INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA INFORMÁTICA




Descoberta de Conhecimento em Dádivas de Sangue

FÁBIO XAVIER CASTRO SILVA
Outubro de 2015

POLITÉCNICO
DO PORTO

Instituto Superior de Engenharia do Porto

Descoberta de Conhecimento em Dádivas de Sangue

Fábio Xavier Castro Silva

Dissertação para obtenção do Grau de Mestre em Engenharia Informática, Área de Especialização em Tecnologias de Conhecimento e Apoio à Decisão

Orientador: Professor Doutor Paulo Oliveira

Júri:

Presidente:

Vogais:

Porto, Outubro 2015

Dedicatória

Quero dedicar este trabalho à minha namorada, família e amigos por toda a força e alento proporcionado para a conclusão do mestrado.

Resumo

Atualmente, são geradas enormes quantidades de dados que, na maior parte das vezes, não são devidamente analisados. Como tal, existe um fosso cada vez mais significativo entre os dados existentes e a quantidade de dados que é realmente analisada. Esta situação verifica-se com grande frequência na área da saúde. De forma a combater este problema foram criadas técnicas que permitem efetuar uma análise de grandes massas de dados, retirando padrões e conhecimento intrínseco dos dados.

A área da saúde é um exemplo de uma área que cria enormes quantidades de dados diariamente, mas que na maior parte das vezes não é retirado conhecimento proveitoso dos mesmos. Este novo conhecimento poderia ajudar os profissionais de saúde a obter resposta para vários problemas.

Esta dissertação pretende apresentar todo o processo de descoberta de conhecimento: análise dos dados, preparação dos dados, escolha dos atributos e dos algoritmos, aplicação de técnicas de mineração de dados (classificação, segmentação e regras de associação), escolha dos algoritmos (C5.0, CHAID, Kohonen, TwoSteps, K-means, Apriori) e avaliação dos modelos criados. O projeto baseia-se na metodologia CRISP-DM e foi desenvolvido com a ferramenta Clementine 12.0. O principal intuito deste projeto é retirar padrões e perfis de dadores que possam vir a contrair determinadas doenças (anemia, doenças renais, hepatite, entre outras) ou quais as doenças ou valores anormais de componentes sanguíneos que podem ser comuns entre os dadores.

Palavras-Chave: Mineração de dados, CRISP-DM, Saúde, Clementine 12.0, Classificação, Segmentação, Regras de Associação

Abstract

Currently, enormous quantities of data are generated which are often not properly analyzed. As such, there is a significant and increasing ditch between the existing data and the quantity that is actually analyzed. This occurs mostly on the healthy area. In order to combat this problem, techniques were created that allow to perform an analysis of a big quantity of data. These techniques permit to retrieve patterns and knowledge intrinsic on data.

The healthy area is an example of an area that creates a lot of data but in most cases it does not retrieves useful knowledge. This new knowledge could help the healthy professionals to obtain answers to various problems.

This dissertation intends to present all the process of data mining: analysis of data, preparation of data, choice the attributes and algorithms, application of data mining techniques (classification, clustering and association rules), choice of algorithms (C5.0, CHAID, Kohonen, TwoSteps, K-means, Apriori) and evaluation of created models. The project is based on CRISP-DM methodology and was developed with Clementine 12.0 tool. The main objective of this project is to retrieve patterns and profiles of givers that can come to suffer some diseases (anemia, renal diseases, hepatitis, among others) or which diseases or anomalous values of sanguineous components that can be common between givers.

Keywords: Data Mining, CRISP-DM, Healthy, Clementine 12.0, Classification, Clustering, Association Rules

Agradecimentos

A realização deste projeto não teria sido possível sem a ajuda de todos aqueles que, de alguma forma, contribuíram para um processo de desenvolvimento apoiado e um final de sucesso, e aos quais gostaria de expressar o meu mais profundo agradecimento.

Gostaria de agradecer ao meu orientador do ISEP, Professor Paulo Oliveira, por todo o apoio ao longo da elaboração da minha dissertação.

Queria expressar também o meu agradecimento a todos os que indiretamente contribuíram para a finalização da minha dissertação, em especial aos meus pais, irmãos, namorada e amigos.

Índice

Lista de Figuras

Lista de Tabelas

Acrónimos e Símbolos

Lista de Acrónimos

Ag-HBs	Antigénio de Superfície do Vírus da Hepatite B
ALT	Alanina Aminotransferase
Anti-HBc	Anticorpo contra o antigénio da superfície do vírus da Hepatite B
Anti-HCV	Anticorpo para o vírus da hepatite C
CRISP-DM	*Cross Industry Standard Process for Data Mining*
ECD	Extração de Conhecimento de Dados
FN	Falsos Negativos
FP	Falsos Positivos
FPR	Taxa de Falsos Positivos
HCT	Hematócrito
HGB	Hemoglobina
HIV	Vírus da Imunodeficiência Humana
IAM	Inteligência Artificial Médica
MCH	Hemoglobina Globular Média
MCHC	Concentração de Hemoglobina Globular Média
MCV	Volume Globular Médio
NEU	Neutrófilos
PLT	Plaquetas
RH	Rhesus
RNAs	Redes Neuronais Artificiais
ROC	*Receiver Operating Characteristic*
SP	Sistemas Periciais
SIDA	Síndrome de ImunoDeficiência Humana Adquirida

TN	Negativos Verdadeiros
TP	Positivos Verdadeiros
TPR	Taxa de Verdadeiros Positivos
VDRL	*Veneral Disease Research Laboraroty*
WBC	*White Blood Cell Count*

Lista de Símbolos

µL	Milímetro cúbico
mmHg	Miligramas de mercúrio

1 Introdução

Na computação existem diversos problemas que são resolvidos através da escrita de um algoritmo. Contudo, existem inúmeros acontecimentos do quotidiano que não podem ser transcritos para um algoritmo, como por exemplo, o reconhecimento de expressões faciais ou o estado de espírito das pessoas [Gama et al., 2012].

A crescente complexidade dos problemas e o elevado volume de dados levou à necessidade de desenvolvimento de ferramentas computacionais que diminuíssem a dependência humana. Para tal, foram criadas técnicas baseadas na estatística, recuperação de dados e inteligência artificial. Estas devem ser capazes de criar autonomamente uma hipótese ou função, a partir da experiência passada, para a resolução de problemas. As técnicas de extração de conhecimento de dados têm sido bastante utilizadas com sucesso num grande número de problemas reais.

