UNIVERSIDADE FEDERAL DO MARANHÃO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
DEPARTAMENTO DE INFORMÁTICA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

Jefferson Ferreira Sá

MINERAÇÃO DE DADOS USANDO O ALGORITMO CHAID

São Luís
2012

Jefferson Ferreira Sá

MINERAÇÃO DE DADOS USANDO O ALGORITMO CHAID

Monografia apresentada ao curso de Ciência da Computação da Universidade Federal do Maranhão, como parte dos requisitos necessários para obtenção do grau de Bacharel em Ciência da Computação.

Orientadora: Prof. Msc. Simara Vieira da Rocha

São Luís

2012

Sá, Jefferson Ferreira

Mineração de dados usando o algoritmo CHAID / Jefferson Ferreira Sá. – São Luís, 2012.

69f.

Impresso por computador (Fotocópia).

Orientadora: Simara Vieira da Rocha.

Monografia (Graduação) – Universidade Federal do Maranhão, Curso de Ciência da Computação, 2012.

1.Algoritmo – Informática 2.KDD 3.Classificação 4.Método CHAID 5.Árvore de decisão I. Título.

CDU 004.421

Jefferson Ferreira Sá

MINERAÇÃO DE DADOS USANDO O ALGORITMO CHAID

Monografia apresentada ao curso de Ciência da Computação da Universidade Federal do Maranhão, como parte dos requisitos necessários para obtenção do grau de Bacharel em Ciência da Computação.

Aprovada em: 13/07/2012

BANCA EXAMINADORA

Simara Vieira da Rocha

Prof. Msc. Simara Vieira da Rocha (Orientadora)
Universidade Federal do Maranhão

Portela

Prof. Msc.Carlos Eduardo Portela Serra de Castro
Universidade Federal do Maranhão

Maria Auxiliadora Freire

Prof. Msc. Maria Auxiliadora Freire
Universidade Federal do Maranhão

AGRADECIMENTOS

Primeiramente a Deus, por ter me dado forças durante toda essa trajetória.

Aos meus pais e meus irmãos que contribuíram com grande incentivo em toda minha jornada acadêmico.

À minha orientadora Prof. Msc. Simara Vieira da Rocha por aceitar a coordenação deste trabalho de conclusão de curso, pois sem sua orientação, dedicação e auxílio, o estudo aqui apresentado seria praticamente impossível.

A todos aqueles que de alguma forma tenham contribuído para realização deste trabalho, em especial a Fábio Batista, Márcio Frazão, Antônio Júnior, Pablo Fernandes, Valéria Priscilla e Nadson Timbó por todo apoio e incentivo.

A todos os meus amigos e familiares indispensáveis nessa jornada.

“A mente que se abre a uma nova ideia, jamais volta ao seu tamanho original”

(Albert Einstein)

RESUMO

O processo de Descoberta de Conhecimento em Bases de Dados ou *Knowledge Discovery in Database* (KDD) é a técnica de identificar padrões úteis em dados não processados. Essa técnica possui diversas etapas, sendo o processo de Mineração de Dados a mais importante na extração de conhecimento. Entre os diversos algoritmos usados nessa etapa, explicamos a fundo uma das técnicas de classificação de dados mais antigas, o método CHAID. Esse trabalho, além de esclarecer o algoritmo CHAID, buscou aplicá-lo a um estudo de caso que usando os registros presentes na base de dados da Pesquisa Nacional por Amostra de Domicílios (PNAD), realizada em todo Brasil pelo Instituto Brasileiro de Geografia e Estatística (IBGE) no ano de 2008, gerou uma Árvore de Decisão que traçou o perfil socioeconômico dos entrevistados relacionando o nível de escolaridade com a situação econômica dos mesmos. Com os resultados percebemos que entre os entrevistados com as menores rendas, a maioria possui apenas o ensino fundamental, sendo que o aumento do rendimento mensal das pessoas está diretamente relacionado ao grau de instrução delas, onde aquelas que apresentam os maiores rendimentos do Brasil, grande parte tem nível superior completo.

Palavras-chave: KDD, Classificação, Método CHAID, Árvore de decisão.

ABSTRACT

The process of Descoberta de Conhecimento em Bases de Dados or Knowledge Discovery in Database (KDD) is the technique to identify useful patterns in unprocessed data. This technique has several stages, the process of Data Mining in the most important knowledge extraction. Among the various algorithms used in this step, we explain the background of the techniques for data classification older method CHAID. This work, besides clarifying the CHAID algorithm, we tried to apply it to a case study using the records present in the database from the Pesquisa Nacional por Amostra de Domicílios (PNAD), conducted throughout Brazil by the Instituto Brasileiro de Geografia e Estatística (IBGE) in 2008, generated a Decision Tree that traced the socioeconomic profile of respondents relating to educational level with the economic situation of the same. With the results we see that among those with the lowest incomes, most have only primary education, and increasing the monthly income of people is directly related to schooling of them, where those with the highest yields in Brazil, a large part has college degrees.

Keywords: KDD, Classification, CHAID method, Decision tree.

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

LISTA DE SIGLAS

AID - *Automatic Iteration Detector*

CART - *Classification and Regression Trees*

CHAID - *Chi-square Automatic Interaction Detection*

DM - *Data Mining*

DW - *Data Warehounsing*

GL - Graus de Liberdade

IBGE - Instituto Brasileiro de Geografia e Estatística

ID3 - *Iterative Dichotomizer* 3

KDD - *Knowledge Discovery in Database*

MD - Mineração de Dados

PNAD - Pesquisa Nacional por Amostra de Domicílios

QUEST - *Quick, Unbiasied, Efficient Statistical Tree*

SUMÁRIO

1. INTRODUÇÃO

Nos últimos anos, em que a maioria das operações e atividades das instituições privadas e públicas é registrada computacionalmente e acumula-se em grandes bases de dados, existe a necessidade de extrair conhecimento dessas fontes de dados, a fim de descobrir relações ocultas, padrões e regras para predizer e correlacionar dados, que podem ajudar as instituições nas tomadas de decisões (PACHECO et al., 1999).

O conjunto de técnicas para descobrir padrões ou extrair conhecimento em dados não processados é chamada de Descoberta de Conhecimento em Bases de Dados, ou em inglês *Knowledge Discovery in Database* (KDD). O KDD é apoiado em técnicas de Mineração de Dados, ou *Data Mining* (DM), que transforma dados em informação (MANNILA, 1997 apud KLEINSCHMIDT, 2007).

Hoje, a informação e o conhecimento são prerrogativas estratégicas e imprescindíveis na busca de maior autonomia nas ações das empresas, controle social e na tomada de decisão com prazos cada vez mais curtos. Por isso, diversas empresas nacionais e internacionais de produção, consumo, mercado financeiro e instituições de ensino já adotaram nas suas rotinas o DM para monitorar arrecadações, consumo de clientes, prevenir fraudes além da previsão de riscos do mercado, dentre outras (DIAS, 2002).

Os algoritmos e os métodos usados no processo de KDD provêm de diversas áreas, como por exemplo, a Estatística. Este processo envolve o uso de algumas tarefas de KDD, tais como: Classificação, Associação e Agrupamento. Tais tarefas utilizam técnicas de DM baseadas em Redes Neurais Artificiais, Árvores de Decisão, Algoritmos Genéticos, Métodos Bayesianos, entre outras (CARVALHO, 2001).

Entre as diversas técnicas e algoritmos usados no processo de KDD, o método CHAID foi escolhido, pois é um método utilizado quando a segmentação é definida em termos de características demográficas aceitando variáveis categóricas nominais ou ordinais como variáveis dependentes. Normalmente, este tipo de variável é utilizada em pesquisas tradicionais que possuem questões demográficas como sexo, faixa-etária e renda salarial.

O método CHAID é um dos algoritmos mais antigos que usa testes de qui-quadrado de *Pearson* em uma tabela de contingência entre as categorias da variável

dependente e as categorias das variáveis independentes. É um método estatístico extremamente eficiente para a classificação de dados (KASS, 1980 apud RODRIGUES, 2005).

Além da vantagem do método CHAID utilizar variáveis dependentes e independentes em diversas formas como nominais, ordinais e categóricas, o algoritmo se trata de Árvores de Decisão onde seus resultados gráficos são de fácil interpretação o que melhora a compreensão do conhecimento obtido.

Sendo assim, essa monografia tem como objetivo geral, explicar o algoritmo CHAID e aplicá-lo a um estudo de caso. Como objetivos específicos:

- Estudar o processo de *Knowledge Discovery in Database* (KDD), a fim de entender todas as fases necessárias para a extração de conhecimento em base de dados.
- Estudar a Mineração de Dados, pois é a etapa mais importante no processo de KDD.
- Estudar a tarefa de Classificação do KDD, a fim de compreender essa tarefa a qual o método CHAID pertence.
- Estudar a técnica de Árvores de Decisão, pois sua compreensão é essencial para interpretar os resultados gerados pelo CHAID.
- Estudar o algoritmo CHAID, pois esse é o objeto de estudo dessa monografia.
- Aplicar o algoritmo a um estudo de caso para mostrar como interpretar seus resultados.

Nessa monografia, utilizaremos no estudo de caso, as informações da base de dados da Pesquisa Nacional por Amostra de Domicílios (PNAD), realizada em todo Brasil pelo Instituto Brasileiro de Geografia e Estatística (IBGE) no ano de 2008, que possui diversas informações sobre os entrevistados e suas famílias, a fim de traçar um perfil socioeconômico relacionando o nível de escolaridade das pessoas com a situação econômica da mesma.

O presente trabalho será composto de mais três capítulos, conforme descrição sumária a seguir:

Capítulo 2 – apresenta a fundamentação teórica com os conceitos de KDD, suas fases e principais tarefas, além da técnica de Árvore de Decisão. É nesse capítulo que apresentamos também de forma detalhada o método CHAID.

Capítulo 3 – nele mostramos o nosso estudo de caso que usa os registros presentes na base de dados PNAD, o procedimento seguido para utilizar o método CHAID e os resultados encontrados pelo algoritmo.

Capítulo 4 – apresenta as conclusões levantadas com o conhecimento minerado e as sugestões para trabalhos futuros.

2. FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão abordados temas importantes para compreensão deste trabalho, onde partiremos da descoberta de conhecimento em bases de dados, passando pela mineração de dados até chegarmos às Árvores de Decisão e, enfim, o método *Chi-square Automatic Interaction Detection* (CHAID) que é o objeto de estudo dessa monografia.

2.1 Knowledge Discovery in Database (KDD)

A técnica de descobrir padrões úteis em dados não processados recebeu diversos nomes, dentre eles *Knowledge Discovery in Database* (KDD) ou em português, Descoberta de Conhecimento em Bases de Dados, *Data Mining* (DM), ou Mineração de Dados, extração de conhecimento, descoberta de informação, processamento de padrões de dados. O termo KDD foi criado em 1989 para referenciar o processo de descoberta de conhecimento em dados e, principalmente, a etapa de mineração de dados, esta que transforma dados em informação (MANNILA, 1997 apud KLEINSCHMIDT, 2007).

O KDD é o processo não trivial de identificar em dados padrões que sejam válidos, novos (ainda não identificados), potencialmente úteis e compreensíveis, visando melhorar o entendimento de um problema ou um procedimento de tomada de decisão (FAYYAD et al. 1996 apud KLEINSCHMIDT, 2007).

Também é definido como uma técnica que possibilita analisar grandes conjuntos de dados, utilizando métodos aproximados. A metodologia do KDD constitui-se em duas etapas principais, o armazenamento dos dados e o DM. Primeiro faz-se necessário criar uma base de dados organizada e com dados suficientes sobre o assunto a analisar, em seguida utilizar métodos aproximados que permitem minerar os dados, para a descoberta das relações contidas em tais dados (COLLAZOS; BARRETO, 2003 apud KLEINSCHMIDT, 2007).

A técnica de KDD possui ferramentas poderosas para a exploração eficiente de informações em grandes bancos de dados, na intenção de auxiliar na tomada de decisão, sendo o DM uma das fases do processo de maior importância.

Toda a metodologia pode ser divida em uma seqüência de cinco etapas: o *Data Warehounsing* (DW), o Pré-processamento, o Enriquecimento, a Mineração de Dados e o Pós-processamento. A Figura 2.1 representa o processo de KDD de forma hierárquica destacando fases e tarefas (AURÉLIO et al., 1999).

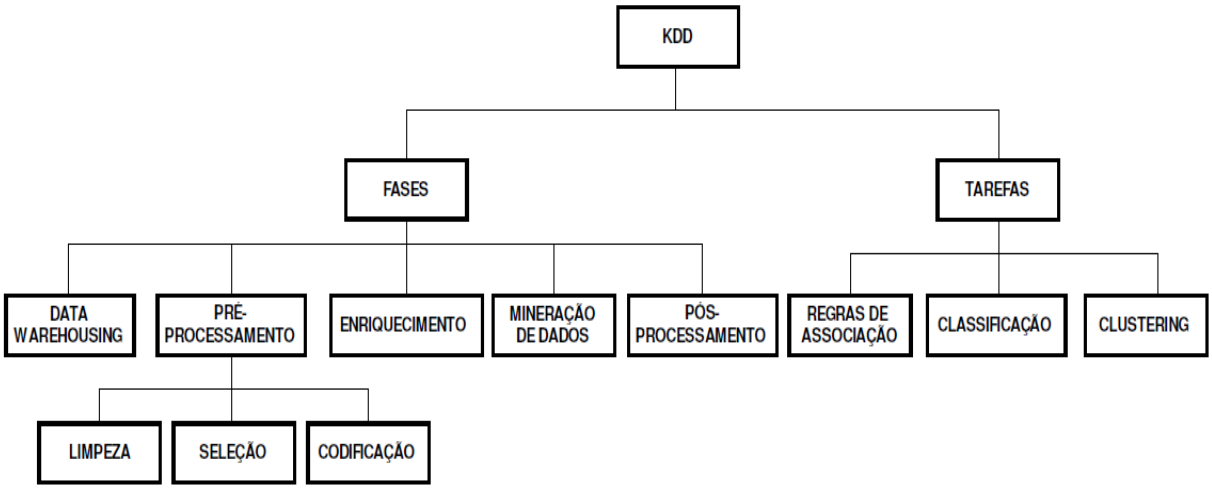


Figura 2.1- Visão hierárquica do processo de KDD (AURÉLIO et al., 1999)

2.1.2 *Data Warehounsing* (DW)

O *Data Warehounsing* (DW) é um banco de dados desenvolvido com a finalidade de dar suporte ao processo de decisão, onde os dados contidos nele são obtidos através dos bancos de dados dos aplicativos disponíveis em um ambiente corporativo que tem por objetivo fornecer uma imagem única da realidade do negócio.

Além disso, é uma arquitetura que organiza, totaliza e armazena os dados permitindo traçar estratégias baseadas nos assuntos da empresa de maneira confiável dando suporte ao nível gerencial nas tomadas de decisão.

Ele também pode ser definido como um repositório integrado, orientado para análise, com dados destinados para serem utilizados como base de dados para suporte a decisão.

O DW surgiu com a intenção de suprir as carências nos sistemas de bancos de dados tradicionais, no que diz respeito à exploração e análise dos dados, integrando e consolidando bases de diferentes acervos.

2.1.3 Pré-Processamento

Esta etapa tem a intenção de consolidar as informações relevantes para o algoritmo minerador, com o objetivo de reduzir a complexidade do problema em questão (PACHECO et al., 1999). Este processo pode ser dividido em três etapas:

- Limpeza de dados: etapa responsável por fazer a correção de possíveis erros existentes e eliminação de valores nulos e redundantes. Ela melhora a base de dados eliminando consultas desnecessárias que seriam executadas pelo algoritmo de mineração e que afetariam o seu processamento.
- Seleção de dados: fase que tem por objetivo escolher os atributos mais importantes em todo o conjunto de atributos existentes na base de dados.
- Codificação dos dados: etapa que divide os valores contínuos dos atributos em uma lista de intervalos, convertendo valores quantitativos em valores categóricos, cujo o objetivo é facilitar a qualidade de resultados.

2.1.4 Enriquecimento dos dados

Têm como objetivo agregar mais informações aos registros existentes, enriquecendo os dados, podendo ser realizadas pesquisas para acrescentar conhecimento, como consultas a bases de dados externas, entre outras técnicas (DIAS, 2002).

A tarefa de enriquecimento dos dados consiste melhorar a informação contida nos registros dos bancos de dados através da criação de novos atributos a partir dos já existentes, agregando novas informações. A geração de totalizadores em variáveis numéricas, a criação de faixa ou classes de valores para atributos contínuos e a generalização de valores de atributos são exemplos de enriquecimento dos dados (CARVALHO, 2001).

