**UNIVERSIDADE DE CAXIAS DO SUL**

**CENTRO DE COMPUTAÇÃO E TECNOLOGIA DA INFORMAÇÃO**

**CURSO DE BACHAREL EM CIÊNCIA DA COMPUTAÇÃO**

**MARINA BELLENZIER**

# MINERAÇÃO DE DADOS NO PORTAL DE ATENDIMENTO E URA DA FOCCO.

Trabalho de Conclusão de Curso do Centro de Computação e Tecnologia – Curso de Bacharel em Ciência da Computação da Universidade de Caxias do Sul

Orientadora: Helena G. Ribeiro.

**Caxias do Sul, RS**
**2013**

# RESUMO

Este trabalho visa mostrar a importância da Descoberta de Conhecimento em Banco de Dados e como este processo pode ser utilizado para tomada de decisão por organizações através de um estudo de caso realizado na empresa Focco Sistemas de Gestão. Ele descreve os conceitos e etapas deste assunto, enfocando na etapa de mineração de dados onde são apresentadas as principais técnicas, métodos, algoritmos e algumas ferramentas que podem ser utilizadas para este tipo de processo. Além disso, no estudo de caso sobre o problema a ser explorado, são descritos todos os processos da empresa Focco Sistemas de Gestão e todos os passos realizados na mineração de dados sobre a base de dados desta empresa.

**Palavras-chaves:** Conhecimento, Decisão, Mineração.

# LISTA DE FIGURAS

# LISTA DE TABELAS

# SUMÁRIO

# 1. INTRODUÇÃO

O conhecimento tem sido reconhecido como um dos mais importantes recursos de uma organização, tornando possíveis ações inteligentes nos planos organizacionais. O processo de gestão do conhecimento abrange toda a forma de gerar, armazenar, distribuir e utilizar o conhecimento, tornando necessária a utilização da tecnologia da informação para facilitar o processo, devido ao grande aumento no volume de dados (ARAUJO e PEREIRA, 2011).

Ao longo do tempo, percebeu-se que a velocidade de coleta de informações era muito maior do que a velocidade de processamento ou análise das mesmas, o que gera um problema onde as organizações possuem uma grande quantidade de dados, porém essas informações de nada servem se não forem analisadas de forma correta e em tempo hábil.

Com o crescimento acelerado da tecnologia e das organizações na atualidade, torna-se necessária a aplicação de técnicas e ferramentas automáticas que agilizem o processo de extração de informações relevantes de grandes volumes de dados. Uma metodologia que tenta solucionar este problema é a descoberta de conhecimento em banco de dados, que é um processo que possibilita a descoberta eficiente do conhecimento sobre um grande conjunto de dados (PUHL, 2011).

A Mineração de Dados é uma técnica que faz parte de uma das etapas da descoberta de conhecimento em banco de dados. Ela é capaz de revelar, automaticamente, o conhecimento que está implícito em grandes quantidades de informações armazenadas nos bancos de dados de uma organização (FAYYAD, PIATETSKY-SHAPIRO e SMYTH, 1996).

Percebida a importância da mineração de dados o enfoque deste trabalho se dará em estudar as técnicas e métodos de mineração de dados para apoiar o processo de tomada de decisões nas equipes de Suporte e Manutenção da organização Focco Sistemas de Gestão.

A organização Focco Sistemas de Gestão possui um crescente aumento na sua carteira de clientes, e cada vez mais torna-se necessário ações rápidas e

eficientes no atendimento destes clientes. Atualmente existe uma base de dados consolidada que armazena todas as informações contendo as dificuldades e necessidades dos clientes, e é sobre este conjunto de dados que serão extraídas informações úteis para que a equipe estratégica da empresa possa utilizá-las como base para determinadas ações em relação ao trabalho da equipe e aperfeiçoamento dos processos e serviços da empresa.

A organização do texto está estruturada conforme descrito a seguir. No capítulo 2, está descrita a conceituação sobre a Descoberta do Conhecimento em banco de dados e relatada suas etapas. O capítulo 3 apresenta uma das etapas da Descoberta do Conhecimento, a Mineração de Dados, de forma mais aprofundada, relacionando as possíveis técnicas, métodos e algoritmos para a realização do processo. No capítulo 4 se encontra toda a compreensão do estudo de caso, informações sobre a empresa Focco Sistemas de Gestão e seus processos, descrição sobre a base de dados que será utilizada para extração dos dados e estudo sobre a ferramenta que será utilizada para o processo de Mineração de Dados, chamada de *Oracle Data Mining*. No capítulo 5 se encontra a descrição de todo o desenvolvimento da parte prática do projeto, a descoberta do conhecimento. Todos os passos são relatados e ilustrados para melhor compreensão. E por fim, no capítulo 6 estão descritas todas as considerações finais sobre o projeto realizado.

# 2. DESCOBERTA DO CONHECIMENTO EM BANCO DE DADOS

Com o aumento da informatização dos processos houve um grande crescimento no armazenamento de informações e nas bases de dados das organizações.

O método tradicional de transformar informações em conhecimento sempre exigiu análises manuais de especialistas de cada área. Com o crescente aumento do volume das informações esse tipo de análise começou a se tornar inviável, surgindo a necessidade de utilizar recursos computacionais para auxiliar as pessoas na busca por informações úteis em grandes conjuntos de dados (FAYYAD, PIATETSKY-SHAPIRO e SMYTH, 1996).

O OLAP (*On-Line Analytical Processing*) é uma classe de sistemas cuja tecnologia permite aos analistas de negócio e gestores a análise dinâmica e multidimensional de dados consolidados de uma organização. Caracteriza-se por fornecer subsídio para tomadas de decisão a partir de análises realizadas sobre bases de dados históricas, resultando em dados resumidos. Desta forma, ainda assim, existe a necessidade de buscar algo mais detalhado e desconhecido, algum padrão oculto que não seja evidente aos olhos dos usuários e analistas, e então as organizações necessitam realizar o que é chamado de descoberta de conhecimento (FAYYAD, PIATETSKY-SHAPIRO e SMYTH, 1996).

A Descoberta de Conhecimento em Banco de Dados (DCBD), ou *Knowlegde of Discovery in Databases* (KDD), é um processo contínuo e cíclico que possibilita a descoberta eficiente do conhecimento sobre um grande conjunto de dados (PUHL, 2011). É um processo não trivial que identifica em dados padrões que sejam válidos, novos e previamente desconhecidos visando melhorar o entendimento de algum problema ou um procedimento de tomada de decisão (FAYYAD, PIATETSKY-SHAPIRO e SMYTH, 1996).

Este processo é importante na medida em que as informações extraídas possibilitam a melhoria dos processos e a tomada de decisões dentro das organizações. Mas não é utilizado para extrair informações específicas que

poderiam ser buscadas através de uma linguagem de consulta usual como SQL, ele deve ser utilizado para buscar as informações úteis, que estão imperceptíveis nas bases de dados (ANDREAZZA, 2009).

O processo da descoberta do conhecimento consiste em diversas fases, que conforme (FAYYAD, PIATETSKY-SHAPIRO e SMYTH, 1996) são: compreensão do negócio, seleção dos dados, limpeza e pré-processamento dos dados, codificação ou transformação dos dados, mineração dos dados, interpretação e avaliação dos dados. A figura 2.1 apresenta essas fases em sequência.

A DCBC possui duas características relevantes no processo: é interativo e iterativo. Interativo pois o usuário pode controlar e intervir nas etapas do processo, e iterativo pois alguma fase pode ser repetida diversas vezes . Na sequência será apresentada uma visão geral de cada etapa deste processo.

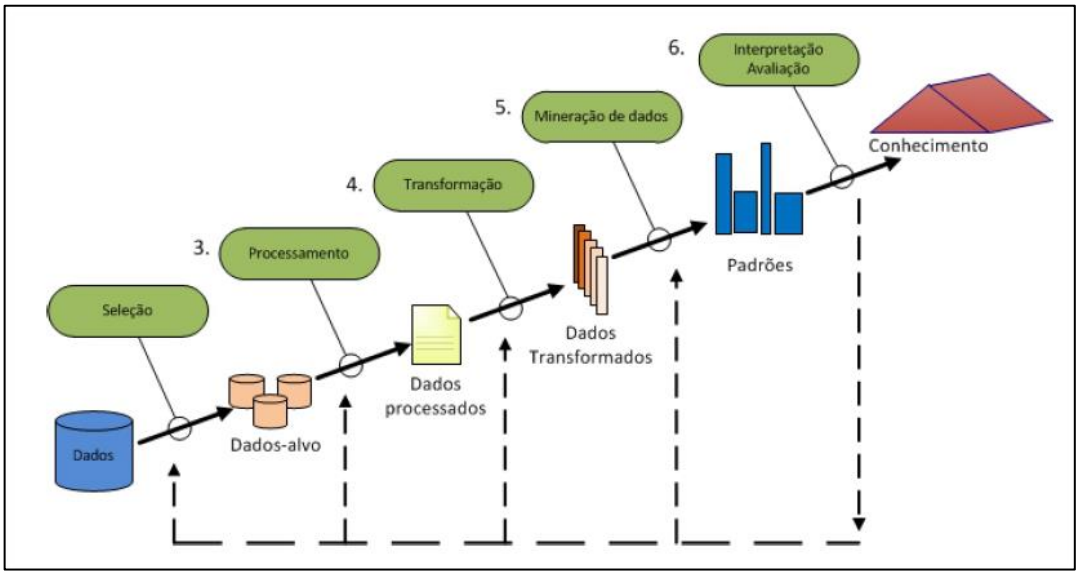



Figura 2.1 - Etapas do processo de KDD (FAYYAD et al., 1996)

## 2.1. COMPREENSÃO DO NEGÓCIO

A primeira etapa da descoberta de conhecimento de banco de dados (DCBD) consiste em entender todo o domínio da aplicação, os requisitos relevantes e os principais objetivos do processo. Esta fase resulta em uma definição do problema de mineração de dados.

## 2.2. SELEÇÃO DOS DADOS

A segunda etapa do processo é utilizada para escolher e criar o melhor conjunto de dados para se realizar a extração de padrões (PIMENTEL e OMAR, 2006). Esta definição da estrutura dos dados pode ser complexa pois os dados podem estar armazenados de diferentes formas, como por exemplo planilhas, XMLs, tabelas em banco de dados, etc, e podem se encontrar em diferentes formatos. "*Esse passo possui impacto significativo sobre a qualidade do resultado do processo*" (PUHL, 2011).

## 2.3. PROCESSAMENTO DOS DADOS

Esta etapa também conhecida como Limpeza dos Dados, é uma etapa muito importante do processo, pois tem por objetivo garantir a qualidade dos dados coletados a fim de melhorar a eficiência dos algoritmos da mineração de dados, porém pode ser bastante demorada.

Nesta fase são realizadas tarefas para eliminar dados inconsistentes e redundantes, recuperar dados incompletos, corrigir formatos inadequados e avaliar dados fora do padrão, como por exemplo, sexo = "F" e "M" ou sexo = "M" e "H".

A disponibilidade de dados de alta qualidade é um dos fatores mais importantes para adquirir sucesso na descoberta de conhecimento. Portanto, quanto pior for a qualidade dos dados informados no processo de DCBD, pior será a qualidade dos modelos de conhecimento gerados (GOLDSCHMIDT, 2005).

## 2.4. TRANSFORMAÇÃO DOS DADOS

Após os dados serem selecionados e limpos, eles necessitam ser armazenados e formatados adequadamente para que sejam aplicados os algoritmos de mineração (PUHL, 2011).

Esta etapa visa disponibilizar os dados de maneira usável e navegável. Com a transformação, o número efetivo de variáveis consideradas é reduzido, ou até mesmo atributos podem ser incluídos com a finalidade de conter informações agregadas contidas implicitamente nos dados (FAYYAD, PIATETSKY-SHAPIRO e SMYTH, 1996).

## 2.5. MINERAÇÃO DOS DADOS

A etapa de mineração de dados consiste em aplicar o algoritmo efetivamente sobre os dados coletados, e essa aplicação pode ser realizada de forma iterativa, ou seja, diversas vezes. É a etapa que utiliza mais recursos computacionais, pois é nesta fase que são realizadas as consultas no grande volume de dados procurando abstrair conhecimento (GOLDSCHMIDT, 2005).

O objetivo desta fase é com a aplicação de métodos e algoritmos identificar padrões e regras sobre os dados que foram coletados nas etapas anteriores.

No capítulo 3 serão descritos de forma aprofundada as técnicas e os métodos deste processo.

## 2.6. INTERPRETAÇÃO E AVALIAÇÃO DOS DADOS

Nesta última fase, os padrões obtidos através das técnicas de mineração de dados devem ser analisados e interpretados, e transformados em conhecimento para que seja alcançado o objetivo final do processo.

Caso o resultado não tenha sido satisfatório, pode-se voltar às etapas anteriores e realizar o processo novamente, até mesmo recomeçar trocando o conjunto de dados e os algoritmos utilizados.

Conforme (FAYYAD, PIATETSKY-SHAPIRO e SMYTH, 1996) pode ser escolhido um modo de visualização diferenciado para que as informações sejam exibidas de uma maneira mais amigável.

# 3. MINERAÇÃO DE DADOS

A mineração de dados é a etapa mais importante do processo de KDD e consiste na aplicação de análise de dados e algoritmos de descoberta que, sob aceitáveis limitações de eficiência computacional, produz uma enumeração específica de padrões (ou modelos) sobre os dados (FAYYAD U.;PIATETSKY-SHAPIRO, 1996).

A mineração de dados pode ser aplicada de duas formas: como um processo de verificação e como um processo de descoberta. Os resultados obtidos com a mineração de dados podem ser usados no gerenciamento de informação, no processamento de pedidos de informação, na tomada de decisão, no controle de processos e em muitas outras aplicações (DIAS, 2002).

Qualquer aplicação de mineração de dados é dependente do contexto em que ocorre. As novas informações descobertas, ou padrões, devem possuir algum grau de certeza para que sejam consideradas válidas, e precisam ser apresentadas de forma clara para que seja compreendida pelos usuários (PIMENTEL e OMAR, 2006).

Para realizar a mineração de dados são utilizados algoritmos baseados em técnicas e métodos que tem origem várias áreas do conhecimento, como por exemplo: a matemática, a estatística e a inteligência artificial (ANDREAZZA, 2009). É importante distinguir o que é uma técnica e o que é um método de mineração. A técnica consiste na especificação **do que** estamos querendo buscar nos dados, que tipo de regularidades ou categoria de padrões temos interesse em encontrar. Já o método de mineração consiste na especificação de procedimentos que nos garantam como descobrir os padrões que nos interessam.

No final deste capítulo serão apresentadas, na Tabela 3.2, as relações entre as técnicas, os métodos e os algoritmos detalhados a seguir.

## 3.1. TÉCNICAS DE MINERAÇÃO DE DADOS

Segundo (DIAS, 2002), não existe uma técnica que resolva todos os problemas de mineração de dados. Diferentes métodos são utilizados para diferentes propósitos, e cada método possui suas vantagens e desvantagens. As funcionalidades diferem entre cada técnica. A seguir serão relatadas algumas técnicas de mineração de dados, assim como suas características e comportamentos.

### 3.1.1. Associação

A Associação é uma técnica que tem como objetivo encontrar relacionamentos ou padrões em uma coleção de dados, ou seja, buscar as ocorrências simultâneas de determinados itens.

A relação entre as ocorrências dos itens são chamadas de “Regras de Associação”, onde estas regras representam um padrão entre os dados da aplicação. (ORACLE, 2012). Por exemplo, em dois conjuntos de atributos X e Y, a regra de associação tem a forma de X → Y, o que significa que os registros que possuem X tendem a consequentemente possui Y (PIMENTEL e OMAR, 2006). Estas regras são muito utilizadas na área de *marketing* para definir o perfil dos consumidores e quais produtos são frequentemente adquiridos juntos. Um exemplo bastante utilizado nesta técnica é o de um supermercado, onde são analisados todos os itens comprados e relacionados os que sempre se encontram juntos, por exemplo, cerveja e salgadinhos, neste caso então a estratégia dos supermercados é colocar estes produtos próximos nas bancadas.

De acordo com (WITTEN e FRANK, 1999) é necessário definir duas propriedades para a utilização desta técnica: suporte e confiança.

- Suporte: é o número mínimo de ocorrências nas quais o atributo aparece no conjunto de dados para que a associação seja considerada frequente. Este número pode ser representado em percentual.

- Confiança: é o número mínimo de vezes que uma associação deve aparecer para ser considerada válida.

A técnica de regra de associação implementa este tipo de tarefa através de algoritmos como o Apriori, DHP, ABS, Sampling entre outros (PIMENTEL e OMAR, 2006). É possível visualizar na Figura 3.1 a maneira que o método de associação funciona. Cada paralelepípedo representa um registro de dados, e o conjunto desses paralelepípedos representa a coleção de dados. A partir desta coleção, são identificadas mesmas ocorrências de dados, representadas por quadrados, círculos e triângulos que ocorrem em cada coleção de dados, se definem por “Regras de Associação”.

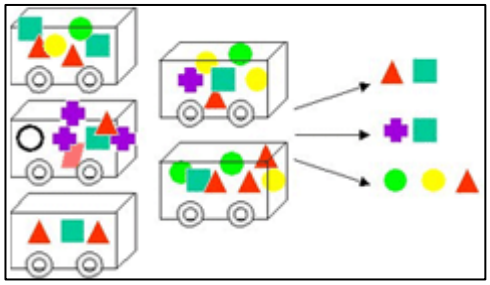


Figura 3.1: Representação da prática da técnica de Associação (ORACLE, 2012)

### 3.1.2. Classificação

A técnica da Classificação busca identificar propriedades comuns entre os registros a fim de associá-los em categorias ou classes já definidas. Estas classes podem ser binárias (por exemplo, 0 ou 1, sim ou não) ou multi-valores. O modelo analisa o conjunto de dados, também chamado de dados de treino, com cada registro já contendo a indicação à qual classe pertence, a fim de “aprender” como classificar um novo registro. (ORACLE, 2012) (ANACLETO, 2009).

Esta técnica pode precisar ser precedida por uma relevante análise, para tentar identificar os atributos que são significativamente importantes para a classificação. Tais atributos serão selecionados para a classificação e outros atributos, que são irrelevantes, podem então ser excluídos (HAN, KAMBER e PEI, 2012).

Um exemplo de método de classificação pode ser sobre o perfil dos colaboradores de uma empresa, utilizando categorias como: perfil técnico, perfil de negócio, perfil de gerência. O modelo poderá analisar os registros e classificar um novo colaborador a partir das suas características.

Diferentes métodos de classificação usam técnicas diferentes para criação dos modelos, e podem ser representados de várias formas. Os mais comuns são: regras de classificação, árvores de decisão e redes neurais (ORACLE, 2012).

A figura 3.2 apresenta o modo como a técnica de Classificação trabalha. Os triângulos e os círculos representam os dados e estes dados são classificados e separados em grupos, neste caso foram necessários apenas dois grupos para distinguir os dados, cada um com características diferentes.

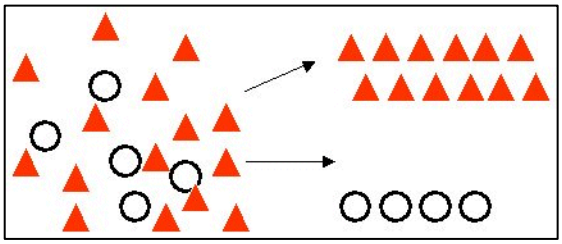


Figura 3.2: Representação da prática da técnica de Classificação (ORACLE, 2012)

### 3.1.3. Regressão

A técnica da Regressão, também conhecida como predição funcional, é semelhante a da Classificação, porém trabalha com valores contínuos e numéricos. Busca explicar uma ou mais variáveis contínuas de interesse em função de outras. A

partir de um modelo criado baseado no histórico dos dados, é possível utilizá-lo para realizar predições e calcular probabilidades (PUHL, 2011).

Os dados históricos para um projeto de regressão é tipicamente dividido em dois conjuntos de dados: um para construir o modelo e outro para realizar os testes em cima deste modelo (ORACLE, 2012).

Um exemplo é predizer o valor de um imóvel em função da sua idade, da sua localização, do seu padrão de construção e da sua área. Outros exemplos podem ser utilizados em análises de tendências, planejamento empresarial, marketing, previsão financeira, ou pesquisas médicas e ambientais (ORACLE, 2012).

A Figura 3.3 ilustra o funcionamento da técnica de Regressão, exemplificando que esta percorre os dados históricos, simbolizados com os triângulos vermelhos e círculos pretos, para melhor predizer os valores, baseando-se em outros valores de situações semelhantes. A forma da técnica percorrer os dados é simbolizada pela flecha em preto.

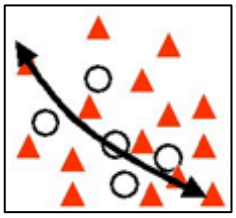


Figura 3.3: Representação da prática da técnica de Regressão (ORACLE, 2012)

### 3.1.4. Clusterização

Clusterização é a técnica que é utilizada para separar os dados ou o conjunto de dados em subconjuntos, chamados de *clusters*. O objetivo desta é maximizar a semelhança entre os elementos de dentro de cada *cluster* e minimizar a similaridade entre elementos de diferentes *clusters* (HAN, KAMBER e PEI, 2012).

No uso desta técnica não existe um conjunto de dados pré-definidos, ou seja, existe uma autonomia do algoritmo em identificar automaticamente os grupos de dados que serão criados (FAYYAD, PIATETSKY-SHAPIRO e SMYTH, 1996).

Um exemplo de utilização deste método é uma loja que deseja abrir uma filial e possui um cadastro de clientes e seus endereços. A partir da implementação deste método pode-se identificar, através dos subconjuntos, onde será o local mais estratégico para a abertura desta filial.

Para implementar esta tarefa podem ser utilizados algoritmos tais como: KMeans, K-Modes, K-Prototypes, K-Medoids, Kohonen e outros (ORACLE, 2012). Na figura 3.4 é apresentado a forma de como a técnica de Clusterização funciona: os registros são representados por círculos, triângulos, quadrados e cruzes coloridas, e a clusterização consegue identificar a semelhança entre os registros e separa-os em quantos grupos forem necessários.

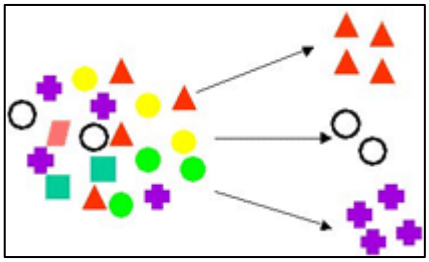


Figura 3.4: Representação da prática da técnica de Clusterização (ORACLE, 2012)

### 3.1.5. Sumarização

A técnica de sumarização consiste em buscar características comuns entre a coleção de dados e representá-los de forma compacta. O resultado deste método normalmente é utilizado em conjunto com outras técnicas, como Clusterização e Classificação. Um exemplo onde esta técnica pode ser utilizada é na contagem de ligações feitas de aparelhos celulares de cada estado do país, a partir disso será

possível realizar uma sumarização e até mesmo a criação de gráficos e estatísticas (PUHL, 2011) (ANDREAZZA, 2009).

### 3.1.6. Detecção de Desvios

A técnica de Detecção de Desvios tem por objetivo encontrar dados que não possuem um comportamento geral como o da maioria. Estes dados fora do padrão são chamados de *outliers* (exceções). Alguns métodos de mineração de dados, como Associação, descartam estes *outliers* como sendo dados indesejados. Porém, em algumas aplicações, tais como detecção de desvios, estas exceções podem ser mais interessantes do que eventos que ocorrem regularmente. Por exemplo, no uso fraudulento de cartões de crédito ao descobrir que certos clientes efetuaram compras de valor extremamente alto, fora de seu padrão habitual de gastos (ANDREAZZA, 2009) (PUHL, 2011).

## 3.2. MÉTODOS DE MINERAÇÃO DE DADOS

Existem diferentes métodos para utilizar as diferentes técnicas de Mineração de Dados. A seguir serão listados e explicados alguns dos principais métodos:

### 3.2.1. Árvore de Decisão

O método de Árvore de Decisão é baseado em probabilidades condicionais que geram regras, ou seja, a partir do conjunto de dados são geradas declarações condicionais que podem ser facilmente interpretadas pelos usuários e utilizadas para gerar um conjunto de registros (ORACLE, 2012).

O funcionamento de uma Árvore de Decisão é realizado em cima de uma estrutura de árvore ou fluxograma que utiliza uma hierarquia de "*if-then*" para classificar os dados. A forma de execução é feita da seguinte forma: dado um conjunto de dados cabe ao usuário escolher uma das variáveis como objeto de saída. A partir daí, o algoritmo encontra o fator mais importante relacionado com a variável de saída e determina-o como o primeiro ramo (chamado de raiz), os demais fatores subsequentemente são classificados como nós até que se chegue ao último nível, a folha.

Um valor alvo é previsto a partir de uma sequência de perguntas. Num determinado estado a pergunta depende da resposta que foi obtida no estado anterior. Desta forma, um problema complexo é decomposto em problemas mais simples. Recursivamente, a mesma estratégia é aplicada a cada subproblema (WITTEN e FRANK, 1999).