O processo de mineração de dados consiste na exploração de grandes volumes de dados à procura de padrões válidos, consistentes, novos e potencialmente úteis (Gama et al., 2012). Este processo é bastante usual em bases de dados de grande dimensão e o resultado final pode ser representado sob a forma de regras, árvores de decisão, etc. O conhecimento adquirido pode ter diversas finalidades, tais como gestão de informação ou tomada de decisão das organizações (previsão de comportamentos e tendências futuras) [Côrtes et al., 2002].

Apesar da área da saúde gerar enormes quantidades de dados diariamente, estes nem sempre são devidamente analisados. Assim sendo, é criado um fosso cada vez mais significativo entre a quantidade de dados existente e a porção de dados que é analisada e compreendida ao longo do tempo. Diversos estudos indicam que a cada vinte meses a quantidade de dados armazenada em todos os repositórios do mundo duplica [Gama et al., 2012]. A utilização deste processo tem contribuído significativamente na elaboração de diagnósticos mais precisos na área da saúde.

1.1 Contextualização

Este documento insere-se no âmbito do desenvolvimento da tese/dissertação do ramo de mestrado de Tecnologias de Conhecimento e Apoio à Decisão, inerente ao curso de Engenharia Informática do Instituto Superior de Engenharia Informática do ano 2014/2015, e pretende documentar toda a análise efetuada ao tema "Extração de Conhecimento num conjunto de dados de dádivas de sangue". Esta análise processa-se pela descoberta de conhecimento inerente nos dados relativos a dádivas de sangue, recolhidos por brigadas móveis em centros de colheita Portugueses. A aquisição de conhecimento é feita através da aplicação de técnicas de mineração de dados, nomeadamente classificação, segmentação e regras de associação. Seguidamente, o conhecimento obtido é avaliado e interpretado para que seja possível chegar e responder aos objetivos delineados.

1.2 Motivação

A informatização dos meios produtivos levou à geração de elevados volumes de dados através de transações eletrónicas, novos equipamentos de observação, dispositivos de armazenamento em massa, entre outros. Assim sendo, sabendo que o conhecimento é poder e que o aproveitamento da informação influencia significativamente o ganho de competitividade, foi necessária a criação de ferramentas que auxiliassem no processo de análise de dados. Aqui surge o processo de extração de conhecimento nos dados. Este processo veio permitir:

- Maior facilidade na análise de enormes bases de dados;
- Descoberta de padrões válidos e, consequentemente, informações inesperadas;
- Novas formas de recolha de dados;
- Modelos de fácil compreensão.

O conjunto de dados relativo a dádivas de sangue a analisar e trabalhar permitirá retirar conhecimento que poderá ser ou não interessante para a área da saúde. Este conhecimento poderá ser analisado e utilizado por profissionais de saúde e, assim sendo, poderá tornar-se útil no diagnóstico e no tratamento de pacientes.

Os dados são cada vez mais importantes para as organizações. Este projeto representa uma oportunidade na exploração e utilização do processo de mineração dos dados na área da saúde, tendo esta área um fosso tão significativo na exploração de dados. Como tal, as motivações pessoais e académicas para este trabalho são bastante elevadas.

1.3 Objetivos

Com esta análise, através da utilização da metodologia CRISP-DM (*Cross-Industry Standard Process for Data Mining)*, pretende-se aplicar técnicas de mineração de dados num conjunto de dados relativo a dádivas de sangue para aquisição de novo conhecimento. Este conhecimento permitirá ajudar a responder a determinados cenários e objetivos traçados, tais como:

- Definição de perfis de dadores que poderão vir a sofrer de possíveis problemas no fígado, doenças autoimunes, entre outras;
- Definição de parâmetros com valores elevados que podem estar associados num boletim analítico;
- Definição da sequência temporal de valores anormais, ou seja, quais os parâmetros de um boletim analítico que podem levar a que outros parâmetros se tornem anormais no futuro.

A aquisição do conhecimento relativo aos objetivos assumidos poderá ter influência nos centros de colheita de sangue através da identificação de padrões importantes. Estes padrões podem melhorar previsões e auxiliar na deteção e correção de anomalias que acontecem frequentemente nos hospitais e centros de colheita.

Uma dissertação é um processo bastante dinâmico onde novos conhecimentos e curiosidades vão surgindo. Assim sendo, outros objetivos secundários foram surgindo ao longo do desenvolvimento do trabalho.

1.4 Metodologia

A utilização de uma metodologia facilita a compreensão, implementação e desenvolvimento do processo de mineração de dados. Para o desenvolvimento de modelos preditivos especializados no ramo da saúde, este estudo baseou-se na metodologia CRISP-DM [IBM, 2011].

A ferramenta utilizada para este estudo denomina-se Clementine 12.0 e utiliza a metodologia mencionada. Importa salientar que é uma ferramenta bastante poderosa e versátil, com uma interface muito intuitiva. Esta metodologia (Figura 1) permite obter uma visão geral do projeto de mineração de dados, baseando-se num ciclo de seis fases [IBM, 2011]:

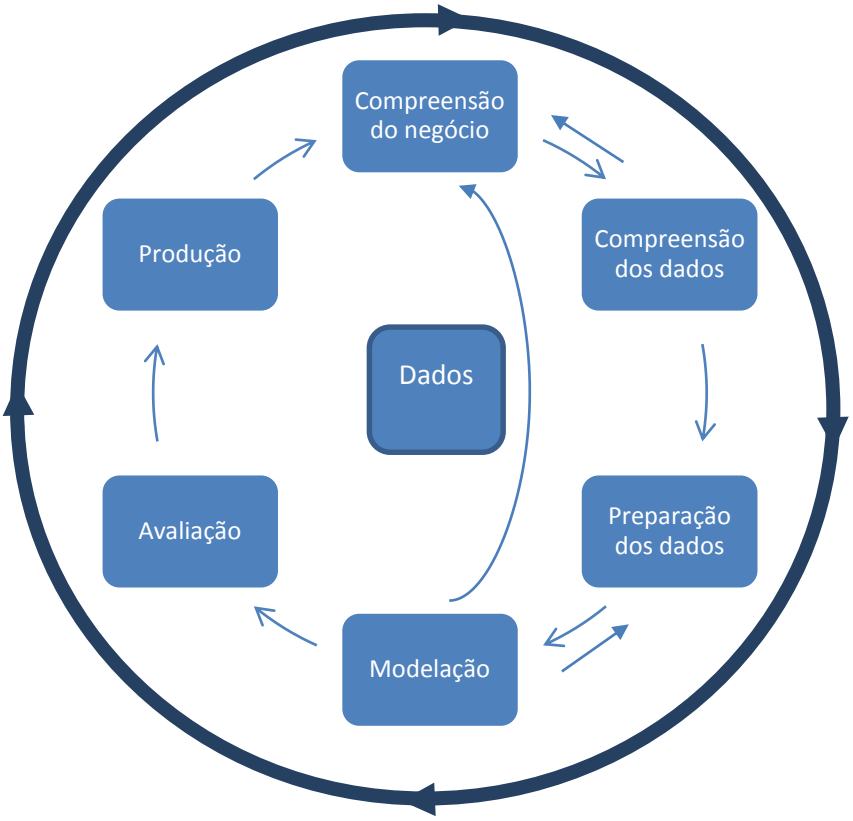

Figura 1. Metodologia CRISP-DM.