2.1.5 Mineração de Dados (MD)

É uma área multidisciplinar que incorpora técnicas utilizadas em diversas áreas como Inteligência Artificial (especialmente a aprendizagem de máquina),

Banco de Dados (recursos para manipular grandes volumes de dados) e Estatística (na avaliação e validação dos dados).

O objetivo principal é descobrir os relacionamentos entre dados e fornecer uma fonte para que possa ser feita uma previsão de tendências futuras baseadas em dados históricos (DIAS, 2002).

Além disso, consiste em obter informações através de uma base de dados existente, usando seus atributos para extrair informações que não são triviais e que precisam ser trabalhadas para serem úteis na tomada de decisão, através da utilização de algoritmos para identificar padrões nos dados analisados (SILVEIRA, 2003).

Pode ser definida também como a utilização de técnicas automáticas de exploração de grandes quantidades de dados de forma a reconhecer novos padrões e relações que, devido ao volume de informações, não seriam facilmente descobertos (CARVALHO, 2001).

A MD é de grande importância em inúmeros tipos de aplicações, como por exemplo, na área de segurança para detectar fraudes em cartões de créditos, análises financeiras e de investimentos, detecção e predição de erros em grandes empresas, análise de informações, limpeza em bases de dados, *marketing* e melhoria no processo industrial entre outros.

Também tem como objetivos a previsão e descrição de modelos. A previsão pode ser obtida através da utilização de variáveis contidas no banco de dados para prever valores desconhecidos ou futuros. A descrição envolve a descoberta de padrões interpretáveis pelos humanos. Dentro do processo de KDD descrever modelos possui maior importância que prever os mesmos. A previsão e a descrição dos modelos são conseguidas selecionando as tarefas, algoritmos e técnicas de extração de dados (FAYYAD et al., 1996 apud KLEINSCHMIDT 2007).

Os algoritmos e as técnicas usados para se criar modelos a partir de dados, provém de diversas áreas como Reconhecimento de Padrões e Estatística. Estas técnicas, muitas vezes, podem ser combinadas para se obter melhores resultados (SILVA, 2003).

As tarefas de KDD são: Classificação, Associação e Agrupamento. Estas tarefas podem utilizar técnicas de MD baseadas em Redes Neurais Artificiais, Árvores de Decisão, Estatística, entre outras.

O Quadro 2.1 mostra as principais tarefas de KDD e as técnicas mais utilizadas para a Mineração de Dados.

Quadro 2.1 - Tarefas e Técnicas de KDD

Tarefas de KDD	Técnicas
Associação	Estatística e Teoria dos Conjuntos
Classificação	Árvores de Decisão, Redes Neurais e Algoritmos Genéricos
Agrupamento ou *Clustering*	Redes Neurais e Estatística
Previsão de Séries Temporais	Lógica Nebulosa e Redes Neurais

Fonte: AURÉLIO et al., 1999.

Como podemos observar no Quadro 2.1, cada tarefa está diretamente relacionada ao domínio da aplicação e interesse do usuário e também que cada uma possui um conjunto de técnicas. Cada tarefa de KDD extrai um tipo diferente de conhecimento do banco de dados, com isso, necessita de um algoritmo diferente para realizá-la.

2.1.5.1 Associação

A tarefa de associação permite relacionar a ocorrência de um determinado conjunto de itens com as ocorrências de outro conjunto de itens.

As regras de associação procuram determinar que fatos ocorram simultaneamente com probabilidade razoável ou que itens estão presentes juntos com certa chance (CARVALHO, 2001).

Em outras palavras, as regras de associação identificam afinidades entre registros de um subconjunto de dados. Sendo essas afinidades/associações expressas na forma de regras (BAPTISTA; CARVALHO, 2003).

Tais regras caracterizam o quanto a presença de um conjunto de itens nos registros de uma base de dados implica na presença de algum outro conjunto distinto de itens nos mesmos registros (AGRAVAL; SRIKANT, 1994 apud DOMINGUES, 2004).

Como exemplo podemos citar as redes varejistas que usam regras de associação para planejar a disposição dos produtos nas prateleiras das lojas, na

intenção de chamar a atenção do cliente, colocando os itens que geralmente são adquiridos na mesma compra, próximos um do outro.

2.1.5.2 Classificação

É a tarefa mais estudada em KDD e tem como objetivo descobrir um conhecimento que possa ser utilizado para prever a classe de um registro (AURÉLIO et al., 1999).

É uma técnica que tem por objetivo descobrir um relacionamento entre um atributo meta, pré-definido, e um conjunto de atributos, buscando classificar uma população de registros através da aplicação em um conjunto menor de dados, a fim de desenvolver um modelo de classificação (BAPTISTA; CARVALHO, 2003).

Regras do tipo SE... ENTÃO... , representa uma forma simbólica de classificação e possuem o seguinte formato:

- SE <antecedente> ENTÃO <conseqüente>

O antecedente é composto por expressões condicionais envolvendo atributos do domínio da aplicação existentes no banco de dados. Já o conseqüente, é formado por uma expressão que evidencia algum valor para um atributo meta, descoberto em função dos valores contidos nos atributos que compõem o antecedente (ROMÃO, 2002).

Sendo assim, as regras de classificação podem ser interpretadas como: SE os atributos preditivos de uma tupla satisfazem as condições no antecedente da regra, ENTÃO a tupla tem a classe indicada no conseqüente da regra.

As Árvores de Decisão são consideradas pela comunidade científica, como uma importante técnica usada na tarefa de classificação, devido a sua representação simples, intuitiva e de fácil compreensão.

Para um melhor entendimento sobre os vários algoritmos utilizados em classificação, podemos observar a Figura 2.2 que esquematiza as relações de funcionamento e utilização de vários métodos e modelos.

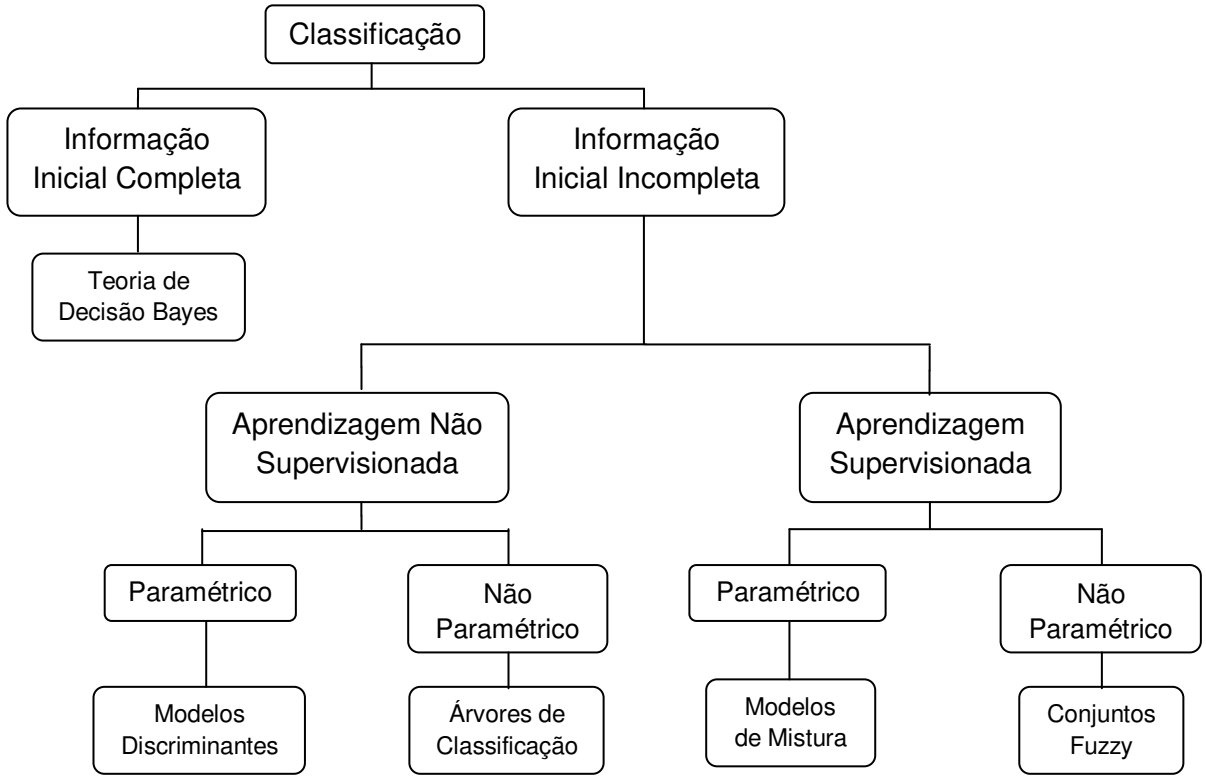


Figura 2.2 - Hierarquia de Classificação (RODRIGUES, 2005)

A classificação é um processo de discriminação de unidades concretas ou abstratas em classes ou categorias, ou de forma abreviada, efeito ou processo de distribuir por classes. Se estas classes estiverem definidas e existir informação sobre a probabilidade de um determinado objeto pertencer a uma classe, entende-se como informação inicial completa, caso contrário, esta informação inicial estará incompleta. Com base no tipo de informação disponível o passo seguinte passa pela utilização e escolha de um método. Os métodos podem ser classificados em paramétricos e não paramétricos (RODRIGUES, 2005).

- Métodos paramétricos: nos métodos paramétricos a distribuição da população tem uma dada forma e as inferências, condicionadas por esse pressuposto, dizem respeito a um ou a vários parâmetros, por exemplo, a Regressão Linear e os Modelos Discriminantes.
- Métodos não-paramétricos: nesse método, a forma da distribuição da população não é conhecida e as inferências processam-se em quadro muito menos restrito e muitas vezes não envolvem parâmetros, por exemplo, as Árvores de Decisão e Análise de *Cluster*.

A principal distinção em relação ao paradigma de aprendizagem, válido para todo tipo de sistemas com capacidade de adaptação, é a aprendizagem supervisionada e aprendizagem não supervisionada.

Na aprendizagem supervisionada, cada exemplo de treino está acompanhado por um valor alvo. Sendo assim, utiliza-se uma variável dependente com informação sobre as classes a que pertencem cada uma das entidades da amostra de treino. Neste conceito, incluem-se técnicas da estatística multivariada como a regressão, Análise Discriminante e as Árvores de Decisão.

O processo de aprendizagem não supervisionada é feito por meio de observação e descoberta. O número de categorias ou classes não estar definido inicialmente, assim o método tem que encontrar atributos estatísticos relevantes buscando descobrir padrões ou semelhanças entre os dados a fim de agrupá-los, ou seja, na aprendizagem não supervisionada os algoritmos assumem sempre que não se conhece a que classe pertence a coleção de dados e tenta uni-los de acordo com semelhanças encontradas. Enquadram-se neste conceito, os algoritmos de análise de grupos, os modelos de mistura e as redes neuronais não supervisionadas.

2.1.5.3 Agrupamento

Esta tarefa tem como principal característica descobrir classes utilizando a similaridade dos valores de seus atributos como fator de decisão. O agrupamento é um método que busca, baseado em medidas de semelhança, definir quantas e quais classes existem em um conjunto de entidades (CARVALHO, 2001).

É uma tarefa que tem como objetivo segmentar os dados formando grupos homogêneos. O agrupamento é aplicado quando ainda não é conhecida nenhuma classe e sua função é produzir uma segmentação do conjunto de registros de entrada de acordo com algum critério estabelecido (SILVA, 2003).

Tem como principal meta gerar classes através de partições do banco de dados em conjunto com tuplas. Essa partição é feita agrupando tuplas com valores de atributos semelhantes em uma mesma classe. Após a criação destas classes, é possível aplicar algoritmos de classificação para produzir regras para as mesmas (PACHECO et al., 1999).

Através da tarefa de agrupamento é possível dividir os dados em subconjuntos homogêneos fáceis de descrever e visualizar. Estes dados podem ser exibidos para o usuário em vez de tentar mostrar todos os dados, o que resultaria na perda de padrões embutidos (FAYYAD, 1997 apud ROMÃO, 2002).

O agrupamento pode ser usado, por exemplo, em um banco de dados escolar, relacionando disciplinas e alunos. Onde uma regra do tipo, 80% dos alunos inscritos em Linguagem de Programação também estão inscritos em Teoria da Computação, pode ser usada pela direção ou secretaria no planejamento do currículo anual, ou alocar recursos como sala de aula e professores (SCHENATZ, 2005).

2.1.5.4 Previsão de Séries Temporais

É definida como a classe de fenômenos cujo processo observacional e consequente quantificação numérica gera uma sequencia de dados distribuídos no tempo (SOUZA, 1989 apud MUELLER 1996). A natureza de uma série temporal e a estrutura de seu mecanismo gerador estão relacionadas com o intervalo de ocorrência das observações no tempo (ANDERSON, 1971 apud MUELLER 1996).

Uma previsão é uma manifestação relativa a sucessos desconhecidos em um futuro determinado. A previsão não constitui um fim em si, mas um meio de fornecer informações e subsídios para uma consequente tomada de decisão, visando atingir determinados objetivos (MORETTIN, 1981 apud MUELLER 1996).

Tem como objetivo a realização de inferências sobre as propriedades ou características básicas do mecanismo gerador do processo estocástico das observações da série. Assim, através da abstração de regularidades contidas nos fenômenos observáveis de uma série temporal existe a possibilidade de se construir um modelo matemático como uma representação simplificada da realidade (BARBANCHO1970 apud MUELLER 1996).

Após a formulação do modelo matemático, obtido pela seleção entre as alternativas de classes de modelos identificadas como apropriadas para essa representação e subsequente estimação de seus parâmetros, é possível utilizá-lo para testar alguma hipótese ou teoria a respeito do mecanismo gerador do processo

estocástico e realizar a previsão de valores futuros da série temporal (GRANGER1977 apud MUELLER 1996).

2.1.5.5 Técnicas de Mineração de Dados

Existem diferentes técnicas de mineração e algoritmos que possibilitam a busca por padrões desconhecidos nos dados. Possuir certo conhecimento sobre essas técnicas, ajuda muito no momento da escolha de uma delas de acordo com os problemas apresentados (SILVEIRA, 2003).

O Quadro 2.2 mostra as técnicas, uma breve descrição, as tarefas associadas a cada uma e alguns dos principais algoritmos.

Quadro 2.2 - Técnicas de KDD e algoritmos

Técnicas	Descrição	Tarefas	Exemplos
Regras de Associação	Estabelece uma correlação estatística entre atributos de dados e conjuntos de dados.	Associação	Apriori, AprioriTid, AprioriHybrid, AIS, SEM e DHP
Árvores de Decisão	Hierarquização dos dados, baseada em estágios de decisão e na separação de classes e subconjuntos.	Classificação e Regressão	CART, CHAID, C5.0, ID-3, QUEST, SLIQ e SPRINT
Raciocínio Baseado em Casos ou MBR	Baseado no método do vizinho mais próxima, combina e compara atributos para estabelecer hierarquia de semelhança.	Classificação e Segmentação	BIRCH, CLARANS e CLIQUE
Algoritmos Genéticos	Métodos gerais de busca e otimização, inspirados na Teoria da Evolução.	Classificação e Segmentação	Algoritmo Genético Simples, Genitor, CHC, GA-Nuggets e GA-PVMINER
Redes Neurais Artificiais	Modelos baseados na fisiologia do cérebro, onde o conhecimento é fruto do mapa das conexões neurais.	Classificação e Segmentação	Perceptron, Rede MLP, Redes de Kohonen, Rede Hopfield, Rede BAM, Rede IAC e Rede LVQ

Fonte: DIAS, 2002.

2.1.6 Pós-Processamento

É a etapa de avaliação do processo de descoberta, que busca melhorar a compreensão e selecionar o conhecimento descoberto que seja mais relevante para o objetivo pretendido.

O principal objetivo desta fase é melhorar a compreensão do conhecimento encontrado pelo algoritmo minerador, por meio da análise dos dados por um especialista. Muitas vezes, a mineração não gera bons resultados, sendo essencial uma nova etapa de enriquecimento, a fim de conseguir agregar mais informações de forma que contribuam no processo de descoberta de conhecimento (PACHECO et al., 1999).

A fase de pós-processamento inicia-se com a avaliação dos padrões realmente interessantes, que representem conhecimento útil, seguido da apresentação do conhecimento extraído para o usuário final, através de técnicas de visualização e representação do conhecimento (SILVA, 2003).

2.2 Árvores de Decisão

O método CHAID, que é o objeto de estudo dessa monografia, é uma técnica de Árvores de Decisão que busca classificar uma população grande de dados em grupos menores de acordo com regras estipulada. Para melhorar a compreensão deste método devemos entender o que são e como funciona esse processo de divisão da informação.