Por exemplo, na Tabela 3.1 temos um conjunto de dados de uma universidade referente a alunos que cursam diferentes cursos e diferentes matérias, e para identificar se estão aprovados ou não foi criado o modelo de Árvore de Decisão conforme a Figura 3.5, que demonstra que para que os alunos sejam aprovados é necessário que tenham nota acima de 6 e frequência maior ou igual que 75%; caso o aluno não atinja a nota 6 mas tenha frequência maior ou igual a 75% é possível realizar recuperação.

| **Aluno** | **Curso** | **Disciplina** | **Frequência** | **Nota** |
|---|---|---|---|---|
| Pedro | Administração de Empresas | Teoria Geral da Administração | 80% | 6,0 |
| Camila | Ciências Contábeis | Estatística Geral | 87% | 8,0 |
| Sandra | Educação Física | Anatomia Geral | 90% | 5,5 |
| Paulo | Matemática | Lógica Matemática | 45% | 7,0 |
| João | Sistemas de Informação | Sistemas Operacionais | 60% | 7,2 |

Tabela 3.1: Conjunto de dados de uma universidade.


Figura 3.5: Árvore de Decisão criada a partir dos conjuntos de dados da universidade

### 3.2.2. Redes Neurais

Redes Neurais é um método de mineração de dados que procura reproduzir de maneira simplificada as conexões do sistema biológico neural. As redes neurais foram originalmente projetadas por psicólogos e neurobiologistas que procuravam desenvolver um conceito de neurônio artificial análogo ao neurônio natural.

Toda rede neural é um conjunto de unidades do tipo *input* e *output*. Essas unidades representam um neurônio e são conectadas umas às outras, cada conexão possui um peso associado. A rede é composta por diversas camadas. Conforme mostrado na Figura 3.6, a primeira linha refere-se à camada de *Input*, esta corresponde aos atributos de classe. A última linha consiste na camada *Output* que corresponde aos possíveis valores dos atributos de classe. As demais linhas referenciam camadas intermediárias, em uma rede neural existe pelo menos uma camada intermediária. Casa nó de uma camada deve possuir ligação com todos os nós da camada seguinte.

A partir de um conjunto de treinamento, procura-se aprender padrões gerais que possam ser aplicados à classificação ou à predição de dados. A função básica

de cada neurônio é avaliar valores de entrada, calcular o total para valores de entrada combinados, comparar o total com um valor limiar e determinar o valor de saída.

A principal vantagem deste método é a sua capacidade de “aprender” a classificar os padrões através de exemplos ou tentativas e erros. Ele pode ser usado quando você pode ter pouco conhecimento das relações entre atributos e classes (HAN, KAMBER e PEI, 2012).

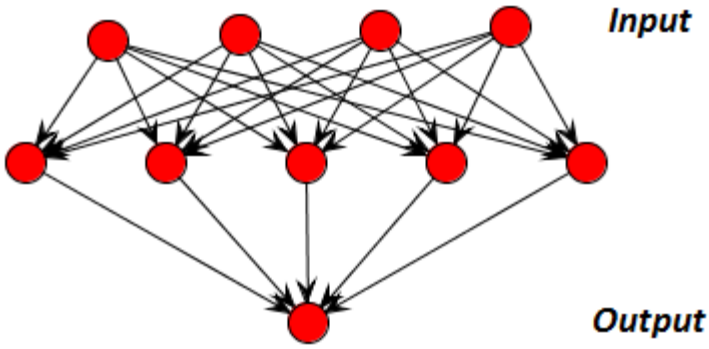



Figura 3.6: Arquitetura de uma rede neural.

## 3.3. ALGORITMOS DE MINERAÇÃO DE DADOS

De acordo com vários métodos de Mineração de Dados, existem diversos algoritmos baseados nestas técnicas. Cada um possui características e funcionamentos distintos e para realizar a escolha de qual utilizar tudo depende do objetivo e do tipo de tarefa que deseja-se realizar.

O processo de mineração pode utilizar dois tipos de aprendizagem: supervisionada e não supervisionada. Na sequencia serão descritos cada tipo de aprendizagem bem como suas características e exemplos de algoritmos que o utilizam.

### 3.3.1. Aprendizagem Supervisionada

Na Aprendizagem Supervisionada é necessária a análise de um especialista que compreenda os objetivos e o contexto do problema para identificar os atributos e variáveis que serão utilizados na execução dos algoritmos (HAN, KAMBER e PEI, 2012).

Nestes tipos de algoritmos, o objetivo de mineração é definido no início do processo de execução do algoritmo e os dados devem ser classificados de acordo com esta definição. A partir disso, a mineração dos dados é realizada para tentar encontrar os padrões e as relações entre os atributos independentes e os atributos dependentes (ORACLE, 2012).

A seguir será apresentado um dos principais algoritmos de Aprendizagem Supervisionada.

#### *3.3.1.1. Naive Bayes*

O algoritmo de classificação *Naive Bayes*, também conhecido como Classificador Bayesiano Ingênuo é baseado em probabilidades condicionais, ou seja, encontra a probabilidade de um evento ocorrer, dada a probabilidade de outro evento que já ocorreu. Ele utiliza o Teorema de *Bayes*, uma fórmula que calcula a probabilidade contando a frequência e combinações de valores nos dados históricos (ORACLE, 2012).

Este algoritmo é denominado ingênuo (*naive*) por assumir que os atributos são condicionalmente independentes, ou seja, a informação de um evento não é informativa sobre nenhum outro.

Naive Bayes cria um modelo padrão de acordo com os dados históricos e calcula a probabilidade dividindo a percentagem de ocorrências de pares por a percentagem de ocorrências únicas. Se estas percentagens são muito pequenas

para um dado preditor, provavelmente não vai contribuir para a eficácia do modelo. Ocorrências abaixo de um certo limite podem ser geralmente ignoradas.

Como exemplo de utilização deste algoritmo, podemos relacionar a um pronto atendimento que gostaria de avaliar e identificar a probabilidade dos pacientes terem possíveis doenças a partir das suas características e sintomas. Para cada paciente, o algoritmo responde a pergunta definindo a acurácia da informação, ou seja, classifica o paciente conforme as classes, utilizando para isso alguns dados relevantes do paciente e apresenta qual a probabilidade da resposta estar correta. Neste caso, uma possível resposta seria: homens com mais que 45 anos e com sintomas de necessidade frequente de urinar e dificuldades na micção possuem uma probabilidade de 70% de terem alguma doença na próstata.

### 3.3.2. Aprendizagem Não Supervisionada

Em algoritmos de Aprendizagem Não Supervisionada não existe um agente externo indicando a resposta desejada para os padrões de entrada e saída. Este tipo de aprendizado também é conhecido como aprendizado auto supervisionado ou de auto-organização por que não requer saída desejada e/ou não precisa usar supervisores para seu treinamento. Ou seja, não possuem exemplos e o número típico de classes é desconhecido. Neste caso, é preciso identificar como os objetos podem ser agrupados por classes, sempre com base em atributos dos mesmos. O sistema de reconhecimento deverá deduzir o número de classes e quais objetos pertencem a quais classes (HAN, KAMBER e PEI, 2012).

Tarefas sem supervisão, tais como agrupamento também são frequentemente utilizados para explorar e caracterizar o conjunto de dados antes de executar uma tarefa de aprendizagem supervisionada (CHAPMAN e HALL, 2009). A seguir serão exibidos dois principais algoritmos de Aprendizagem Não Supervisionada.

#### *3.3.2.1. Apriori*

O algoritmo Apriori é baseado em Regras de Associação, tem como objetivo encontrar um conjunto de itens que aparecem frequentemente no conjunto maior de dados. O suporte para um conjunto de itens é a razão entre o número de transações em que o conjunto de itens aparece para o número total de operações (CHAPMAN e HALL, 2009).

Neste algoritmo, a mineração de dados pode ser dividida em duas condições, a primeira é encontrar todos os conjuntos frequentes que satisfazem o mínimo do fator suporte e a segunda, é que a partir deste conjunto de frequências são criadas as regras de associação para satisfazerem o fator mínimo de confiança (ORACLE, 2012).

Embora o algoritmo realize diversas buscas sucessivas em toda uma base de dados e mantenha um ótimo desempenho em termos de tempo de processamento, este algoritmo não é recomendado para encontrar associações raras num grande volume de dados. É bastante utilizado para encontrar situações específicas, por exemplo, um supermercado pode utilizar o algoritmo Apriori para identificar quais as relações de compras dos clientes.

#### *3.3.2.2. K-Means*

O algoritmo *K-Means*, também chamado de K-Médias, faz parte da técnica de Clusterização e tem como objetivo fornecer uma classificação (*cluster*) de informações de acordo com os próprios dados. Esta classificação é baseada em análises e comparações entre os valores dos dados e desta maneira o algoritmo vai fornecer de forma automática estes grupos de classificações sem nenhuma supervisão humana, por isso este algoritmo pertence a aprendizagem não supervisionada (ORACLE, 2012).

O algoritmo *K-Means* é um algoritmo de agrupamento iterativo, simples para implementar e executar, relativamente rápido, fácil de se adaptar, e comum na prática. É historicamente um dos algoritmos mais importantes em mineração de dados.

Para a execução do algoritmo deve ser informado pelo usuário o parâmetro k, que indica o número de classes que será criado, daí o nome *K-means*. Ele divide o conjunto de dados em k *clusters* de tal forma que objetos dentro de um grupo, tendem a ser mais semelhantes uns aos outros, em relação a objetos pertencentes a diferentes grupos.

Para gerar as classes e classificar as ocorrências, o algoritmo faz uma comparação entre cada valor por meio de uma distância calculada. Geralmente utiliza-se a distância euclidiana para calcular o quão 'longe' uma ocorrência está da outra. Após o cálculo das distâncias o algoritmo calcula centróides para cada uma das classes. Conforme o algoritmo vai iterando, o valor de cada centróide é ajustado pela média dos valores de cada atributo de cada ocorrência que pertence a este centroide (CHAPMAN e HALL, 2009).

Na figura 3.7 são exibidos os passos realizados pelo algoritmo *K-Means*: na primeira figura podemos visualizar que o algoritmo começa a distinguir os grupos com características semelhantes, na segunda separa mais ainda os dados identificando os centroides e na última já é possível visualizar o resultado final com os grupos de clusterização bem formados e com centroides bem definidos.

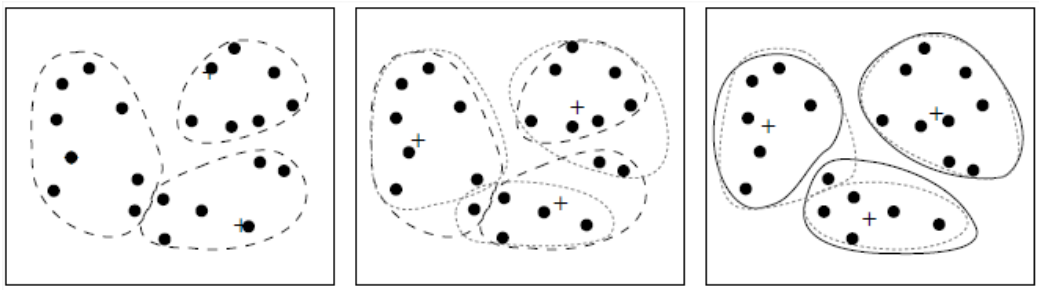


Figura 3.7: Exemplo do algoritmo *K-means* (HAN, KAMBER e PEI, 2012)

| Técnicas | Métodos | Algoritmos |
|---|---|---|
| Associação | | Apriori |
| Classificação | Árvore de Decisão<br>Redes Neurais | Naive Bayes |
| Regressão | Árvore de Decisão | |
| Clusterização | | K-means |
| Sumarização | | |
| Detecção de Desvios | | |

Tabela 3.2: Relação entre as técnicas, métodos e algoritmos apresentados.

## 3.4. FERRAMENTAS DE MINERAÇÃO DE DADOS

Diferentes tipos de ferramentas de mineração de dados estão disponíveis no mercado, cada um com suas próprias características. É necessário que se esteja ciente dos diferentes tipos de ferramentas de mineração de dados disponíveis para analisar a possibilidade de atender a necessidade do processo. Na sequência serão apresentadas algumas ferramentas de mineração de dados.

### 3.4.1. Weka

O software Weka é um software livre, para mineração de dados e foi desenvolvido na linguagem Java por um grupo de pesquisadores de uma universidade na Nova Zelândia. Ao longo do tempo se consolidou como o mais

utilizado em ambiente acadêmico para técnicas de mineração de dados (WITTEN e FRANK, 1999).

O Weka aplica os métodos de aprendizagem sobre a base de dados, extrai e analisa as informações dos padrões. Seu ponto forte é o método de classificação, mas também é capaz de minerar regras de associação e *clustering*. Os dados podem ser visualizados em forma de histogramas, e a apresentação de resultados através de árvores de decisão ou regras, entre outros e para isso utiliza uma interface chamada Weka Explorer (Waikato, 2012).

O software suporta arquivos no formato de ARFF que é o formato padrão do software, e também os formatos CSV e C45. Por tratar-se de um software livre, com código aberto, novos algoritmos podem ser desenvolvidos e integrados a ele.

### 3.4.2. Microsoft Analysis Services

Trata-se de um grupo de serviços OLAP (On Line Analytical Processing) e de mineração de dados disponíveis no Microsoft SQL Server. Ele permite projetar, criar e visualizar modelos de mineração de dados que são construídos. É integrado à plataforma .NET, da Microsoft, e tem implementado diversos algoritmos de mineração de dados. Além disso, define uma linguagem de consulta específica para mineração de dados dentro do SQL Server, a DMX (Data Mining Extensions to SQL language) (JACOBSON, MISNER e CONSULTING, 2005).

Microsoft Analysis Services fornece vários tipos de algoritmos para serem usados nas soluções de mineração de dados, por exemplo: algoritmos de classificação, algoritmos de regressão, algoritmos de associação, entre outros.

### 3.4.3. IBM Intelligent Miner

O banco de dados da IBM, chamado de DB2, é um conjunto que produtos que combinam a administração de dados, com uma poderosa infraestrutura de inteligência corporativa, pois permite a construção de *data warehouses* e a utilização do componente *Intelligent Miner*, cuja o site da ferramenta (IBM, 2012) a classifica com uma poderosa ferramenta para análise de dados integrada. Nesta ferramenta para mineração dos dados, são suportadas as tradicionais técnicas de mineração de dados: análise de agrupamentos, análise de afinidades, classificação, estimativa e previsão. Além disso, ótimos componentes de apresentação estão disponíveis para possibilitar uma análise visual dos resultados (IBM, 2012).

### 3.4.4. Oracle Data Mining

O Oracle Data Mining é uma ferramenta incluída no Banco de Dados Oracle para execução de técnicas de mineração nos dados. É um poderoso *software* que disponibiliza técnicas para encontrar padrões, identificar atributos chave e associações escondidas ou valores anormais num conjunto de dados e que se encontram no *Oracle Database*. Por possuir total integração com o banco de dados, a ferramenta maximiza o desempenho e disponibiliza um eficiente aproveitamento dos recursos do sistema eliminando a necessidade da extração dos dados.

No estudo de caso deste trabalho será utilizada esta ferramenta, pois além de possuir funcionalidades robustas possui total integração com o banco de dados Oracle, o que faz com que o desempenho seja muito melhor, e está disponível para a realização deste trabalho. Em capítulos posteriores será apresentado de forma mais aprofundada as características e funcionalidades desta ferramenta.

# 4. ESTUDO DE CASO

Hoje em dia, pode-se perceber que novas empresas surgem a todo o momento de forma acelerada. No contexto atual, as empresas são obrigadas a criar estratégias para se diferenciar de sua concorrência e se destacarem no mercado.

Atualmente, um fator primordial para as empresas garantirem seu sucesso é a conquista da satisfação de seus clientes. Muitas empresas já perceberam este novo cenário, em que o foco principal da mesma se dá primeiramente no cliente e depois no produto ou serviço. Para conquistar essa satisfação é necessário conhecer o cliente e saber quais são as suas necessidades.

Para isso é necessário que todos os dados e informações do cliente e da empresa sofram uma profunda análise e a partir disso sejam tomadas decisões estratégicas pelos gestores. Atualmente existem empresas que constroem todo o seu negócio com base na capacidade de analisar estas informações.

Neste capítulo serão apresentadas as definições que servirão como base para a compreensão do desenvolvimento do estudo de caso e do processo de mineração de dados.

## 4.1. COMPREENSÃO PARA O ESTUDO DE CASO

Este trabalho foi realizado sobre a base dados gerada pelo "Portal de Atendimento da Focco" e a URA (Unidade de Resposta Audível) utilizada pela empresa. Para um melhor entendimento, nesta seção serão descritas as informações referentes à empresa, ao portal, à URA e aos processos das áreas relacionadas.

### 4.1.1. Informações da Empresa

A Focco Sistemas de Gestão foi fundada no dia primeiro de novembro de 1989 pelos sócios os Srs. Evandro Bállico e Joel Caetano Polesello. Iniciaram os trabalhos com o objetivo de desenvolverem sistemas de informação através do produto MFoccus, desenvolvido sobre a base tecnológica Mumps, onde controlava Contabilidade, Folhas de Pagamentos e Livros Fiscais.

Com a necessidade do mercado e o avanço das tecnologias em 1993 surgiu a necessidade de criar um novo sistema, este para realizar a gestão e voltado totalmente para o segmento industrial. O novo *software* era chamado de ZFoccus, feito na linguagem Zim que ao longo do tempo foi adquirindo mais funcionalidades e se incorporando.

Em 1998 iniciou um projeto que se deu devido ao surgimento de bancos de dados relacionais, e a Focco começava a criação do *software* Focco3i em cima da tecnologia *Oracle*. O Focco3i foi implantado no primeiro cliente em 1998 e hoje este é o principal produto da Focco Sistemas de Gestão, comercializado com o nome de FoccoERP pois trata-se de um ERP para indústrias e comércios.

Atualmente a Focco possui mais de 200 clientes, dos segmentos Moveleiro, Estofados, Colchões, Encarroçadoras, Implementos rodoviários, Máquinas alimentícias, Guindastes, Peças automobilísticas, Moinhos e Bebidas, com um total de 10 mil usuários conectados ao FoccoERP. Conta com uma estrutura de 180 funcionários e 3 filiais: Arapongas no Paraná, São José em Santa Catarina e São Leopoldo no Rio Grande do Sul. Além do produto FoccoERP, possui mais o FoccoLOJAS que é um sistema de gestão criado para  lojas de móveis planejados.

### 4.1.2. Portal de Atendimento da Focco

O "Portal de Atendimento Focco" é um canal de relacionamento disponibilizado pela empresa Focco Sistemas de Gestão para que os clientes possam incluir solicitações de dúvidas e problemas nos produtos, necessidades de customizações ou melhorias nos produtos adquiridos, solicitação de consultoria e serviços, e necessidade de alterações legais. Na figura 4.1 pode-se visualizar a tela inicial do portal.

O portal foi escrito na linguagem Genexus, e utiliza o banco de dados Oracle como repositório de dados. O Genexus é um gerador de código totalmente procedural, possibilitando pouquíssima reutilização de código. O grande ganho dele é a produtividade, isto é, a rapidez no desenvolvimento das soluções. Está estruturado basicamente em uma camada, pois ele é totalmente orientado ao desenvolvimento da interface com o usuário. O Genexus é uma plataforma proprietária, desenvolvida pela Artech, uma empresa uruguaia sediada em Montevidéu.

Inicialmente, o portal chamava-se WP, e havia sido desenvolvido por uma empresa terceira, e com a necessidade de muitas customizações a Focco adquiriu o produto, e atualmente pode realizar qualquer alteração.


Figura 4.1: Tela inicial do Portal de Atendimento da Focco.

#### *4.1.2.1. Solicitações*

As solicitações são protocolos abertos pelos clientes através do “Portal de Atendimento da Focco” onde ficam registrados todos os dados e informações desde a sua abertura até a sua finalização. Existe um histórico para cada solicitação onde são informadas todas as ações que são feitas: em caso de troca de área, troca de responsável, troca de status, troca de priorização, etc.

As solicitações abertas pelo cliente podem estar em áreas diferentes da Focco: Suporte, Manutenção, Consultoria, Infraestrutura, Customização, Desenvolvimento. E dentro de cada área pode estar com um responsável diferente, que são os atendentes destas solicitações, ou seja, as pessoas responsáveis pelo atendimento, solução e retorno ao cliente.

Cada solicitação possui um status, que pode ser alterado conforme as ações sobre esta solicitação forem modificadas. Segue a listagem de possíveis status:

- Incluída: solicitações que forem incluídas por algum usuário no Portal de Atendimento da Focco.

- Em Atendimento: solicitações que possuem algum responsável atendendo a mesma.

- Devolvida: solicitações que forem devolvidas ao usuário que a incluiu para que sejam informados mais dados, ou solicitados testes.

- Entregue: solicitações que forem entregues com a solução para o usuário que inclui a mesma.

- Fechada: solicitações que forem encerradas pelo usuário que incluiu a mesma.

- Reincluída: solicitações que foram entregues com uma solução, porém o usuário que a incluiu não concorda com o retorno.

Existe uma priorização que deve ser realizada pelos atendentes da Focco quando refere-se às solicitações do Portal de Atendimento. Esta priorização indica qual solicitação deve ser atendida primeiramente. Isto é feito através de uma pontuação calculada pelo portal onde os atendentes respondem algumas perguntas baseadas em situações que podem estar ocorrendo. As solicitações são classificadas da seguinte forma:

- Operações críticas paradas: Problemas que impedem o fluxo principal de trabalho do cliente. A pontuação da priorização será maior que 5 mil e a urgência será “Alta”.

- Operações críticas paradas com situação de contorno: Problemas de operação que comprometem o fluxo de trabalho do cliente, mas possuem operações para contorno. A pontuação da priorização fica entre 1 mil e 5 mil e a urgência pode ser “Alta” ou “Baixa”, dependendo de cada situação.

- Falha no sistema com baixo impacto na operação: Problemas que não comprometem o fluxo de trabalho do cliente e não possuem grande impacto nas operações. A priorização será de 0 a 1 mil e a urgência sempre será “Baixa”.

Esta priorização deve ser seguida por todas as áreas da Focco. O cliente que possui mais urgência sempre deve ser atendido antes. Existem outros fatores que influenciam nesta pontuação, como por exemplo: tempo de dias em que a solicitação foi aberta, curva ABC de clientes, etc. A curva ABC de clientes é uma listagem montada de acordo com o valor que a Focco Sistemas de Gestão fatura sobre cada cliente sobre o total do faturamento.

### 4.1.3. URA

A Focco Sistemas de Gestão utiliza uma URA (Unidade de Resposta Audível), fornecida pela empresa Intelbras, para automatizar e melhorar o atendimento telefônico com os clientes. A URA nada mais é que uma central telefônica automatizada que controla todas ligações recebidas e discadas, cria gravações de espera, separa os ramais em canais para que agilize o atendimento, controla as filas de espera através de “estacionamentos” de forma organizada e armazena informações de cada ligação para futuras estatísticas. Na central telefônica da Focco foram criados diferentes canais para que o cliente selecione a opção com quem deseja falar, que hoje são: 1 para Suporte, 2 para Consultoria *On-line*, 3 para Vendas e 9 para a Telefonista. Por exemplo, o cliente que entra em contato com a equipe de Suporte, simplesmente pressiona o número indicado pela URA e automaticamente um atendente dará o auxílio necessário. Esta central armazena muitas informações como: telefone de origem e destino, data e hora da ligação, duração da ligação e do tempo do toque, etc.

### 4.1.4. Suporte

A equipe de Suporte da Focco Sistemas de Gestão é preparada para realizar o primeiro atendimento a clientes que necessitem de auxílio para dúvidas ou problemas encontrados no *software* FOCCOERP. A equipe conta com mais de quinze analistas de negócio especialistas em diversas áreas para realizar o atendimento completo do produto.

O Suporte também tem uma importante missão que é a de priorizar as solicitações e determinar qual cliente e qual situação necessita de atendimento mais rápido. Isso é feito de acordo com a experiência do atendente que consegue distinguir qual das situações é mais impactante para o processo dos clientes da Focco e quais não são. Esta priorização não é divulgada aos clientes com o intuito de evitar conflitos maiores, porém ela deve ser seguida rigorosamente por todas as equipes internas da Focco.

O fluxo da equipe começa quando o cliente realiza a abertura de uma solicitação, esta solicitação entra diretamente para os atendentes do Suporte identificarem e classificarem tal solicitação de acordo com a priorização necessária. Após todas as solicitações sofrerem priorizações os atendentes devem iniciar os atendimentos de acordo com a ordenação das solicitações: da maior prioridade para a menor prioridade.

O atendente deve identificar a situação relatada pelo cliente e concluir se a mesma refere-se a uma dúvida de processo, dúvida de funcionalidades, erro de cadastro ou erro de parametrização, e entrar em contato com o responsável pelo cliente que incluiu a mesma, esclarecendo a questão e encerrando a solicitação.

Caso seja identificado que o atendimento se refere a um erro do sistema, falha na análise ou inconsistência gerada por algum “*bug*” do produto, a solicitação deve ser encaminhada automaticamente para equipe de Manutenção, responsável pela correção e alteração dos códigos-fontes dos produtos, com a análise e simulação completa do erro encontrado.

