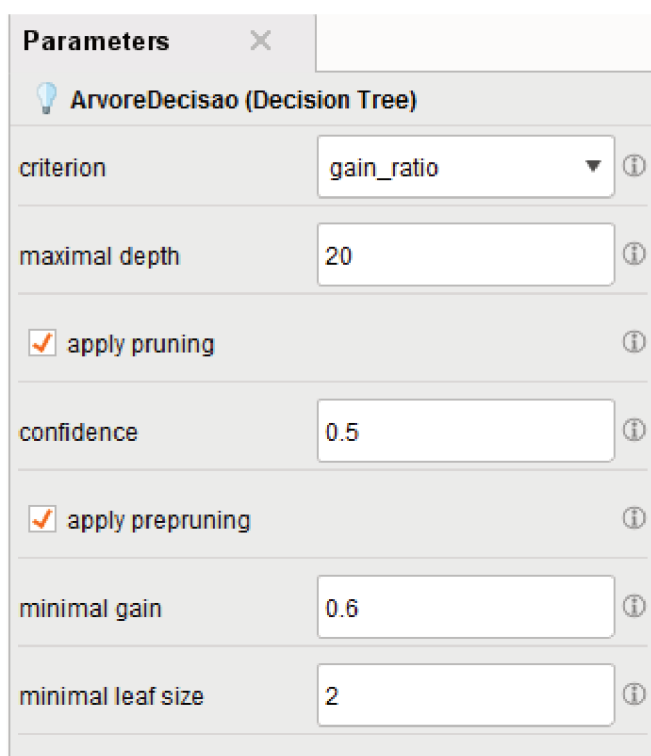
Figura 22 – Saída dos dados dos clientes de 2016.

A Figura 22 representa apenas como os dados de 2016 vão auxiliar na previsão dos valores de 2017, a coluna destacada em verde junto com a coluna faltante do arquivo de 2017 serão as bases para a classificação.

A ferramenta permite que alguns parâmetros sejam definidos pelo usuário, de acordo com a necessidade de cada um. Para o trabalho foram definidos os parâmetros, conforme Figura 23:

- Critério: critério em que os atributos serão selecionados para dividir, no caso foi o critério escolhido foi o *Razão de Ganho*.
- Profundidade máxima da árvore: a profundidade de uma árvore varia dependendo do tamanho e da natureza do seu conjunto de dados. Este parâmetro é usado para restringir o tamanho da Árvore de Decisão.
- Aplicação de poda: quando arestas ou sub-árvores estejam refletindo ruídos ou dados ausentes.
- Confiança: especifica o nível de confiança utilizado para o cálculo do erro pessimista da poda.
- Aplicação de pré-poda: permite que antes da escolha do próximo atributo para um nó de decisão pode ser aplicada uma poda.

- Ganho mínimo: o nó é dividido se o ganho for maior do que o ganho mínimo. O maior valor do ganho mínimo resulta em menos divisões e, portanto, uma árvore menor, porém um valor muito alto impedirá completamente a divisão e uma árvore com um único nó é gerada.
- Tamanho mínimo da folha: 2, isto é, árvore é gerada de tal forma que cada sub-conjunto de nó de folha tem pelo menos o número mínimo de tamanho de folha de instâncias.