As Árvores de Decisão são utilizadas para aprender com as informações geradas, bem como para tomar decisões. O processo de aprendizagem ocorre à medida que observa suas interações com o mundo e seu processo interno de tomada de decisões (POZZER, 2006).

É uma estrutura parecida com um fluxograma, onde cada nó interno denota um teste em um atributo, cada galho representa um resultado do teste, cada nó folha guarda um rótulo de classe onde o nó superior em uma árvore é a raiz (HAN; KAMBER, 2006).

Além disso, são ferramentas poderosas e muito usadas para classificação e predição, sendo seu grande atrativo o fato de representarem regras que podem ser

expressas em linguagem comum, de modo que os humanos possam entendê-las (BERRY; LINOFF, 2004).

Podem ser definidas como uma estrutura usada para dividir uma grande quantidade de registros em conjuntos menores, aplicando-se uma seqüência simples de regras de decisão, onde a cada divisão sucessiva, os membros do subconjunto resultante, tornam-se cada vez mais semelhantes entre si. Portanto, consiste em um conjunto de regras para dividir uma população grande e heterogênea em pequenos grupos homogêneos de acordo com a variável resposta desejada.

O aspecto mais importante de uma Árvore de Decisão é como se faz a divisão dos grupos em grupos menores, de maneira que os novos nós tenham mais pureza que os seus antecessores em relação à variável resposta. O objetivo das divisões é montar uma árvore na qual se associe um novo registro a alguma classe que tenha um determinado comportamento em relação à variável resposta (BERRY; LINOFF, 2004).

O objetivo das divisões é descobrir, em cada nível da árvore, qual é a melhor variável independente que separe os grupos subseqüentes em que se predomine uma única classe, ou seja, o processo de construção da árvore passa por diversas iterações até achar a divisão que leva à maior pureza, sucessivamente, até que não seja mais possível fazer divisões, ou por falta de registros, ou por uma divisão adicional não aumentar a pureza.

Podem ser classificadas como representações simples do conhecimento e um meio eficiente de construir classificadores que predizem classes baseadas nos valores de atributos de um conjunto de dados (GARCIA, 2000).

Uma Árvore de Decisão tem a função de dividir recursivamente um conjunto de treinamento, até que cada subconjunto obtido contenha casos de uma única classe. Para atingir este objetivo, o algoritmo escolhido para a Árvore de Decisão examina e compara a distribuição de classes durante a construção da árvore. Os resultados obtidos são dados organizados de maneira compacta, com a árvore podendo ser utilizada para classificar novos casos (QUILAN, 1993 apud RODRIGUES, 2005).

São construídas baseadas no modelo *Top-Down*, ou seja, do nó raiz em direção às folhas. De uma forma geral, esta filosofia baseia-se na sucessiva divisão

do problema em vários subproblemas de menores dimensões, até que uma solução para cada um dos problemas mais simples seja encontrada.

Assim, são uma sequência de partições de um banco de dados de modo a maximizar diferenças sobre uma variável dependente (HAIR et al., 1995). Uma Árvore de Decisão é um instrumento de apoio à tomada de decisão que apresenta um único nó inicial. Esse nó inicial, que possui a informação total de uma população considerada, é dividido de acordo com um primeiro critério estipulado gerando novos nós. Por sua vez, esses novos nós são subdivididos em um novo conjunto de nós, a partir de um segundo critério estipulado. Assim sendo, cada novo nó conterá como informação uma percentagem de seu nó de origem. As subdivisões ocorrem de maneira sequencial enquanto existirem critérios diferentes que justifiquem novas divisões.

São muito utilizadas na área de *marketing* e outras diversas áreas que a utilizam para fins como, por exemplo, na medicina para determinação de diagnósticos ou para descobrir quais variáveis que contribuem para a melhoria do sistema de saúde, ciência da computação na estruturação de dados, biologia para classificação biológica, na análise de mercados, verificando quais as variáveis associadas com o volume de vendas (preço, geografia, características dos consumidores), entre outros campos da ciência.

Podem ser classificadas em dois tipos (AMO, 2003):

- Árvores de regressão (*regression tree*): usadas para estimar variáveis numéricas, por exemplo, no cálculo do valor de um carro.
- Árvores de classificação (*classification tree*): utilizadas quando a variável sob análise representar uma categoria, como por exemplo, faixa etária dos torcedores do time A.

As Árvores de Decisão podem ser utilizadas com objetivos diferentes, de acordo com o problema que se pretende resolver. Podemos, por exemplo, querer classificar os dados referentes a uma população da forma mais eficiente possível ou descobrir qual é a estrutura de um determinado tipo de problema, compreender quais as variáveis que afetam a sua resolução e construir um modelo que o solucione. Com elas é possível escolher as variáveis explicativas que realmente nos interessam para descrever a situação, deixando de lado as menos relevantes.

A Figura 2.3 representa um exemplo de Árvore de Decisão, onde constam dados que relatam as condições para uma pessoa receber um empréstimo. Nesse

caso existem duas possíveis classes referentes a receber o empréstimo: Sim e Não. Os atributos são montante, salário e conta. O montante pode assumir os valores: Médio, Alto ou Baixo; o atributo salário pode assumir Baixo ou Alto; e o atributo conta pode ser Sim ou Não. Alguns dados são exemplos da classe Sim, ou seja, os requisitos exigidos por um banco a uma pessoa para a concessão de um empréstimo são satisfatoriamente preenchidos. Outros são da classe Não, isto é, os requisitos exigidos não são plenamente satisfeitos. A classificação, nesse caso, resulta em uma estrutura de árvore, que pode ser usada para todos os objetos do conjunto (BRADZIL, 1999 apud RODRIGUES, 2005).

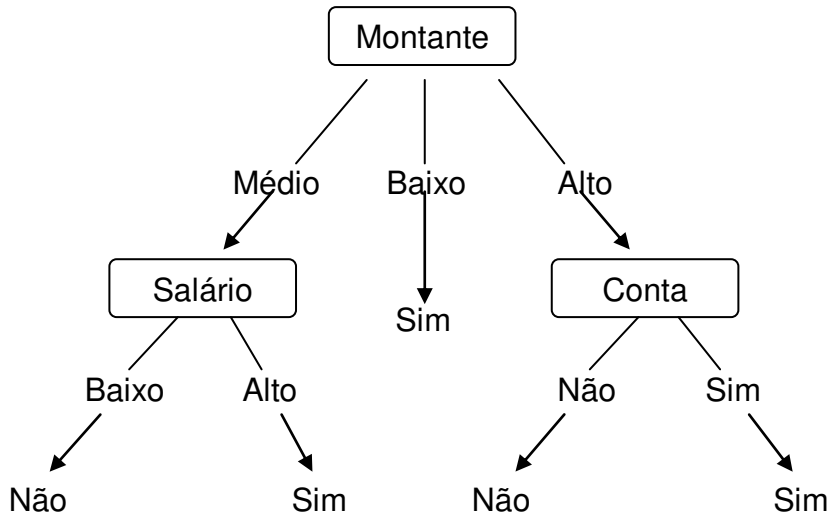


Figura 2.3 Exemplo de Árvores de Decisão (RODRIGUES, 2005).

Devido ao fato de as Árvores de Decisão tenderem a crescer muito, elas são muitas vezes substituídas pelas suas regras. Uma regra pode ser compreendida sem que haja a necessidade de referenciar as outras.

Com base na Árvore de Decisão apresentada na Figura 2.3, pode-se exemplificar a derivação de regras. Dois exemplos de regras obtidas a partir dessa árvore são mostrados a seguir:

- Se montante = médio e salário = baixo então classe = não.
- Se montante = médio e salário = alto então classe = sim.

Existem vários algoritmos de classificação de dados, entre eles está o método *Automatic Iteration Detector* (AID) que é considera o pioneiro nesse ramo, pois é baseado na análise de variância para segmentação das observações em grupos

distintos, para os quais podem ser desenvolvidos modelos de previsão causais por apresentam aplicações especialmente adequadas para grandes volumes de observações e variáveis explicativas em várias escalas de medida. (RODRIGUES, 2005).

Os principais algoritmos de Árvores de Decisão são evoluções deste primeiro algoritmo, AID, entre eles podemos destacar (RODRIGUES, 2005):

- *Iterative Dichotomizer* 3 (ID3) - consiste em um processo de indução, com o objetivo de escolher sempre o melhor atributo para cada nó. É um processo recursivo que após ter escolhido um atributo para um nó, começando pela raiz, aplica o mesmo algoritmo aos descendentes desse nó, até que certos critérios de parada sejam verificados (QUILAN, 1986 apud RODRIGUES, 2005).
- C4.5 – é um método melhorado relativamente ao ID3 que, entre outras melhorias, combate o problema de *overfitting*, que se caracteriza como a redução de algumas sub-árvores a folhas, utilizando uma estratégia de poda de árvore. O princípio orientador deste algoritmo foi criado por *William Occam*, que dá primazia à escolha de hipóteses menos complexas, compatíveis com a realidade observada, semelhante ao conceito de parcimónia da estatística (QUILAN, 1993 apud RODRIGUES, 2005).
- *Classification and Regression Trees* (CART) - este algoritmo é um modelo de regressão não-paramétrico que estabelece uma relação entre as variáveis independentes, com uma única variável dependente também chamada de resposta, ou alvo. O modelo é ajustado mediante sucessivas divisões binárias no conjunto de dados, para tornar os sub-conjuntos de dados da variável resposta cada vez mais homogêneos (BREIMAN et al., 1984 apud RODRIGUES, 2005).
- *Quick, Unbiasied, Efficient Statistical Tree* (QUEST) - do mesmo modo que o algoritmo CART, também é binário. No entanto, separa o processo de seleção das variáveis de partição do processo de busca da melhor segmentação dos dados em classes. Pode ser aplicado a qualquer tipo de variáveis preditivas ou explicativas, mas a variável dependente tem de ser nominal. Se várias das variáveis preditivas ou explicativas possuem o mesmo valor informativo, então todas têm a mesma probabilidade de ser escolhidas (LOH; SHIH, 1997 apud RODRIGUES, 2005).

- *Chi-square Automatic Interaction Detection* (CHAID) - é um dos métodos mais antigos de Árvore de Classificação, o algoritmo tem por base os testes de qui-quadrado de *Pearson* em uma tabela de contingência entre as categorias da variável dependente e as categorias das variáveis independentes. Constitui um método estatístico extremamente eficiente para a segmentação, ou crescimento de uma árvore (KASS, 1980 apud RODRIGUES, 2005). Por ser o objeto de estudo desse trabalho, esse método será melhor explicado na seção 2.3.

2.3 Método CHAID

O método CHAID teve sua origem no AID. Nela, assume-se a utilização de uma variável dependente contínua e variáveis independentes qualitativas ou categorizadas. Esta técnica foi expandida para os casos onde a variável dependente é qualitativa. No CHAID, os dados são divididos, a cada passo do algoritmo, em grupos otimizados, não necessariamente em dois subgrupos, através da maximização da significância da estatística do qui-quadrado.

As categorias das variáveis independentes são unidas se elas mostrarem padrões de comportamento semelhantes (homogeneidade) em relação à variável dependente. Além disto, para cada uma das categorias das variáveis independentes selecionadas, a técnica escolhe a próxima variável que melhor prediz a categoria da variável anterior. Ao terminar, os resultados da análise são mostrados em forma de uma árvore, onde as variáveis independentes aparecem de acordo com a capacidade de prever níveis específicos de outras variáveis independentes. O resultado final do CHAID representa os segmentos da população, que diferem segundo um determinado critério.

É um método para segmentação de uma população de interesse. É geralmente utilizado quando a segmentação é definida em termos de características demográficas ou variáveis categóricas com poder de predição. Nela, associa-se uma probabilidade de resposta para cada segmento, essas probabilidades depois são usadas para organizar os segmentos e selecionar o mais promissor.

Este algoritmo aceita variáveis categóricas nominais ou ordinais como variáveis dependentes. Normalmente, este tipo de variável é utilizada em pesquisas

tradicionais de *marketing*, aplicadas em questões demográficas como sexo, faixa-etária e renda salarial, grau de instrução ou outra variável dependente previamente definida. Quando os preditores são contínuos, eles são transformados em um preditor categorizado para posterior utilização do algoritmo, por exemplo, renda salarial (TURE et al., 2009).

O método CHAID é baseado nos testes de associação qui-quadrado e particiona o conjunto de dados em subconjuntos mutuamente exclusivos que melhor descrevem a variável resposta exaustivamente (TURE et al., 2009).

O algoritmo aplica a Árvore de Classificação para dividir um conjunto de dados em sub-conjuntos que discriminam de maneira diferenciada a variável resposta (dependente) e para combinar categorias que não diferem significativamente entre si.

Ele opera em uma variável dependente de escala nominal ou ordinal e maximiza a significância da estatística qui-quadrado em cada partição, caracterizando-se como uma estrutura de testes de significância.

O procedimento do método CHAID começa com a definição de uma a variável dependente(d) com no mínimo duas categorias (d ≥ 2) e uma preditora(c) para análise, com duas ou mais categorias (c ≥ 2). O que deve ser feito depois é reduzir a tabela de contingência c x d, em uma tabela j x d com associação mais significativa resultante da combinação das categorias do preditor, onde j é as categorias da variável preditora em questão. Para isso, primeiro é preciso calcular a estatística $\boldsymbol{T}_{(j)}^{(i)}$, estatística qui-quadrado para o i-ésimo método de formação de uma tabela j x d (2 ≤ j ≤ c). Então, se $\boldsymbol{T}_{(j)}^{(*)}$ = max $\boldsymbol{T}_{(j)}^{(i)}$, é o maior valor da estatística qui-quadrado encontrado para a tabela j x d, escolhe-se o $\boldsymbol{T}_{(j)}^{(*)}$ como valor de maior significância associada. A intenção do CHAID é pesquisar por um $\boldsymbol{T}_{(j)}^{(*)}$ estatística qui-quadrado máximo, avaliando a entrada de cada variável no modelo e verificando se sua contribuição é significativa ou não, entre as variáveis preditoras (KASS, 1980 apud SILVEIRA, 2010).

A aplicação do método permite a classificação de novos objetos através do conhecimento das categorias das variáveis explicativas, também chamadas de variáveis independentes ou preditoras.

Baseado no Teste de qui-quadrado (χ^2) o método CHAID acumula os desvios quadrados padronizados entre as frequências observadas e as frequências esperadas, sendo calculado pela fórmula 2.1 proposta por *Karl Pearson* (RODRIGUES, 2005):

$$\chi^2 = \sum_i \left[\frac{(O_i - E_i)^2}{E_i} \right] \quad (2.1)$$

Onde:

- i é a célula da tabela de contingência;
- O_i é a freqüência observada na célula;
- E_i é a freqüência esperada da célula.

O cálculo da frequência esperada é feito pela fórmula 2.2.

$$E = \frac{(l \; x \; c)}{n} \quad (2.2)$$

Onde:

- l total marginal da linha;
- c total marginal da coluna;
- n total.

Quanto maior o valor de χ^2 maior será a probabilidade de as frequências observadas estarem divergindo das frequências esperadas. Dessa forma, valores elevados indicam dependência, ou seja interações entre a variável resposta e as variável explicativa.

Com isso, duas hipóteses devem ser observadas:

- Hipótese nula: as freqüências observadas não são diferentes das freqüências esperadas.
- Hipótese alternativa: as freqüências observadas são diferentes das freqüências esperadas.

A estatística qui-quadrado tem distribuição com Graus de Liberdade (GL) ou nível de significância que representa a máxima probabilidade de erro que se tem ao rejeitar uma hipótese é calculado pela formula 2.3:

$$GL = (m - 1) * (n - 1) \quad (2.3)$$

Onde:

- m é número de linhas da tabela de contingência;
- n é número de colunas da tabela de contingência.

As possíveis dependências entre as variáveis explicativas e a resposta podem ser verificadas através do estudo das frequências cruzadas entre elas, caso não exista dependência se espera que a frequência relativa da variável resposta dentro de cada categoria da variável explicativa corresponda as frequências marginais da variável resposta.

A validação da dependência entre as variáveis é feita entre a comparação da estatística de teste com o valor crítico da distribuição χ^2 determinado pelo nível de significância, chamado de *p-value*, normalmente 5%, e pelos Graus de Liberdade da estatística. Se o valor da estatística é maior que o valor crítico, que é o valor tabelado para o qui-quadrado, então a hipótese alternativa que existe dependência entre as variáveis é aceita.

O método CHAID agrupa as categorias homogêneas da variável explicativa submetida ao teste, a partir das categorias prévias da variável sujeita ao CHAID com relação à variável resposta. Como a variável explicativa pode conter um número grande de categorias, a questão é identificar quais categorias podem ser agrupadas entre si. Para fazer essa identificação, o método gera uma tabela cruzada para cada par de categorias prévias da variável explicativa, ou seja, uma combinação das categorias prévias duas a duas, em relação à variável resposta.