- **Compreensão do negócio (Business Understanding)**
  Esta fase principia o entendimento dos objetivos, requisitos e exigências do projeto. É necessário formular o entendimento do problema e perceber qual o seu enquadramento. Seguidamente, é desenhado um plano de implementação para o problema de acordo com o processo de mineração de dados;

- **Compreensão dos dados (Data Understanding)**
  A compreensão dos dados inicia com a identificação das fontes de dados existentes e tem como objetivo aumentar a familiaridade com os dados facilitando a descoberta de caraterísticas principais.

- **Preparação dos dados (Data Preparation)**
  Esta fase agrupa todas as tarefas para a construção do conjunto final de dados:
  - Escolha dos dados a usar para a análise desejada. Os critérios desta escolha devem ter em consideração os objetivos do problema de mineração de dados e determinadas condições técnicas (volume, tipo de dados, entre outras);
  - Limpeza e tratamento dos dados.

- **Fase de Modelação (Modeling)**
  É aqui que os algoritmos são selecionados e os parâmetros são ajustados para otimização dos resultados. Importa salientar que "o ajuste dos parâmetros implica a reconstrução do modelo" (Anacleto, 2009). Isto ocorre até encontrar os resultados

desejados. No final da construção dos modelos são criados procedimentos para validar a qualidade dos modelos.

- **Fase de Avaliação (Evaluation)**
  A fase de avaliação dos modelos é crucial para analisar e estimar de que forma os modelos criados respondem ao problema formulado na fase de compreensão do negócio. No final, deve ser tomada a decisão relativamente à aplicação dos resultados obtidos.

- **Fase de Produção (Deployment)**
  A interpretação dos resultados pode resultar na aplicação dos modelos criados a novos conjuntos de dados, na integração com aplicações externas ou na construção de relatórios.

Apesar de não garantir resultados, esta metodologia permite disciplinar o processo alinhando os objetivos do projeto de mineração dos dados com o negócio.

1.5 Estrutura do documento

Este documento está redigido segundo o novo acordo ortográfico e organiza-se em seis capítulos: Introdução, Análises Sanguíneas, Descoberta de Conhecimento nos Dados, Exploração e Preparação dos Dados, Modelação e Avaliação e Conclusões. Esta secção do documento pretende descrever e sintetizar os seis capítulos declarados anteriormente.

O presente capítulo "**Introdução"** pretende efetuar uma apresentação da análise, aduzindo as respetivas motivações, finalidades e contextualização do trabalho desenvolvido. Também é referida e detalhada a metodologia utilizada para o desenvolvimento do mesmo e a estrutura do documento.

No capítulo **"Análises Sanguíneas"** é focado o tema "parâmetros associados a análises clínicas". Aqui são detalhados os parâmetros avaliados numa análise sanguínea.

Relativamente ao capítulo "**Descoberta de Conhecimento nos Dados",** tem como intuito apresentar um estudo acerca de todas as fases associadas ao processo de descoberta de conhecimento nos dados.

O capítulo "**Exploração e Preparação dos Dados"** apresenta as caraterísticas principais dos dados, ajudando assim a uma maior familiarização com os mesmos. Também tem como objetivo detalhar todo o processo efetuado com a ferramenta *SQL Server Data Tools,* assim como apresentar todas as alterações efetuadas aos dados de forma a diminuir os dados inconsistentes e com ruído. O projeto do tipo *Microsoft SQL Server Integration Services* efetua a migração de uma base de dados desenvolvida em *Access* para *SQL*.

O capítulo denominado **"Modelação e Avaliação"** pretende apresentar os modelos criados de acordo com os algoritmos utilizados e a avaliação dos mesmos.

Finalmente, o capítulo **"Conclusões"** sintetiza toda a investigação realizada neste trabalho de mestrado. Seguidamente, será realizada uma reflexão acerca dos resultados obtidos ao longo da análise efetuada. Para finalizar, será descrito algum trabalho futuro que poderá ser importante para consolidar o trabalho desenvolvido.

2 Análises Sanguíneas

Este capítulo tem como objetivo detalhar todos os parâmetros inseridos no conjunto de dados analisados por cada dádiva efetuada. Isto permitiu uma melhor compreensão dos atributos ao longo da elaboração de todo o processo de descoberta de conhecimento.

Após a dádiva, todas as unidades de sangue colhidas são sujeitas a uma série de provas analíticas e a um processo de separação dos componentes antes de serem administradas aos doentes. O sangue é um fluido circulante constituído pelo plasma, hemácias, leucócitos e plaquetas [Simscience, 2015]. É responsável pelo transporte de nutrientes e oxigénio a todos os órgãos do corpo humano e é classificado em grupos e subgrupos [Centro de Ciência Júnior, 2013]. Os grupos sanguíneos distinguem-se em A, B, AB e O e os subgrupos em Rh positivo e Rh negativo. Em Portugal, segundo um estudo efetuado pelo Instituto Português de Sangue, o grupo sanguíneo mais frequente é o A (46,6%), seguindo-se o grupo O (42,3%), B (7,7%) e AB (3,4%). Relativamente ao fator Rhesus (Rh), o mesmo estudo estima que 85,5% da população Portuguesa possui tipo sanguíneo Rh positivo.

Todos os indivíduos saudáveis, com hábitos de vida salutares e com peso igual ou superior a 50 Kg podem doar sangue [Instituto Português do Sangue e da Transplantação, 2015]. A doação de sangue por parte do dador pode ser espontânea, vinculada ou autóloga. Uma doação espontânea consiste numa ação solidária com o único interesse de ajudar o próximo. Uma doação vinculada destina-se a um paciente específico ao passo que a doação autóloga se destina ao próprio indivíduo.