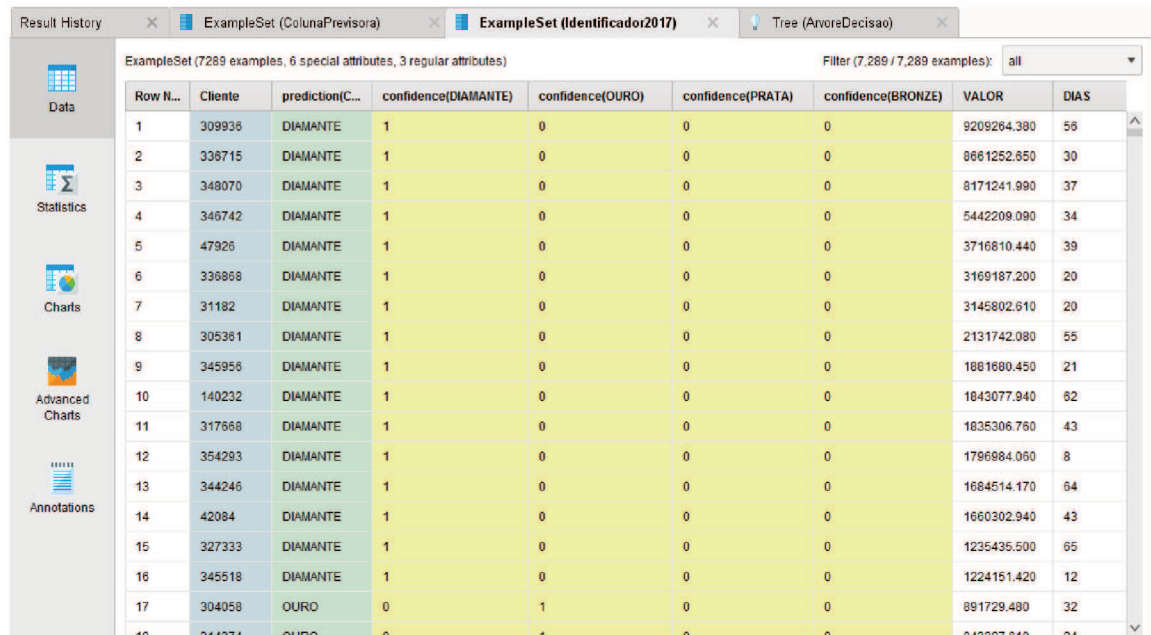


The image shows a 'Parameters' dialog box for a decision tree algorithm. The title bar says 'Parameters' with a close button. The main title is 'ArvoreDecisao (Decision Tree)' with a lightbulb icon. The parameters are listed in a table-like structure:

Parameter	Value	Info
criterion	gain_ratio	?
maximal depth	20	?
☒ apply pruning		?
confidence	0.5	?
☒ apply prepruning		?
minimal gain	0.6	?
minimal leaf size	2	?

Figura 23 – Parâmetros utilizados no algoritmo de árvore de decisão.

As figuras a seguir representam o resultado do cálculo da confiança para cada categoria que um cliente pode receber. A Figura 24 mostra que os clientes das linhas 1 a 16 são classificados como Diamante sem dúvida.



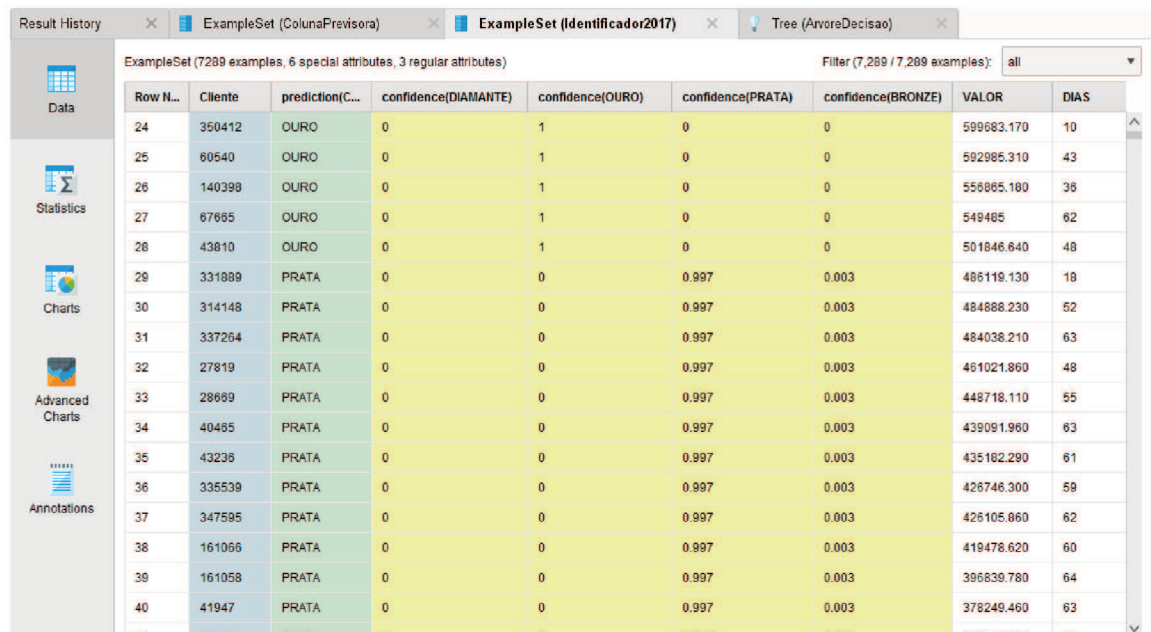
Result History | ExampleSet (ColunaPrevisora) | ExampleSet (Identificador2017) | Tree (ArvoreDecisao)