Tem que ser levado em consideração que existe um limite de combinações dependendo do tipo de variável explicativa. Se for variável nominal o método testa todas as combinações possíveis, mas no caso de variável contínua ou ordinal ele não testa combinações de categorias não adjacentes.

Em seguida, é feito um teste para cada tabela cruzada, calculando-se a estatística χ^2 e o *p-value*, para cada par de categorias em questão. Depois, o CHAID

agrupa o par de categorias que apresentar o maior valor dentro da distribuição χ^2. É preciso lembrar que, se a estatística χ^2 de uma tabela cruzada é estatisticamente significante, ou seja, o valor da estatística de teste maior que o valor crítico, significa que não pode ser aceita a hipótese nula de independência entre as categorias, o que implica dizer que existe uma relação entre as variáveis contidas na tabela cruzada e, portanto, o par de categorias prévias em questão não pode ser agrupado em relação à variável de resposta, já que esse par não revela homogeneidade com esta variável. Podemos concluir que um valor maior que o nível de significância, *p-value*, representa que as duas categorias são homogêneas e podem ser agrupadas em relação à variável resposta.

Feito o agrupamento do par de categorias mais homogêneas, o algoritmo continua agora com as outras categorias da variável, sendo geradas novas combinações de categorias e novas tabelas cruzadas e outros testes χ^2, buscando detectar um novo par de categorias que podem ser agrupadas até o ponto que nenhuma das categorias restantes possa ser considerada homogênea. É importante levar em consideração que o método CHAID permite que o resultado final da categorização seja uma única categoria contendo todas as respostas possíveis da variável explicativa com relação à variável resposta, sendo assim, a variável explicativa em questão não apresenta relação com a variável de resposta, pois uma única categoria resultante guarda todas as respostas possíveis.

Pelo fato de ocorrer sucessivos testes de comparações, pelo cálculo da estatística qui-quadrado, é calculado uma correção na chamada desigualdade de Bonferroni utilizado para obter-se um nível de significância ajustado (SILVEIRA, 2010).

Suponha-se que um campo preditor originalmente tem I categorias, e ele é reduzido a r categorias depois da etapa de divisão. O multiplicador de Bonferroni (B) é o número de maneiras possíveis que uma categoria pode ser dividida em categorias r. Para r = I, B = 1, 2 ≤ r < I. Para cada tipo de variável preditora, o multiplicador é obtido da definição do coeficiente binominal. Assim para variáveis preditoras com categorias ordinais o multiplicador B é definido pela equação 2.4.

$$B = \binom{I-1}{r-1} \tag{2.4}$$

No caso de as categorias da variável preditora ser nominal usa-se a equação 2.5.

$$B = \sum_{v=0}^{r-1} (-1)^v \frac{(r-v)^I}{v!(r-v)!} \quad (2.5)$$

Mas se a variável preditora possuir categorias ordinais com alguma informação desconhecido, calcula-se B pela equação 2.6.

$$B = \binom{I-2}{r-2} + r\binom{I-2}{r-1} \quad (2.6)$$

A intenção dessa correção é provê um conjunto de métodos e níveis de significâncias para determinado grupo resguardando o erro dos sucessivos cálculos da estatística qui-quadrado, não ultrapassar um valor de nível de significância ($\boldsymbol{\alpha}$) estipulado no teste. Dessa forma, pela desigualdade de Bonferroni, $\boldsymbol{\alpha_0}$ é o valor com o qual o grupo de categorias deve ser testado, onde $\boldsymbol{N_B}(\mathrm{I})$ é o fator de ajuste de Bonferroni. Podemos calcular $\boldsymbol{\alpha_0}$ pela equação 2.7.

$$\alpha_0 = \frac{\alpha}{N_B(\mathrm{I})} \quad (2.7)$$

Sendo assim, o funcionamento do método CHAID pode ser resumido em alguns passos:

1. Definir uma variável dependente categórica onde as categorias sejam duas ou mais.

2. Definir as variáveis independentes ou preditoras onde as categorias sejam duas ou mais.

3. Para cada preditor: fazer uma tabela cruzada das categorias do preditor com as categorias da variável dependente. Feito isso calcula-se a significância: $\boldsymbol{\varkappa^2}$ (estatística qui-quadrado) e avalia se é necessário a correção da desigualdade de Bonferroni. Então, é feita a verificação da homogeneidade: se não existir nenhuma diferença estatisticamente significativa, as categorias da variável independente são unidas. Caso contrário, o nó é dividido pelas categorias identificadas.

4. Escolher a variável independente com maior significância na variável dependente. Calculando a significância: χ^2 (estatística qui-quadrado) e verificando se esta variável preditora tem maior significância que as outras, caso tenha, é feita a subdivisão dos dados de acordo com o numero de categorias do preditor seguinte.

Caso o valor de significância seja abaixo do qual uma variável é significativa para ser considerada preditora, a variável é desconsiderada.

Algumas condições de parada para a técnica devem ser observadas como:

- Não existem mais variáveis independentes a serem avaliadas.
- Profundidade máxima da árvore foi alcançada.
- O nó é puro, ou seja, todos os registros no nó têm o mesmo valor para todos os campos preditoras utilizados pelo modelo.

O método CHAID é limitado a variáveis categorizadas e não se restringi a divisões binárias, como ocorre com outros processos de segmentação, e não faz suposição de normalidade para as variáveis. Além dessas vantagens, podem ser levantadas outras (HEINECK; FREITAS, 2008):

- A utilização de variáveis dependentes e independentes em diversas formas como nominais, ordinais e categóricas;
- A possibilidade de os valores faltantes serem tratados, sendo agregados ao grupo mais próximo de acordo com a homogeneidade da distribuição da variável dependente ou isolados em uma categoria distinta;
- O fato de o método se tratar de Árvores de Decisão seus resultados gráficos são de fácil interpretação.

Em contrapartida alguns pontos negativos também podem ser levantados como (SILVEIRA, 2010).

- Necessidade de amostras grandes para que seja alcançado resultados satisfatórios;
- Instabilidade da árvore CHAID, pois as variáveis independentes são consideradas de forma seqüencial, sendo assim se a ordem em que as variáveis preditoras estão dispostas mudar, o algoritmo pode encontrar outros resultados, não garantido uma única solução ótima.

Como exemplo de uso do método CHAID, podemos utilizar o algoritmo para verificar se existe relação entre classes da sociedade e nível superior, assim como mostrado no Quadro 2.3.

Quadro 2.3 – Exemplo de uso do método CHAID

Amostra	Sem Superior	Com Superior	Total
Classe Média Alta	25	45	70
Classe Média	15	25	40
Classe Média Baixa	10	30	40
Total	50	100	150

Calculando a frequência esperada: $E = 70 \times 50 / 150 = 23,3333$

χ^2 parcial $= (O - E)^2 / E = [(25 - 23,3333)^2 / 23,3333)] = 0,1190$

$\chi^2 = 0,1190 + 0,0595 + 0,2083 + 0,1042 + 0,8333 + 0,4167 = 1,7410$

Para GL $= (2 - 1) \times (3 - 1) = 2$ o χ^2 tabelado $= 5,991$

Como, χ^2 calculado(1,7410) $<$ χ^2 tabelado(5,991), então, não existe relação entre classes da sociedade e nível superior.

No processo de análise dos dados, utilizamos o *software* estatístico Clementine, na versão 12.0, para a execução do método CHAID. Nessa ferramenta, todas as etapas são feitas através de nós, como por exemplo, carregar a base de dados para análise. No primeiro nó carregamos a base de dados, o nó utilizado nessa etapa é de acordo com o formato do arquivo da base de dados. O segundo nó é o *Type,* que lista todas as variáveis presentes na base de dados, nele determinamos a variável dependente e as variáveis independentes. O terceiro é o nó de modelagem, no nosso caso o nó CHAID. Esses três nós são conectados em sequência, onde o primeiro se conecta com o segundo e este ao terceiro nó, isso é feito, pois a saída do primeiro nó e entrada para o segundo (nó *Type*) e a saída deste último é entrada para o terceiro (nó CHAID). Ao executar o nó que representa o algoritmo CHAID, ele gera outro nó com a Árvore de Decisão encontrada. Além do método CHAID, o Clementine possui outros algoritmos de Árvore de Decisão, como o CART e QUEST. Existem também outros *softwares* estatísticos que possuem Árvores de Decisão em seus procedimentos, como por exemplo, o SAS e o *Statistica*.

Neste capitulo vimos o processo de KDD, suas fases e tarefas dando destaque a etapa de Mineração de Dados, onde podemos entender algumas técnicas, em especial as Árvores de Decisão que são de fundamental importância

para a compreensão do método CHAID. Vimos como este método faz as divisões da informação em nós de acordo com uma variável dependente, agora iremos ver como o algoritmo CHAID funciona em uma situação prática de uso.

3. ESTUDO DE CASO

Neste capítulo descreveremos o estudo de caso, cujo objetivo é explicar o uso do algoritmo CHAID. Para tanto, serão utilizadas as informações da base de dados da Pesquisa Nacional por Amostra de Domicílios (PNAD), que foi realizada no ano de 2008, a fim de traçar um perfil socioeconômico relacionando o nível de escolaridade das pessoas com a situação econômica da mesma. Desta forma, as seções seguintes descrevem em detalhes a utilização do mesmo.

3.1 Descrição do Contexto

No processo de mineração de informações utilizamos uma parte da base de dados da Pesquisa Nacional por Amostra de Domicílios (PNAD), realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE) no ano de 2008. Embora a pesquisa seja realizada anualmente, seus dados levam certo tempo para serem disponibilizados de forma gratuita, por isso utilizamos os registros da pesquisa realizada no ano de 2008.

A PNAD é um sistema de pesquisas domiciliares, implantado progressivamente no Brasil a partir de 1967, que tem como finalidade a produção de informações básicas para o estudo do desenvolvimento socioeconômico do país (IBGE, 2008).

Trata-se de um sistema de pesquisas por amostra de domicílios que, por ter propósitos múltiplos, investiga diversas características socioeconômicas, umas de caráter permanente nas pesquisas, como as características gerais da população, de educação, trabalho, rendimento e habitação, e outras com periodicidade variável, como as características sobre migração, fecundidade, nupcialidade, saúde, nutrição e outros temas que são incluídos no sistema de acordo com as necessidades de informação para o país (IBGE, 2008).

A PNAD é realizada por meio de uma amostra probabilística de domicílios, obtida em três estágios de seleção: unidades primárias - municípios; unidades secundárias - setores censitários; e unidades terciárias - unidades domiciliares (domicílios particulares e unidades de habitação em domicílios coletivos).

Na seleção das unidades primária e secundária (municípios e setores censitários) da PNAD, foram adotadas a divisão territorial e a malha setorial, que delimita o tamanho dos municípios, vigentes em 1º de agosto de 2000 que também foi a utilizada para a realização do Censo Demográfico do mesmo ano (IBGE, 2008).

No primeiro estágio, as unidades (municípios) foram classificadas em duas categorias: autorrepresentativas (probabilidade 1 de pertencer à amostra) e não autorrepresentativas. Os municípios pertencentes à segunda categoria passaram por um processo de estratificação e, em cada estrato, foram selecionados com reposição e com probabilidade proporcional à população residente obtida no Censo Demográfico 2000 (IBGE, 2008).

No segundo estágio, as unidades (setores censitários) foram selecionadas, em cada município da amostra, também com probabilidade proporcional e com reposição, sendo utilizado o número de unidades domiciliares existentes por ocasião do Censo Demográfico 2000 como medida de tamanho (IBGE, 2008).

No último estágio foram selecionados, com equiprobabilidade, em cada setor censitário da amostra, os domicílios particulares e as unidades de habitação em domicílios coletivos para investigação das características dos moradores e da habitação (IBGE, 2008).

Na PNAD 2008, foram pesquisadas 391.868 pessoas e 150.591 unidades domiciliares distribuídas por todas as Unidades da Federação no período de 21 a 27 de setembro de 2008 e incluiu três levantamentos adicionais: a terceira realização da Pesquisa Suplementar de Saúde, a segunda da Pesquisa Suplementar sobre Acesso à Internet e Posse de Telefone Móvel Celular para Uso Pessoal, permitindo acompanhar a evolução de indicadores-chave da Tecnologia da Informação e Comunicação; e a Pesquisa Especial de Tabagismo (IBGE, 2008).

3.2 Passos seguidos na Mineração dos Dados

Após a escolha da PNAD 2008 como base a ser utilizada em nosso estudo de caso, fizemos os seguintes passos:

- Determinar as variáveis de interesse: a fim de escolher aquelas que são mais importantes para traçar o perfil socioeconômico que relacione o nível de escolaridade das pessoas com a situação econômica da mesma.

- Limpar os dados: para retirar variáveis sem valores informados e não prejudicar a definição do perfil procurado.
- Aplicar o algoritmo CHAID: para que ele gere a árvore de decisão com as variáveis mais importantes na determinação do perfil buscado.
- Analisar os resultados: para verificar o que foi encontrado, levantar discussões e avaliar as implicações.

3.3 Determinação das variáveis de interesse

A base de dados apresenta 424 variáveis de diversos tipos, como idade do entrevistado, quantidade de filhos, cor ou raça entre outras. Primeiro, escolhemos a variável dependente, como vimos na seção 2.3, que na base é a variável V6007, representando o grau de instrução do entrevistado. Escolhemos o grau de instrução porque vamos traçar os perfis dos entrevistados de acordo com o nível de escolaridade de cada um. A variável grau de instrução possui as classes mostradas no Quadro 3.1:

Quadro 3.1 - Variável dependente e suas classes

Variável dependente	Classes
Grau de instrução	Elementar (primário)
	Médio 1º ciclo (ginasial, etc.)
	Médio 2º ciclo (científico, clássico, etc.)
	Regular do ensino fundamental ou do 1º grau
	Regular do ensino médio ou do 2º grau
	Educação de jovens e adultos ou supletivo do ensino fundamental ou do 1º grau
	Educação de jovens e adultos ou supletivo de ensino médio ou do 2º grau
	Superior - graduação
	Mestrado ou doutorado
	Alfabetização de jovens e adultos
	Creche
	Classe de alfabetização - CA
	Maternal, jardim de infância etc.

Após a escolha da variável dependente, o passo seguinte é determinarmos as variáveis independentes, como a nossa intenção é traçar um perfil socioeconômico relacionando o nível de escolaridade das pessoas com a situação econômica que ela se encontra no período de realização da pesquisa nos estados do Brasil,

reduzimos a quantidade inicial de variáveis escolhendo as que continham dados relacionados com renda, como rendimento mensal, condição da propriedade, etc, além de outras relevantes como sexo, tipo de domicílio e unidade da federação, ao final foram escolhidas 16 variáveis independentes que estão listadas no Quadro 3.2.

Quadro 3.2 - Variáveis independentes

Variáveis independentes	Classes	Motivo de escolha
UF	Rondônia, Acre, Amazonas, Roraima, Pará, Amapá, Tocantins, Maranhão, Piauí, Ceará, Rio Grande do Norte, Paraíba, Pernambuco, Alagoas, Sergipe, Bahia, Minas Gerais, Espírito Santo, Rio de Janeiro, São Paulo, Paraná, Santa Catarina, Rio Grande do Sul, Mato Grosso do Sul, Mato Grosso, Goiás, Distrito Federal.	Verificar a relação entre os estados, situação financeira e a escolaridade dos entrevistados.
Idade	Intervalos categorizados pelo *software* de análise	Avaliar a relação das idades com os rendimentos.
Sexo	Masculino Feminino	Verificar se existem diferenças entre sexos.
Posição no trabalho	Empregado Trabalhador doméstico Conta própria Empregador Trabalhador não remunerado de membro da unidade domiciliar Outro trabalhador não remunerado Trabalhador na construção para o próprio uso	Avaliar a relação entre as classes de trabalhadores, os rendimentos e o grau de instrução.
Horas trabalhadas por semana	Até 14 horas 15 a 39 horas 40 a 44 horas 45 a 48 horas 49 horas ou mais	Verificar as relações entre horas trabalhadas, rendimentos e escolaridade, para avaliar, por exemplo, se quem trabalha mais tem melhores rendimentos.
Rendimento mensal	Intervalos categorizados pelo *software* de análise	Variável que contém o rendimento.
Setor do emprego	Privado Público	Verificar se existe relação entre o setor que o entrevistado trabalha com rendimento.
Forma de contratação no emprego	Somente por jornada de trabalho Somente por produção ou comissão Somente por tarefa ou empreitada Por jornada de trabalho e produção ou comissão Outra forma	Avaliar a relação entre forma de contratação e o rendimento mensal.
Tem carteira assinada atualmente	Sim Não	Verificar se existe relação entre possuir carteira assinada e o rendimento mensal.
Local onde trabalha	Loja, oficina, fábrica, escritório, escola, repartição pública, galpão, etc. Fazenda, sítio, granja, chácara, etc. No domicílio em que mora Em domicílio de empregador, patrão, sócio ou freguês	Verificar se existe relação entre o local de trabalho o rendimento mensal e o nível de escolaridade.