Os indivíduos do sexo masculino podem efetuar dádivas de sangue quatro vezes por ano (com intervalos de 3 meses) e os do sexo feminino três vezes por ano (com intervalos de 4 meses).

Todavia, após a dádiva de sangue ser concluída, é feita uma análise sanguínea tendo em consideração diversos parâmetros. Seguidamente, serão descritos em detalhe os parâmetros avaliados numa análise sanguínea.

2.1 Glóbulos vermelhos

Os glóbulos vermelhos, também conhecidos como hemácias ou eritrócitos, são os elementos presentes em maior quantidade no sangue. São constituídos por hemoglobina, iões, glicose, água e enzimas e têm como principal função transportar o oxigénio a todas as células do corpo humano [TodaBiologia, 2015]. A tabela seguinte (Tabela 1) apresenta os valores normais de hemácias no sangue por milímetro cúbico.

Tabela 1. Valores de referência das hemácias no sangue [TodaBiologia, 2015].

Estado	**Hemácias**	
Normal	**Homem**	**Mulher**
	5 milhões/mm^3	4,5 milhões/mm^3
Anormal	Outros valores	

2.2 Volume globular médio

O volume globular médio (MCV) determina o volume médio dos glóbulos vermelhos do sangue. Um valor elevado do MCV normalmente evidencia uma deficiência de vitamina B12 ou de ácido fólico podendo dever-se a doenças hepáticas [Fisioterapia para todos, 2014]. Anemias causadas por carência de ácido fólico contêm hemácias com tamanho elevado, enquanto as anemias por falta de ferro apresentam hemácias com tamanho reduzido [Md.Saúde, 2015a].

A Tabela 2 indica os valores normais e anormais para o volume corpuscular médio.

Tabela 2. Valores de referência para o volume globular médio.

Estado	**MCV**
Normal	80-95 fentolitros
Anormal	Outros valores

2.3 Hematócrito

O hematócrito (HCT) é um exame de diagnóstico que permite avaliar a quantidade de glóbulos vermelhos no volume total de sangue. Um valor elevado indicia desidratação ou policitemia. Pelo contrário, um valor baixo pode indicar anemia, perda sanguínea, hemólise, leucemia, cirrose, entre outras etologias [TuaSaúde, 2014a].

De seguida são apresentados os valores de referência para este parâmetro (Tabela 3).

Tabela 3. Valores de referência para o hematócrito.

Estado	**HCT**
Normal	35%-50%
Anormal	Outros valores

2.4 Hemoglobina

A hemoglobina (HGB) é uma molécula que se encontra ao nível dos glóbulos vermelhos responsável pela oxigenação dos tecidos. A hemoglobina confere aos glóbulos vermelhos a sua forma funcional.

Um baixo valor de hemoglobina correlaciona-se com uma diminuição dos glóbulos vermelhos no sangue e denomina-se anemia. Esta condição pode ter origem diversa, como perdas de sangue (cirurgia, trauma, úlcera gástrica), causas nutricionais (défice de vitamina B12, ferro, ácido fólico), causas relacionadas com o anormal funcionamento da medula óssea ou com a administração de determinados fármacos (exemplo: quimioterapia), falência renal, entre outras [MedicineNet, 2014].Os valores de hemoglobina no sangue podem ser determinados por análise clínica, através de uma reação específica e um método de deteção apropriado. Estes valores são expressos em grama de hemoglobina por decilitro de sangue e são apresentados na Tabela 4.

Tabela 4. Valores de referência para a hemoglobina.

Estado	**HGB**	
Normal	**Homem**	**Mulher**
	14-18 gramas por decilitro	12-16 gramas por decilitro
Grave	< 8	
Anormal	Outros valores	

2.5 Hemoglobina globular média

A hemoglobina globular média (MCH) indica o valor de hemoglobina contida por glóbulo vermelho. É determinada pela razão entre o valor de hemoglobina e o número de glóbulos vermelhos [BioCampello, 2015]. Um valor inferior ao normal (Tabela 7) é indicador de hipocromia (a hemoglobina é responsável pela coloração dos glóbulos vermelhos) e um valor superior verifica-se na presença de macrocitose. Este indicador pode ser afetado pela hiperlipidémia: um valor aumentado de lípidos circulantes pode causar um falso aumento do nível de hemoglobina determinado [Laboratório Pioledo, 2015]. A Tabela 5 apresenta os valores de referência para este parâmetro.

Tabela 5. Valores de referência para hemoglobina globular média.

Estado	**MCH**
Normal	27-32 picogramas
Anormal	Outros valores

2.6 Concentração de hemoglobina globular média

A concentração de hemoglobina globular média (MCHC) indica a concentração média de hemoglobina por glóbulo vermelho. É determinada pela razão entre o valor de hemoglobina e o valor de hematócrito. Um valor baixo de MCHC pode ser consequência de perda de sangue, baixos níveis de ferro ou anemia hipocrómica. Não ocorre hipercromia devido às concentrações mais elevadas reduzirem a solubilidade da hemoglobina. A Tabela 6 apresenta os valores de referência.

Tabela 6. Valores de referência para a concentração de hemoglobina globular média.

Estado	**MCHC**
Normal	32-36 gramas por decilitro
Anormal	Outros valores

2.7 Plaquetas

As plaquetas (PLT) medeiam o processo de coagulação sanguínea. São maioritariamente utilizadas em doenças oncológicas e hematológicas e para transplantes da medula óssea. Normalmente, a contagem de plaquetas é solicitada quando o paciente demora demasiado tempo a parar de sangrar devido a um pequeno corte [Infoescola, 2015a]. Os valores de referência para o número de plaquetas encontram-se sumariados na Tabela 7.

Tabela 7. Valores de referência das plaquetas.

Estado	PLT
Normal	150.000 a 450.000 / µL de sangue
Risco	Inferior a 20.000 / µL de sangue
Anormal	Outros valores

Existem diversas causas para um valor reduzido das plaquetas: doenças autoimunes, úlceras gástricas, neoplasia da medula óssea, células cancerígenas, doenças renais, etc. Em contrapartida, pode observar-se um número elevado de plaquetas em indivíduos com uma doença sanguínea grave (síndrome mieloproliferativo). Devido à quantidade elevada de plaquetas, estas têm a tendência de se unirem umas às outras, originando coágulos e consequentemente, um bloqueio num vaso sanguíneo [TuaSaúde, 2014b].

2.8 Leucócitos

Os leucócitos, também designados por glóbulos brancos, são as células responsáveis pela defesa do organismo [Oliveira, 2011].