ExampleSet (7289 examples, 6 special attributes, 3 regular attributes) Filter (7,289 / 7,289 examples): all

Row N...	Cliente	prediction(C...	confidence(DIAMANTE)	confidence(OURO)	confidence(PRATA)	confidence(BRONZE)	VALOR	DIAS
1	309936	DIAMANTE	1	0	0	0	9209264.390	56
2	338715	DIAMANTE	1	0	0	0	8661252.850	30
3	348070	DIAMANTE	1	0	0	0	8171241.990	37
4	346742	DIAMANTE	1	0	0	0	5442209.090	34
5	47926	DIAMANTE	1	0	0	0	3716810.440	39
6	336868	DIAMANTE	1	0	0	0	3169187.200	20
7	31182	DIAMANTE	1	0	0	0	3145802.610	20
8	305381	DIAMANTE	1	0	0	0	2131742.080	55
9	345956	DIAMANTE	1	0	0	0	1881680.450	21
10	140232	DIAMANTE	1	0	0	0	1843077.940	62
11	317668	DIAMANTE	1	0	0	0	1835306.760	43
12	354293	DIAMANTE	1	0	0	0	1796884.060	8
13	344246	DIAMANTE	1	0	0	0	1684514.170	64
14	42084	DIAMANTE	1	0	0	0	1660302.940	43
15	327333	DIAMANTE	1	0	0	0	1235435.500	65
16	345518	DIAMANTE	1	0	0	0	1224151.420	12
17	304058	OURO	0	1	0	0	891729.480	32
18	344374	OURO	0	1	0	0	843337.540	24

Figura 24 – Cálculo de confiança para os clientes categorizados como Diamante de 2017.

Já nos clientes mostrados na Figura 25 é possível observar que o resultado deste cálculo foi diferente para cada cliente. No resultado tem-se que $\text{CONFIANÇA}(\text{PRATA}) = 0.997$ e $\text{CONFIANÇA}(\text{BRONZE}) = 0.003$, o que significa que esses clientes de acordo com o algoritmo tiveram uma chance de 0.003 de serem categorizados como Bronze.



Result History | ExampleSet (ColunaPrevisora) | ExampleSet (Identificador2017) | Tree (ArvoreDecisao)

ExampleSet (7289 examples, 6 special attributes, 3 regular attributes) Filter (7,289 / 7,289 examples): all

Row N...	Cliente	prediction(C...	confidence(DIAMANTE)	confidence(OURO)	confidence(PRATA)	confidence(BRONZE)	VALOR	DIAS
24	350412	OURO	0	1	0	0	599683.170	10
25	60540	OURO	0	1	0	0	592985.310	43
26	140398	OURO	0	1	0	0	558865.180	36
27	67655	OURO	0	1	0	0	549485	62
28	43810	OURO	0	1	0	0	501846.640	48
29	331889	PRATA	0	0	0.997	0.003	486119.130	18
30	314148	PRATA	0	0	0.997	0.003	484888.230	52
31	337264	PRATA	0	0	0.997	0.003	484038.210	63
32	27819	PRATA	0	0	0.997	0.003	461021.860	48
33	28669	PRATA	0	0	0.997	0.003	448718.110	55
34	40485	PRATA	0	0	0.997	0.003	439091.960	63
35	43238	PRATA	0	0	0.997	0.003	435182.290	61
36	335539	PRATA	0	0	0.997	0.003	426746.300	59
37	347595	PRATA	0	0	0.997	0.003	426105.860	62
38	161066	PRATA	0	0	0.997	0.003	418478.620	60
39	161058	PRATA	0	0	0.997	0.003	396839.780	64
40	41947	PRATA	0	0	0.997	0.003	378249.460	63
41	350412	PRATA	0	0	0.997	0.003	370445.310	67

Figura 25 – Cálculo de confiança para os clientes categorizados como Diamante de 2017.