	Em local designado pelo empregador, cliente ou freguês Em veículo automotor Em via ou área pública Outro	
Tipo de domicílio	Casa Apartamento Cômodo	Avaliar relações entre escolaridade e os tipos de domicílio.
Condição da propriedade	Próprio – já pago Próprio – ainda pagando Alugado Cedido por empregador Cedido de outra forma Outra condição	Verificar se existe relação entre a situação financeira e a condição da propriedade do entrevistado.
Possui telefone celular	Sim Não	Variável que representa o poder aquisitivo dos entrevistados.
Possui máquina de lavar roupa	Sim Não	Variável que demonstra o poder aquisitivo dos entrevistados.
Possui microcomputador	Sim Não	Variável que representa o poder aquisitivo dos entrevistados.
Possui freezer	Sim Não	Variável que mostra o poder aquisitivo dos entrevistados.

3.4 Limpeza dos dados

Após a escolha das variáveis dependente e independentes, fizemos a limpeza dos dados, esse processo foi necessário porque na base de dados existiam muitas variáveis sem valores informados, o que prejudicaria na definição do perfil procurado e também a eficiência do método CHAID. Ao final desse processo, de um total de 391.868 registros, a base ficou com 127.071 registros de todos os estados do Brasil, a serem analisados pelo algoritmo.

3.5 Análise e interpretação dos Dados

Nessa etapa, utilizamos o Clementine, como explicado na seção 2.3, para análise e construção do perfil. Com ele, fizemos as seguintes etapas:

Primeiro acrescentamos o nó para importar a base de dados (Figura 3.1). Nele carregamos a base de dados com as informações de todo Brasil no nó SPSS[1], como podemos observar na Figura 3.2.

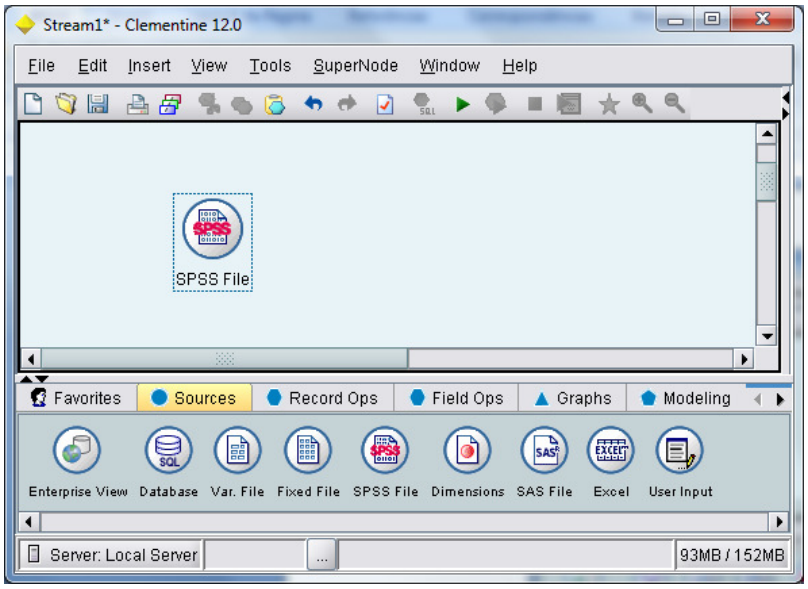


Figura 3.1 - Adicionando nó SPSS

[1] Ferramenta estatística para modelagem de dados.

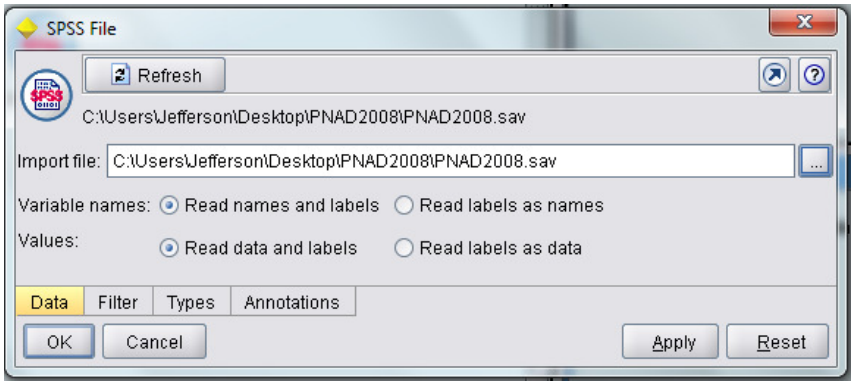


Figura 3.2 Importando a base de dados

Depois acrescentamos o nó *Type* (Figura 3.3), que lista todas as variáveis, usamos ele para separar a variável dependente das variáveis independentes.

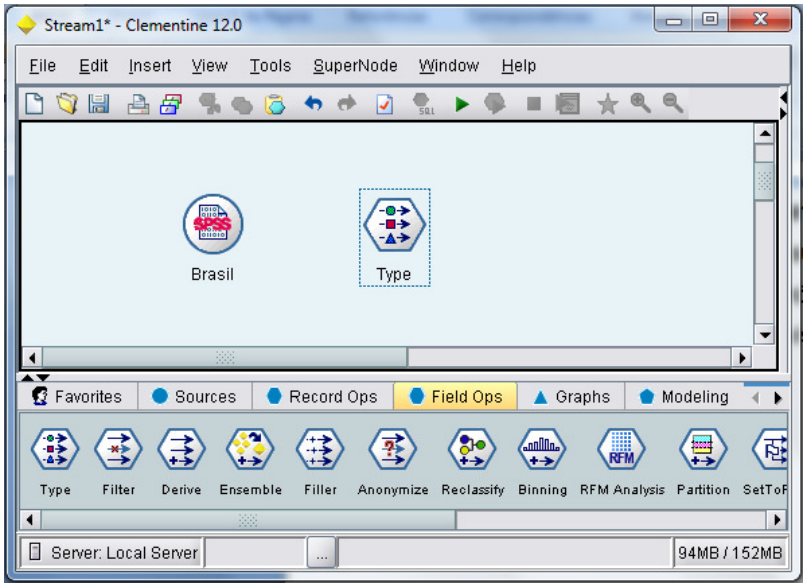


Figura 3.3 - Acrescentando o nó *Type*

Em seguida, adicionamos o nó CHAID (Figura 3.4), que quando executado gera outro nó com a Árvore de Decisão encontrada pelo algoritmo.

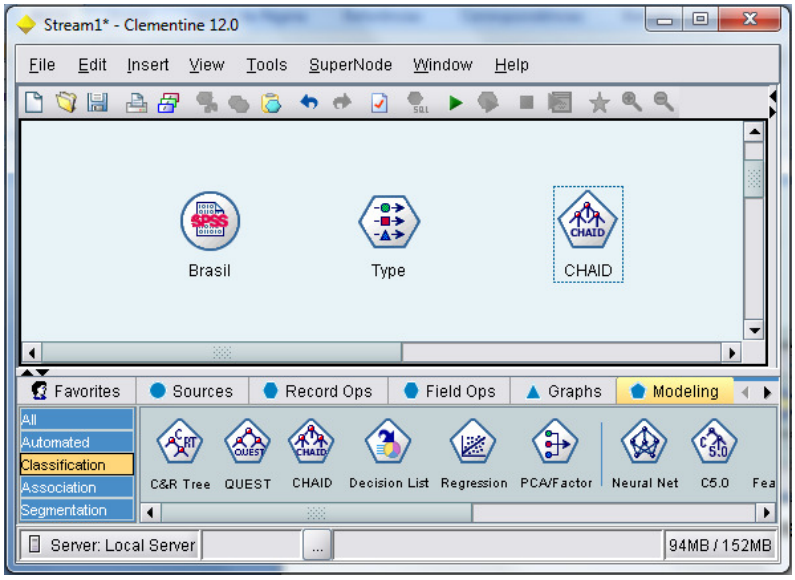


Figura 3.4 - Adicionando o nó CHAID

Por último, conectamos o nó SPSS ao *Type* e este ao CHAID, como na Figura 3.5.

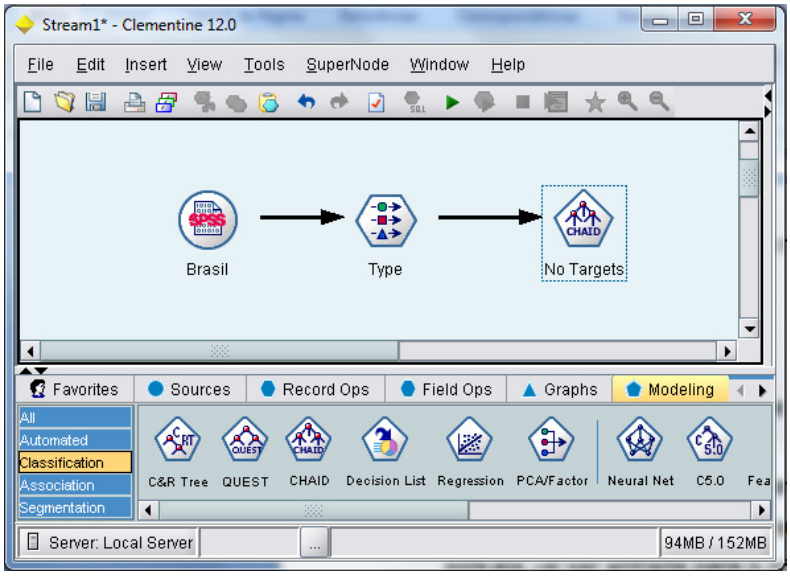


Figura 3.5 - Nós de análise

O nó *Type* possui as variáveis selecionadas, nele determinamos qual é a variável dependente mudando sua direção para saída, como indicado na Figura 3.6, pois ela vai ser entrada para o nó CHAID, no caso a variável V6007, que representa o grau de instrução do entrevistado, as outras marcadas como entrada são as variáveis independentes.

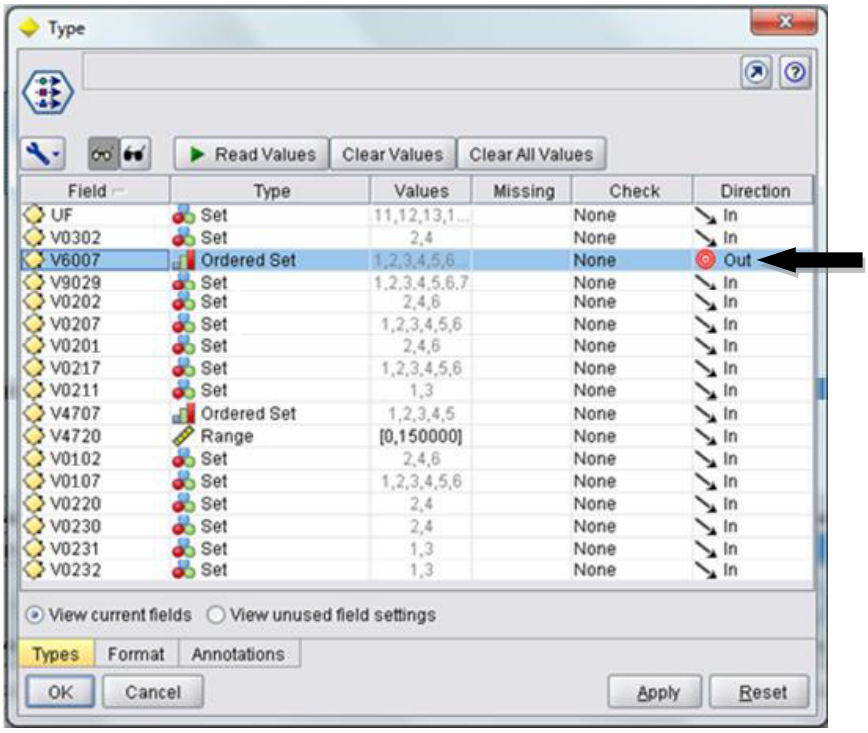


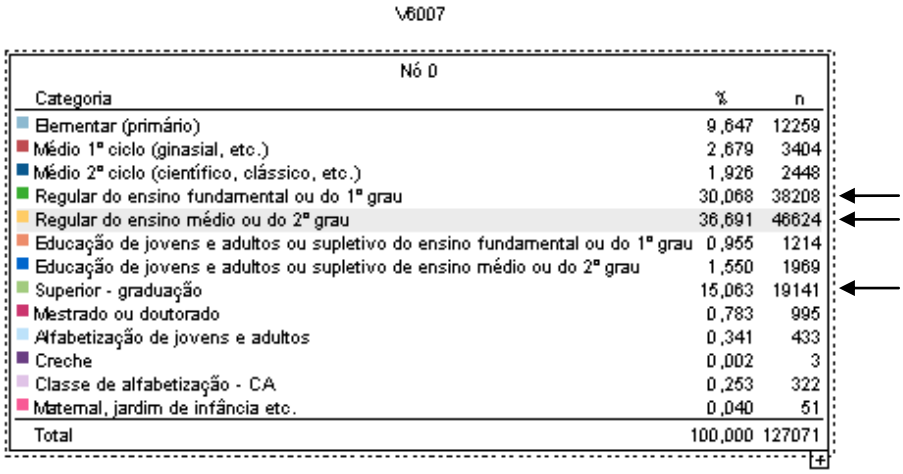
Figura 3.6 Configurando o nó *Type*

Em seguida, executamos o CHAID que cria um outro nó com a Árvore de Decisão gerada pelo algoritmo, nela observamos que foram criados 26 nós. Como a quantidade de nós é grande, a Figura 3.7 foi construída para melhorar a visualização da árvore, por isso distorcemos um pouco sua estrutura.

Pela árvore observamos que o método CHAID encontrou o rendimento mensal como variável mais importante relacionada ao grau de escolaridade que melhor classifica os registros presentes na base de dados. O *software* verificou a homogeneidade das classes da variável, gerando oito categorias para o grau de escolaridade e para cada uma, relacionou a mais importante. Para melhor visualização e interpretação dos resultados, mostraremos separadamente o nó raiz e cada um de seus ramos.

Figura 3.7 - Árvore de Decisão gerada para dados de todo Brasil

Na Figura 3.8 está o nó raiz, onde podemos observar que a maioria dos entrevistados, 36,69%, concluíram o ensino médio; 30,06%, o ensino fundamental e apenas 15,06% possuem o nível superior completo.