O valor de leucócitos no sangue ou *white blood cell count* (WBC) pode ser aferido através da realização de análises clínicas. Em média, existem cerca de 4000 a 11 000 leucócitos por milímetro cúbico de sangue (Tabela 8). Valores superiores (leucocitose) ou inferiores (leucopenia) têm significado clínico: valores de leucócitos superiores a 11 000 por milímetro cúbico de sangue podem significar que o individuo em causa esta a combater algum tipo de infeção, sofreu algum tipo de traumatismo, entre outras causas; um nível inferior a 4000 leucócitos por milímetro cúbico de sangue indica que o sistema imunitário do indivíduo se encontra debilitado, apresentando maior risco de infeção.

Tabela 8. Valores de referência para os leucócitos no sangue.

Estado	WBC
Valores normais	4000-11 000 leucócitos por milímetro cúbico de sangue
Leucocitose	> 11 000 leucócitos por milímetro cúbico de sangue
Leucopenia	< 4000 leucócitos por milímetro cúbico de sangue

Numa análise clínica é determinada a fórmula leucocitária, que consiste na determinação, em forma percentual e absoluta, do valor dos diferentes tipos de leucócitos circulantes - neutrófilos, eosinófilos, basófilos, linfócitos e monócitos [TuaSaúde, 2012]. Os neutrófilos são o tipo de leucócitos mais abundante e a determinação dos seus valores percentuais será abordada de seguida.

2.9 Percentagem de neutrófilos

Os neutrófilos (NEU) são responsáveis pela imunidade ou defesa do organismo. Estes têm a capacidade de eliminar células danificadas e combater infeções como as infeções bacterianas. Diversas situações podem estar na origem de uma percentagem elevada de neutrófilos como infeções, cancro, diabetes ou leucemia mielóide crónica. A diminuição do valor de neutrófilos pode ter na sua origem a cirrose, anemia aplástica ou leucemias agudas [TuaSaúde, 2012].

A percentagem de neutrófilos reflete assim a proporção deste tipo de leucócitos numa amostra sanguínea em função do valor leucocitário total [albanesi.it, 2015]. De seguida são apresentados os valores de referência relativos à percentagem de neutrófilos (Tabela 9).

Tabela 9. Valores de referência para os neutrófilos [Perry and Potter, 2013].

Estado	**NEU%**
Normal	55%-70%
Anormal	Outros valores

2.10 Plasma

O plasma constitui cerca de 55 a 60 % do sangue [Infoescola, 2015b]. O plasma é um componente líquido de cor amarelada que permite que as células sanguíneas fiquem suspensas. É composto por água, gases, nutrientes, hormonas e enzimas. Tem como funções:

- Transporte de nutrientes e medicamentos para todo o corpo;
- Manutenção da pressão osmótica intravascular;
- Proteção do organismo contra infeções e outros distúrbios no sangue.

2.11 Tensão Arterial

A tensão ou pressão arterial consiste na força exercida pelo sangue sobre a parede das artérias após este ter sido bombeado pelo ventrículo cardíaco esquerdo [BIAL, 2009]. É um importante parâmetro fisiológico, garantindo um eficaz aporte sanguíneo a todas as células do organismo.

A tensão arterial expressa-se em mililitros de mercúrio (mmHg) e compreende os valores de pressão sistólica ou máxima e pressão diastólica ou mínima, que correspondem à pressão exercida quando o coração bombeia o sangue através do coração e quando ocorre relaxamento cardíaco, respetivamente [Fundação Portuguesa de Cardiologia, 2014]. A tensão arterial é classificada de acordo com os valores de pressão sistólica e diastólica nas seguintes categorias (Tabela 10):

Tabela 10. Classificação dos valores de pressão arterial com base na Norma nº 020/2011 de 28/09/2011 da Direção-Geral da Saúde [Direção-Geral de Saúde, 2011].

Classificação pressão arterial	Pressão arterial sistólica		Pressão arterial diastólica
Hipotensão	< 90		< 60
Ótima	< 120	ou	< 80
Normal	120 - 129	ou	80 – 84
Normal - alta	130 - 139	ou	85 - 89
Hipertensão Grau I	140 - 159	ou	90 – 99
Hipertensão Grau II	160 - 179	ou	100 - 109
Hipertensão Grau III	≥ 180	ou	≥ 110
Hipertensão sistólica isolada	≥ 140	e	< 90

Antes da dádiva de sangue, os valores de pressão arterial dos dadores são avaliados de forma a aferir a possibilidade da mesma. Caso o dador se apresente hipotenso, a dádiva não deverá ser realizada. O mesmo se aplica para indivíduos não hipertensos com valores de pressão sistólica superior a 180 mmHg ou diastólica superior a 100 mmHg. Indivíduos hipertensos deverão apresentar valores de tensão arterial controlados para que a dádiva possa ser realizada (Hemominas, 2015).

2.12 Alanina aminotransferase

A alanina aminotransferase (ALT) é uma enzima associada ao fígado. Esta enzima é libertada no sangue em elevadas quantidades quando existe dano na membrana do hepatócito (célula hepática) [Medicamentos e Saúde, 2014]. Normalmente, é efetuado o teste ALT em conjunto com outros testes para verificar possíveis doenças no fígado como cirrose, hepatite, eclampsia, colestase, entre outras. A Tabela 11 apresenta os valores de referência convencionados para esta enzima.

Tabela 11. Valores de referência para a alanina aminotransferase.

Estado	ALT	
Normal	**Homem**	**Mulher**
	≤ 55 unidades por litro	≤ 30 unidades por litro
Anormal	Outros valores	

2.13 Marcadores serológicos de Hepatite B

A hepatite B traduz a infeção pelo vírus da hepatite B, responsável por danos graves no fígado. De facto, e segundo dados da organização da organização mundial de saúde, estima-se que mais de 780 000 indivíduos morram por ano devido a complicações relativas à hepatite B, como cirrose e cancro hepático [OMS, 2015a].

A infeção pelo vírus da hepatite B pode manifestar-se de forma crónica ou aguda, e pode ser transmitida através do contacto com sangue (como por exemplo, através de transfusões sanguíneas, partilhas de seringas) ou outros fluidos corporais (como a saliva) de indivíduos infetados, via sexual e da mãe para o filho durante o parto.

Muitos indivíduos são portadores assintomáticos, não apresentando doença hepática ativa. Devido aos factos apresentados, torna-se clara a necessidade de despistar a ausência desta infeção nos dadores de sangue, de forma a evitar que a dádiva resulte na propagação deste vírus.