A ferramenta *RapidMiner* também permite que sejam exibidas algumas estatísticas produzidas em relação aos dados, tais como: valor mínimo e máximo dos atributos, média entre os valores, quantidade de atributos nas categorias (quantos clientes são Diamante, Ouro, Prata e Bronze), e se algum atributo possui valor faltante. A Figura 26 ilustra a tela em que podem ser visualizadas estas estatísticas.

Name	Type	Missing	Statistics		
Id Cliente	Integer	0	Min 104	Max 355327	Average 217888.295
Prediction prediction(CLASSE T)	Polynomial	0	Least OURO (12)	Most BRONZE (6931)	Values BRONZE (6931), PRATA
Confidence_DIAMANTE confidence(DIAMANTE)	Real	0	Min 0	Max 1	Average 0.002
Confidence_OURO confidence(OURO)	Real	0	Min 0	Max 1	Average 0.002
Confidence_PRATA confidence(PRATA)	Real	0	Min 0	Max 0.997	Average 0.045
Confidence_BRONZE confidence(BRONZE)	Real	0	Min 0	Max 1	Average 0.951
VALOR	Real	0	Min 923.940	Max 9209264.380	Average 35034.331
DIAS	Integer	0	Min 1	Max 66	Average 27.738

Figura 26 – Outros cálculos produzidos pelo *RapidMiner*.

Uma aba "*Charts*" pode ser selecionada (ver no lado esquerdo da Figura 27) para que seja apresentado um gráfico de dispersão. Neste contexto, este gráfico pode ser utilizado para comprovar a relação entre os valores simultâneos das duas variáveis relacionadas ao processo, mostrando o que acontece quando uma ou outra se alteram. Assim, o gráfico mostra como os clientes estão dispersos relacionando-se dias de venda e valor bruto da venda. Avaliando-se os dias de venda e o valor bruto das mesmas pelo gráfico apresentado, o diagrama resultante não apresenta uma relação linear entre as duas variáveis.

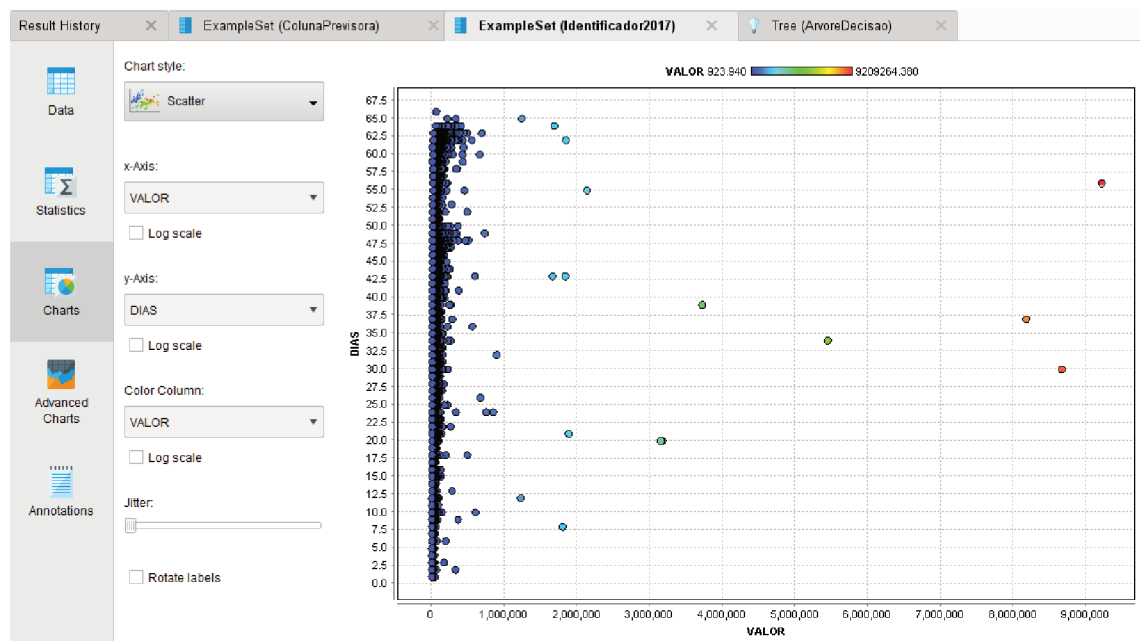


Figura 27 – Gráfico de Dispersão representando a quantidade de dias em relação ao valor bruto por cliente.