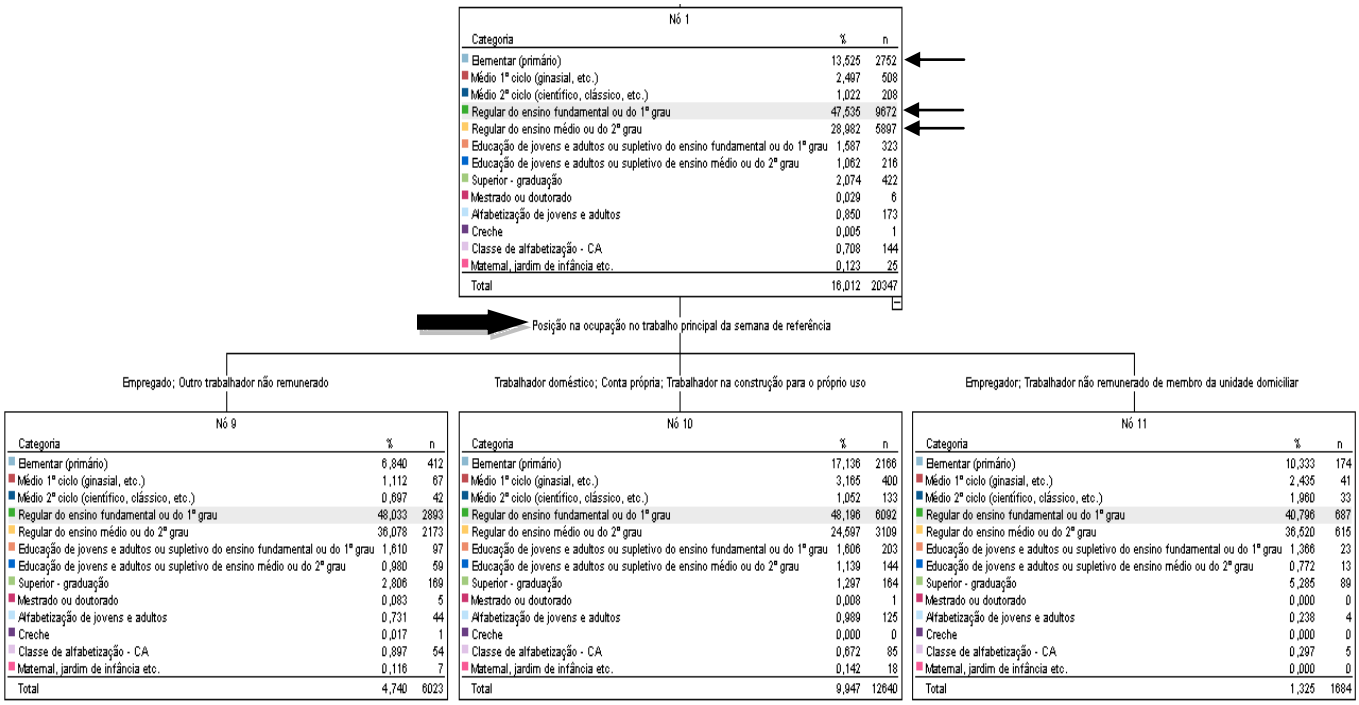

V6007

Nó 0

Categoria	%	n
Elementar (primário)	9,647	12259
Médio 1º ciclo (ginasial, etc.)	2,679	3404
Médio 2º ciclo (científico, clássico, etc.)	1,926	2448
Regular do ensino fundamental ou do 1º grau	30,068	38208
Regular do ensino médio ou do 2º grau	36,691	46624
Educação de jovens e adultos ou supletivo do ensino fundamental ou do 1º grau	0,955	1214
Educação de jovens e adultos ou supletivo de ensino médio ou do 2º grau	1,550	1969
Superior - graduação	15,063	19141
Mestrado ou doutorado	0,783	995
Alfabetização de jovens e adultos	0,341	433
Creche	0,002	3
Classe de alfabetização - CA	0,253	322
Maternal, jardim de infância etc.	0,040	51
Total	100,000	127071

Figura 3.8 - Nó raiz para dados de todo Brasil

Na figura 3.9, vemos que 47,53% dos entrevistados concluiu o ensino fundamental; 28,98% o ensino médio e 13,58% o primário. A variável mais importante para o ramo é a posição no trabalho.


Nó 1

Categoria	%	n
Elementar (primário)	13,525	2752
Médio 1º ciclo (ginasial, etc.)	2,497	508
Médio 2º ciclo (científico, clássico, etc.)	1,022	208
Regular do ensino fundamental ou do 1º grau	47,535	9672
Regular do ensino médio ou do 2º grau	28,982	5897
Educação de jovens e adultos ou supletivo do ensino fundamental ou do 1º grau	1,587	323
Educação de jovens e adultos ou supletivo de ensino médio ou do 2º grau	1,062	216
Superior - graduação	2,074	422
Mestrado ou doutorado	0,029	6
Alfabetização de jovens e adultos	0,850	173
Creche	0,005	1
Classe de alfabetização - CA	0,708	144
Maternal, jardim de infância etc.	0,123	25
Total	16,012	20347

Posição na ocupação no trabalho principal da semana de referência

Empregado; Outro trabalhador não remunerado

Nó 9

Categoria	%	n
Elementar (primário)	6,840	412
Médio 1º ciclo (ginasial, etc.)	1,112	67
Médio 2º ciclo (científico, clássico, etc.)	0,697	42
Regular do ensino fundamental ou do 1º grau	48,033	2893
Regular do ensino médio ou do 2º grau	36,078	2173
Educação de jovens e adultos ou supletivo do ensino fundamental ou do 1º grau	1,610	97
Educação de jovens e adultos ou supletivo de ensino médio ou do 2º grau	0,980	59
Superior - graduação	2,806	169
Mestrado ou doutorado	0,083	5
Alfabetização de jovens e adultos	0,731	44
Creche	0,017	1
Classe de alfabetização - CA	0,897	54
Maternal, jardim de infância etc.	0,116	7
Total	4,740	6023

Trabalhador doméstico; Conta própria; Trabalhador na construção para o próprio uso

Nó 10

Categoria	%	n
Elementar (primário)	17,136	2166
Médio 1º ciclo (ginasial, etc.)	3,165	400
Médio 2º ciclo (científico, clássico, etc.)	1,052	133
Regular do ensino fundamental ou do 1º grau	48,196	6092
Regular do ensino médio ou do 2º grau	24,597	3109
Educação de jovens e adultos ou supletivo do ensino fundamental ou do 1º grau	1,606	203
Educação de jovens e adultos ou supletivo de ensino médio ou do 2º grau	1,139	144
Superior - graduação	1,297	164
Mestrado ou doutorado	0,008	1
Alfabetização de jovens e adultos	0,989	125
Creche	0,000	0
Classe de alfabetização - CA	0,672	85
Maternal, jardim de infância etc.	0,142	18
Total	9,947	12640

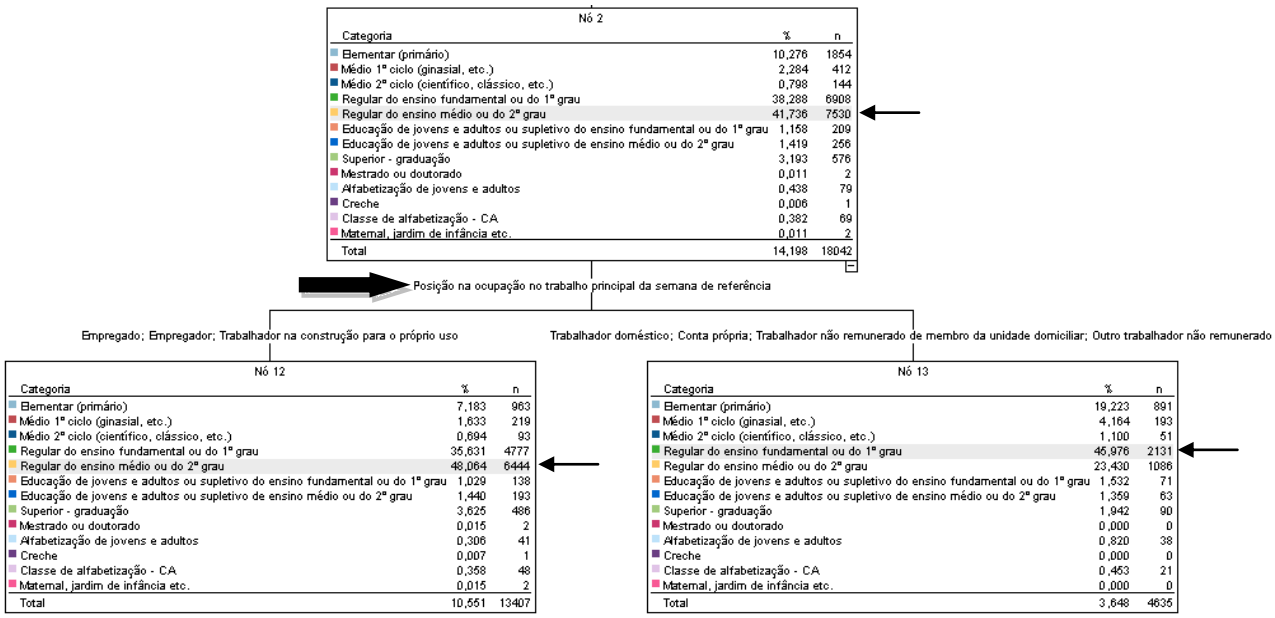
Empregador; Trabalhador não remunerado de membro da unidade domiciliar

Nó 11

Categoria	%	n
Elementar (primário)	10,333	174
Médio 1º ciclo (ginasial, etc.)	2,435	41
Médio 2º ciclo (científico, clássico, etc.)	1,960	33
Regular do ensino fundamental ou do 1º grau	40,796	687
Regular do ensino médio ou do 2º grau	36,520	615
Educação de jovens e adultos ou supletivo do ensino fundamental ou do 1º grau	1,366	23
Educação de jovens e adultos ou supletivo de ensino médio ou do 2º grau	0,772	13
Superior - graduação	5,285	89
Mestrado ou doutorado	0,000	0
Alfabetização de jovens e adultos	0,238	4
Creche	0,000	0
Classe de alfabetização - CA	0,297	5
Maternal, jardim de infância etc.	0,000	0
Total	1,325	1684

Figura 3.9 - Ramo onde o rendimento é menor ou igual a R$ 414,00

Na Figura 3.10, observamos que a maioria 41,73% tem o ensino médio e a variável mais importante é a posição no trabalho onde para as categorias empregado, empregador e trabalhador da construção 48,06% possui o ensino médio e para a classe de trabalhadores domésticos, autônomos e não remunerados, 45,97% tem apenas o ensino fundamental completo.



Nó 2

Categoria	%	n
Elementar (primário)	10,276	1854
Médio 1ª ciclo (ginasial, etc.)	2,284	412
Médio 2º ciclo (científico, clássico, etc.)	0,798	144
Regular do ensino fundamental ou do 1º grau	38,288	6908
Regular do ensino médio ou do 2º grau	41,736	7530
Educação de jovens e adultos ou supletivo do ensino fundamental ou do 1º grau	1,158	209
Educação de jovens e adultos ou supletivo de ensino médio ou do 2º grau	1,419	256
Superior - graduação	3,193	576
Mestrado ou doutorado	0,011	2
Alfabetização de jovens e adultos	0,438	79
Creche	0,006	1
Classe de alfabetização - CA	0,382	69
Maternal, jardim de infância etc.	0,011	2
Total	14,198	18042

Posição na ocupação no trabalho principal da semana de referência

Empregado; Empregador; Trabalhador na construção para o próprio uso

Nó 12

Categoria	%	n
Elementar (primário)	7,183	963
Médio 1ª ciclo (ginasial, etc.)	1,633	219
Médio 2º ciclo (científico, clássico, etc.)	0,694	93
Regular do ensino fundamental ou do 1º grau	35,631	4777
Regular do ensino médio ou do 2º grau	48,064	6444
Educação de jovens e adultos ou supletivo do ensino fundamental ou do 1º grau	1,029	138
Educação de jovens e adultos ou supletivo de ensino médio ou do 2º grau	1,440	193
Superior - graduação	3,625	486
Mestrado ou doutorado	0,015	2
Alfabetização de jovens e adultos	0,306	41
Creche	0,007	1
Classe de alfabetização - CA	0,358	48
Maternal, jardim de infância etc.	0,015	2
Total	10,551	13407

Trabalhador doméstico; Conta própria; Trabalhador não remunerado de membro da unidade domiciliar; Outro trabalhador não remunerado

Nó 13

Categoria	%	n
Elementar (primário)	19,223	891
Médio 1ª ciclo (ginasial, etc.)	4,164	193
Médio 2º ciclo (científico, clássico, etc.)	1,100	51
Regular do ensino fundamental ou do 1º grau	45,976	2131
Regular do ensino médio ou do 2º grau	23,430	1086
Educação de jovens e adultos ou supletivo do ensino fundamental ou do 1º grau	1,532	71
Educação de jovens e adultos ou supletivo de ensino médio ou do 2º grau	1,359	63
Superior - graduação	1,942	90
Mestrado ou doutorado	0,000	0
Alfabetização de jovens e adultos	0,820	38
Creche	0,000	0
Classe de alfabetização - CA	0,453	21
Maternal, jardim de infância etc.	0,000	0
Total	3,648	4635

Figura 3.10 - Ramo onde o rendimento é maior que R$ 414,00 e menor ou igual a R$ 450,00

Na Figura 3.11, 41,78% possui o ensino médio, a variável mais importante também é a posição no trabalho sendo que para os empregados 48,69% possui o ensino médio.

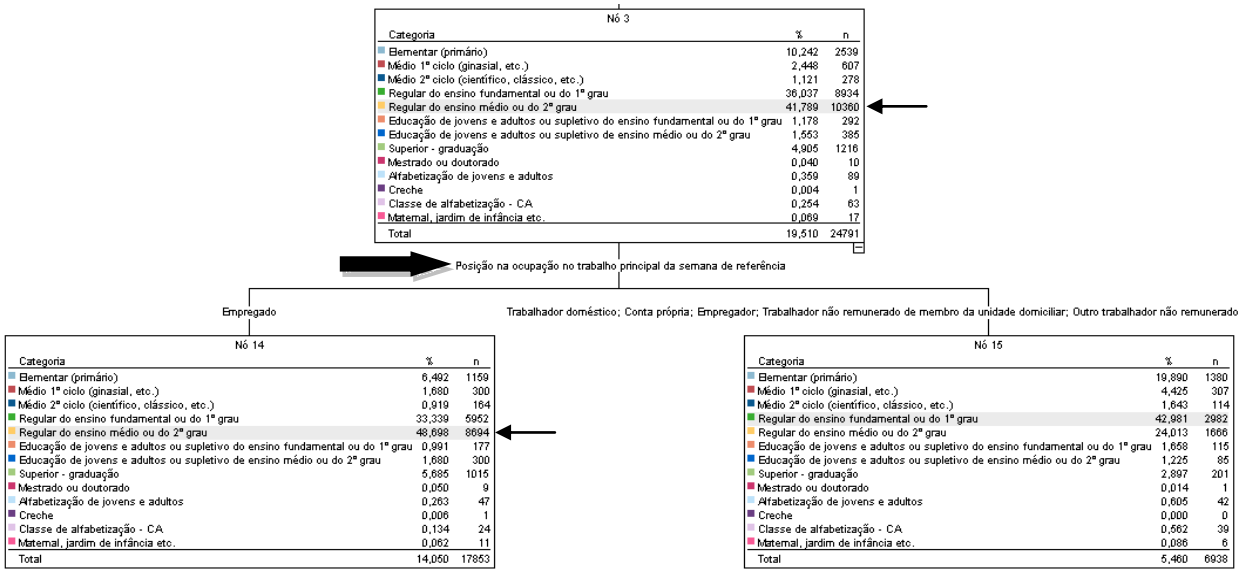


Figura 3.11 - Ramo para rendimento maior que R$ 450,00 e menor ou igual a R$ 649,00

Para os entrevistados com rendimento maior que R$ 649,00 e menor ou igual a R$ 800,00, Figura 3.12, 43,85% tem o ensino médio sendo que para as pessoas com ocupação de empregado ou empregador cerca de 49% possui o segundo grau completo.

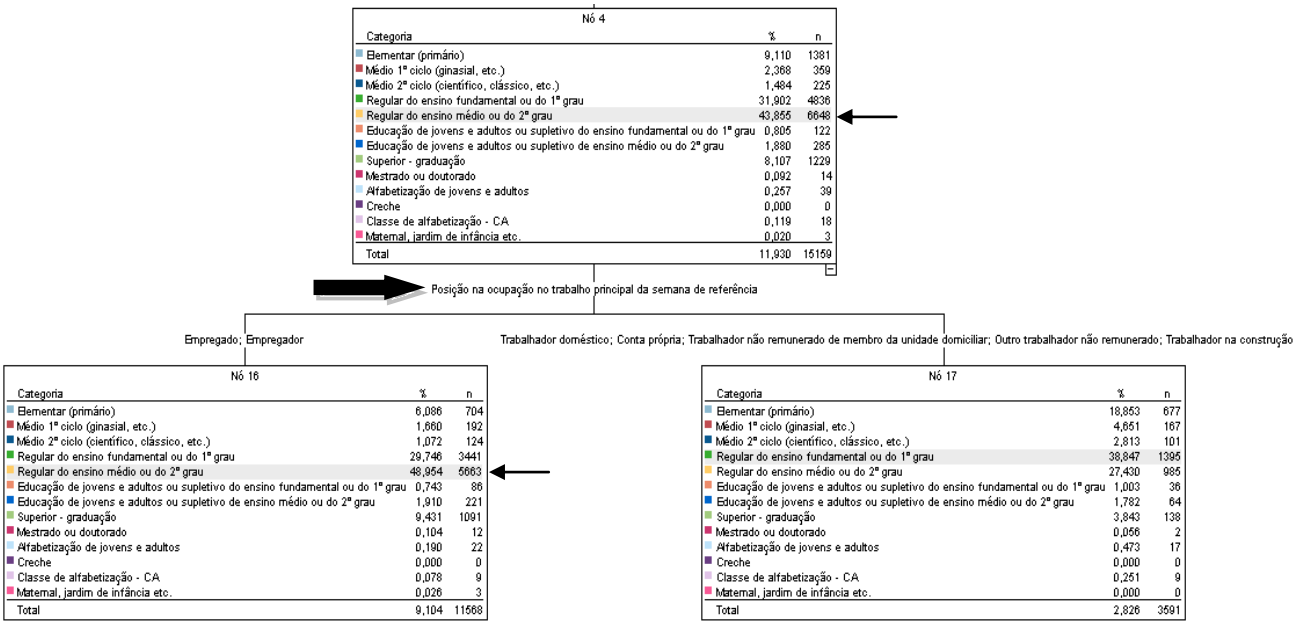


Figura 3.12 - Ramo para rendimento maior que R$ 649,00 e menor ou igual a R$ 800,00

Já no ramo mostrado na Figura 3.13, aproximadamente 41% possui o ensino médio, sendo que a variável mais importante é a tem computador, onde dos que apresentam, 44,20% completou o segundo grau e 22,94% o nível superior.