Na figura 4.2 pode-se visualizar o fluxo da equipe de Suporte relatado no parágrafo anterior.

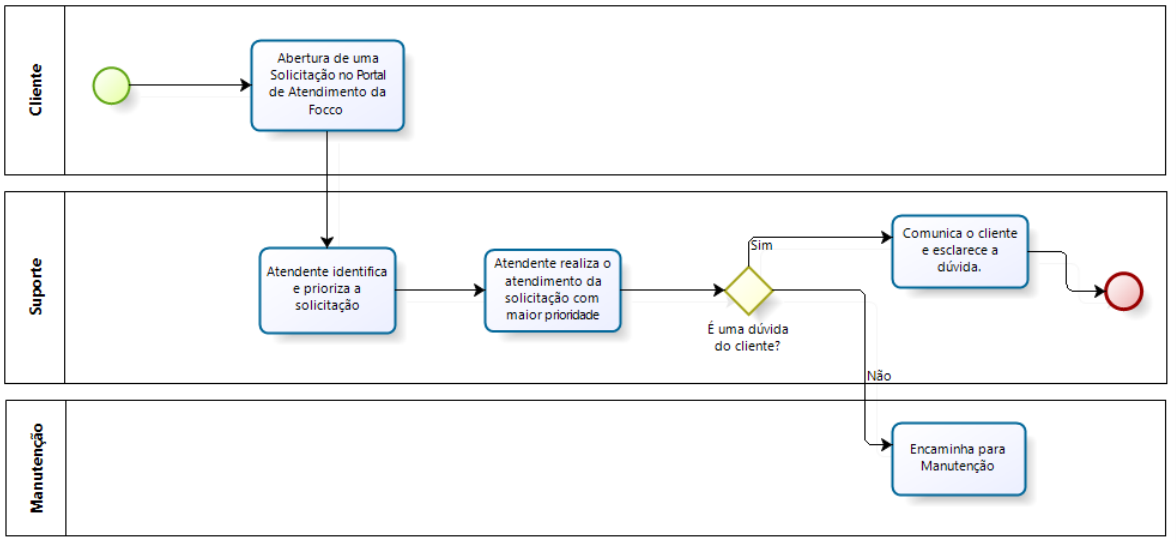



Figura 4.2: Fluxo de atendimento da equipe de Suporte da Focco Sistemas de Gestão.

### 4.1.5. Manutenção

A Manutenção é a equipe na Focco Sistemas de Gestão capacitada para realizar manutenções no produto FOCCOERP, solucionando erros e *bugs* relatados pelos clientes através das solicitações. Esta equipe é constituída por pouco menos de dez pessoas entre analistas de sistemas e desenvolvedores.

Esta equipe tem o importante papel de alterar os códigos-fontes do produto com muita responsabilidade sobre qualquer futuro problema que possa ocorrer em diferentes situações e em diferentes clientes. É necessário um grande conhecimento sobre o módulo e os programas a fim de garantir a integridade das correções.

O fluxo da Manutenção inicia quando a equipe de Suporte identifica que existe um erro no *software* e encaminha o chamado para equipe de Manutenção corrigir o problema. A solicitação já chega priorizada de acordo com o processo do

Suporte, e os desenvolvedores e analistas devem seguir a sequência de priorizações para realizar o atendimento das solicitações.

Caso os analistas identifiquem que a situação que se encontra para ser corrigida não se trata de um erro de sistema, mas é um problema de cadastros incorretos ou falta de parametrizações que o Suporte não conseguiu identificar, a solicitação é Devolvida para equipe de Suporte para que seja reavaliada.

As demais solicitações devem ser corrigidas, alterando o código-fonte dos programas nas bases do cliente e solicitando os devidos testes ao responsável do cliente que incluiu a solicitação. Após a confirmação de solução do erro, são replicadas tais correções nas bases internas da Focco para que sejam disponibilizadas em versões futuras do *software*; e então a solicitação é encerrada.

Na figura 4.3, visualiza-se o fluxo completo do atendimento que é realizado pela equipe de Manutenção.

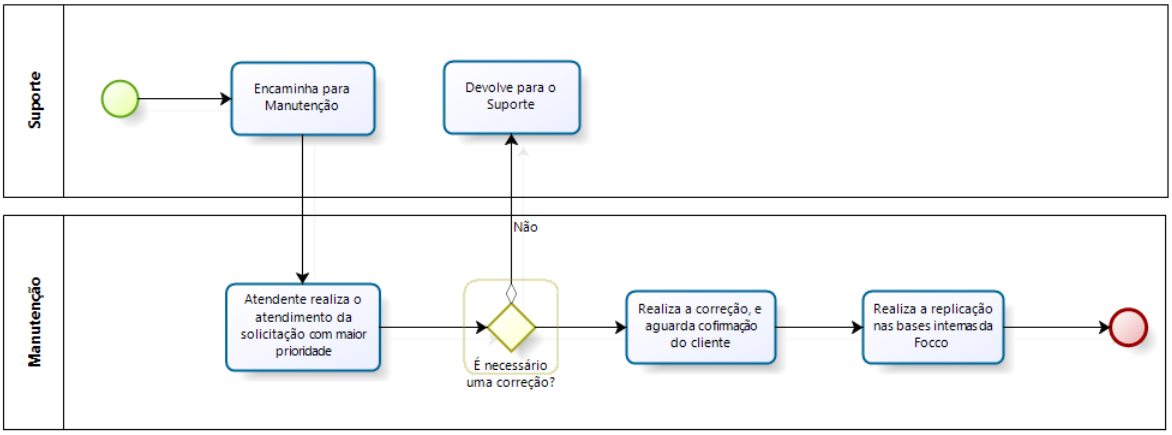



Figura 4.3: Fluxo de atendimento da equipe de Manutenção da Focco Sistemas de Gestão.

## 4.2. CENÁRIO PARA O ESTUDO DE CASO

Após realizar um levantamento com a equipe estratégica da Focco Sistemas de Gestão, verificou-se diversas situações que poderiam ser exploradas através da

mineração de dados para a obtenção de algumas informações relevantes para a tomada de decisão. A seguir estão descritas:

- Relação entre a quantidade de solicitações que os clientes abrem e seu segmento.

- Relação entre as datas de abertura das solicitações e as características dos clientes.

- Relação entre o número de solicitações abertas e o número de usuários que utilizam o produto simultaneamente nos cliente.

- Relação entre as ligações que entram pela URA, solicitações de dúvidas do Portal de Atendimento e número de usuários dos clientes.

- Relação entre os módulos para os quais são abertos os chamados e os segmentos dos clientes.

- Relação entre o valor pago pelos clientes e o esforço gasto para estes.

Porém no estudo de caso deste trabalho foi explorada apenas uma das relatadas acima, a seguinte situação: Relação entre as ligações que entram pela URA, os chamados do Portal de Atendimento e número de usuários dos clientes. Esta foi escolhida devido a necessidade de tomar ações mais imediatas, e segundo o gestor da área esta relação seria mais proveitosa.

### 4.2.1. Pesquisa de Estudos de Caso

Foram pesquisados alguns estudos de caso que utilizam de técnicas de mineração de dados para identificar relações de informações e mapear perfil de usuários com o objetivo de auxiliar a decisão de escolher qual técnica e qual algoritmo utilizar. Abaixo uma breve descrição da problemática desses estudos e as técnicas aplicadas.

#### *4.2.1.1. Utilização da Mineração de Dados no "Portal de Atendimento Focco"*

O estudo de caso (ANDREAZZA, 2009) teve como objetivo identificar o custo que os clientes de curva ABC da empresa de software Focco Sistemas de Gestão representam, avaliando o valor pago e o esforço realizado pela empresa para o atendimento. Foi criado um *Data Warehouse* e utilizada a ferramenta da Oracle para realizar a mineração de dados, e as técnicas de Classificação e de Clusterização. Segundo o autor, estas técnicas foram escolhidas pois melhor se adaptavam ao cenário proposto do estudo de caso.

#### *4.2.1.2. Estudo de Perfil de Usuários de Redes Sociais através da Mineração de Dados*

No trabalho de conclusão de curso (PUHL, 2011) a autora descreve um estudo de caso que identifica o perfil dos usuários de redes sociais para auxiliar no marketing da empresa de software Procad. Foram extraídos os dados da rede social *Facebook* e incluídos em um arquivo texto onde o software *Weka* realizou a mineração dos dados. As técnicas utilizadas foram: Classificação, Clusterização e Regras de Associação. Conforme descrito no artigo, tais técnicas foram escolhidas pois possuem melhor funcionalidades para identificar perfil de usuários.

#### *4.2.1.3. Aplicação de Técnicas de Data Mining em Logs de Servidores Web*

Na dissertação (CHIARA, 2003) o objetivo foi identificar a relação entre as interações entre um navegador e o servidor *WEB* através de *logs*. Para o estudo de caso desta dissertação foi alimentado um *Data Warehouse* através de uma fonte de

dados que consiste da sequência de cliques que um usuário efetua num site, ou seja, o registro de *log* produzidos pelo servidor Web toda vez que uma requisição HTTP é completada. Foi utilizada a técnica de Clusterização para minerar estes dados, com a conclusão que após o agrupamento e o discernimento dos dados foi possível obter uma melhor compreensão sobre os exemplos que pertencem a cada cluster.

#### *4.2.1.4. Evasão no Ensino Superior*

O estudo de caso da dissertação (SOUZA, 2008) trata o problema de evasão do ensino superior dos jovens brasileiros. A autora separa em grupos os tipos de jovens que se preparam para entrar no ensino superior e quais as características semelhantes entre si, e através da técnica de Classificação busca regras que podem perceber em qual grupo o jovem se classifica e identificar as dificuldades que este irá adquirir para iniciar o ensino superior.

De acordo com os artigos lidos e as técnicas que foram utilizadas, neste estudo de caso serão trabalhadas as seguintes técnicas no processo de mineração de dados: Clusterização e Classificação, e os algoritmos K-means e Naive Bayes.

## 4.3. PERFIL DE USUÁRIO

Conhecer o perfil dos usuários e dos clientes traz uma série de benefícios para a instituição, o principal deles é a capacidade de melhorar a qualidade de seus serviços prestados. Conhecendo o público alvo é possível montar uma melhor estratégia de *marketing* e com isto obter resultados mais significativos com a venda de produtos e/ou serviços.

Existem vários estudos que utilizam técnicas de mineração de dados para a identificação de perfil de usuário em diversas áreas, com o objetivo de estudar e compreender o comportamento do usuário. Pode-se citar dois estudos que foram apresentados neste trabalho que apresentam exatamente isso, “Estudo de Perfil de Usuários de Redes Sociais através da Mineração de Dados” (PUHL, 2011) e “Evasão no Ensino Superior” (SOUZA, 2008).

Neste estudo de caso iremos tratar o perfil do usuário como perfil do cliente, onde cada cliente será distinguido pelas suas características extraídas através das informações armazenadas pelo Portal de Atendimento da Focco e pela URA.

Os clientes serão classificados a partir de três grupos:

- Iniciantes: clientes que possuem um nível baixo de conhecimento sobre os produtos da Focco. Pode ser mensurado através do percentual de solicitações abertas por este cliente que são atendidas como dúvidas dos usuários. Nestas situações serão evidenciadas necessidades de oferecer um serviço de treinamento aos usuários do sistema.

- Intermediários: clientes que possuem um bom nível de conhecimento sobre os produtos da Focco, onde existem tanto solicitações de dúvidas sobre o funcionamento dos produtos como solicitações de correções.

- Experientes: clientes que possuem um ótimo nível de conhecimento sobre os produtos da Focco, e incluem um percentual muito baixo de solicitações de dúvidas. Clientes experientes não necessitam de treinamentos sobre o produto adquirido, pois já possuem uma maturidade na utilização dos sistemas.

## 4.4. CARACTERÍSTICAS DA BASE DE DADOS

Todos os dados do Portal de Atendimento da Focco estão armazenados em um Sistema Gerenciador de Banco de Dados *Oracle*. Nesta estrutura pode-se encontrar todas as informações sobre todas as solicitações, bem como seus

históricos, informações de clientes, e tudo que possa envolver o Portal de Atendimento.

Na figura 4.4 pode-se visualizar uma parte do modelo conceitual desta base de dados. Somente foram exibidas as tabelas que serão utilizadas neste estudo de caso.

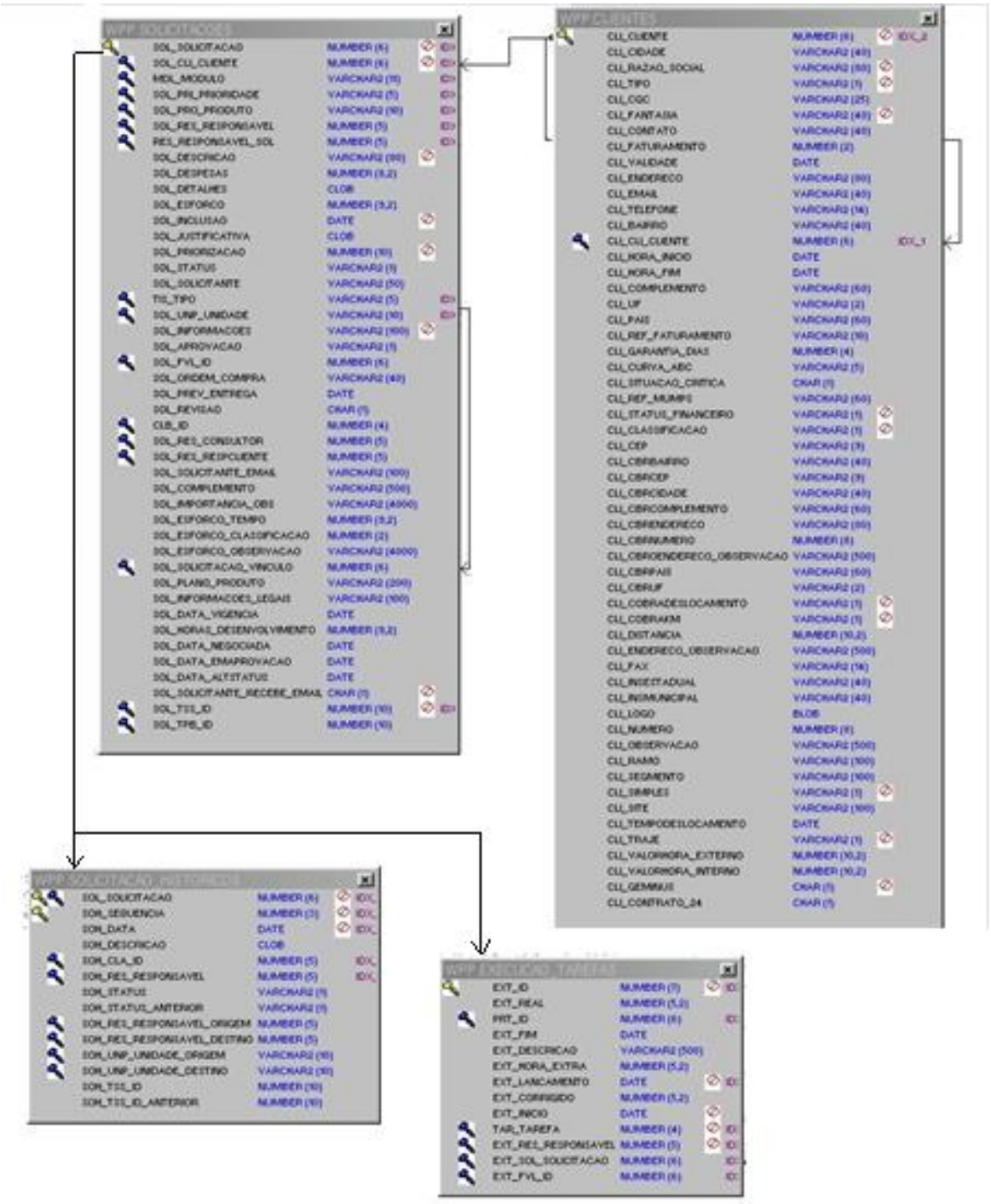


Figura 4.4: Parte do Modelo Relacional da Base de Dados do Portal de Atendimento da Focco

A tabela CLIENTES é utilizada para armazenar os principais dados do cliente que poderá acessar o Portal de Atendimento da Focco. Atualmente esta tabela possui cerca de 360 registros.

A tabela SOLICITACOES armazena todas as informações referente a solicitação que o cliente abriu, e ao longo do seu atendimento alguns campos como status, responsável, unidade, etc, podem mudar. Esta, atualmente, possui mais de 116 mil registros.

A tabela SOLICITACAO_HISTORICO serve para armazenar todos os históricos da solicitação, existirá um registro para cada ação tomada sobre a solicitação, por exemplo, para caso de troca de área, troca de responsável, troca de status, troca de unidade, troca de priorização, etc. Atualmente esta tabela possui quase 7 milhões de registros.

A tabela EXECUCAO_TAREFAS é utilizada para armazenar informações de tempos utilizados para os atendimentos. Para cada solicitação o atendente deve apontar as horas e os minutos utilizados para realização do atendimento. Esta tabela, atualmente, possui quase 520 mil registros.

Os dados da URA podem ser visualizados por um software específico da central telefônica, e extraídos em um arquivo texto. Neste arquivo texto as informações são separadas em colunas.

- 1ª coluna: informa se a ligação foi originada pela Focco (O), se foi recebida externamente (R) ou se foi interna (I);

- 2ª coluna: informa o número de origem da ligação;

- 3ª coluna: informa o número de destino da ligação;

- 4ª coluna: informa um número identificador da chamada para possíveis rastreamentos;

- 5ª coluna: informa a data e a hora que ocorreu a ligação;

- 6ª coluna: informa a duração das ligações atendidas;

- 7ª coluna: informa a duração dos *rings*, ou seja, o tempo que ficou tocando sem ninguém atender;

- 8ª coluna: informa se a ligação foi atendida (ATD), não atendida (NAT) ou IDT.

Na figura 4.5 podemos visualizar um exemplo deste arquivo texto com todas as colunas descritas acima.

Para realizar a seleção e o pré-processamento dos dados para este estudo de caso será necessário utilizar um *Data Warehouse*.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| I | 9040 | 9044 | | 03/08/2012 10:37 | 00:00:31 | 00:00:00 | ATD |
| O | 9067 | 5433586700 | 8931 | 03/08/2012 10:32 | 00:05:02 | 00:00:10 | ATD |
| I | 9040 | 9016 | | 03/08/2012 10:37 | 00:00:18 | 00:00:00 | ATD |
| O | 9063 | 5435111111 | 8932 | 03/08/2012 10:37 | 00:01:00 | 00:00:29 | ATD |
| I | 9114 | 9028 | | 03/08/2012 10:35 | 00:04:18 | 00:00:00 | ATD |
| R | 1734261234 | DISA | 8931 | 03/08/2012 10:40 | 00:00:00 | 00:00:00 | IDT |
| R | 1734261234 | Grupo1 | 8931 | 03/08/2012 10:40 | 00:00:00 | 00:00:00 | IDT |
| R | 1734261234 | DISA | 8931 | 03/08/2012 10:40 | 00:00:10 | 00:00:00 | ATD |
| R | 1734261234 | 9064 | 8931 | 03/08/2012 10:40 | 00:00:00 | 00:00:00 | IDT |
| I | 9025 | 9064 | | 03/08/2012 10:40 | 00:00:00 | 00:00:03 | NAT |
| R | 1734261234 | 9018 | 8931 | 03/08/2012 10:41 | 00:00:00 | 00:00:00 | IDT |
| R | 1734261234 | 9025 | 8931 | 03/08/2012 10:40 | 00:00:45 | 00:00:00 | ATD |
| R | 5432183700 | DISA | 8934 | 03/08/2012 10:41 | 00:00:00 | 00:00:00 | IDT |
| R | 5432183700 | Grupo1 | 8934 | 03/08/2012 10:41 | 00:00:00 | 00:00:00 | IDT |
| R | 5432183700 | DISA | 8934 | 03/08/2012 10:41 | 00:00:27 | 00:00:00 | ATD |
| I | 9000 | 9016 | | 03/08/2012 10:41 | 00:00:11 | 00:00:00 | ATD |
| O | 9019 | 4732741600 | 8933 | 03/08/2012 10:40 | 00:01:14 | 00:00:17 | ATD |
| R | 5432183700 | 9064 | 8934 | 03/08/2012 10:41 | 00:00:32 | 00:00:00 | ATD |
| I | 9057 | 9035 | | 03/08/2012 10:41 | 00:00:30 | 00:00:00 | ATD |
| O | 9040 | 1939354412 | 8932 | 03/08/2012 10:40 | 00:02:00 | 00:00:11 | ATD |
| I | 9019 | 9049 | | 03/08/2012 10:42 | 00:00:28 | 00:00:00 | ATD |
| R | 1734261234 | 9018 | 8931 | 03/08/2012 10:41 | 00:03:12 | 00:00:00 | ATD |
| R | 1734261234 | 9062 | 8931 | 03/08/2012 10:44 | 00:00:00 | 00:00:00 | IDT |
| I | 9018 | 9062 | | 03/08/2012 10:43 | 00:00:45 | 00:00:00 | ATD |
| I | 9000 | 9036 | | 03/08/2012 10:44 | 00:00:14 | 00:00:00 | ATD |
| I | 9010 | 9025 | | 03/08/2012 10:44 | 00:00:18 | 00:00:00 | ATD |
| R | 5432918600 | DISA | 8932 | 03/08/2012 10:45 | 00:00:00 | 00:00:00 | IDT |
| R | 5432918600 | 9000 | 8932 | 03/08/2012 10:45 | 00:00:00 | 00:00:00 | IDT |

Figura 4.5: Exemplo de um arquivo com os dados extraídos da URA.

## 4.5. ORACLE DATA MINING

*Oracle Data Mining* é uma ferramenta da *Oracle* disponível opcionalmente somente para banco de dados *Oracle Enterprise Edition*. Esta tecnologia permite extrair grandes quantidades de informações de uma ou mais fontes de dados com o objetivo de gerar novas oportunidades de negócio. Com base nos dados extraídos é possível que uma empresa compreenda melhor seus clientes podendo assim improvisar e organizar campanhas de marketing, vendas e suporte aos clientes.

O ODM suporta técnicas de aprendizagem supervisionadas e não supervisionadas e inclui algoritmos de Classificação e Regressão, Regras de Associação, Modelos de *Clustering*, Seleção e Importância de Atributos incluídos no banco de dados *Oracle*. A maioria destes algoritmos podem ser aplicados a dados estruturados ou não estruturados.

Possui uma interface chamada *Oracle Data Miner*, que realiza o pré-processamento dos dados, disponibiliza configurações e otimizações para todos os parâmetros pré-definidos e auxilia na construção e avaliação de modelos e na interpretação dos resultados.

A partir de julho de 2010, divulgado no site da *Oracle*, a interface *Oracle Data Miner* passou a ser integrada no *SQL Developer*, ferramenta criada pela própria *Oracle* para realizar o gerenciamento do banco de dados.

Neste estudo de caso foi utilizada a ferramenta *SQL Developer* com as funcionalidades do *Oracle Data Miner*. Foi instalada uma máquina virtual com sistema operacional *SUSE Linux Enterprise Server 11 SP1* para a instalação do Banco de Dados *Oracle 11gR2 Enterprise Edition* com a opção *Oracle Data Mining*. Após foi instalada a ferramenta *SQL Developer*, configuradas as conexões para a base de dados e a partir do mesmo criado o repositório de *Data Miner* onde será realizado o processo de Mineração.

Após a conexão realizada, foi realizado um teste com uma base de dados fictícia para estudo da ferramenta e das funcionalidades. Então criou-se um *Data Source* que aponta as tabelas de origem para mineração através de um atalho conforme Figura 4.7.