5.2 Análise dos Resultados Obtidos

Mesmo que a ferramenta utilizada para a mineração de dados forneça esta quantidade de informações bem intuitivas, os usuários da área de marketing não estão familiarizados com ela. Portanto, após os dados percorrerem todo o processo, o resultado foi extraído para uma planilha eletrônica e esses dados utilizados para desenvolver o relatório para os usuários finais dessa informação.

A ideia final é disponibilizar o que foi estudado e analisado para a área de marketing, que daqui em diante irá analisá-los em uma nova perspectiva, para isso usaremos a ferramenta *QlikView* para o desenvolvimento de um relatório para facilitar a análise da área. A visão solicitada pela equipe de marketing foi bastante simples de desenvolver, já que é somente para um assunto específico, ter noção quanto monetariamente os clientes compram, a quantidade que eles compram e os dias que os produtos são adquiridos.

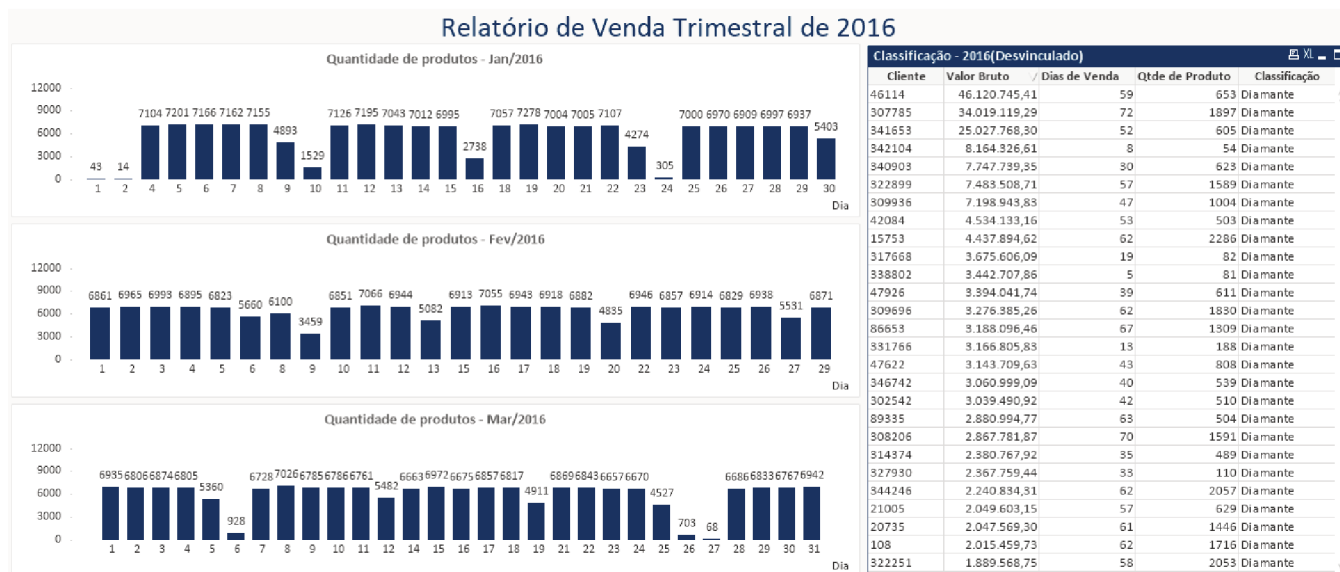


Figura 28 – Relatório de venda trimestral de 2016.