O diagnóstico laboratorial permite confirmar a infeção, assim como se a doença se encontra na fase crónica ou aguda. Vários marcadores são utilizados para este despiste [Roche, 2015a]. Habitualmente, o despiste realiza-se essencialmente com base no antigénio HBs e nos anticorpos anti-HBc total e anti-HBs. De seguida serão abordados em detalhe dois marcadores utilizados na realização do presente projeto, o antigénio HBs e o anticorpo anti-HBc.

O antigénio de superfície do vírus da hepatite B (Ag-HBs) é o primeiro marcador a surgir, indicando a ocorrência de infeção. Caso este marcador esteja presente no soro por mais de 6 meses, a doença é diagnosticada como crónica, sendo indicador do risco de desenvolvimento de doença hepática crónica e cancro hepático numa fase mais tardia [OMS, 2015a]. De forma a inferir se se trata de um portador ativo ou inativo deverão ser analisados os valores detetados para as transaminases AST e ALT: valores elevados revelam dano hepático e, neste caso específico, indicam que o portador tem doença ativa [Roche, 2015a].

O anticorpo contra o antigénio da superfície do vírus da hepatite B (anti-HBc) é um marcador serológico que indica a presença de infeção pelo vírus da hepatite B. Existem dois tipos deste anticorpo:

- Anti-HBc IgM (imunoglobulina M): surge na fase aguda da doença. Funciona como marcador para a Hepatite fulminante e a sua ocorrência no soro verifica-se em simultâneo com a elevação dos valores das transaminases [Hermes Pardini, 2015, CDC, 2015]. Os valores deste marcador tendem a decrescer até aos 6 meses de infeção.
- Anti-HBc IgG (imunoglobulina G): surge geralmente por volta dos 2 meses de infeção e persiste ao longo da vida do indivíduo infetado [Hermes Pardini, 2015]. A sua presença em associação com a de outros marcadores é critério de diagnóstico, podendo indicar:
    - Infecção recente: quando Ag-HBs e anti-HBs são negativos (período de "Janela Imunológica");

- Infecção crônica: quando Ag-HBs apresenta já níveis muito baixos ou indetetáveis.

2.14 Marcadores serológicos de Hepatite C

A hepatite C é uma doença hepática causada pelo vírus da hepatite C. Este pode causar doença aguda ou crónica e transmite-se através do sangue (partilha de seringas, transfusões de sangue infetado, entre outras), durante o parto e sexualmente [OMS, 2015b].
Dados da organização mundial de saúde estimam que 130 a 150 milhões de indivíduos sejam portadores de hepatite crónica, e que 500 000 indivíduos morram por ano devido às complicações relacionadas com esta doença.
O primeiro passo para o diagnóstico da infeção passa pela pesquisa do anticorpo para o vírus da hepatite C (anti-HCV). Caso a pesquisa deste anticorpo seja positiva, a exposição ao vírus é confirmada. No entanto, é necessário despistar se se trata de um caso de doença recente ou doença já tratada, uma vez que este anticorpo persiste por tempo indefinido mesmo após a cura [OMS, 2015b] [Roche, 2015a]. Para despistar estes dois casos, é necessário pesquisar a presença de material genético viral (ácido ribonucleico do vírus da Hepatite C): caso este esteja presente, a doença é recente; se ausente, a doença está curada.

2.15 Marcadores serológicos do Vírus da Imunodeficiência Humana (HIV)

O Vírus da Imunodeficiência Humana (HIV) é transmitido através de relações sexuais não protegidas, pelo contacto com sangue infetado e de mãe para filho no momento do parto ou durante a amamentação [Roche, 2015b]. Este vírus é responsável pela destruição das células que conferem proteção ao organismo (glóbulos brancos), levado ao desenvolvimento do síndrome de imunodeficiência humana adquirida (SIDA). Assim, o indivíduo infetado torna-se suscetível ao desenvolvimento de doenças designadas por oportunistas, que são causadas por agentes como bactérias e fungos incapazes de causar doença em indivíduos imunocompetentes. A presença do vírus indica que o indivíduo é portador do mesmo, podendo transmiti-lo a outros. O diagnóstico de infeção pelo HIV realiza-se através da pesquisa do anticorpo anti-HIV no sangue.

O diagnóstico da infeção por HIV realiza-se através da pesquisa do anticorpo anti-HIV no sangue [Labluxor, 2012]. Apenas indivíduos expostos ao vírus desenvolvem estes anticorpos. No entanto, estes anticorpos só começam a ser produzidos cerca de 3 meses após a exposição vírica, um período designado por janela imunológica, em que os anticorpos ainda não são detetáveis no sangue apesar do indivíduo se encontrar infetado. Por esta razão, deve repetir-se o teste 6 meses depois no caso de indivíduos expostos a fatores de risco de infeção com resultado negativo ao teste anti-HIV.

2.16 Marcador serológico da sífilis

A sífilis é uma doença bacteriana causada pela espiroqueta *Treponema pallidum*. Pode ser transmitida sexualmente ou de mãe para filho. O seu diagnóstico, numa fase já mais avançada/tardia pode basear-se no teste designado por Veneral Disease Research Laboratory (VDRL) [Md.Saúde, 2015b].

Este teste baseia-se na pesquisa de anticorpos contra a bactéria *Treponema pallidum*, os quais surgem no sangue cerca de 6 semanas após a exposição à bactéria. Os resultados do teste expressam-se através da maior diluição que é possível realizar na amostra de sangue para que ainda se consigam detetar estes anticorpos. Assim, por exemplo, se o resultado for 1/16 significa que a amostra pode ser diluída no máximo até 16 vezes para que a deteção destes anticorpos seja possível. Quanto maior o valor de diluição possível, maior a quantidade de anticorpos presente no sangue.

No entanto, a presença destes anticorpos no sangue pode ser verificada noutras situações distintas à sífilis, como nos casos de hepatite, lúpus e artrite reumatoide. O diagnóstico de sífilis é considerado positivo para valores de diluição no teste VDRL superiores a 32 (1/32) (Tabela 12).

Tabela 12. Valores de diluição do teste de VDRL para diagnóstico de sífilis.

Estado	VDRL
Diagnóstico de sífilis	≥ 32 diluições (1/32) com anticorpo detetável
Presença do anticorpo, com diagnóstico de sífilis negativo	< 32 diluições (1/32) com anticorpo detetável

3 Descoberta de Conhecimento nos Dados

Este capítulo apresenta um conjunto de informação importante e relevante para o desenvolvimento do presente projeto, fazendo referência a ideias e a trabalhos que foram realizados até ao momento no processo de descoberta de conhecimento nos dados. Este processo, permite a obtenção de conhecimento intrínseco aos dados. Como tal, torna-se bastante útil nas mais diversas áreas, nomeadamente na área da saúde, como por exemplo, na interpretação dos dados resultantes de análises sanguíneas e seus componentes.