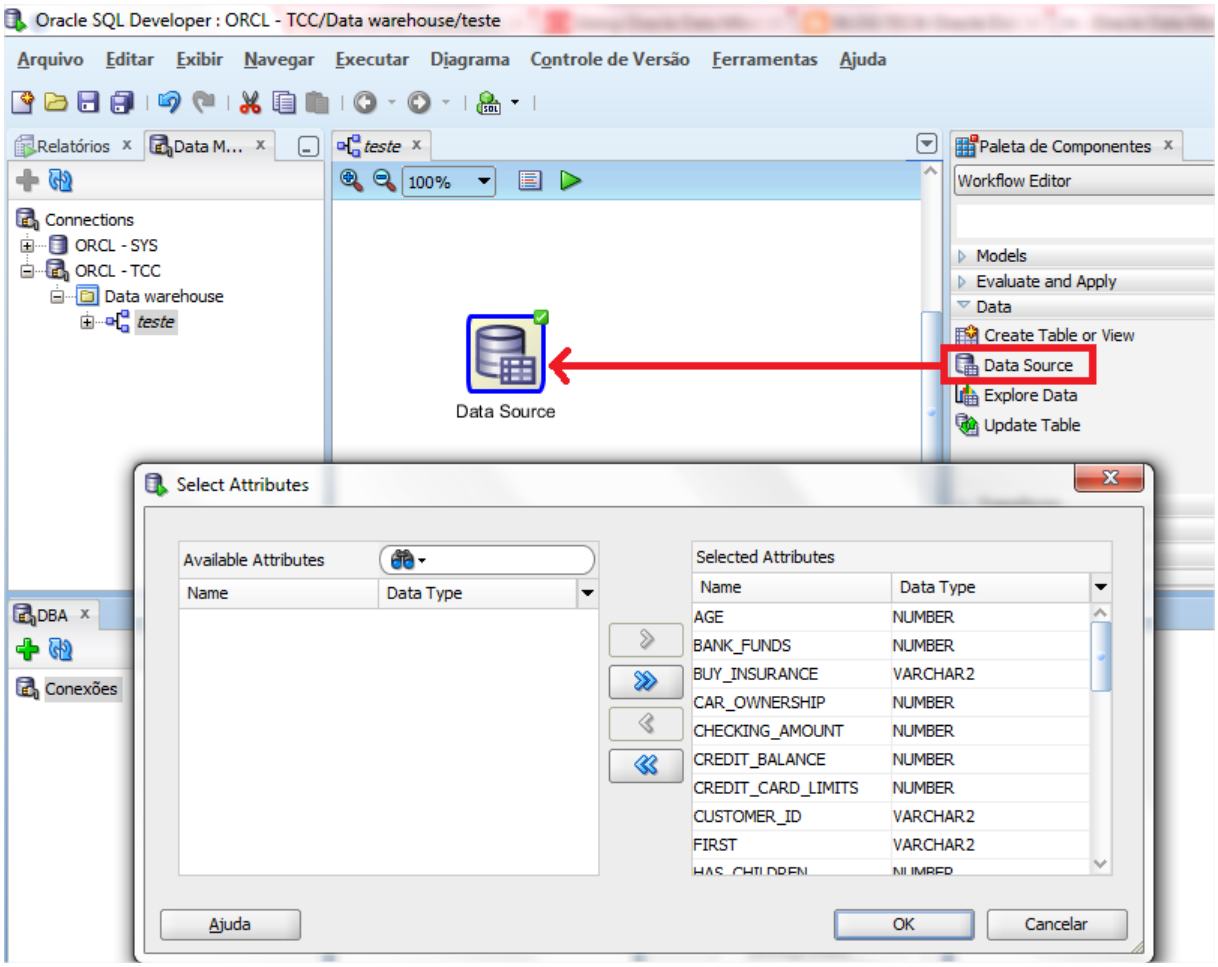



Figura 4.7: Forma de criação do *Data Source* para Mineração de Dados.

A partir do *Data Source* é possível incluir um *Explore Data*, onde pode ser escolhido um atributo para realizar o agrupamento. Na Figura 4.8 podemos visualizar onde seleciona-se o nodo que conecta-se com a base de dados.

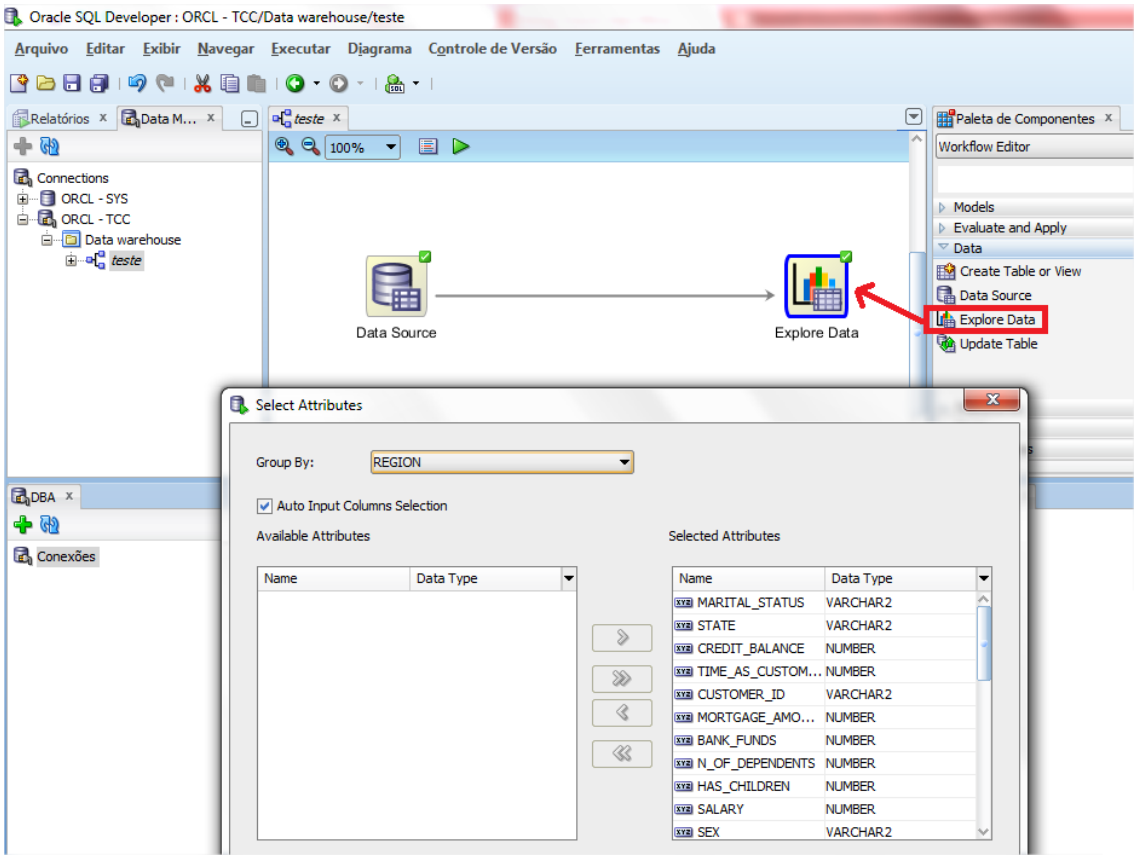



Figura 4.8: Conexão com o *Explore Data* utilizado para gerar estatísticas dos dados.

Desta forma será possível selecionar as técnicas e os algoritmos de Mineração de Dados que o *Oracle Data Miner* suporta através da aba *Models* conforme mostra a figura 4.10.

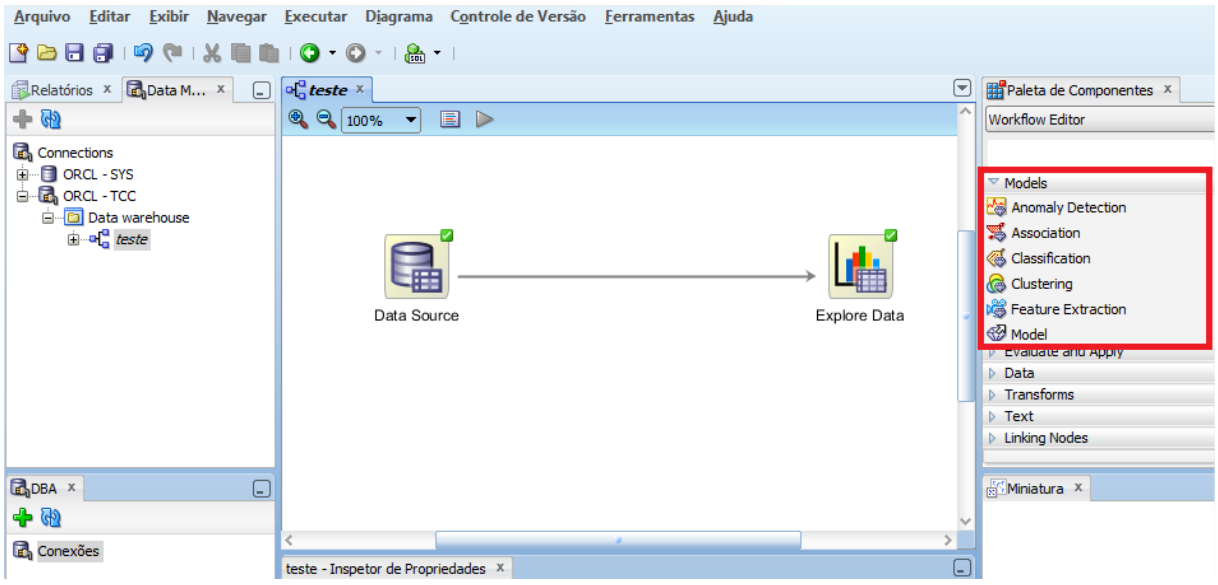



Figura 4.10: Aba *Models*, onde se apresentam as técnicas de Mineração de Dados.

Para as técnicas que serão utilizadas na execução da Mineração de Dados deste trabalho existem os seguintes algoritmos na ferramenta Oracle Data Mining:

- Classificação: Modelos Lineares Generalizados, Máquina de Vetores de Suporte, Naive Bayes (apresentado neste trabalho) e Árvore de Decisão.

- Clusterização: K-means (apresentado neste trabalho) e outro algoritmo chamado O-Cluster.

A *Oracle* aponta que nos últimos meses aumentou a busca pela solução *Oracle Data Mining*. “Isso aconteceu pela mudança de mercado de exigir respostas mais rápidas às suas dúvidas e também pela maior maturidade das empresas em questões de análises dos dados” (ORACLE, 2012).

# 5. DESENVOLVIMENTO DO ESTUDO DE CASO

Nos dias atuais, a informação e o conhecimento são considerados um bem intangível para as organizações, e para as áreas que possuem o contato com o cliente que podem gerar uma base de informações a ser utilizadas para auxiliar no relacionamento. Conhecer o cliente é fundamental para a empresa.

Para gerar novas informações que apoiem a equipe estratégica da Focco Sistemas de Gestão na tomada de decisões e corroborar os conceitos apresentados neste trabalho foi realizado um estudo de caso no "Portal de Atendimento Focco". Neste capítulo estão detalhadas todas as etapas deste estudo de caso, desde a definição da base de dados e as fases da descoberta do conhecimento que são: seleção, limpeza, transformação, mineração e avaliação dos resultados.

## 5.1. SELEÇÃO DOS DADOS

Conforme já descrito no capítulo 2.2 esta etapa é uma das mais importantes na DCBD, pois conforme for a qualidade na seleção de dados e no pré-processamento, melhor será o resultado.

Para realizar a seleção de dados e o pré-processamento para este estudo de caso, foi necessário utilizar um Data Warehouse.

### *5.1.1. Data Warehouse*

Um *Data Warehouse* consiste em um conjunto de dados consolidado e organizado que mantém dados históricos para análise a fim de dar suporte à tomada de decisões estratégicas da empresa. Surgiu da necessidade de integrar dados corporativos espalhados em diferentes máquinas, softwares e sistemas

operacionais, para que fosse possível tornar os dados acessíveis a todos os usuários (VIEIRA, 2012).

Como a construção de um *Data Warehouse* envolve um contínuo pré-processamento dos dados (limpeza, transformações, integração, dentre outras ações), este conjunto de dados é uma fonte potencial de informação para ser submetida ao processo de descoberta de conhecimento em bases de dados (ORACLE, 2012).

Para a modelagem de um *Data Warehouse* é necessário primeiramente definir de que forma os dados serão apresentados, após deve ser definido de que forma estes dados serão extraídos e carregados na base. O processo de carregar os dados inclui tarefas como limpeza, padronização e transformação (VIEIRA, 2012).

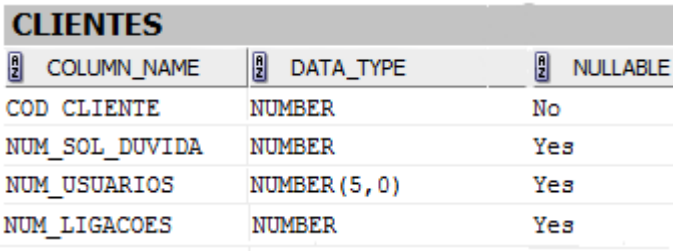
Neste estudo de caso optou-se por criar um *Data Warehouse* para concentrar as informações das diversas tabelas do Portal de Atendimento Focco e os dados que a URA armazena, com o intuito de facilitar a análise e obtenção dos dados no processo da Mineração de Dados.

Foi criada uma tabela chamada de CLIENTES para armazenar todos os dados já compactados por cliente para facilitar no momento do processo da mineração dos dados. Esta tabela foi preenchida a partir da base de dados apresentada no capítulo 4.4 guardando informações como total de dúvidas, ligações e números de usuários de cada cliente. No próximo capítulo poderemos visualizar como os dados foram utilizados.

### *5.1.2. Dados utilizados no Estudo de Caso*

Para as informações que seriam utilizadas no estudo de caso, foi necessário criar a seguinte uma tabela chamada CLIENTES que armazena as informações referentes aos clientes (atributo COD_CLIENTE), número de solicitações abertas num determinado período (atributo NUM_SOL_DUVIDA), número de licenças compradas pelo cliente (atributo NUM_USUARIOS) e o número de ligações da URA

(atributo NUM_LIGACOES) para o período. É possível visualizar a estrutura desta tabela na Figura 5.1. O período escolhido para filtrar as solicitações e as ligações da URA foi o de 01/08/2012 a 31/12/2012, devido as informações da URA só terem sido gravadas de forma correta a partir desta do mês de agosto de 2012.



**CLIENTES**

| COLUMN_NAME | DATA_TYPE | NULLABLE |
|---|---|---|
| COD_CLIENTE | NUMBER | No |
| NUM_SOL_DUVIDA | NUMBER | Yes |
| NUM_USUARIOS | NUMBER(5,0) | Yes |
| NUM_LIGACOES | NUMBER | Yes |

Figura 5.1: Estrutura da tabela de CLIENTES.

Para incluir as informações na tabela de clientes foi necessário buscar as informações da base de dados onde as informações do “Portal de Atendimento da Focco” são armazenadas, no banco Oracle, e através do documento texto que a central telefônica disponibiliza com todas as ligações realizadas e recebidas.

Na base de dados do “Portal de Atendimento da Focco” foram extraídas as informações da tabela de CLIENTES, onde buscou-se o código do cliente e o segmento, e da tabela de SOLICITAÇÕES, onde contou-se o número de solicitações de dúvidas incluídas no período para cada cliente. Para obter o número de usuários e o segmento dos clientes precisou-se solicitar a equipe Comercial da empresa para que analisasse os contratos e identificasse o número de licenças vendidas definindo em que categoria o cliente se enquadra. Estas informações foram repassadas através de uma planilha.

Para extrair as informações do documento texto foi necessário inclui-las no programa Microsoft Excel, para que fossem separadas as colunas e armazenadas em uma tabela temporária. Esta tabela temporária foi criada dentro da base de dados do “Portal de Atendimento da Focco” apenas para contagem das ligações. Como na tabela de CLIENTES existia a informação do número do telefone do cliente, buscou-se o confronto entre o número do cliente e o número do registro da URA e então descobriu-se a quantidade de registros de ligações que a Focco

Sistemas de Gestão recebeu de cada cliente. Foi levado em consideração também números com ramais diferentes, realizando o confronto também sem os últimos três números, que representavam os diferentes ramais.

## 5.2. PROCESSAMENTO DOS DADOS

Por ser uma base de dados que atende diversos processos dentro da empresa Focco Sistemas de Gestão, foi necessária uma limpeza dos dados.

Primeiramente identificamos juntamente com a equipe Comercial da empresa, através das datas dos contratos dos clientes aqueles que possuíam data maior que 01/08/2012 e retiramos da consulta, pois nestes casos deixariam as estatísticas distorcidas. Esta informação foi repassada através de uma planilha e como eram poucos os registros que era necessário ignorar, eles foram apagados das tabelas de forma manual.

A informação do segmento do cliente estava de diversas formas, alguns estavam nulos e para os outros não existia um padrão. Por exemplo, cliente do segmento metal mecânico tinham respostas como “METAL” ou “METALMECANICO”. Para os clientes que possuíam a informação do segmento em branco foi solicitada à equipe Comercial da Focco que identificasse nos contratos e repassasse a informação correta, estes foram alterados manualmente através do banco de dados e os outros foram alterado para que a existisse um padrão dos dados, definindo as seguintes respostas:

- MOVELEIRO
- METALMECANICO
- BEBIDAS/MOINHOS/DISTRIBUIDORES

Na tabela onde se armazenou os registros de ligações da URA os números dos telefones foram armazenados da seguinte forma: apenas dígitos onde os dois primeiros referenciavam o DDD da cidade de origem e os outros oito o número de telefone do estabelecimento. Porém ao confrontar com os telefones da tabela do

“Portal de Atendimento” com o documento texto gerado pela URA algumas inconsistências foram identificadas, na base de dados não existia um padrão para armazenar o número do telefone do cliente. Por exemplo, existiam telefones com a seguinte sintaxe (51) 3662.3505 ou (51)36623505. Foi padronizado para que ficassem apenas números: 5136623505.

## 5.3. MINERAÇÃO DOS DADOS

No Cenário do Estudo de Caso, o objetivo era verificar se existia alguma relação entre o número de solicitações de dúvida abertas pelo cliente com o número de ligações realizadas pelo mesmo e o número de licenças que o mesmo havia adquirido.

Analisando o cenário proposto, podemos perceber o interesse em identificar os grupos de clientes e seus perfis.

Como não existe uma técnica em específico que solucione todos os problemas foi necessário dividir em quatro etapas, são elas:

- Novo agrupamento de clientes conforme número de solicitações de dúvida.
- Novo agrupamento de clientes conforme número de registro de ligações.
- Novo agrupamento de clientes conforme número de licenças adquiridas.
- Relação entre os agrupamentos.

Na figura 5.2 podemos visualizar de forma mais clara a sequência das etapas do processo de mineração dos dados, e qual técnica será aplicada para cada etapa: para as etapas 1, 2 e 3 será aplicada a técnica de Clusterização com o algoritmo *K-means*, e para a etapa 4, que juntará os resultados das 3 etapas anteriores, será utilizada a técnica de Classificação com o algoritmo *Naive Bayes*.

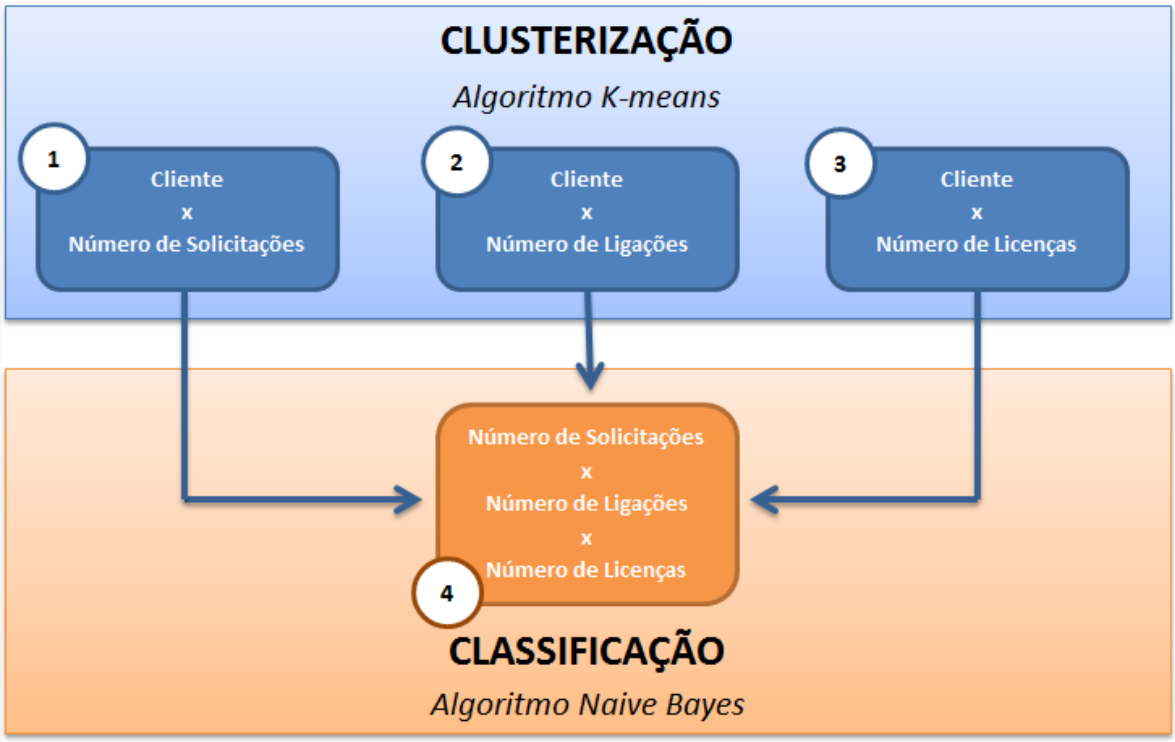



Figura 5.2: Etapas que serão realizadas na mineração de dados deste caso de uso.

### 5.3.1. Etapa 1: novo agrupamento de clientes conforme número de solicitações de dúvida

A Etapa 1 tem como objetivo criar novos grupos de clientes conforme o número de solicitações de dúvida abertas no período de agosto a dezembro de 2012. Os grupos serão nomeados como “INICIANTE”, ”INTERMEDIARIO” e “EXPERIENTE”, determinando o perfil do cliente.

Para esta etapa optou-se por utilizar a técnica de Clusterização e o algoritmo *K-means,* pois estes são os que se adaptam na identificação de afinidades sem que haja um conhecimento prévio dos grupos de clientes que serão formados.

Os passos para a aplicação desta técnica e algoritmo na ferramenta escolhida, *Oracle Data Mining*, encontram-se detalhados nas seções 5.3.1.1, 5.3.1.2, 5.3.1.3 e 5.3.1.4.

### 5.3.1.1. *Data Source*

O passo inicial para realizar a mineração de dados na ferramenta da *Oracle* é a seleção do *Data Source* (tabela de origem dos dados). Selecionou-se a tabela CLIENTES, criada no *Datawarehouse*, e selecionou-se apenas os atributos COD_CLIENTE e NUM_SOL_DUVIDA, conforme mostra a Figura 5.3.

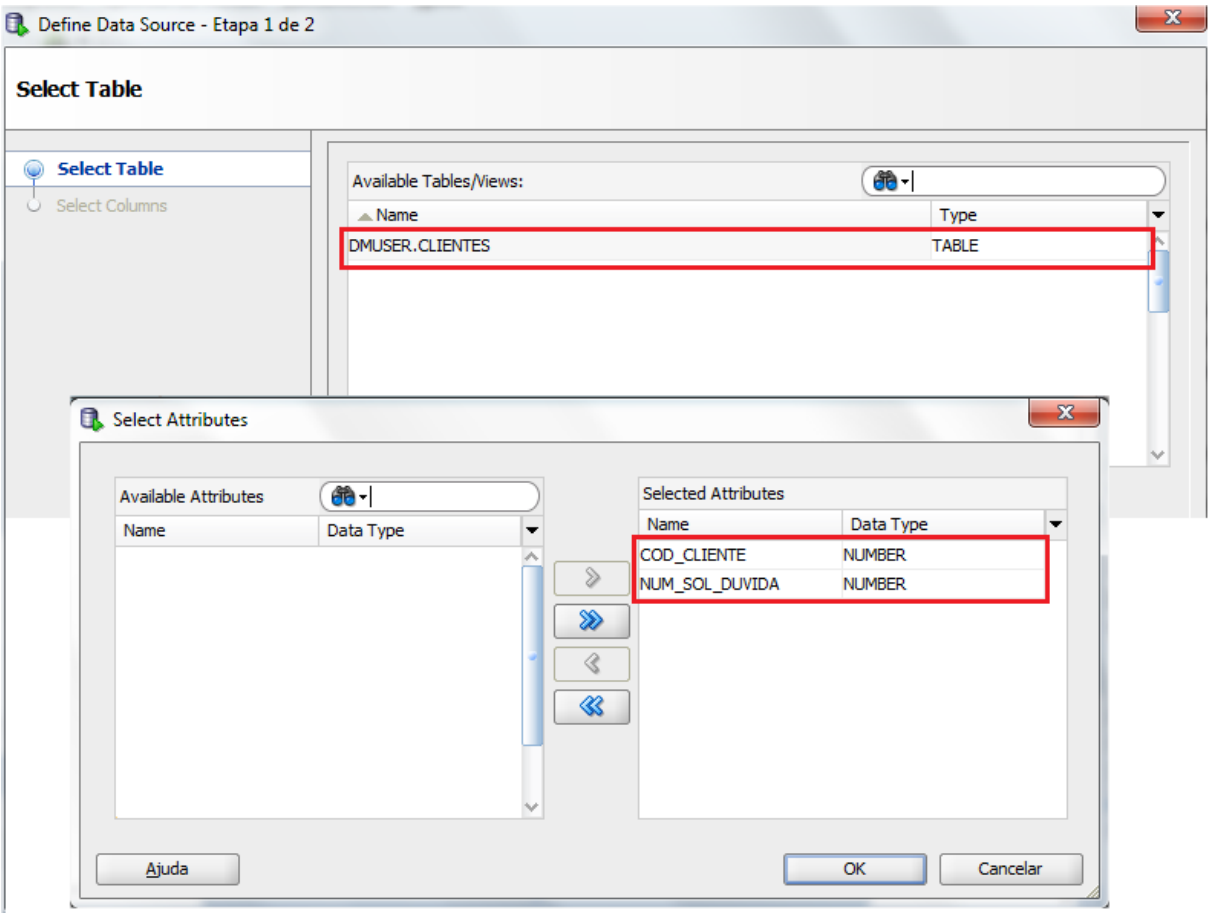



Figura 5.3: Seleção dos atributos da Etapa 1 que serão utilizados no processamento do algoritmo.