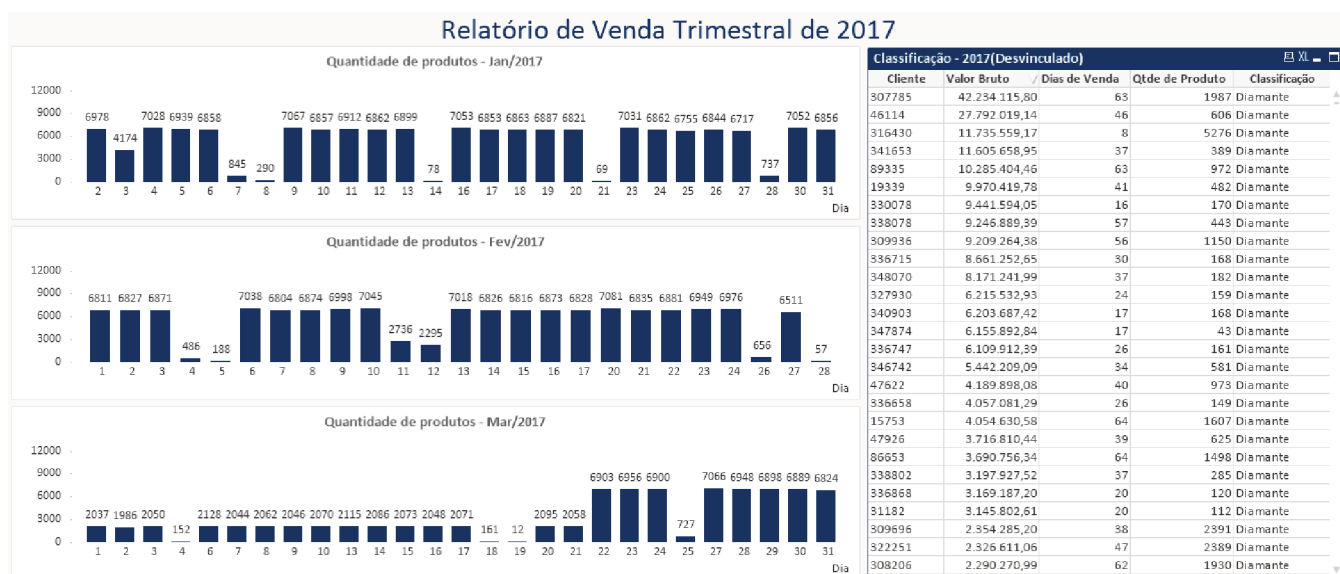


Figura 29 – Relatório de venda trimestral de 2017.

Os usuários desejam simplicidade para explorar os relatórios, por isso foram solicitadas tabelas com as mesmas informações que foram exploradas no início do estudo, e um conjunto de três gráficos um para cada mês separados pelo ano. O eixo vertical dos gráficos representa a quantidade de produtos e o eixo horizontal representa o dia do mês, permitindo que sejam analisados quantos produtos são vendidos por dia em cada um dos meses.

Após uma semana depois da entrega dos relatórios para a área de marketing, foi feita uma análise como eles estavam utilizando as informações e se obtiveram resultados

positivos. O resultado obtido foi que eles puderam analisar quais são os dias que possui uma maior quantidade de produtos vendidos, o que é importante para saber em qual época do mês eles podem fazer algum tipo de promoção para os clientes. Os clientes são identificados após a seleção de algum intervalo do mês, e a partir disso são direcionados diferentes tipos de promoções de acordo com a categoria do cliente. Outro ponto que será analisado por eles, é se houve alguma queda de venda para algum cliente, ou se este cliente mudou de alguma categoria, isso é importante pois, caso algum cliente tenha baixado suas compras na distribuidora, eles possam entrar em contato com o cliente e fazer novos acordos.

Portanto, todo o estudo feito ao longo deste trabalho teve um resultado positivo e correto, tanto para a área de marketing quanto para a distribuidora como um todo, que poderá analisar seus clientes de uma forma mais rápida e segura.

6 Conclusão

A mineração de dados emerge como uma importante ferramenta na descoberta de conhecimento a partir de dados provenientes do processo empresarial, e representa uma oportunidade de ampliar significativamente a capacidade de transformar grandes volumes de dados em informação útil. Neste contexto, este trabalho teve a finalidade de aplicar um conjunto de métodos que permitiram a análise dos dados de vendas de produtos farmacêuticos, utilizando um algoritmo para a classificação dos clientes. Foi realizada uma revisão dos temas atuais da área da descoberta de conhecimento em bancos de dados e algoritmos de classificação visando suportar a proposição do trabalho.