Hoje em dia, muitas entidades deparam-se com demasiados dados mas retiram pouca informação proveniente dos mesmos. Com o crescimento substancial das bases de dados, as organizações deparam-se com o problema em utilizar, convenientemente, os dados disponíveis [Anacleto, 2009]. Como resultado da necessidade de novas ferramentas para extração de conhecimento a partir da grande quantidade de dados disponível, surge o processo de descoberta de conhecimento nos dados. Segundo Fayyad, o processo de descoberta de conhecimento nos dados é iterativo e interativo e define-se como o processo de identificar nos dados, padrões ou modelos válidos, novos, compreensíveis e potencialmente úteis [Fayyad et al., 1996].

O conhecimento no âmbito da área da mineração dos dados é uma relação existente nos dados que é interessante e útil num domínio específico. Para obter o conhecimento é necessário conjugar uma série de etapas. Antes da aplicação de algoritmos de extração de conhecimento de dados (ECD) a um conjunto de dados, é importante que estes sejam analisados. Essa análise, que pode ser realizada por meio de técnicas baseadas na estatística, permite uma melhor compreensão da distribuição dos dados e permite selecionar as técnicas

mais apropriadas para o pré-processamento dos dados, tendo como consequência o aperfeiçoamento da construção dos modelos preditivos [Dantas et al., 2008].

A descoberta de conhecimento engloba várias fases, designadamente a compreensão dos dados, a preparação dos dados, a aplicação de algoritmos de mineração de dados e a avaliação do conhecimento descoberto. Estas fases estão apresentadas na Figura 2 e serão abordadas ao longo deste subcapítulo.

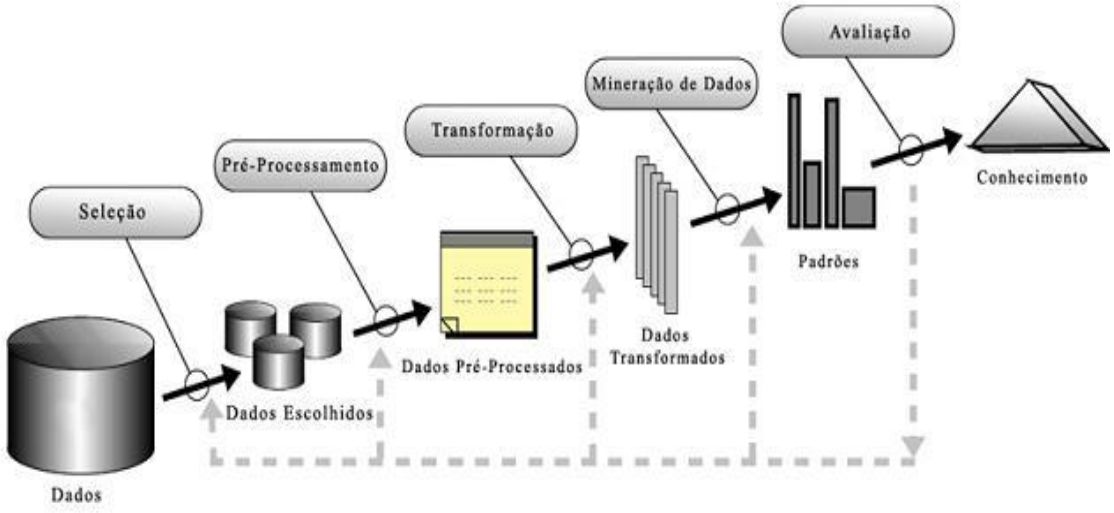

Figura 2. Etapas da extração de conhecimento nos dados [Devmedia, 2015].

3.1 Fase de Seleção

A fase de seleção é a primeira fase do processo de extração de conhecimento nos dados. Aqui são decididos quais os conjuntos de dados que serão relevantes para que seja retirada a informação útil pretendida [Bueno e Viana, 2012]. Para tal, é necessário deter conhecimento da área em estudo. Nesta fase, pode ser efetuada a ordenação dos atributos com base num critério de importância, a redução da dimensionalidade do espaço de busca de atributos, entre outras ações [Ribeiro, 2010].

Quando é dito que "quanto maior o número de atributos numa base de dados, maior o poder discriminatório do classificador e maior a facilidade de extração de modelos de conhecimento da base de dados", em termos práticos nem sempre corresponde à realidade dado que:

- Podem existir problemas de dimensionalidade, encadeando um aumento significativo do tempo computacional de acordo com o número de atributos presentes;
- Atributos irrelevantes podem gerar confusão ao algoritmo de aprendizagem e, assim sendo, esconder as distribuições de pequenos conjuntos de atributos realmente relevantes;
- Muitas das vezes os atributos escolhidos não são os melhores, apresentando problemas de redundância.

Existem medidas de informação que possibilitam determinar o ganho de informação a partir de um atributo. Um atributo Xi é preferido ao atributo Xj se o ganho de informação a partir do atributo Xi for maior que a partir do atributo Xj. Esta verificação é possível através do cálculo da entropia (Equação 1). A entropia H da variável aleatória discreta X mede a incerteza associada ao valor de X [Gama et al., 2012].

(1) $$H(X) = - \sum_{x \in X} p(x) * \log(p(x))$$

Uma vez que os atributos são as variáveis desta etapa, o conhecimento da escala e do tipo de um atributo permite uma identificação mais adequada da forma de preparação e modelação dos dados. O tipo de um atributo pode representar quantidades (quantitativo ou numérico) ou qualidades (qualitativo, simbólico ou categórico). Os atributos quantitativos podem ser contínuos ou discretos, sendo que os valores contínuos são normalmente representados por números reais e os atributos discretos têm como conteúdo um número contável de valores. Em relação aos atributos qualitativos, estes geralmente são representados por um número finito de símbolos ou nomes. A escala define as operações que podem ser realizadas sobre os valores do atributo e classifica os atributos como:

- Nominais: classifica as unidades em classes ou categorias com base na caraterística que representa e não estabelece qualquer relação de grandeza ou ordem. Se apenas podemos dizer que um objeto é diferente de outros então o objeto tem uma escala nominal [Sistema Galileu de Educação Estatística, 2015]. Exemplos de atributo: sexo, código postal, entre outros;
- Ordinais: além de manter as caraterísticas da escala nominal, é possível estabelecer uma ordenação das categorias onde os dados são classificados de acordo com uma sequência com significado. Podemos, por exemplo, indicar se um valor é igual, maior ou menor que outro;
- Intervalares: os dados são diferenciados e ordenados por números demonstrados numa escala de origem arbitrária. A alanina aminotransferase é um atributo intervalar dado que é representado por números que variam dentro de um intervalo.