### 5.3.1.2. *Técnica*

Na sequência é necessário selecionar a técnica que irá ser aplicada, neste caso selecionamos a de clusterização e informamos qual o atributo refere-se ao

identificador dos dados, neste caso selecionou-se o atributo COD_CLIENTE, conforme Figura 5.4.

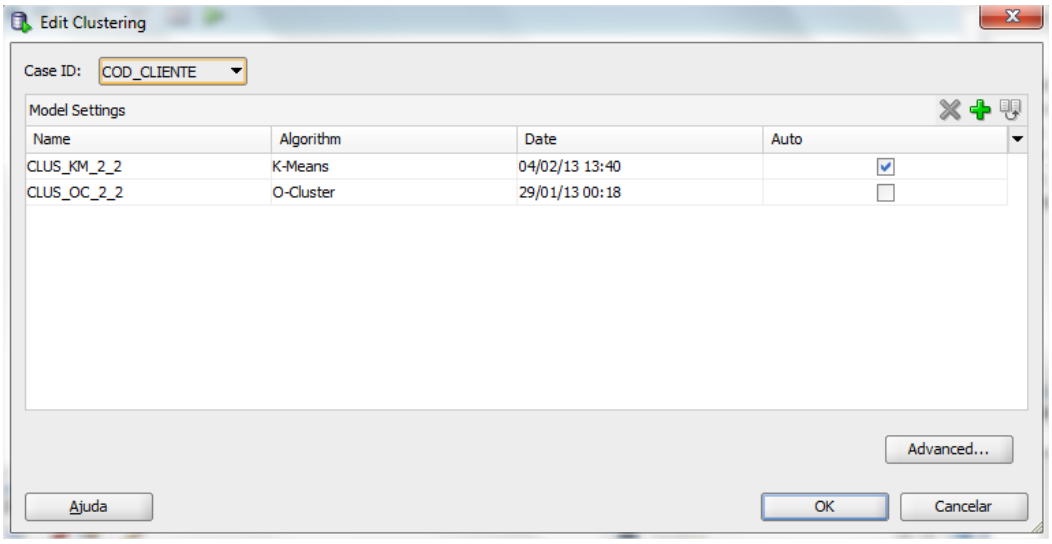



Figura 5.4: Seleção do atributo identificador na Etapa 1.

A ferramenta utilizada sempre executa as técnicas para todos os algoritmos que a mesma disponibiliza, pois possui uma funcionalidade para comparar os algoritmos. No caso desta etapa do estudo de caso iremos analisar somente o algoritmo escolhido, o *K-means*. Sendo assim podemos selecionar as configurações deste algoritmo para possível alterações. Os parâmetros alterados foram os destacados na Figura 5.5 e relacionados abaixo:

- *Number of Clusters*: número de clusters, ou seja, o número de grupos que serão criados a partir da base de dados informada.

- *Distance Function*: define qual função será utilizada para agrupar os clusters. A resposta default é e a Distância Euclidiana. Esta distância é uma das medidas de similaridade entre pontos mais utilizadas na prática. Quanto menor o valor da Distância Euclidiana entre dois pontos, maior a similaridade entre eles.

A Distância Euclidiana é definida como a soma da raiz quadrada da diferença entre x e y em suas respectivas dimensões, visualizada pela fórmula(WITTEN e FRANK, 1999): $\sqrt{((x1 - x2)^2 + (y1 - y2)^2)}$

- *Number of Histogram Bins*: parâmetro utilizado para definir o número de barras exibidas nos histogramas para cada atributo do cluster.

- *Split Criterion*: parâmetro que define a criação de um novo cluster. A resposta default é *Variance* que divide o cluster que produz a maior variância em relação ao original.

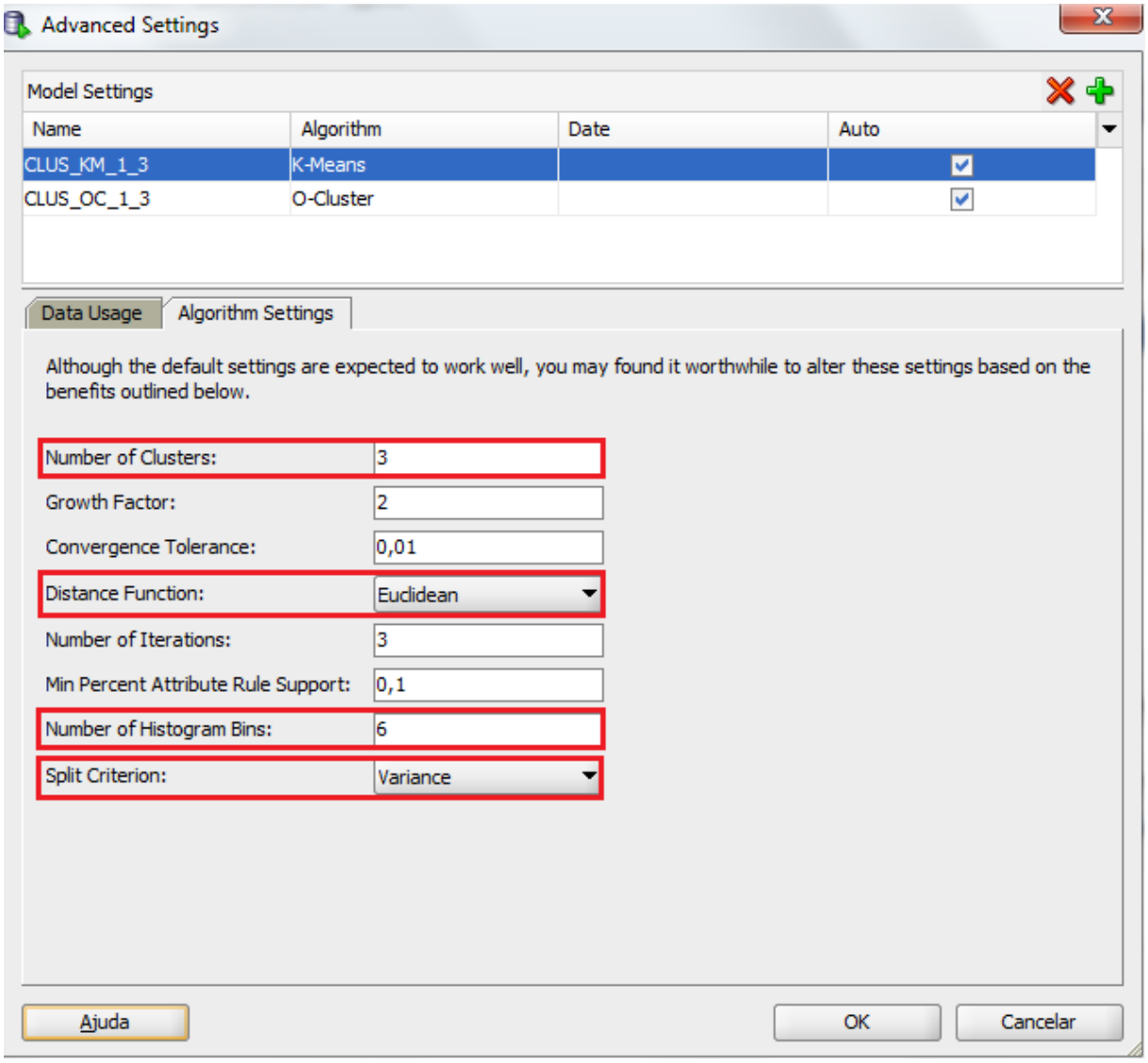



Figura 5.5: Configuração dos parâmetros do algoritmo K-means para Etapa 1.

### *5.3.1.3. Resultados*

Somente é possível a visualização dos resultados da construção do modelo de mineração de dados após a finalização do processamento. Para esta visualização a ferramenta do *Oracle* disponibiliza alguns recursos que facilitam a análise dos resultados de acordo com cada algoritmo.

Para o algoritmo *K-means* a ferramenta permite a visualização dos resultados através de histogramas, regras e de uma árvore para dividir e subdividir os *clusters*.

Para a Etapa 1 do estudo de caso podemos identificar 5 *clusters* que foram criados pelo algoritmo (1, 2, 3, 4 e 5), sendo que os clusters 1 e 2 são intermediários e foram utilizados para a criação dos demais, conforme ilustra a Figura 5.6.

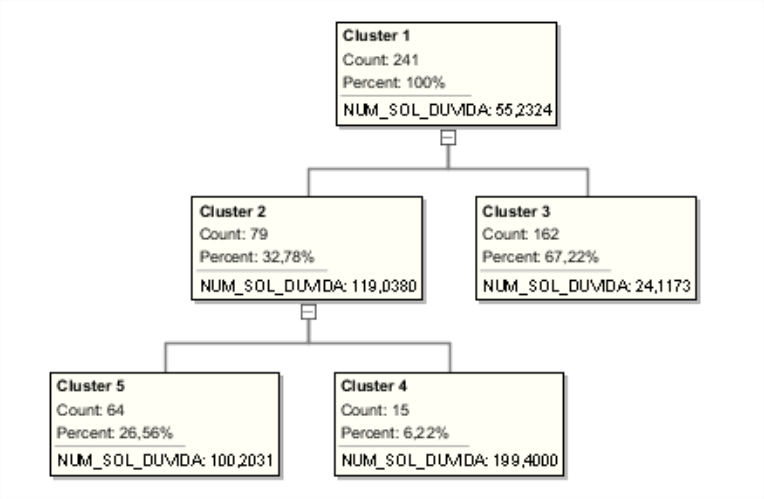



Figura 5.6: Clusters criados para a Etapa 1 visualizados pela ferramenta do Oracle

A ferramenta também permite visualizar os detalhes de cada cluster e desta forma podemos identificar quais foram os centróides definidos para o atributo na criação dos clusters utilizados no trabalho de mineração de dados, como mostra a Figura 5.7, e representado através do gráfico da Figura 5.8.

Attributes: 1 out of 1

| Attribute | Centroid(3) | Centroid(4) | Centroid(5) |
|---|---|---|---|
| NUM_SOL_DUVIDA | 24,375 | 199,4 | 100,203125 |

Figura 5.7: Centroides dos 3 clusters criados no processamento da técnica na Etapa 1.

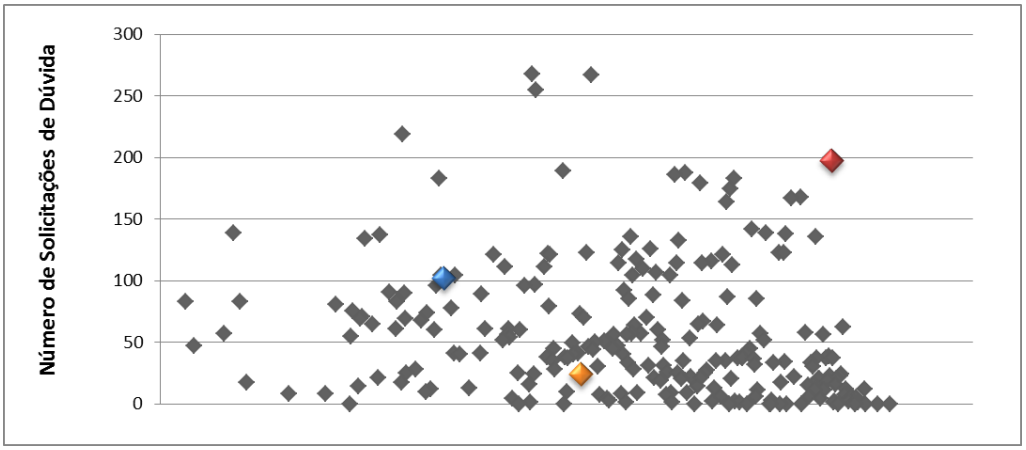



Figura 5.8: Gráfico que representa a dispersão dos dados e os centroides destacados.

### *5.3.1.4. Interpretação*

Ao analisar os detalhes de cada *cluster* criado e as regras definidas pelo algoritmo, foram realizadas as seguintes interpretações:

- Quando o número de solicitações for menor que 44,6 o cliente se enquadra no *cluster* 3 representado pelo perfil de “Experiente”, conforme ilustrado nas figuras 5.9 e 5.10.

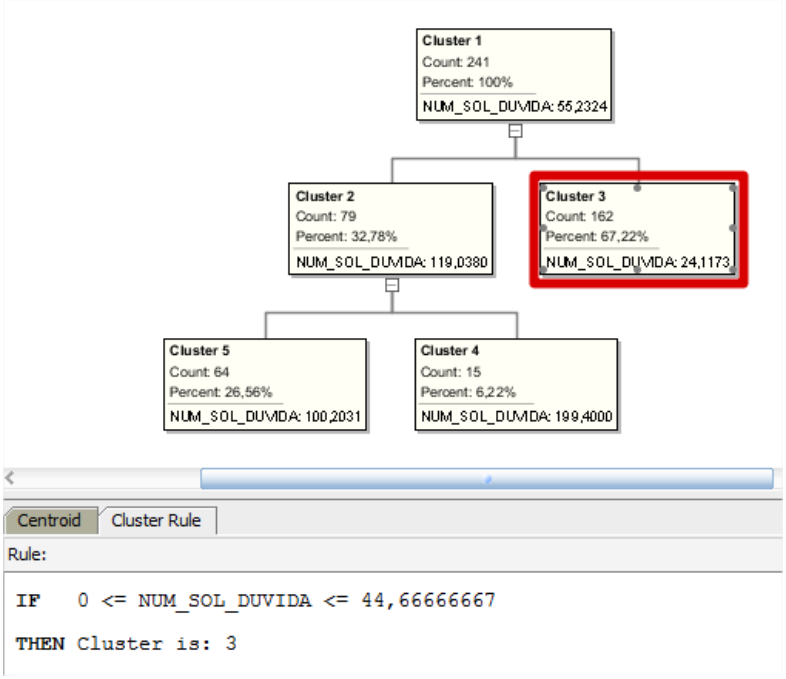



Figura 5.9: Regras geradas para o cluster 3 na Etapa 1.

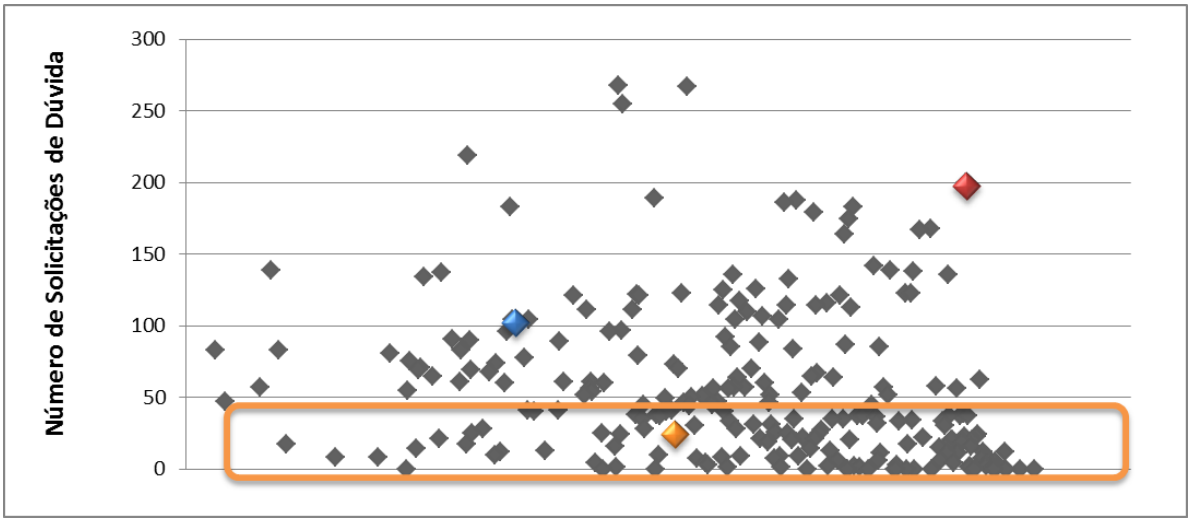



Figura 5.10: Conjunto de dados que pertencem ao cluster 3 definidos na Etapa 1.

- Quando o número de solicitações for maior que 44,6 e menor que 134 o cliente se enquadra no *cluster* 5 representado pelo perfil de “Intermediário”, conforme ilustrado nas figuras 5.11 e 5.12.

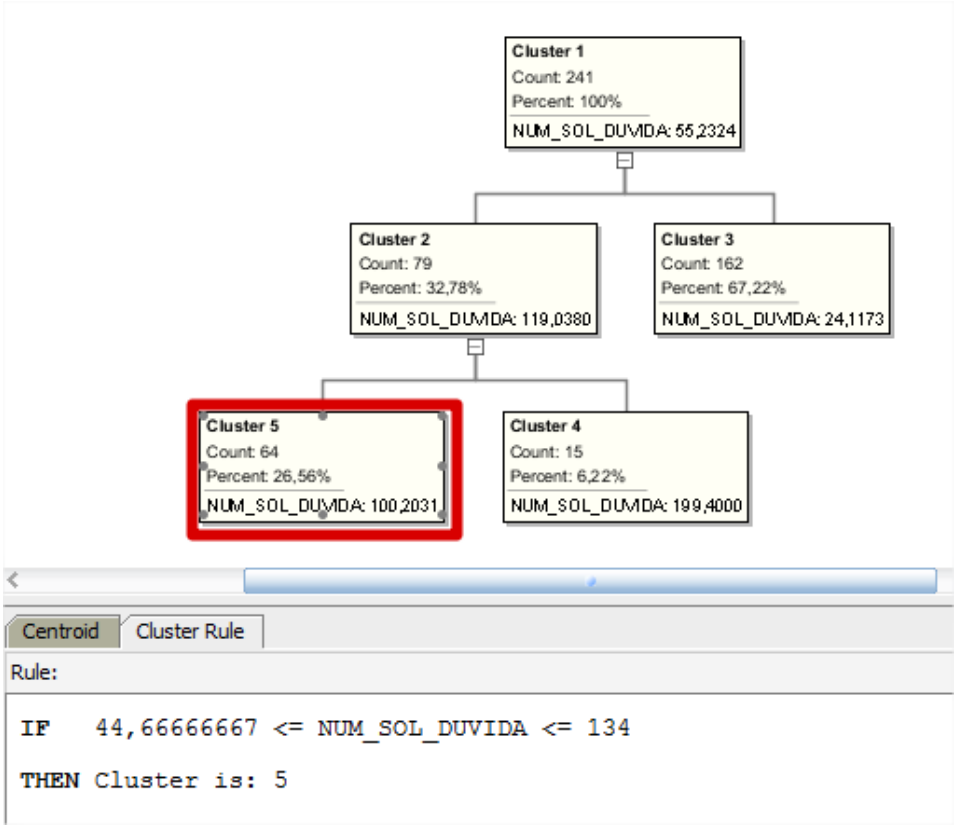



Figura 5.11: Regras geradas para o cluster 5 na Etapa 1.

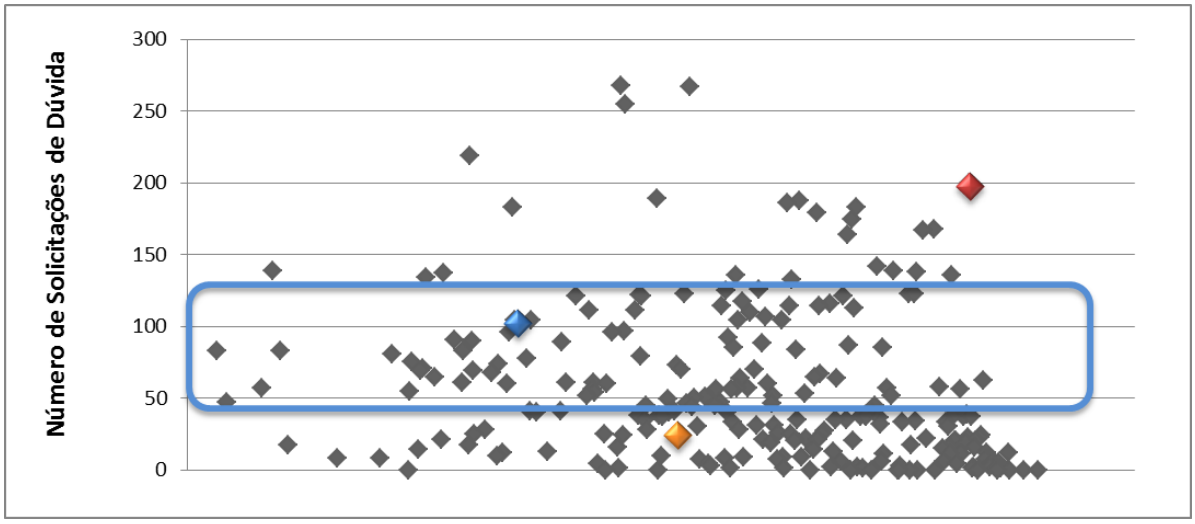



Figura 5.12: Conjunto de dados que pertencem ao cluster 5 definidos na Etapa 1.

- Quando o número de solicitações for maior que 134 e menor que 268 o cliente se enquadra no *cluster* 4 representado pelo perfil de “Iniciante”, conforme ilustrado nas figuras 5.13 e 5.14.

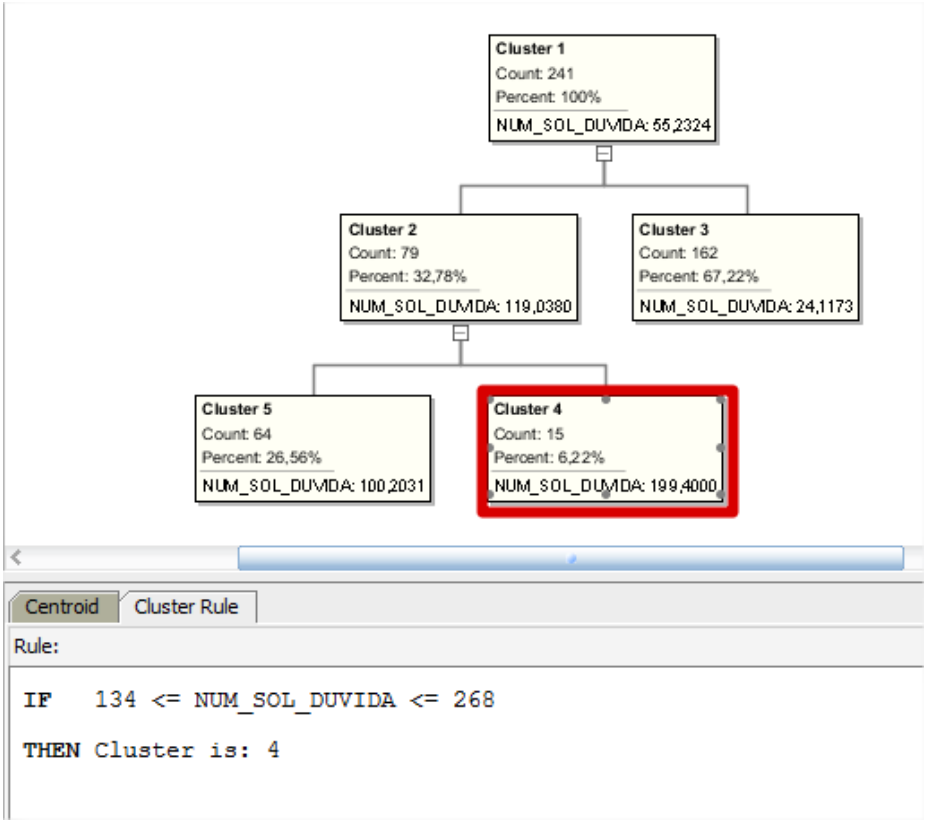



Figura 5.13: Regras geradas para o cluster 4 na Etapa 1.

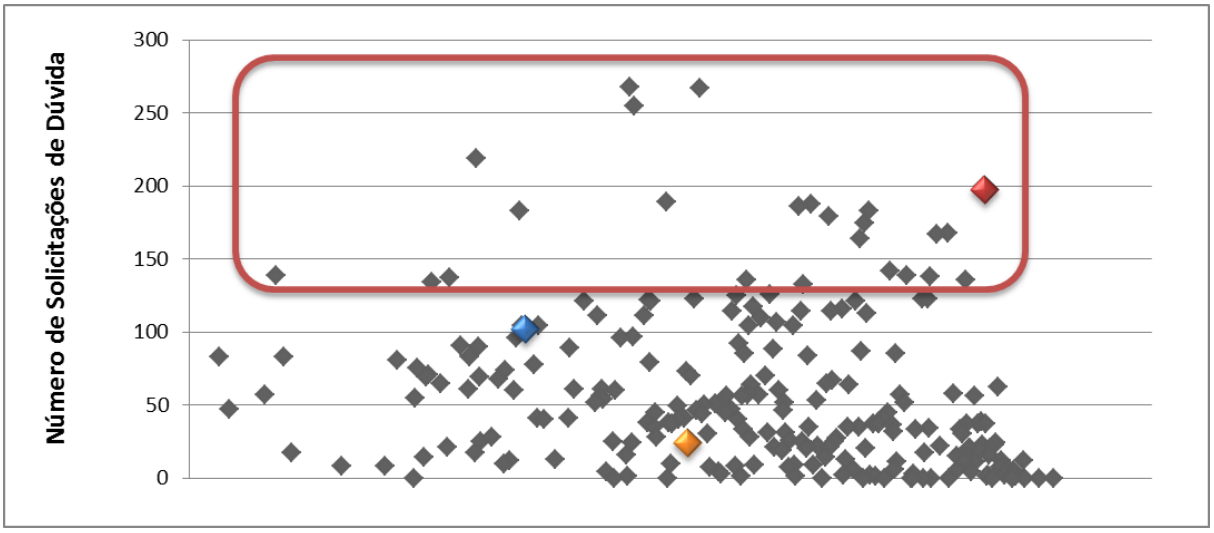



Figura 5.14: Conjunto de dados que pertencem ao cluster 4 definidos na Etapa 1.