O algoritmo aplicado baseou-se em Árvores de Decisão, por ser simples e adequado à tarefa de classificação de clientes em categorias. Esse processo foi desenvolvido na ferramenta *RapidMiner*. A aplicação desenvolvida atendeu às expectativas gerando resultados satisfatórios e permitindo a produção de análises sobre os clientes, que puderam ser expostas através de relatórios compostos por gráficos, tabelas, entre outros.

De acordo com a área de marketing, esses relatórios auxiliaram na análise que anteriormente era realizada manualmente, venda por venda, o que demandava muito esforço da equipe. Assim, as informações de valor de venda de cada cliente, relacionadas aos dias que eles compraram e quais os diferentes produtos que foram obtidos, facilitam a compreensão do conjunto de vendas através de tabelas intuitivas e gráficos para comparação.

Ao longo deste trabalho surgiram novas possibilidades que não foram desenvolvidas, como utilizar regras de associação para descobrir quais são os produtos que são mais vendidos em conjuntos com outros produtos. Esta análise pode proporcionar que a equipe de logística possa organizar os produtos de uma maneira mais eficiente nos galpões de estoque. Na mesma linha do presente trabalho, poderia-se propor a utilização de métodos de regressão para que além de dizer qual é a categoria dos clientes, fosse possível prever qual seria o valor bruto de venda para esses clientes no mesmo intervalo de tempo, porém no ano seguinte. Ainda, outro método que poderia ser utilizado neste trabalho no lugar de classificação, seria usar a técnica de clusterização para criar agrupamentos de clientes seguindo aspectos de similaridade entre eles. A partir dos resultados desta análise, seria possível estudar a oferta de promoções levando-se em conta as questões de similaridade.

Referências

- AFONSO, M. *Aprendendo ETL*. 2015. ETL. Disponível em: <<http://aprendendoetl.com.br/index.php?id=61>>. Acesso em: 18 dez. 2015. Citado na página 16.
- CAMILO, C. O.; SILVA, J. C. da. *Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas*. 2009. Descoberta de conhecimento. Disponível em: <http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_001-09.pdf>. Acesso em: 04 jul. 2017. Citado na página 18.
- DAVILA, V. H. L. *Estatística Descritiva*. 2010. IME. Disponível em: <<http://www.ime.unicamp.br/~hlachos/estdescr1.pdf>>. Acesso em: 23 dez. 2015. Citado na página 11.
- DIAS, M. M. *UM MODELO DE FORMALIZAÇÃO DO PROCESSO DE DESENVOLVIMENTO DE SISTEMAS DE DESCOBERTA DE CONHECIMENTO EM BANCO DE DADOS*. 2001. Descoberta de conhecimento. Disponível em: <http://www.din.uem.br/~intersul/intersul_arquivos/documentos/tese_Madalena.pdf>. Acesso em: 04 jul. 2017. Citado na página 17.
- ELMASRI, R.; NAVATHE, S. *Fundamentals of Database Systems*. 6th. ed. USA: Addison-Wesley Publishing Company, 2010. ISBN 0136086209, 9780136086208. Citado na página 8.
- FIGUEIREDO, C. X. *Descoberta de Conhecimento em Banco de Dados: Um Estudo de Caso da Pesquisa Científica na Universidade Federal de Lavras*. 2004. Descoberta de conhecimento. Disponível em: <http://repositorio.ufla.br/bitstream/1/9570/3/ARTIGO_Descoberta_de_conhecimento_em_banco_de_dados_um_estudo_de_caso_da_pesquisa_ci.pdf>. Acesso em: 04 jul. 2017. Citado na página 13.
- FILHO, A. M. S. *A Era da Informação*. 2001. Era da informação. Disponível em: <http://www.espacoacademico.com.br/002/02col_mendes.htm>. Acesso em: 18 dez. 2015. Citado na página 7.
- GARCÍA, B. H. M. *QlikView 11 for Developers*. [S.l.]: Packt Publishing Ltd, 2012. ISBN 1482205491, 9781482205497. Citado na página 20.
- GRUPOALVO. *Estatística – Variáveis contínuas e discretas*. 2009. Estatística – Variáveis contínuas e discretas. Disponível em: <<https://grupoalvo.wordpress.com/2009/06/08/estatistica-variaveis-continuas-e-discretas/>>. Citado na página 27.
- KIMBALL, M. R. R. *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. [S.l.]: John Wiley and Sons, 2013. ISBN 1482205491, 9781482205497. Citado 2 vezes nas páginas 22 e 23.
- LEMONS, E. P.; STEINERAND, M. T. A.; NIEVOLA, J. C. *Análise de crédito bancário por meio de redes neurais e árvores de decisão: uma aplicação simples de data mining*. 2005. Indução de Regras e Árvores de Decisão. Disponível em: <<http://www.redalyc.org/html/2234/223417392002/>>. Acesso em: 07 jul. 2017. Citado na página 27.