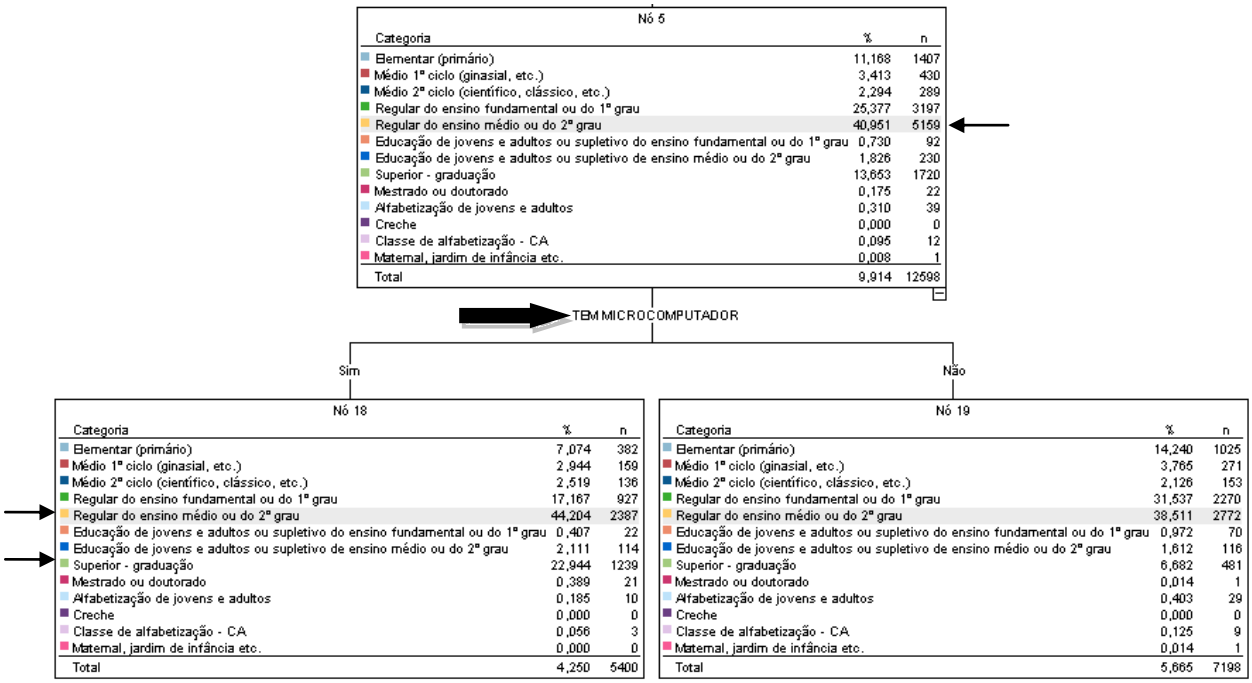


Figura 3.13 - Ramo com rendimento maior que R$ 800,00 e menor ou igual a R$ 1.000,00

Para o ramo da Figura 3.14, a variável mais importante também é a posição no trabalho, sendo que 38,15% possui o ensino médio e os que são empregados um pouco mais de 42% possui o segundo grau.

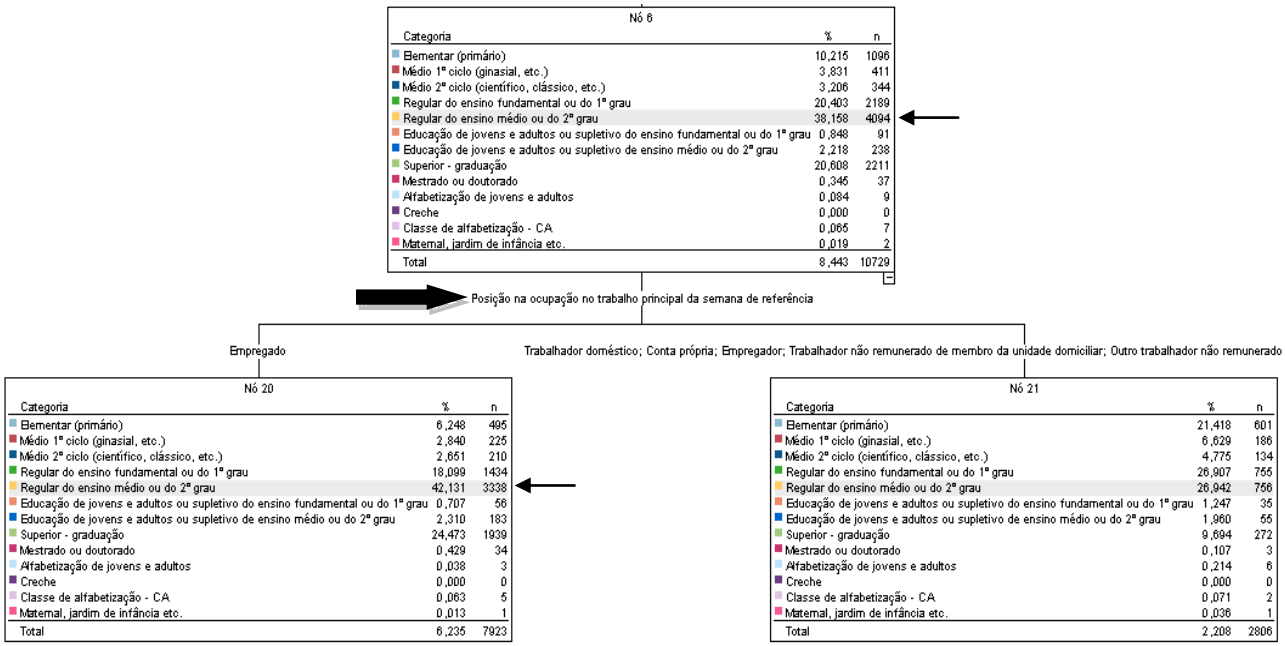


Figura 3.14 - Ramo onde o rendimento é maior que R$ 1.000,00 e menor ou igual a R$ 1.480,00

Para pessoas com rendimento maior que R$ 1.480,00 e menor ou igual a R$ 2.498,00, Figura 3.15, a maioria possui nível superior, 35,17%, e a variável mais importante é o sexo, sendo que para os homens 39,9% completou o ensino médio, 23,27% o ensino superior e das mulheres 57,89% possui o ensino superior completo.

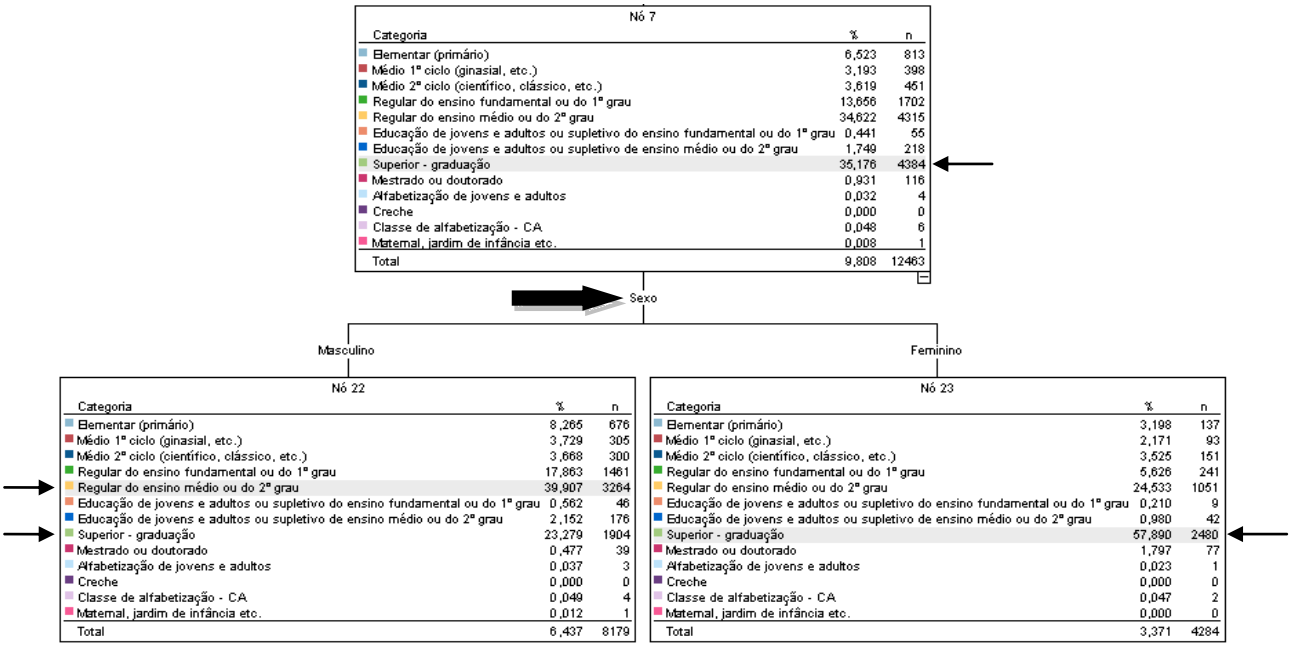


Figura 3.15 - Ramo com rendimento maior que R$ 1.480,00 e menor ou igual a R$ 2.498,00

Já para o ramo com pessoas que apresentam os maiores rendimentos mensais do Brasil, acima de R$ 2.498,00 como mostrado na Figura 3.16, cerca de 57% possui superior completo sendo a variável mais determinante a posição no trabalho, onde dos empregados, 65,33% possui nível superior e das outras categorias, 44,27% tem o terceiro grau completo.

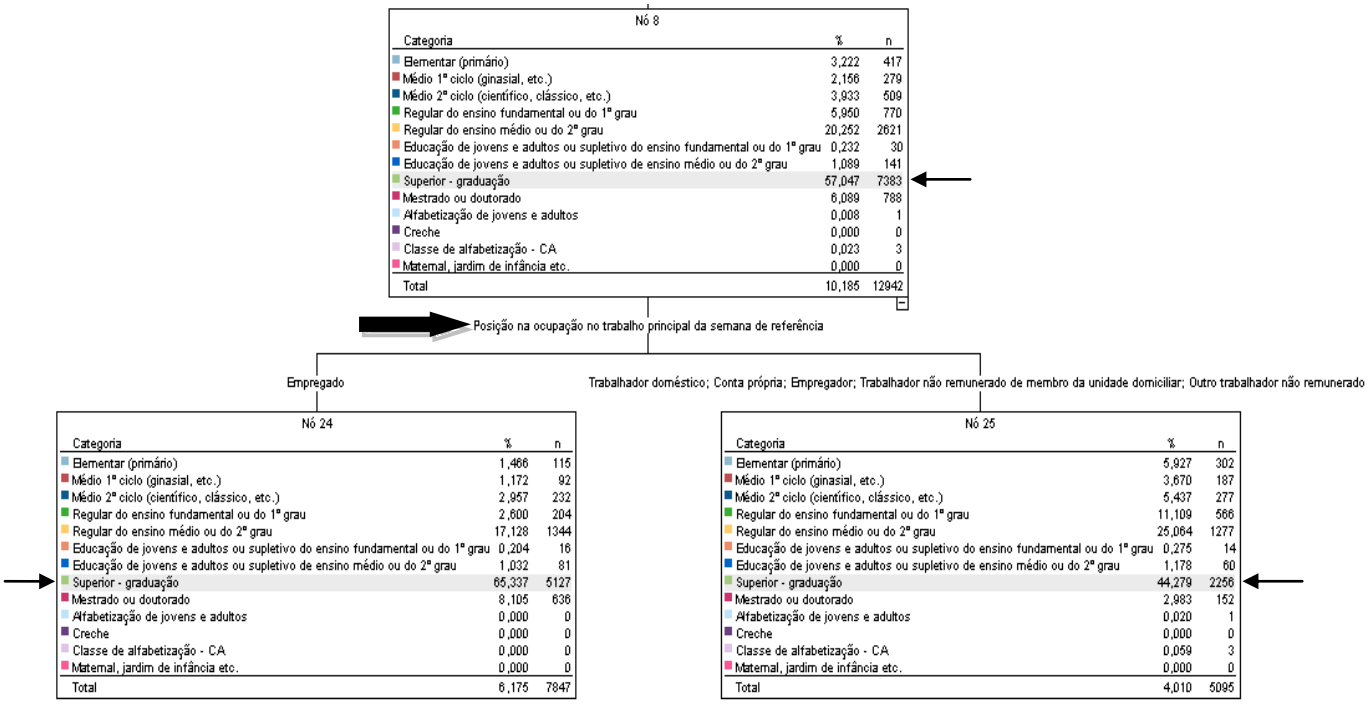


Figura 3.16 - Ramo onde o rendimento é maior que R$ 2.498,00

3.6 Discussão

Com o método CHAID foi possível levantar as seguintes conclusões com os dados do PNAD de todo Brasil, no ano de 2008:

O método CHAID após verificar todas as variáveis, avaliou que a variável mais importante ligada ao grau de instrução dos entrevistados que melhor classifica os registros presentes na base é a renda mensal.

Segundo os resultados encontrados pelo CHAID, a variável que mais influencia na definição da renda mensal em seis dos oito ramos gerados pelo algoritmo é a posição no trabalho.

Entre os entrevistados que possuem as menores rendas, ou seja, rendimento menor ou igual a R$ 414,00, 47,53% concluiu apenas o ensino fundamental. Já para as pessoas com renda mensal maior que R$ 800,00 e menor ou igual a R$ 1.000,00, a variável mais importante foi possui computador sendo que dos que tem um microcomputador, 44,20% completou o segundo grau e 22,94% o nível superior. Pode parecer estranho esse resultado, mas temos que levar em questão, entre outros fatores, que o governo tem dado incentivos as famílias adquirirem computadores o que pode explicar esse conhecimento minerado.

Dos consultados pela pesquisa com rendimento maior que R$ 1.480,00 e menor ou igual a R$ 2.498,00, a variável mais determinante na classificação dos registros pertencentes a essa classe de valores encontrado pelo método CHAID, foi o sexo, onde dos homens, 39,9% completou o ensino médio e das mulheres 57,89% possui o ensino superior, o que pode levantar a questão se existe diferenças salariais entre homens e mulheres com o mesmo grau de instrução, pois mulheres com mais estudo do que homens estão na mesma faixa de rendimentos mensais.

Entre os entrevistados que possuem as maiores rendas do Brasil, 57,04% concluiu o ensino superior, sendo a posição no trabalho um dos fatores determinantes onde da categoria de empregados, 65,33% tem o terceiro grau completo.