A partir da definição de quais seriam as divisões dos grupos formados em relação ao número de solicitações de dúvida abertas no período pelos clientes, foi necessário incluir o atributo PERFIL na tabela CLIENTES que enquadra o cliente nestes grupos criados. As informações deste atributo foram cadastradas conforme o script ilustrado pela Figura 5.15.

```
DECLARE
   v_perfil VARCHAR(15);
BEGIN
   FOR s IN (SELECT num_sol_duvida
                  , cod_cliente
               FROM clientes)
   LOOP
      IF s.num_sol_duvida <= 45 THEN
         v_perfil := 'EXPERIENTE';
      ELSIF s.num_sol_duvida > 45 AND s.num_sol_duvida <= 134 THEN
         v_perfil := 'INTERMEDIARIO';
      ELSIF s.num_sol_duvida > 134 THEN
         v_perfil := 'INICIANTE';
      END IF;

      UPDATE clientes
         SET perfil = v_perfil
       WHERE cod_cliente = s.cod_cliente;
   END LOOP;
END;
```

Figura 5.15: Script executado para popular o campo PERFIL da tabela CLIENTES.

Após as interpretações e com as novas informações extraídas no trabalho de mineração de dados armazenadas no atributo PERFIL da tabela CLIENTES, a Etapa 1 do Estudo de Caso estava concluída.

### *5.3.2. Etapa 2: novo agrupamento de clientes conforme número de registro de ligações*

A Etapa 2 tem como objetivo criar novos grupos de clientes conforme o número de registros de ligações armazenados pela URA no período de agosto a dezembro de 2012. Os grupos serão nomeados como “INICIANTE”, ”INTERMEDIARIO” e “EXPERIENTE”, determinando o perfil do cliente.

Assim como na Etapa 1, nesta etapa também optou-se por utilizar a técnica de Clusterização e o algoritmo K-means. Os passos para a aplicação desta técnica e algoritmo encontram-se detalhados nas seções 5.3.2.1, 5.3.2.2, 5.3.2.3 e 5.3.2.4.

#### *5.3.2.1. Data Source*

Para a Etapa 2, selecionou-se a tabela CLIENTES no *Data Source*, porém selecionou-se os atributos COD_CLIENTE e NUM_LIGACOES, conforme mostra a Figura 5.16.

#### *5.3.2.2. Técnica*

No momento do processo onde seleciona-se a técnica que será utilizada, assim como na Etapa 1, a técnica escolhida é a de Clusterização, e o atributo identificador foi o COD_CLIENTE.

Para o algoritmo escolhido, o *K-means,* foram alteradas as seguintes parametrizações, conforme Figura 5.17:

- *Number of Cluster, Distance Function*, *Number of Histogram Bins* e *Split Criterion*: já explicados no capítulo 5.3.1.2.

- *Number of Iterations:* utilizados para definir até quantas iterações serão realizadas, ou seja, até quantas vezes serão recalculados os centroides.

- *Growth Factor:* parâmetro utilizado para controlar a utilização da memória durante o processo de mineração de dados.

- *Convergence Tolerance*: tolerância mínima de erro, utilizada para definir como a hierarquia dos *clusters* será formada.

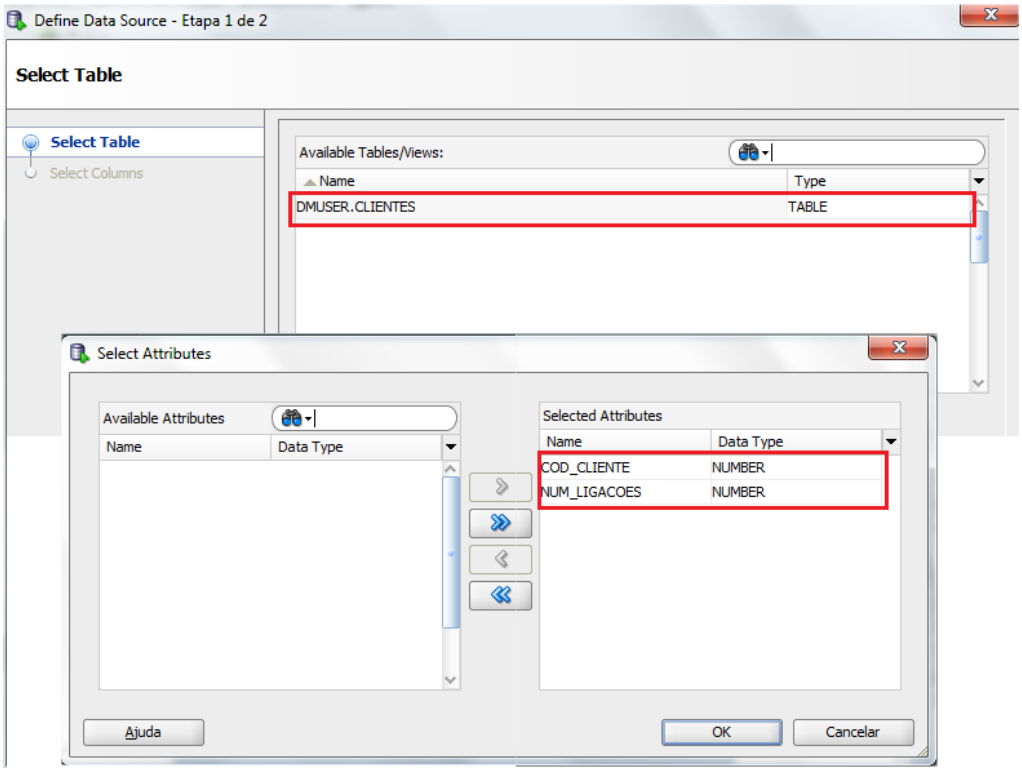



Figura 5.16: Seleção dos atributos da Etapa 2 que serão utilizados no processamento do algoritmo.

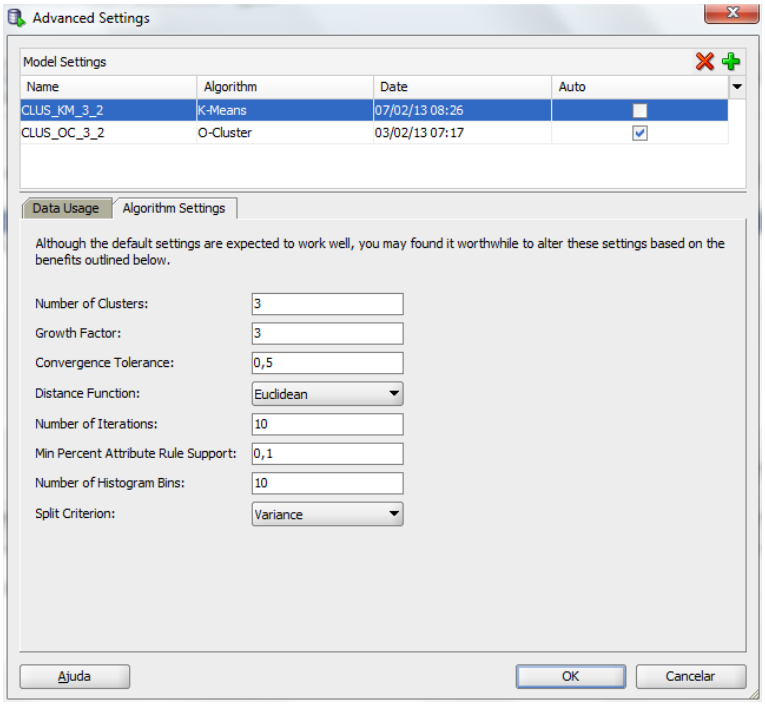



Figura 5.17: Configuração dos parâmetros do algoritmo *K-means* para Etapa 2.

### *5.3.2.3. Resultados*

Para a Etapa 2 do estudo de caso também podemos identificar 5 *clusters* criados pelo algoritmo (1, 2, 3, 4 e 5), sendo que os *clusters* 1 e 2 são intermediários e foram utilizados para a criação dos demais, conforme ilustra a Figura 5.18.

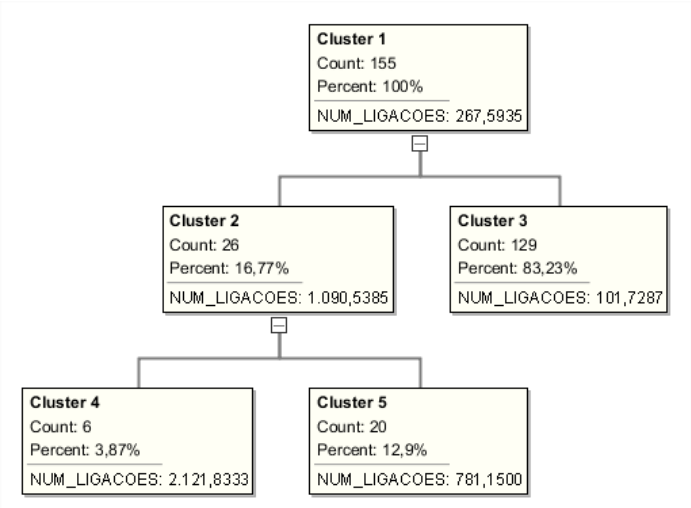



Figura 5.18: *Clusters* criados para a Etapa 2 visualizados pela ferramenta do Oracle.

Podemos visualizar os centróides definidos para o atributo na criação dos clusters através da Figura 5.19, e representado pelo gráfico da Figura 5.20.

Attributes: 1 out of 1

| Attribute | Centroid(3) | Centroid(4) | Centroid(5) |
|---|---|---|---|
| NUM_LIGACOES | 101,72868217 | 2.121,83333333 | 781,15 |

Figura 5.19: Centroides dos 3 *clusters* criados no processamento da técnica na Etapa 2.

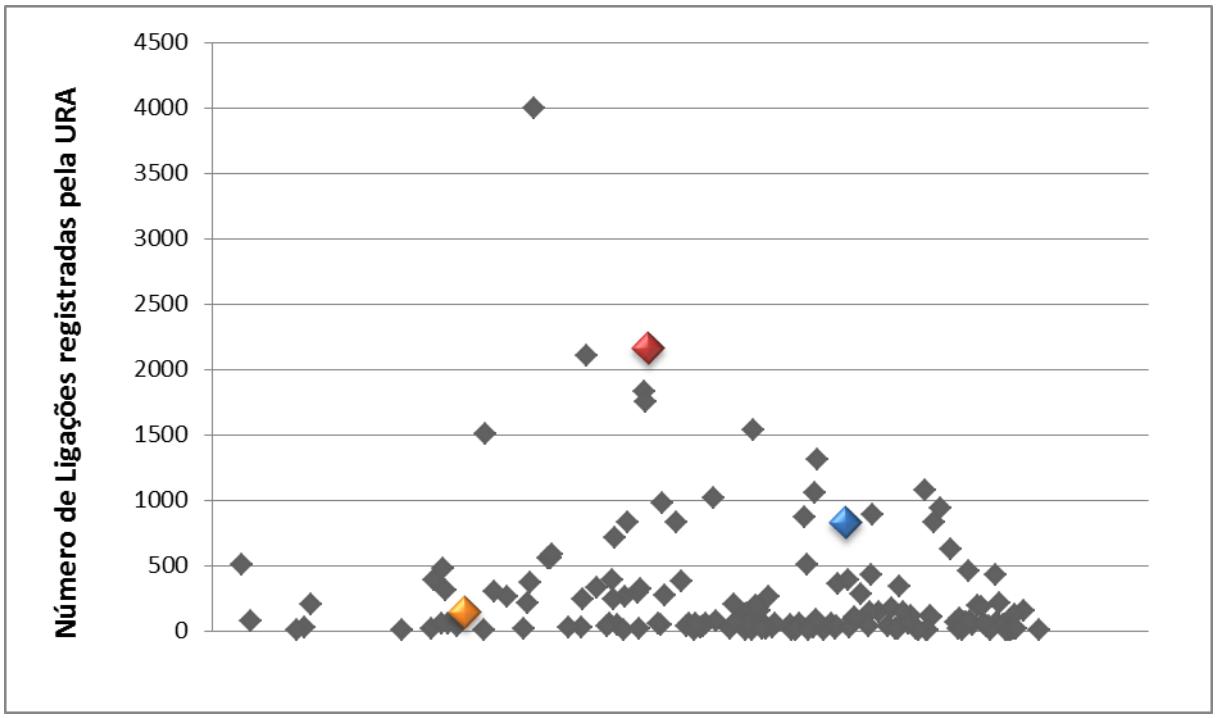



Figura 5.20: Gráfico que representa a dispersão dos dados e os centroides destacados na Etapa 2.

### *5.3.2.4. Interpretação*

Ao analisar os detalhes de cada cluster criado e as regras definidas, foram realizadas as seguintes interpretações:

- Quando o número de ligações for maior ou igual a 4 e menor ou igual a 403 o cliente se enquadra no cluster 3 representado pelo perfil de “Experiente”, conforme ilustrado nas figuras 5.21 e 5.22.

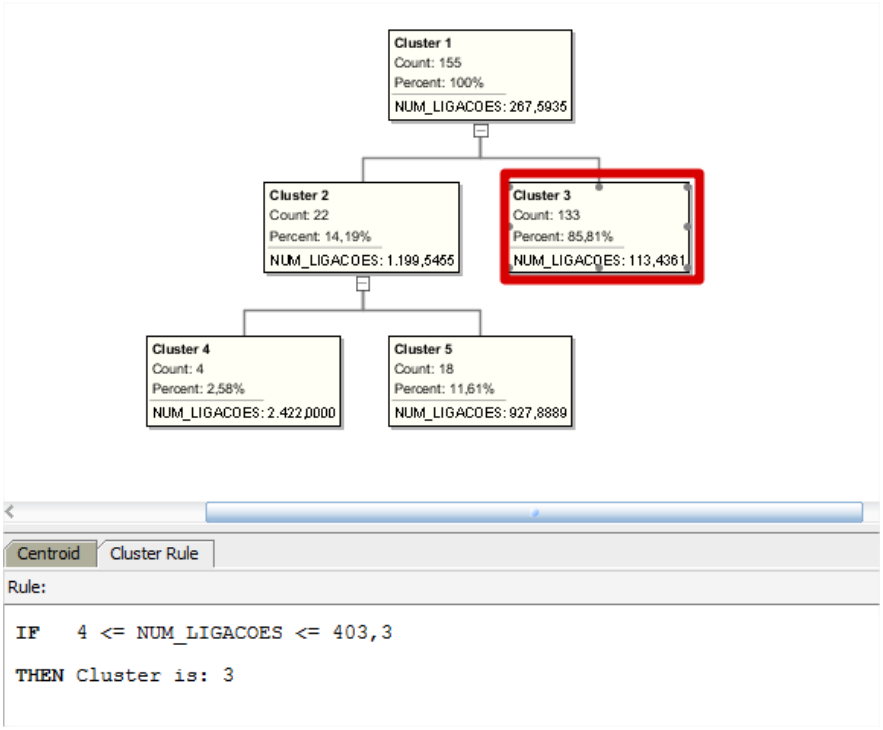



Figura 5.21: Regras geradas para o *cluster* 3 na Etapa 2.

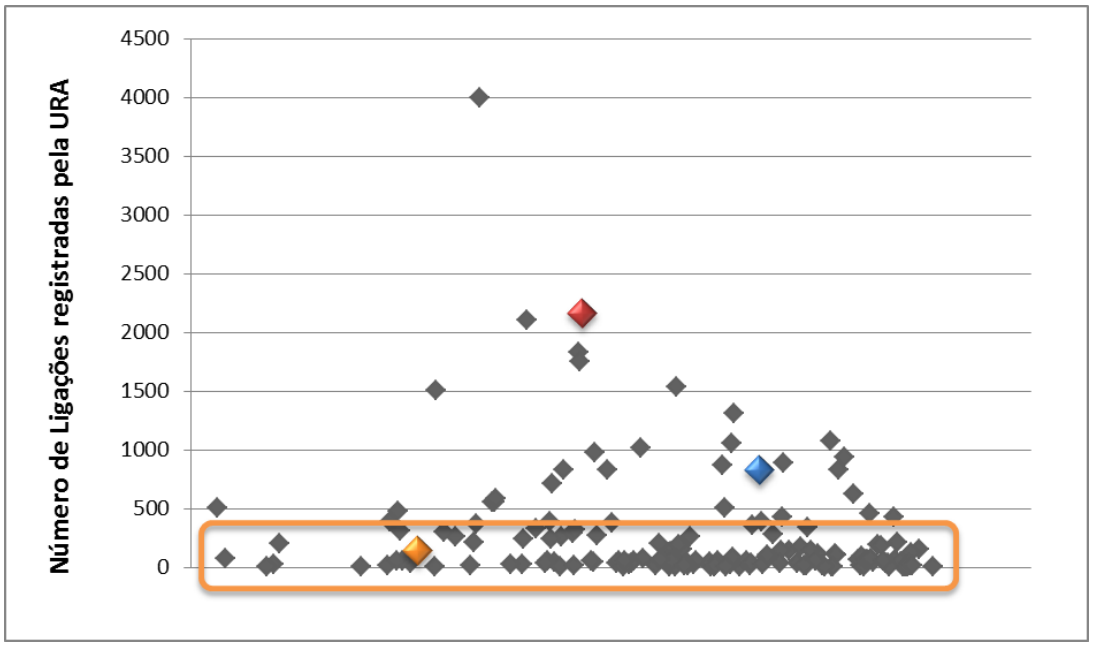



Figura 5.22: Conjunto de dados que pertencem ao *cluster* 3 definidos na Etapa 2.

- Quando o número de ligações for maior que 403 e menor que 1201 o cliente se enquadra no *cluster* 5 representado pelo perfil de “Intermediário”, conforme ilustrado nas figuras 5.23 e 5.24.

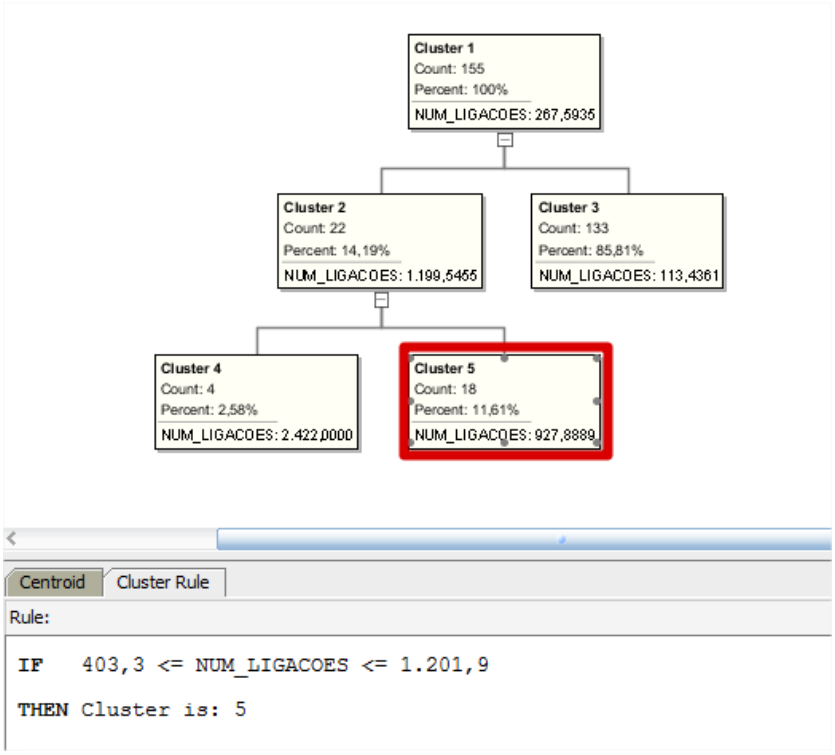



Figura 5.23: Regras geradas para o cluster 5 na Etapa 2.

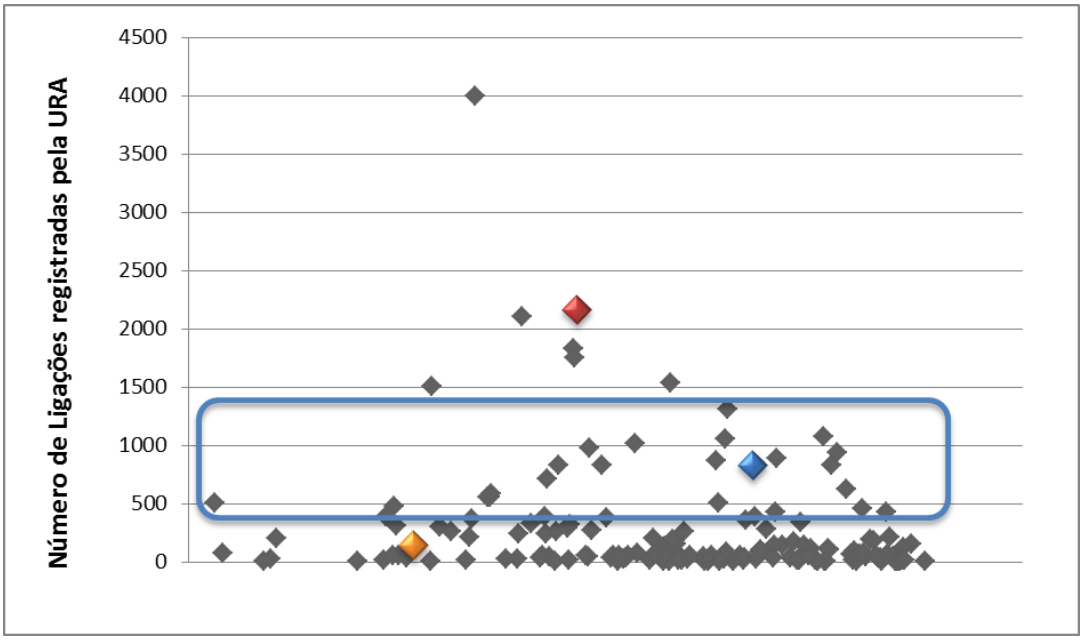



Figura 5.24: Conjunto de dados que pertencem ao *cluster* 5 definidos na Etapa 2.

- Quando o número de ligações for maior que 1201 e menor que 3997 o cliente se enquadra no *cluster* 4 representado pelo perfil de “Iniciante”, conforme ilustrado nas figuras 5.25 e 5.26.

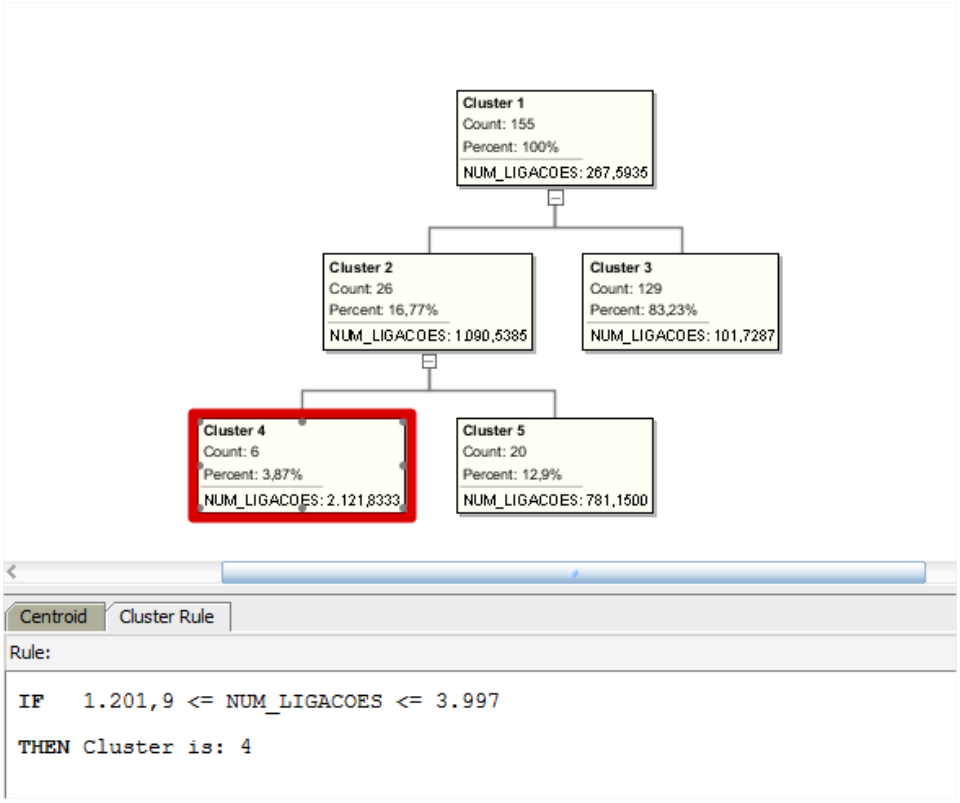



Figura 5.25: Regras geradas para o *cluster* 4 na Etapa 2.