MARIA, E. *Um estudo sobre as ferramentas OLAP*. 2007. OLAP. Disponível em: <<http://www.devmedia.com.br/um-estudo-sobre-as-ferramentas-olap/6691>>. Acesso em: 04 jul. 2017. Citado na página 19.

MATHEUS, R. F.; PARREIRAS, F. S. Inteligência empresarial versus business intelligence: abordagens complementares para o apoio à tomada de decisão no Brasil. In: *Congresso Anual da Sociedade Brasileira de Gestão do Conhecimento*. São Paulo, Brasil: [s.n.], 2004. p. 1–15. Disponível em: <http://www.fernando.parreiras.nom.br/publicacoes/ie_bi.pdf>. Citado na página 7.

MONARD, M. C.; BARANAUSKAS, J. A. *Indução de Regras e Árvores de Decisão*. 2003. Indução de Regras e Árvores de Decisão. Disponível em: <<http://dcm.ffclrp.usp.br/~augusto/publications/2003-sistemas-inteligentes-cap5.pdf>>. Acesso em: 07 jul. 2017. Citado 3 vezes nas páginas 4, 28 e 29.

ORACLE. *Oracle Help Center*. 2017. Descoberta de conhecimento. Disponível em: <<https://docs.oracle.com/en/>>. Acesso em: 04 jul. 2017. Citado na página 14.

PRASS, F. S. *What is data warehouse?* 2014. Data warehouse. Disponível em: <<http://fp2.com.br/blog/index.php/2012/um-visao-geral-sobre-fases-kdd/>>. Acesso em: 04 abr. 2016. Citado 2 vezes nas páginas 11 e 17.

RAPIDMINER. *RapidMiner Studio Manual*. [S.l.], 2014. Disponível em: <<https://docs.rapidminer.com/downloads/RapidMiner-v6-user-manual.pdf>>. Citado 3 vezes nas páginas 9, 18 e 29.

REDBOOKS, I. *IBM InfoSphere DataStage Data Flow and Job Design*. IBM Redbooks, 2008. (IBM redbooks). ISBN 9780738431116. Disponível em: <<https://books.google.com.br/books?id=anvDMgEACAAJ>>. Citado 2 vezes nas páginas 9 e 11.

Sá, T. et al. *Uma Análise Comparativa entre as Ferramentas OLAP como Apoio a Soluções de BI nas Empresas*. 2011. OLAP. Disponível em: <http://infobrasil.inf.br/userfiles/15-S1-1-97127-Uma%20An%C3%A1lise%20Comparativa____.pdf>. Acesso em: 04 jul. 2017. Citado na página 20.

TOFFOLO, T.; CARVALHO, M. A. *Divisão e Conquista*. 2017. Divisão e Conquista. Disponível em: <http://www.decom.ufop.br/toffolo/site_media/uploads/2011-1/bcc402/slides/08._divisao_e_conquista.pdf>. Citado na página 27.

WIKI. *Peter Drucker*. 2010. Peter Drucker. Disponível em: <https://en.wikipedia.org/wiki/Peter_Drucker>. Acesso em: 23 dez. 2015. Citado na página 7.

ZUBEN, F. J. V.; ATTUX, R. R. F. *Árvores de Decisão*. 2010. Árvores de Decisão. Disponível em: <ftp://vm1-dca.fee.unicamp.br/pub/docs/vonzuben/ia004_1s10/notas_de_aula/topico7_IA004_1s10.pdf>. Citado na página 28.