Por outro lado, apartir destas conclusões podemos também gerar as seguintes regras para os perfis de pessoas:

1) Regra para rendimento mensal <= R$ 414,00
    - Se posição no trabalho = empregado ou outro trabalhador não remunerado então
        - 48,03% de chance de possuir o ensino fundamental
        - 36,07% de chance de ter o ensino médio
    - Se posição no trabalho = trabalhador doméstico ou conta própria ou trabalhador da construção então
        - 48,19% de chance de possuir o ensino fundamental
        - 24,59% de chance de ter o ensino médio
    - Se posição no trabalho = empregador ou trabalhador não remunerado de outro membro da família então
        - 40,79% de chance de possuir o ensino fundamental
        - 36,52% de chance de ter o ensino médio
2) Regra para rendimento mensal > R$ 414,00 e rendimento mensal <= R$ 450,00
    - Se posição no trabalho = empregado ou empregador ou trabalhador da construção então
        - 48,06% de chance de possuir o ensino médio
        - 35,63% de chance de ter o ensino fundamental

- Se posição no trabalho = trabalhador doméstico ou conta própria ou trabalhador não remunerado de outro membro da família ou outro trabalhador não remunerado então
    - 45,97% de chance de possuir o ensino fundamental
    - 23,43% de chance de ter o ensino médio

3) Regra para rendimento mensal > R$ 450,00 e rendimento mensal <= R$ 649,00
    - Se posição no trabalho = empregado então
        - 48,69% de chance de possuir o ensino médio
        - 33,33% de chance de ter o ensino fundamental
    - Se posição no trabalho = trabalhador doméstico ou conta própria ou trabalhador não remunerado de outro membro da família ou outro trabalhador não remunerado ou empregador ou trabalhador da construção então
        - 42,98% de chance de possuir o ensino fundamental
        - 24,03% de chance de ter o ensino médio
4) Regra para rendimento mensal > R$ 649,00 e rendimento mensal <= R$ 800,00
    - Se posição no trabalho = empregado ou empregador então
        - 48,95% de chance de possuir o ensino médio
        - 29,74% de chance de ter o ensino fundamental
    - Se posição no trabalho = trabalhador doméstico ou conta própria ou trabalhador não remunerado de outro membro da família ou outro trabalhador não remunerado ou trabalhador da construção então
        - 38,84% de chance de possuir o ensino fundamental
        - 27,43% de chance de ter o ensino médio
5) Regra para rendimento mensal > R$ 800,00 e rendimento mensal <= R$ 1.000,00
    - Se tem computador = sim então
        - 44,29% de chance de possuir o ensino médio
        - 22,94% de chance de ter nível superior
    - Se tem computador = não então
        - 38,51% de chance de possuir o ensino médio
        - 31,53% de chance de ter o ensino fundamental

6) Regra para rendimento mensal > R$ 1.000,00 e rendimento mensal <= R$ 1.480,00
    - Se posição no trabalho = empregado então
        - 42,13% de chance de possuir o ensino médio
        - 24,47% de chance de ter o nível superior
    - Se posição no trabalho = trabalhador doméstico ou conta própria ou trabalhador não remunerado de outro membro da família ou outro trabalhador não remunerado ou empregador ou trabalhador da construção então
        - 26,94% de chance de possuir o ensino médio
        - 26,90% de chance de ter o ensino fundamental
7) Regra para rendimento mensal > R$ 1.480,00 e rendimento mensal <= R$ 2.498,00
    - Se o sexo = masculino então
        - 39,90% de chance de possuir o ensino médio
        - 23,27% de chance de ter nível superior
    - Se o sexo = feminino então
        - 57,89% de chance de possuir nível superior
        - 24,53% de chance de ter o ensino médio
8) Regra para rendimento mensal > R$ 2.498,00
    - Se posição no trabalho = empregado então
        - 65,33% de chance de possuir nível superior
        - 17,12% de chance de ter o ensino médio
    - Se posição no trabalho = trabalhador doméstico ou conta própria ou trabalhador não remunerado de outro membro da família ou outro trabalhador não remunerado ou empregador ou trabalhador da construção
        - 44,27% de chance de possuir nível superior
        - 25,06% de chance de ter o ensino médio

Para melhor visualização das regras encontrados, elaboramos o Quadro 3.3.

Quadro 3.3 - Regras encontradas

Rendimento Mensal	Variável mais importante	Classes da variável	Resultado encontrado
[R$ 0,00 - R$ 414,00]	Posição no trabalho	Empregado ou outro trabalhador não remunerado	48,03% o ensino fundamental 36,07% o ensino médio
		Trabalhador doméstico ou conta própria ou trabalhador da construção	48,19% o ensino fundamental 24,59% o ensino médio
		Empregador ou trabalhador não remunerado de outro membro da família	40,79% o ensino fundamental 36,52% o ensino médio
(R$ 414,00 - R$ 450,00]	Posição no trabalho	Empregado ou empregador ou trabalhador da construção	48,06% o ensino médio 35,63% o ensino fundamental
		Trabalhador doméstico ou conta própria ou trabalhador não remunerado de outro membro da família ou outro trabalhador não remunerado	45,97% o ensino fundamental 23,43% o ensino médio
(R$ 450,00 - R$ 649,00]	Posição no trabalho	Empregado	48,69% o ensino médio 33,33% o ensino fundamental
		Trabalhador doméstico ou conta própria ou trabalhador não remunerado de outro membro da família ou outro trabalhador não remunerado ou empregador ou trabalhador da construção	42,98% o ensino fundamental 24,03% o ensino médio
(R$ 649,00 - R$ 800,00]	Posição no trabalho	Empregado ou empregador	48,95% o ensino médio 29,74% o ensino fundamental
		Trabalhador doméstico ou conta própria ou trabalhador não remunerado de outro membro da família ou outro trabalhador não remunerado ou trabalhador da construção	38,84% o ensino fundamental 27,43% o ensino médio
(R$ 800,00 - R$ 1.000,00]	Tem computador	Sim	44,29% o ensino médio 22,94% o nível superior
		Não	38,51% o ensino médio 31,53% o ensino fundamental
(R$ 1.000,00 - R$ 1.480,00]	Posição no trabalho	Empregado	42,13% o ensino médio 24,47% o nível superior
		Trabalhador doméstico ou conta própria ou trabalhador não remunerado de outro membro da família ou outro trabalhador não remunerado ou empregador ou trabalhador da construção	26,94% o ensino médio 26,90% o ensino fundamental
(R$ 1.480,00 - R$ 2.498,00]	Sexo	Masculino	39,90% o ensino médio 23,27% o nível superior
		Feminino	57,89% o nível superior 24,53% o ensino médio

(R$ 2.498,00)	Posição no trabalho	Empregado	65,33% o nível superior 17,12% o ensino médio
		Trabalhador doméstico ou conta própria ou trabalhador não remunerado de outro membro da família ou outro trabalhador não remunerado ou empregador ou trabalhador da construção	44,27% o nível superior 25,06% o ensino médio

Nesse capitulo vimos o método CHAID aplicado a um estudo de caso de forma detalhada. Nele mostramos como usamos o algoritmo CHAID para a base de dados PNAD, realizada no ano de 2008, para traçar o perfil socioeconômico relacionando o nível de escolaridade dos entrevistados com a situação econômica dos mesmos. Explicamos como foi feito o processo de escolha das variáveis de interesse, a limpeza dos dados, a análise e interpretação dos resultados, além de levantarmos discussão sobre o conhecimento minerado. Em seguida, vamos apresentar algumas conclusões, levantar as dificuldades encontradas, avaliar o que foi desenvolvido e explicar quais estensões do mesmo.

4. CONCLUSÃO

Essa monografia apresentou as etapas e técnicas para se chegar ao conhecimento através da Mineração de Dados em grandes repositórios de dados, onde foi mostrado a importância de sua utilização nas mais variadas aplicações. Pode-se concluir que a extração de conhecimento é feita através de um conjunto complexo de etapas, mas que no final ajuda a descobrir informações que estavam implícitas na base de dados, pois com a utilização de técnicas e ferramentas de análise de dados é possível obter uma visão mais detalhada sobre o que o repositório de dados possui de mais relevante, auxiliando na tomada de decisão e ajudando no aumento da produtividade das empresas públicas e privadas.

Apresentamos todas as etapas de KDD necessárias para a extração de conhecimento, além de algumas de suas técnicas. Mostramos como utilizar e interpretar as Árvores de Decisão e quais são as principais técnicas. Entre essas técnicas, mostramos de forma detalhada o método CHAID e o aplicamos em um estudo de caso.

Nesse sentido, aplicamos o algoritmo CHAID na base de dados da PNAD realizada pelo IBGE no ano de 2008, para traçar um perfil socioeconômico relacionando o nível de escolaridade dos entrevistados com a situação econômica dos mesmos. Sendo assim, encontramos o perfil pretendido com a aplicação do método e concluímos que variável mais importante ligada ao grau de instrução dos entrevistados que melhor classifica os registros presentes na base é a renda mensal. Percebemos também, que a posição no trabalho é variável que mais influencia na definição da renda mensal na maioria ramos gerados pelo algoritmo.

Entre os entrevistados com as menores rendas, concluímos que a maioria possui apenas o ensino fundamental, sendo que o aumento do rendimento mensal das pessoas está diretamente relacionado ao grau de instrução delas, onde aquelas que apresentam os maiores rendimentos do Brasil a maioria tem nível superior completo.

Para trabalhos futuros, propomos utilizar para a base de dados PNAD, outros algoritmos de classificação que também usem informações socioeconômicas como entrada e tenham como resultado uma Árvores de Decisão, a fim de compararmos com o método CHAID, o desempenho e o conhecimento minerado pelos algoritmos.

Propomos também, usar o algoritmo CHAID e outras bases de dados de pesquisas com informações semelhantes as da PNAD realizadas, por exemplo, em países desenvolvidos e subdesenvolvidos, a fim de traçarmos o perfil socioeconômico relacionando o nível de escolaridade dos entrevistados com a situação econômica dos mesmos, para analisarmos as diferenças entre os perfis encontrados nesses países e no Brasil.

Outro trabalho que também pode ser desenvolvido é a utilização da PNAD para desenvolver um modelo estatístico que trace o perfil socioeconômico dos entrevistados. Para isso, parte da base de dados seria utilizada para treinamento e outra para testes, a fim de avaliar os modelos gerados.

REFERÊNCIAS

AMO, S. Curso de Data Mining do Programa de Mestrado em Ciência da Computação da Universidade Federal de Uberlândia, 2003. Disponível em: <http://www.deamo.prof.ufu.br/CursoDM.html>. Acesso em: março de 2012.

AURÉLIO, M.; VELLASCO, M.; LOPES C. H. Descoberta de conhecimento e mineração de dados apostila, 1999. Disponível em: < http://www.ica.ele.puc-rio.br/cursos/download/DM-apostila1.pdf>. Acesso em: setembro de 2011.

BABTISTA J.; CARVALHO D. R. Data mining como apoio a decisão em projetos públicos. In: CONGRESSO BRASILEIRO DE COMPUTAÇÃO, 3., 2003, Itajaí. Anais...Itajaí:Univali-CTTMar, 2003. Disponível em: < http://www.ppgia.pucpr.br/~silla /publications/index.html>. Acesso em: janeiro de 2012.

BERRY, M. J. A.; LINOFF, G. S. Data mining techniques: for marketing, sales, and customer support. In: Michael J. A. Berry, Gordon Linoff, New York: JohnWiley, 2004. Disponível em: < http://www.wiley.com/WileyCDA/WileyTitle/productCd-0471470643. html>. Acesso em: setembro de 2011.

CARVALHO, L. A. V. Datamining: a Mineração de Dados no Marketing, Medicina, Economia, Engenharia e Administração. São Paulo: Érica, 2001.

COLLAZOS, K.; BARRETO, J. KDD ferramenta para análise de dados epidemiológicos. In: Anais do III Congresso Brasileiro de Computação – Workshop de Informática aplicada à Saúde - CBCOMP'2003, Itajaí, 2003.

DIAS, M. M. Parâmetros na escolha de técnicas e ferramentas de mineração de dados. Departamento de Informática, Maringá, Paraná, 2002. Disponível em: <http://eduem.uem.br/ojs/index.php/ActaSciTechnol/article/ download/2549/1569>. Acesso em: março de 2012.

DOMINGUES, M. A. Generalização de regras de associação. Dissertação de Mestrado, ICMC-USP, São Carlos, São Paulo, 2004. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-10082004-154242/pt-br.php > . Acesso em: março de 2012.

GARCIA, S.C. O uso de árvores de decisão na descoberta de conhecimento na área da saúde. In: SEMANA ACADÊMICA, 2000. Rio Grande do Sul: Universidade Federal do Rio Grande do Sul, 2000. Disponível em: <http://www.lume.ufrgs.br/handle/10183/4703>. Acesso em: março de 2012.

HAIR, J. F.; ANDERSON, R. E.; TATHAM R. L.; BLACK, W. C. Multivariates Data Analysis. 4. ed. New Jersey: Prentice Hall, 1995. Disponível em: <http://books.google.com.br/books?id=S1gZAQAAIAAJ&hl=pt-BR&source=gbs_book_other_versions>. Acesso em: março de 2012.

HAN, J.; KAMBER, M. Data Mining: Concepts and Techniques, In: Academic Press, USA, 2006. Disponível em: <http://www.cs.uiuc.edu/homes/hanj/bk2/toc.pdf>. Acesso em: março de 2012.

HEINECK, L. F. M.; FREITAS, A. A. F. Segmentação de mercado: proposta de uma metodologia de associação entre clientes e produtos no contexto do mercado imobiliário. Universidade Federal do Ceará, 2008. Disponível em: <http://seer.ufrgs.br/ambienteconstruido/article/view/5181>. Acesso em: março de 2012.

IBGE. Pesquisa Nacional por Amostra de Domicílios, vol. 29, Brasil, 2008. Disponível em: < http://www.ibge.gov.br/home/estatistica/populacao/trabalhoerendimento/ pnad2008 /brasilpnad2008.pdf>. Acesso em: maio de 2012.

KLEINSCHMIDT, M. Mineração de dados para avaliação do perfil de usuários do Sistema de Informação da Academia da UNIVALI. Trabalho de Conclusão de Curso, Itajaí: UNIVALI, 2007. Disponível em: <http://siaibib01.univali.br/pdf/Marlon%20 Kleinschmidt.pdf>. Acesso em: março de 2012.

MUELLER, Alessandro. Uma Aplicação de Redes Neurais Artificiais na Previsão do Mercado Acionário. Dissertação de Mestrado, UFSC, Florianópolis. 1996. Disponível em: <http://www.eps.ufsc.br/disserta96/mueller/index/index.htm#sumario>. Acesso em: maio de 2012.

MORGAN, J.N. e SONQUIST, J.A. Problems in the analysis of survey data and a proposal, Journal of the American Statistical Association, v. 58, 1963.

PACHECO, M. A.; VELLASCO, M.; LOPES, C. H. Descoberta de conhecimento e mineração de dados, Notas de Aula em Inteligência Artificial. Rio de Janeiro, ICA-Laboratório de Inteligência Computacional Aplicada, departamento de Engenharia elétrica – PUC-RIO, 1999. Disponível em: <www.ica.ele.puc-rio.br/cursos/download/ DMapostila1.pdf>. Acesso em: setembro de 2011.

POZZER, C. T. Aprendizado por Árvores de Decisão. In: Universidade Federal de Santa Maria, Notas de Aula da Disciplina de Programação de Jogos 3D, Rio Grande do Sul, 2006. Disponível em: <http://www-usr.inf.ufsm.br/~pozzer/disciplinas/ pj3d_decisionTrees.pdf>. Acesso em: março de 2012.

RODRIGUES, M. A. S. Árvores de Classificação. Trabalho de Conclusão de Curso, In: Universidade dos Açores, departamento de Matemática, Ponta Delgada, 2005. Disponível em: < http://www.amendes.uac.pt/monograf/ monograf05arvoreClass.pdf >. Acesso em: março de 2012.

ROMÃO, W. Descoberta de Conhecimento Relevante em Banco de Dados sobre Ciência e Tecnologia. Trabalho de Pós-Graduação. Florianópolis-SC : UFSC, 2002. Disponível em: < http://www.din.uem.br/~intersul/intersul_arquivos/documentos/Tese %20Wesley.pdf>. Acesso em: março de 2012.

SCHENATZ, B. N. Utilização de Data Mining em um sistema de informação gerencial para o diagnóstico da formação de professores da graduação. 2005. Dissertação de Mestrado, UFSC, Florianópolis. 2005. Disponível em: < http://aspro02.npd.ufsc.br/arquivos/220000/224900/ 18_224929.htm>. Acesso em: março de 2012.

SILVA, S. G. Estudo de técnicas e utilização de mineração de dados em uma base de dados da saúde pública. Trabalho de Conclusão de Curso, Universidade Luterana do Brasil, Canoas, 2003. Disponível em: < http://pt.scribd.com/doc/56537190/Estudo-de-Tecnicas-e-Utilizacao-de-Mineracao-de-Dados-em-uma-base-de-dados-da-saude-publica-gercely-da-silva>. Acesso em: março de 2012.

SILVEIRA, R. F. Mineração de dados aplicada a definição de índices em sistemas de raciocínio baseado em casos. Dissertação de Mestrado, Universidade Federal do Rio Grande do Sul, Porto Alegre. 2003. Disponível em: < http://www.inf.ufrgs.br/bdi/administrator/components/com_jresearch/files/publications/Websis-RosemariSilveira.pdf >. Acesso em: março de 2012.

SILVEIRA, M. M. Estratégias de aplicação de análise estatística multivariada no desenvolvimento de novos produtos. Dissertação de Mestrado. Universidade Federal do Rio Grande do Sul, Porto Alegre. 2010. Disponível em: < http://www.lume.ufrgs.br/bitstream/handle/10183/28793/000770246.pdf?sequence=1 >. Acesso em: março de 2012.

TURE M.; TOKATLI F.; KURT I. Using Kaplan-Meier Analysis together with Decision Tree Methods (C&RT, CHAID, QUEST, C4.5 and ID3) in Determining Recurrence-free Survival of Breast Cancer Patients. Science Direct. Expert System with Applications 36. 2009. Disponível em: < http://dl.acm.org/citation.cfm?id=1465032>. Acesso em: março de 2012.