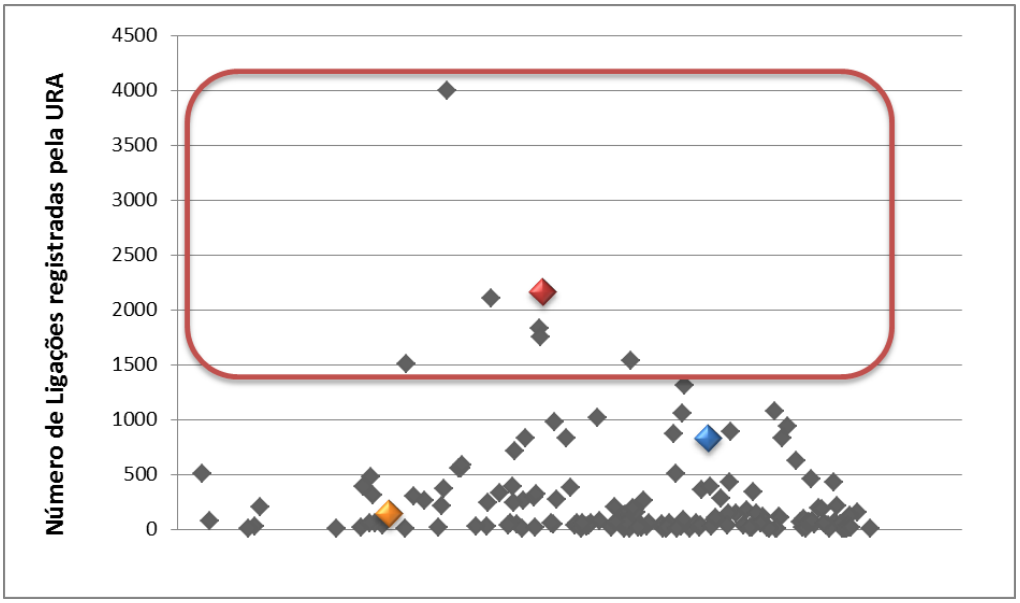



Figura 5.26: Conjunto de dados que pertencem ao *cluster* 4 definidos na Etapa 2.

A partir da definição de quais seriam as divisões dos grupos formados em relação ao número de ligações registradas no período pelos clientes, foi necessário incluir o atributo PERFIL_URA na tabela CLIENTES que enquadra o cliente nestes grupos criados. As informações deste atributo foram cadastradas conforme o script ilustrado pela Figura 5.27.

```
DECLARE
   v_perfil_ura VARCHAR(15);
BEGIN
   FOR s IN (SELECT num_ligacoes
                  , cod_cliente
               FROM clientes)
   LOOP
      IF s.num_ligacoes <= 403 THEN
         v_perfil_ura := 'EXPERIENTE';
      ELSIF s.num_ligacoes > 403 AND s.num_ligacoes <= 1201 THEN
         v_perfil_ura := 'INTERMEDIARIO';
      ELSIF s.num_ligacoes > 1201 THEN
         v_perfil_ura := 'INICIANTE';
      END IF;

      UPDATE clientes
         SET perfil_ura = v_perfil_ura
       WHERE cod_cliente = s.cod_cliente;
   END LOOP;
END;
```

Figura 5.27: Script executado para popular o campo PERFIL_URA da tabela CLIENTES.

Após as interpretações e com as novas informações extraídas no trabalho de mineração de dados armazenadas no atributo PERFIL_URA da tabela CLIENTES, a Etapa 2 do Estudo de Caso estava concluída.

### *5.3.3. Etapa 3: novo agrupamento de clientes conforme número de licenças adquiridas*

A Etapa 3 tem como objetivo criar novos grupos de clientes conforme o número de licenças adquiridas no momento do contrato dos clientes. Os grupos serão nomeados como “BAIXO”, ”MÉDIO” e “ALTO”, determinando a quantidade de licenças de cada cliente.

Assim como na Etapa 1 e 2, nesta etapa também optou-se por utilizar a técnica de Clusterização e o algoritmo K-means. Os passos para a aplicação desta técnica e algoritmo encontram-se detalhados nas seções 5.3.3.1, 5.3.3.2, 5.3.3.3 e 5.3.3.4.

#### *5.3.3.1. Data Source*

Para a Etapa 3, selecionou-se a tabela CLIENTES no Data Source, porém selecionou-se os atributos COD_CLIENTE e NUM_USUARIOS, conforme mostra a Figura 5.28.

#### *5.3.3.2. Técnica*

No momento do processo onde seleciona-se a técnica que será utilizada, assim como na Etapa 1 e 2, a técnica escolhida é a de Clusterização, e o atributo identificador foi o COD_CLIENTE.

Para o algoritmo escolhido, o K-means, foram alteradas as seguintes parametrizações, conforme Figura 5.29:

- *Number of Cluster, Distance Function, Number of Histogram Bins* e *Split Criterion*: já explicados no capítulo 5.3.1.2.

- *Number of Iterations*, *Growth Factor* e *Convergence Tolerance*: já explicados no capítulo 5.3.2.2.

### *5.3.3.3. Resultados*

Assim como nas Etapas 1 e 2 do estudo de caso também podemos identificar 5 *clusters* criados pelo algoritmo (1, 2, 3, 4 e 5), sendo que os *clusters* 1 e 2 são intermediários e foram utilizados para a criação dos demais, conforme ilustra a Figura 5.30.

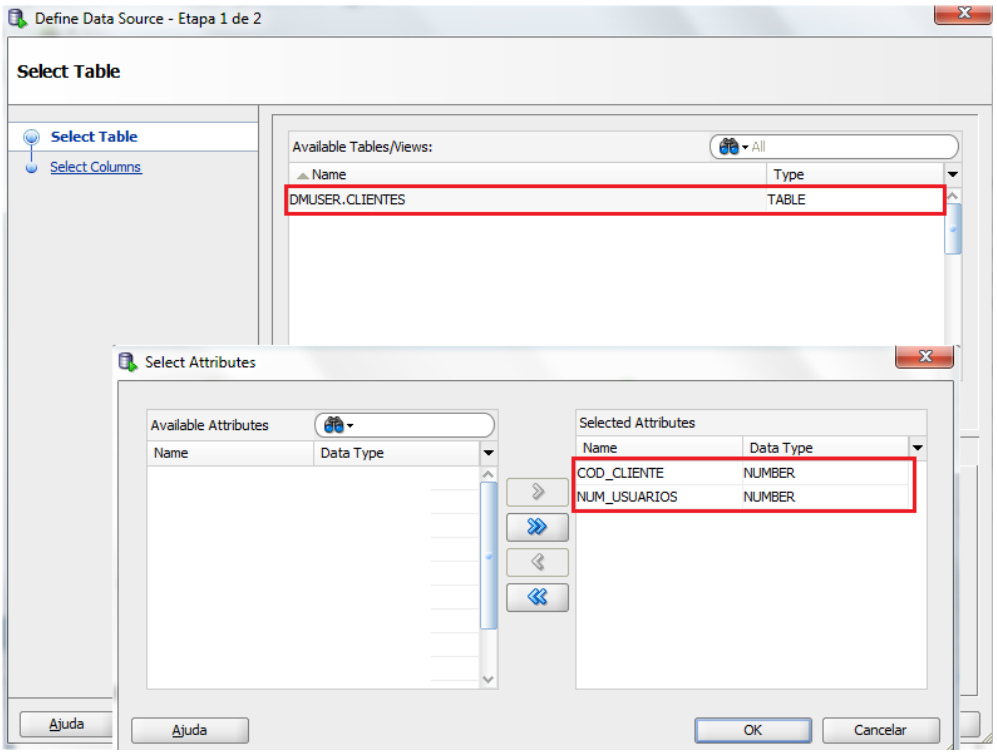



Figura 5.28: Seleção dos atributos da Etapa 3 que serão utilizados no processamento do algoritmo.

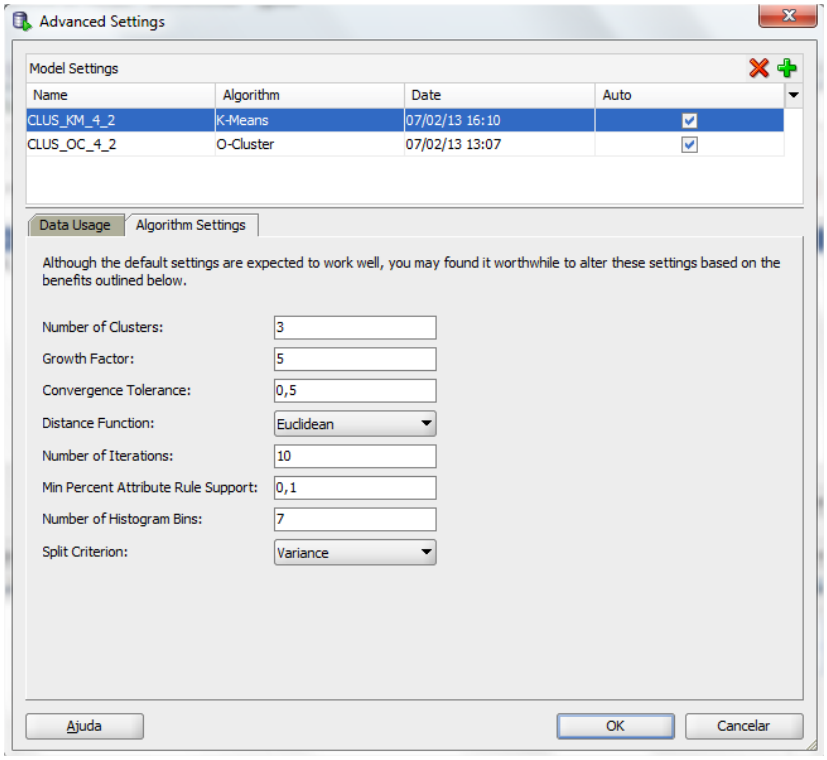



Figura 5.29: Configuração dos parâmetros do algoritmo K-means para Etapa 3.

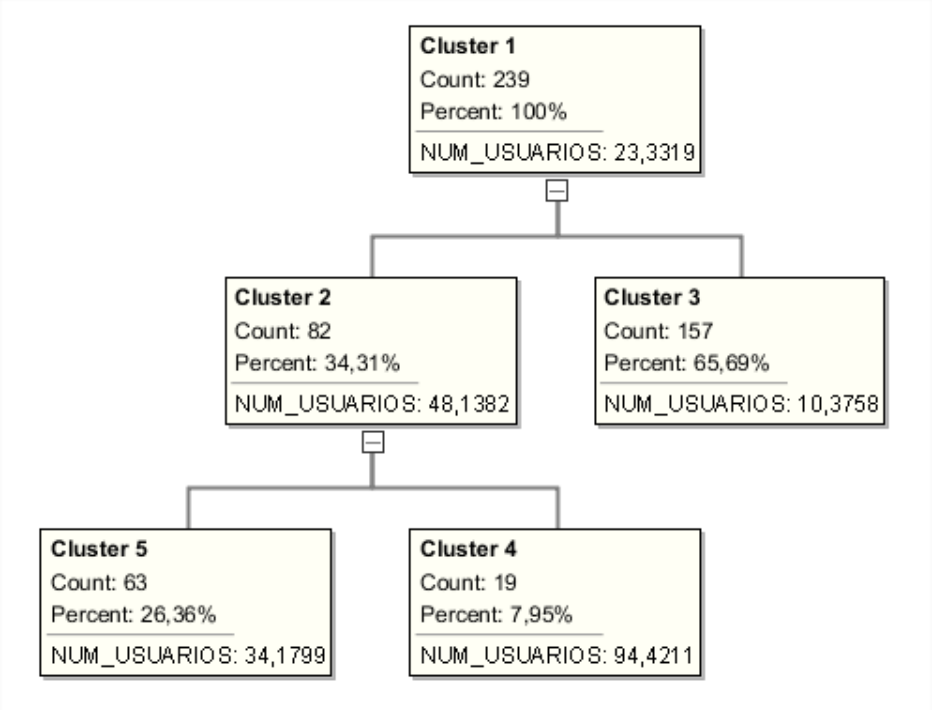



Figura 5.30: *Clusters* criados para a Etapa 3 visualizados pela ferramenta do Oracle.

Podemos visualizar os centróides definidos para o atributo na criação dos clusters através da Figura 5.31, e representado pelo gráfico da Figura 5.32.

Attributes: 1 out of 1

| Attribute | Centroid(3) | Centroid(4) | Centroid(5) |
|---|---|---|---|
| NUM_USUARIOS | 10,37579618 | 94,42105263 | 34,17987195 |

Figura 5.31: Centroides dos 3 *clusters* criados no processamento da técnica na Etapa 3.

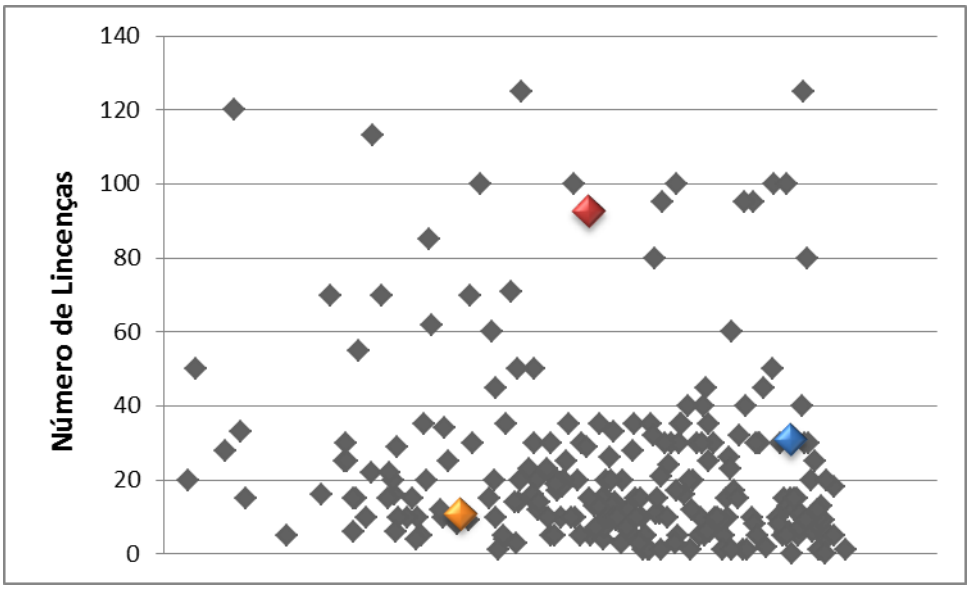



Figura 5.32: Gráfico que representa a dispersão dos dados e os centroides destacados.

### *5.3.3.4. Interpretação*

Ao analisar os detalhes de cada cluster criado e as regras definidas, foram realizadas as seguintes interpretações:

- Quando o número de licenças for menor que 17 o cliente se enquadra no cluster 3 representado pelo grupo “BAIXO”, conforme ilustrado nas figuras 5.33 e 5.34.

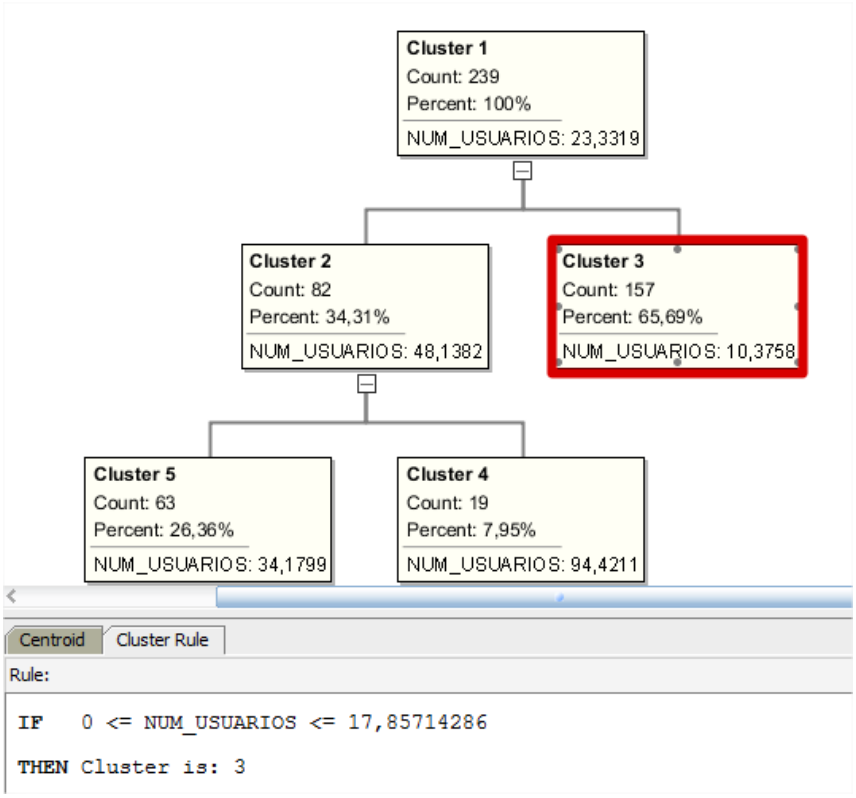



Figura 5.33: Regras geradas para o cluster 3 na Etapa 3.

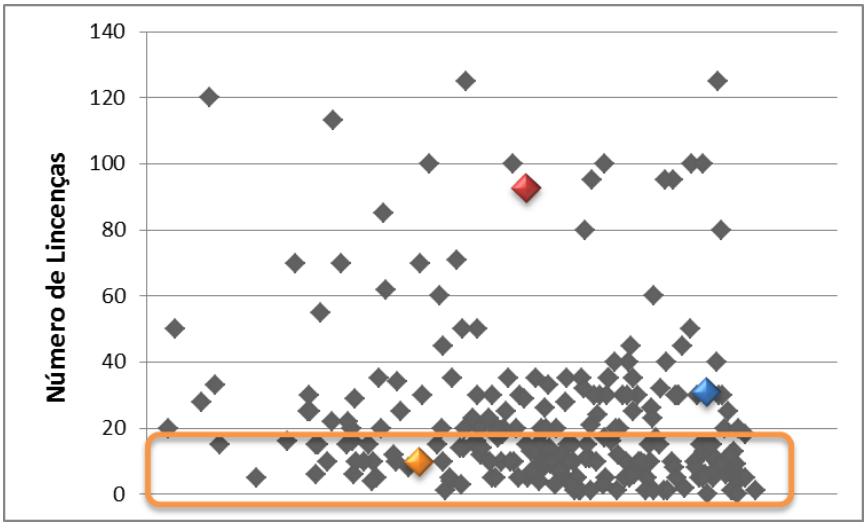



Figura 5.34: Conjunto de dados que pertencem ao cluster 3 definidos na Etapa 3.

- Quando o número de licenças for maior que 17 e menor que 53 o cliente se enquadra no cluster 5 representado pelo grupo “MÉDIO”, conforme ilustrado nas figuras 5.35 e 5.36.

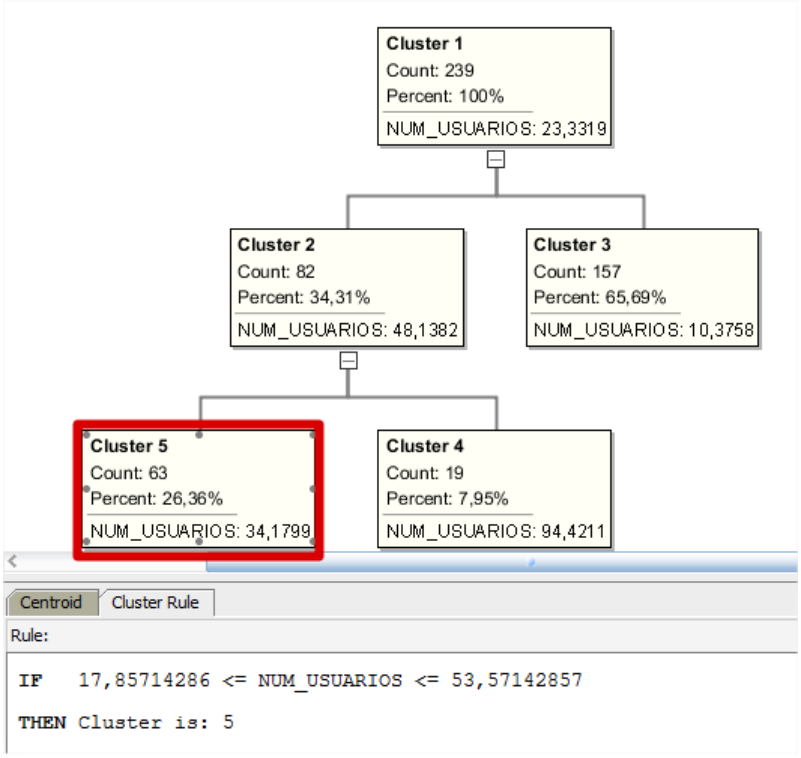



Figura 5.35: Regras geradas para o cluster 5 na Etapa 3.

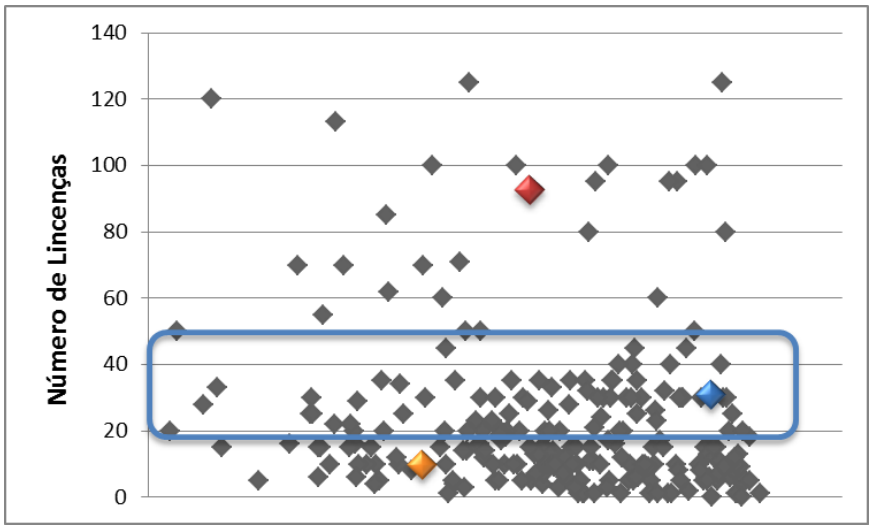



Figura 5.36: Conjunto de dados que pertencem ao cluster 5 definidos na Etapa 3.

- Quando o número de licenças for maior que 53 e menor que 125 o cliente se enquadra no cluster 4 representado pelo grupo “ALTO”, conforme ilustrado nas figuras 5.37 e 5.38.

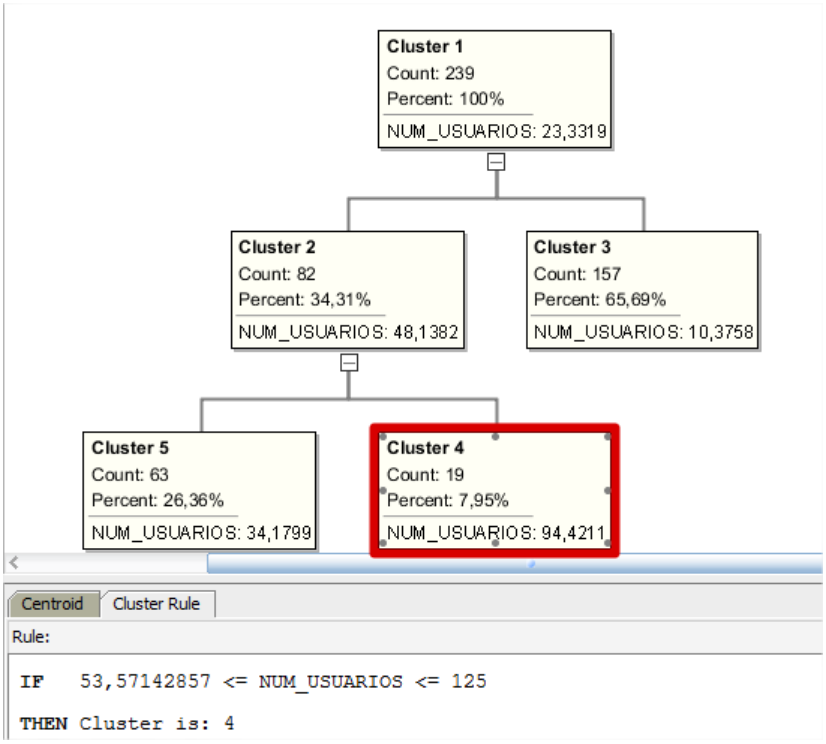



Figura 5.37: Regras geradas para o cluster 4 na Etapa 3.

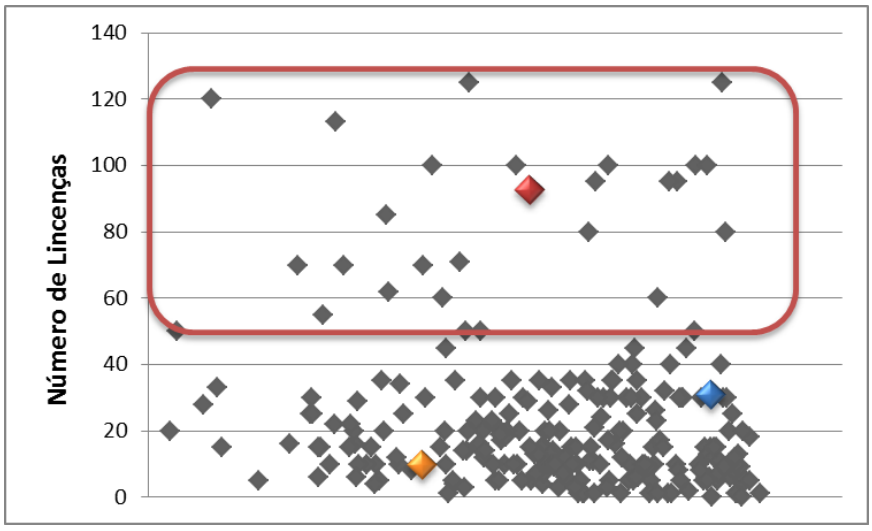



Figura 5.38: Conjunto de dados que pertencem ao cluster 4 definidos na Etapa 3.

A partir da definição de quais seriam as divisões dos grupos formados em relação ao número de licenças adquiridas no momento do contrato com os clientes, foi necessário incluir o atributo PERFIL_LICENCAS na tabela CLIENTES que enquadra o cliente nestes grupos criados. As informações deste atributo foram cadastradas conforme o script ilustrado pela Figura 5.39.

```
DECLARE
   v_perfil_licencas VARCHAR(15);
BEGIN
   FOR s IN (SELECT num_usuarios
                  , cod_cliente
               FROM clientes_dm)
   LOOP
      IF s.num_usuarios <= 17 THEN
         v_perfil_licencas := 'BAIXO';
      ELSIF s.num_usuarios >17 AND s.num_usuarios <= 53 THEN
         v_perfil_licencas := 'MEDIO';
      ELSIF s.num_usuarios > 53 THEN
         v_perfil_licencas := 'ALTO';
      END IF;

      UPDATE clientes_dm
         SET perfil_licencas = v_perfil_licencas
       WHERE cod_cliente = s.cod_cliente;
   END LOOP;
END;
```

Figura 5.39: Script executado para popular o campo PERFIL_LICENCAS da tabela CLIENTES.

Após as interpretações e com as novas informações extraídas no trabalho de mineração de dados armazenadas no atributo PERFIL_LICENCAS da tabela CLIENTES, a Etapa 3 do Estudo de Caso estava concluída.

### *5.3.4. Etapa 4: relação entre os agrupamentos*

A Etapa 4 do Estudo de Caso tem por objetivo realizar uma avaliação para verificar se existe alguma relação entre os grupos criados nas Etapa 1, 2 e 3.

Para realizar esta etapa foi utilizada a técnica de Classificação e o algoritmo *Naive Bayes*, conforme decisão tomada a partir dos casos de uso estudados no capítulo 4.2.1.

O funcionamento e as características para utilização desta técnica e algoritmo na construção do modelo de mineração de dados para esta etapa estão descritos a seguir:

### *5.3.4.1. Data Source*

Na tela de seleção do *data source*, foi selecionada a tabela CLIENTES e os atributos COD_CLIENTE, PERFIL, PERFIL_URA e PERFIL_LICENSAS, estes últimos representando, respectivamente, os resultados das Etapas 1, 2 e 3, conforme ilustra a Figura 5.40.

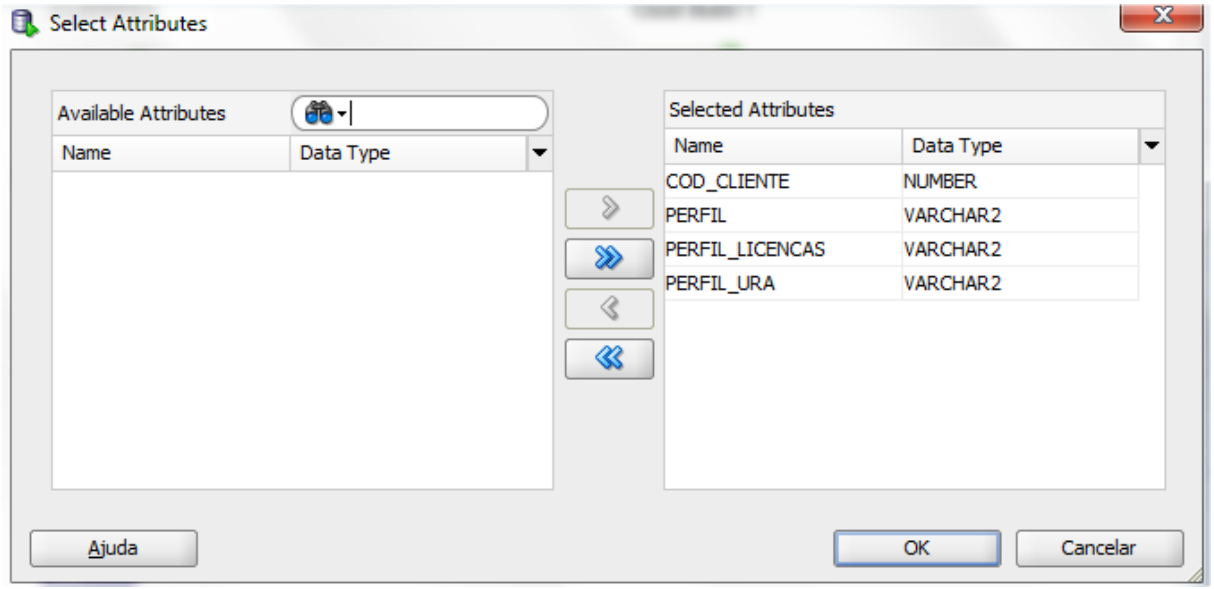



Figura 5.40: Seleção dos atributos da Etapa 4 que serão utilizados no processamento do algoritmo.

### *5.3.4.2. Técnica*

No momento do processo onde seleciona-se a técnica que será utilizada, é feita a escolha da Classificação, e é neste momento que deve ser definido qual o atributo que será o alvo do algoritmo (target) e que servirá de base para as

classificações. O atributo escolhido é o PERFIL_LICENCAS, conforme mostra a figura 5.41.

Da mesma forma que para o algoritmo de Clusterização, a ferramenta executa a técnicas para todos os algoritmos que a mesma disponibiliza, no caso desta etapa do estudo de caso iremos analisar somente o algoritmo Naive Bayes.

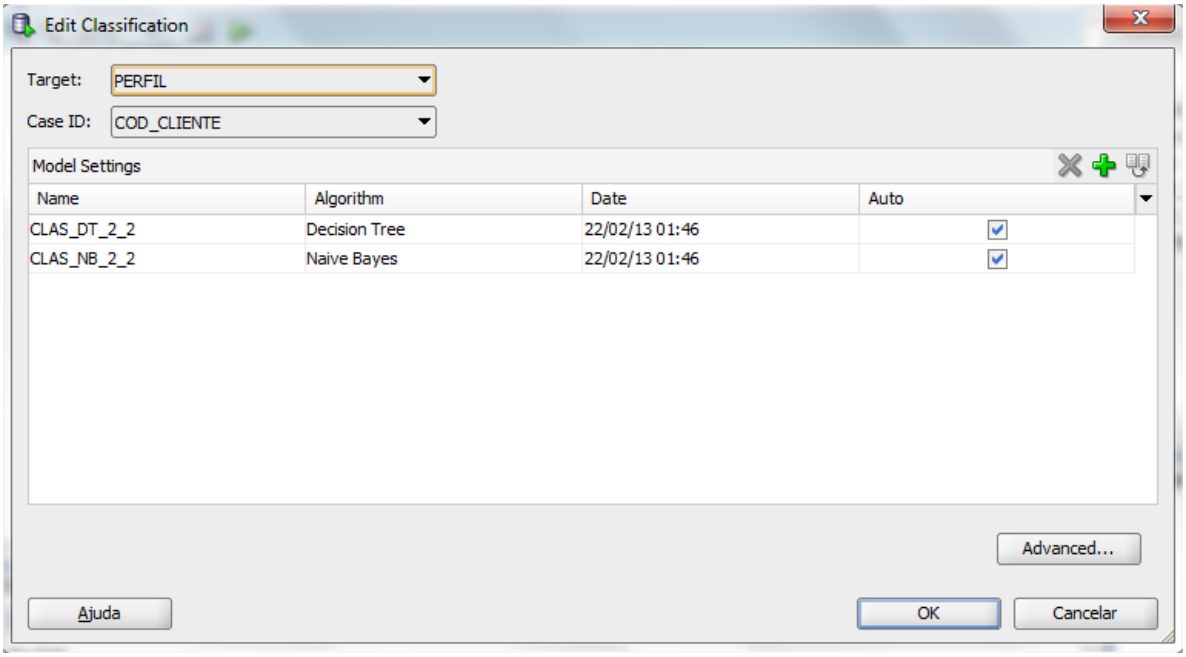



Figura 5.41: Atributo alvo escolhido para execução da Etapa 4.

### *5.3.4.3. Resultados*

Após o processamento, a ferramenta disponibiliza as telas com os resultados do algoritmo Naive Bayes, onde é possível identificar para cada classe definida, qual a probabilidade da combinação dos atributos acontecer.

Para a Etapa 4 do Estudo de Caso foi possível visualizar os seguintes resultados:

- Para os clientes com Perfil de Solicitações classificados como INICIANTE, existe a probabilidade de 74% ter um perfil ALTO ou MÉDIO de Licenças e uma probabilidade de 51% ter um perfil INICIANTE ou INTERMEDIARIO em relação a quantidade de ligações realizada para a empresa Focco. Conforme a figura 5.42.

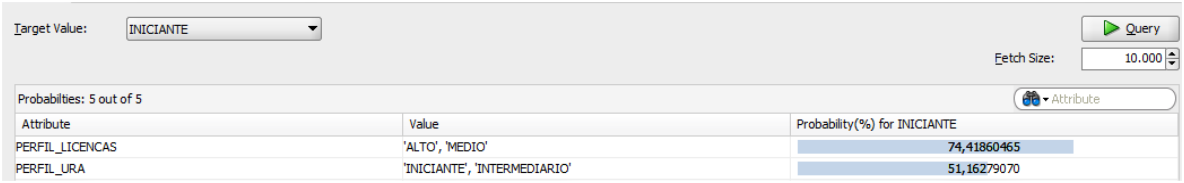



Figura 5.42: Probabilidade em relação ao Perfil de Solicitações INICIANTE dos clientes.

- Para os clientes com Perfil de Solicitações classificados como INTERMEDIARIO, existe a probabilidade de 80% ter um perfil BAIXO de Licenças e uma de 100% ter um perfil EXPERIENTE em relação a quantidade de ligações realizada para a empresa Focco. Conforme a figura 5.43.

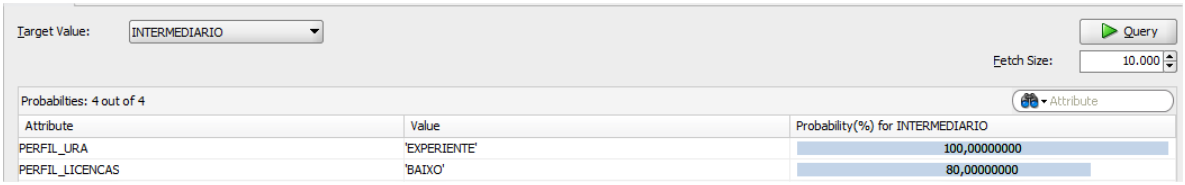



Figura 5.43: Probabilidade em relação ao Perfil de Solicitações INTERMEDIARIO dos clientes.

- Para os clientes com Perfil de Solicitações classificados como EXPERIENTE, existe a probabilidade de 76% ter um perfil BAIXO de Licenças e 100% ter um perfil EXPERIENTE em relação a quantidade de ligações realizada para a empresa Focco. Conforme a figura 5.44.

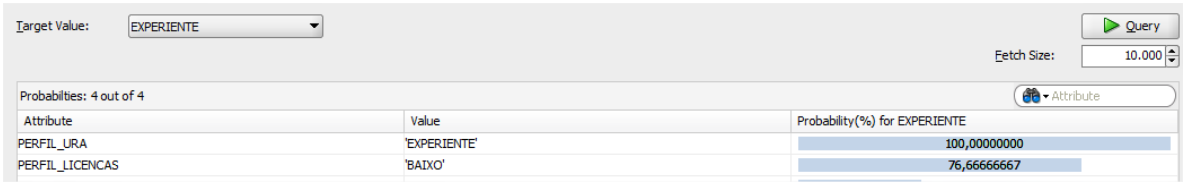



Figura 5.44: Probabilidade em relação ao Perfil de Solicitações EXPERIENTE dos clientes.

## 5.4. INTERPRETAÇÃO E AVALIAÇÃO DOS DADOS

A partir dos resultados e das interpretações encontradas nas quatro etapas do processo de mineração de dados sobre a base de dados da Focco Sistemas de Gestão, foi possível encontrar o gráfico da Figura 5.45 como conclusão.

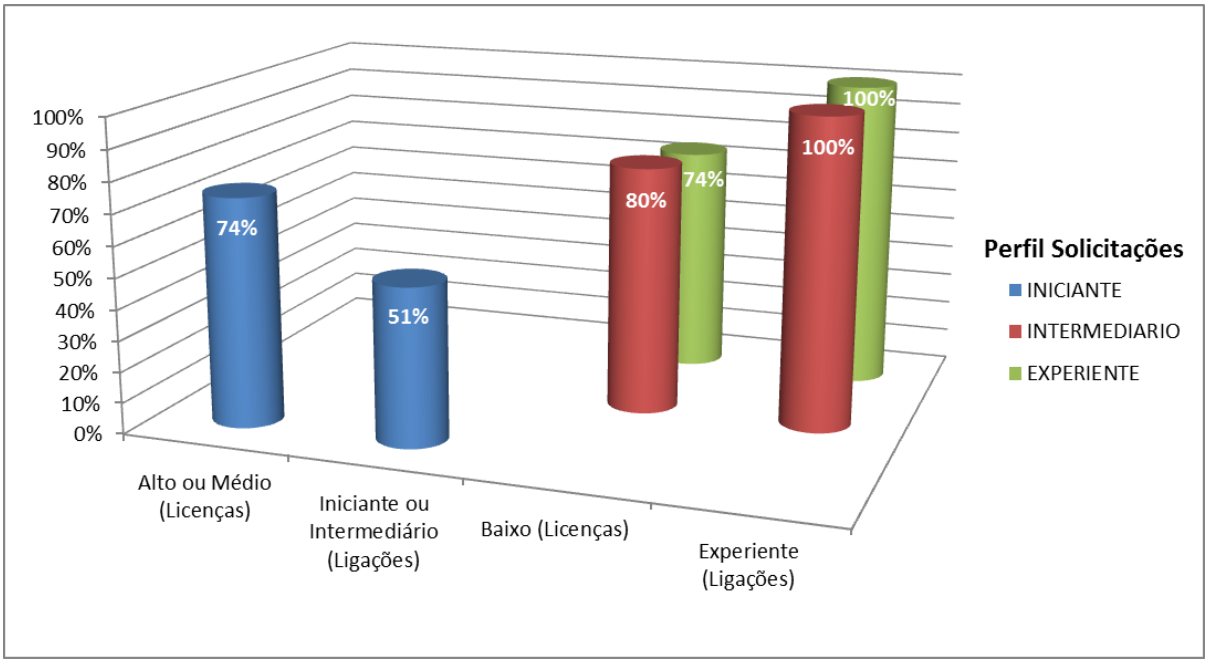



Figura 5.45: gráfico dos resultados do processo de mineração de dados sobre o estudo de caso.

Neste gráfico podemos visualizar a relação entre a quantidade de clientes com determinados Perfil de Solicitações e Perfil de Licenças e Ligações da URA. É possível diretamente encontrar a semelhança do comportamento de clientes com Perfil de Solicitações INTERMEDIÁRIO e EXPERIENTE, e que realmente o que se destaca é o Perfil de Solicitações INICIANTE.

Analisando este tipo de perfil, podemos entender que são clientes que possuem um Perfil de Licenças ALTO ou MÉDIO, ou seja, um número de licenças maior que 17 (conforme relatado no capítulo 5.3.3).

Desde então a Focco Sistemas de Gestão nunca teve nenhuma estatística ou informação relevante sobre seus clientes e seus atendimentos e nunca tomou

alguma ação baseada em estatísticas, sempre supriu as necessidades conforme a demanda sem preparar nem planejar as equipes para as devidas ações.

Contudo, foi conversado com um especialista que conhece todo o processo da Focco e coordena todas as equipes envolvidas neste cenário. Foi apresentado ao qual o cenário do estudo de caso e todos os métodos e as técnicas utilizadas para encontrar os resultados. Algumas conclusões foram extraídas e convertidas em possíveis ações para a empresa, com o objetivo final de melhorar o entendimento sobre seus clientes e melhorar o atendimento com eles.

**Conclusão:** Com tais informações foi possível identificar as tendências do comportamento dos clientes e o que seria o desejável para a Focco. Conforme a experiência do especialista, o ideal para a Focco seriam clientes com um Perfil de Licenças ALTO ou MEDIO porém com um Perfil de Solicitações e de Ligações INTERMEDIARIOS ou EXPERIENTES. Pois pensando em custo para a empresa, quanto mais licenças maior o faturamento com estes clientes e quanto menor a quantidade de solicitações de dúvida e ligações para o Suporte da Focco menor o custo para a empresa.

Clientes com um número alto de licenças possuem uma tendência de custar mais para a Focco devido a grande rotatividade de usuários das empresas. A cada novo usuário utilizando o produto, novas dúvidas e novas ligações ocorrerão, porém para chegar no Perfil de Usuário ideal a empresa deve tomar algumas ações.

Como estratégia para a Focco, as ações poderiam ser tomadas primeiramente para clientes que possuem um Perfil de Ligações INICIANTE, pois no momento da ligação existe a necessidade de ter um recurso exclusivo e urgente para cada cliente/usuário, e no caso das solicitações existe um tempo de espera do cliente que pode fazer com que os recursos da equipe de Suporte sejam subdivido.

**Ação 1:** Para estes clientes que possuem uma rotatividade de funcionários alta a solução é a documentação do software. Atualmente os produtos da Focco possuem pouca documentação sobre os processos e os programas dos produtos. A ação a ser tomada neste caso seria um investimento no HELP do produto e em E-Learning para novos usuários.

**Ação 2:** Deverá ser realizado um trabalho sobre clientes com Perfil de Licenças ALTO ou MEDIO afim que estes possuam um Perfil EXPERIENTE de Solicitações e Ligações. Deverão ser relacionados e elencados os clientes com perfil de INICIANTE, identificando maiores dificuldades de cada cliente, agendando treinamentos com a equipe de Consultoria afim de sanar dúvidas.

**Ação 3:** As ações 1 e 2 podem ser previamente planejadas no momento em que a Focco Sistemas de Gestão realiza uma venda. Identificando que esta venda atinge um número de licenças superior a 17, os treinamentos com a Consultoria e serviços de E-learning já devem ser oferecidos, até mesmo incluindo estes serviços em contratos ou acordos no momento do fechamento.

# 6. CONCLUSÕES

Com o término do Trabalho de Conclusão onde foi realizada primeiramente uma pesquisa bibliográfica referente a Descoberta do Conhecimento em Banco de Dados e após isto a aplicação do conceito no estudo de caso sobre o "Portal de Atendimento" da Focco Sistemas de Gestão, foi possível extrair algumas conclusões.

Uma conclusão extraída após a finalização deste trabalho é que as empresas tem o costume de armazenar um grande volume de dados em banco de dados porém não sabem como extrair informações de forma inteligente e confiável. Muitas vezes acabam utilizando força bruta o que não resulta em grandes conclusões e possíveis ações. Para esta tarefa existem ferramentas que auxiliam e que visam descobrir um conhecimento até então não percebido.

Como existem diversas técnicas e algoritmos que possuem características e funcionalidades diferentes, antes de realizar o trabalho de Mineração de Dados foi necessário definir bem o cenário que será utilizado para a descoberta do conhecimento, pois foi a partir dele que optou-se por escolher tais técnicas e algoritmos.

A ferramenta escolhida, que já se encontra disponível, foi totalmente pensada na melhor desempenho dos algoritmos e melhor integração com o banco de dados já utilizado na base de dados de origem. Além, é claro, de ser sempre muito bem relacionada como uma das melhores e mais seguras ferramentas para mineração de dados em todas as bibliografias pesquisadas.

Um fato importante que vale ressaltar é que iniciar um processo de Descoberta de Conhecimento em Banco de Dados pode ser considerado relativamente fácil, pois o grande desafio e a grande dificuldade encontrada na resolução deste trabalho é encontrar uma finalidade de acordo com as expectativas iniciais. Os passos da Descoberta de Conhecimento que envolviam procedimentos de extração, limpeza e preparação dos dados, foram executados com muito cuidado pois já se sabia que estes poderiam interferir muito no resultado final. E na etapa da

mineração de dados foi-se gasto muito tempo realizando testes, e em cada teste identificando todos os parâmetros das técnicas utilizados.

Para trabalhos futuros novas informações estratégicas para a Focco Sistemas de Gestão podem ser descobertas. Como, por exemplo, confrontando o resultado deste trabalho com os segmentos dos clientes. Ou também identificando qual segmento gera mais lucro e custo para a empresa.

Como conclusão final temos o fato que embora se tenha dificuldades ao realizar um processo de DCBD, este tipo de trabalho nas bases de dados deve ser incorporado ao dia a dia das organizações para principalmente dar suporte na tomada de decisões pelas equipes estratégicas, como foi realizado neste estudo de caso com a empresa Focco Sistemas de Gestão.

# 7. BIBLIOGRAFIA


ANACLETO, A. C. D. S. Aplicação de Técnicas de Data Mining em Extracção de Elementos de Documentos Comerciais. **Tese de Mestrado em Análise de Dados e Sistemas de Apoio à Decisão**, Universidade do Porto - Faculdade de Economia, 2009.

ANDREAZZA, S. P. Utilização da Mineração de Dados no "Portal de Atendimento Focco". **Trabalho de Conclusão de Curso (Bacharel em Sistemas de Informação)** , Universidade de Caxias do Sul, 2009.

ARAUJO, J. ; PEREIRA, S. E. Aplicação de Data Mining na área de CRM como ferramenta gerencial para tomada de decisão em empresas modernas. **Revista UNIABEU**, p. 144-160, 2011.

CHAPMAN; HALL. **The Top Ten Algorithms in Data Mining**. [S.l.]: [s.n.], 2009.

CHIARA, R. Aplicação de técnicas de data mining em logs de servidores web. **Instituto de Ciências Matemáticas e de Computação - USP**, 2003.

DIAS, M. M. Parâmetros na escolha de técnicas e ferramentas de mineração de dados. **Acta Scientiarum**, Maringá, v. 24, p. 1715-1725, 2002.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, v. 17, p. 38-54, 1996.

GOLDSCHMIDT, R. **Data mining:** um guia prático. Rio de Janeiro: [s.n.], 2005.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining:** Concepts and Techniques. [S.l.]: Morgan Kaufmann, 2012.

IBM. **IBM DB2 Intelligent Miner**. Disponivel em: <http://www.ibm.com/developerworks/data/library/tutorials/iminer>. Acesso em: 2012.

JACOBSON, ; MISNER , ; CONSULTING,. **MICROSOFT SQL SERVER 2005 ANALYSIS SERVICES - Passo a Passo**. [S.l.]: Bookman, 2005.

ORACLE. **Oracle Data Mining Mining Techniques and Algorithms**. Disponivel em: <http://www.oracle.com/technology/products/bi/odm/>. Acesso em: 2012.

PIMENTEL, E. P.; OMAR, N. Descobrindo Conhecimentos em Dados de Avaliação da Aprendizagem com Técnicas de Mineração de Dados. **Anais do XXVI Congresso da SBC**, Campo Grade, MS, p. 147-155, Julho 2006.

PUHL, E. Estudo de Perfil de Usuários de Redes Sociais através da Mineração de Dados. **Trabalho de Conclusão de Curso (Bacharel em Sistemas de Informação)**, Universidade de Caxias do Sul, 2011.

SOUZA, S. L. D. Evasão no Ensino Superior: Um estudo utilizando a mineração de dados como ferramenta de gestão do conhecimento em um Banco de Dados referente a Graduação de Engenharia. **Universidade Federal do Rio de Janeiro**, Rio de Janeiro, 2008.

VIEIRA, R. D. S. Implantação de Business Intelligence no sistema NL gestão. **Trabalho de Conclusão de Curso (Bacharel em Sistemas de Informação)**, 2012.

WAIKATO. **Data Mining with Open Source Machine Learning Software in Java**. Disponivel em: <http://www.cs.waikato.ac.nz/ml/weka/>. Acesso em: Setembro 2012.

WITTEN, I. H.; FRANK, E. **Data Mining:** Practical Machine Learning Tools and Techniques with Java Implementations. [S.l.]: Morgan Kaufmann, 1999.