Na seleção dos atributos podem ser utilizados dois tipos de redução [Ribeiro, 2010]:

- Redução de Dados Horizontal: caraterizada pela escolha de casos, em que as operações mais utilizadas são:

    - Segmentação da base de dados: é efetuada uma escolha dos atributos para direcionar o processo de segmentação. O conjunto de dados resultante é o considerado posteriormente no processo de extração de conhecimento de dados;
    - Eliminação direta de casos: especificação dos casos que devem ser eliminados da base de dados;
    - Amostragem aleatória: consiste em retirar um número preestabelecido de registos da base de dados aleatoriamente, de forma a originar um conjunto com uma quantidade menor que a original;

- Agregação de informação: os dados com maior nível de detalhe são consolidados, sendo gerada nova informação com melhor detalhe.

- Redução de Dados Vertical: é uma operação de pré-processamento importante no processo de ECD e consiste na eliminação ou substituição dos atributos de um conjunto de dados. O principal objetivo desta redução é encontrar um conjunto mínimo de atributos e, ao mesmo tempo, preservar a informação original. As abordagens mais utilizadas são:

    - Abordagem independente de modelo (*Filter)*: a seleção de atributos é realizada sem a consideração do algoritmo de mineração de dados que será aplicado nos atributos selecionados;
    - Abordagem dependente de Modelo (*Wrapper)*: o algoritmo de mineração de dados é testado para cada conjunto de atributos e, seguidamente, os resultados são avaliados.

Importa salientar a existência de três estratégias clássicas para seleção do conjunto de atributos que podem ser utilizados em qualquer das abordagens acima descritas:

- Seleção sequencial para frente (*Forward Selection)*: cada atributo é iterativamente adicionado a um subconjunto de atributos candidatos tendo em conta uma medida de qualidade;
- Seleção Sequencial para Trás (*Backward Selection)*: cada atributo é retirado do conjunto inicial de acordo com uma medida de qualidade;
- Combinação das estratégias anteriores: a seleção para frente e a seleção para trás podem ser combinadas onde, a cada passo do algoritmo, o algoritmo seleciona o melhor atributo e remove o pior da mesma forma.

Para finalizar, é de realçar que esta é uma fase crucial para que as fases posteriores ocorram da forma esperada.

3.2 Pré-processamento de dados

O desempenho dos algoritmos de ECD é afetado pelo estado dos dados. Diferentes conjuntos de dados podem apresentar diferentes caraterísticas, dimensões ou formatos. Os valores dos atributos podem estar limpos ou conter imperfeições (inconsistência, duplicação ou ausência de valores). Apesar desta etapa consumir mais de metade do tempo de todo o processo de mineração de dados [Freitas, 2006], o pré-processamento torna-se fundamental dado melhorar a performance dos algoritmos de análise.

Este subcapítulo pretende detalhar algumas técnicas de pré-processamento frequentemente utilizadas para o aumento da qualidade dos dados, para a construção de modelos mais fiáveis e para a adequabilidade dos dados a um algoritmo de ECD. As técnicas mais utilizadas são:

1. **Eliminação manual de atributos**

   Os atributos devem ser eliminados quando não contribuem para a estimativa do valor do atributo alvo, sendo estes considerados irrelevantes. Um atributo pode ser irrelevante quando contém o mesmo valor para todos os objetos e, assim sendo, não há informação que consiga distinguir os objetos, ou quando não deve formar um conjunto de dados uma vez que não faz parte do domínio do problema de decisão [Gama et al., 2012].

2. **Integração de dados**

   Atualmente, pode ser necessária a integração de diferentes bases de dados para que sejam utilizadas numa aplicação de ECD. Para tal, é necessária a identificação dos objetos que estão presentes nos diferentes conjuntos de dados. A identificação da entidade é realizada através da procura de atributos comuns nas diferentes bases de dados. Este processo pode ser complicado devido à existência de atributos equivalentes mas com nomes diferentes.

3. **Amostragem de dados**

   Uma elevada quantidade de objetos numa base de dados pode dificultar o desempenho dos algoritmos de ECD. Quanto maior a quantidade de dados utilizada, maior tende a ser a taxa de acerto do modelo e menor a eficiência computacional do processo indutivo. Como ponto de equilíbrio, um algoritmo de ECD não necessita de utilizar todo o conjunto de dados, podendo utilizar apenas uma parte do conjunto original.

   Importa salientar que a existência de classes com um número de exemplos expressivamente maior que as demais pode levar à indução de classificadores tendenciosos para as classes maioritárias. Nesta situação é concluído que o conjunto de dados está desbalanceado.

4. **Dados desbalanceados**

   Em vários conjuntos de dados é normal o número de objetos variar entre as diferentes classes. Estes dados são nomeados dados desbalanceados e a maior parte dos algoritmos de extração de conhecimento pioram o seu desempenho na presença destes dados uma vez que tendem a valorizar classes predominantes e a ignorar classes de menor representação. Por vezes, quando a classe de interesse é uma classe rara (classe com menor representação) os classificadores que foram gerados a partir de bases de treino desbalanceadas apresentam elevadas taxas de falsos negativos [Machado e Ladeira, 2007]. Caso não haja possibilidade de gerar novos dados através do mesmo processo que gerou o conjunto atual, existe a possibilidade de utilizar técnicas que procuram balancear artificialmente o conjunto de dados:

   - Redefinição do tamanho do conjunto de dados: a utilização desta técnica pode levar a um acréscimo de objetos à classe minoritária, assim como a eliminação de objetos da classe maioritária. Os novos objetos podem representar situações que

nunca ocorrerão induzindo um modelo inadequado para os dados. Além disso, pode ocorrer *overfitting* (modelo é superajustado aos dados de treino) ou pode verificar-se a eliminação de objetos da classe maioritária, o que se designa por underfitting (o modelo induzido não se ajusta aos dados de treino) [Gama et al., 2012];

- Utilização de diferentes custos de classificação para as diferentes classes: a existência de classes maioritária e minoritária leva à utilização de custos de classificação;
- Indução de um modelo para uma dada classe: utilização de técnicas de classificação para uma classe, onde a classe minoritária, ou a classe maioritária ou ambas as classes são aprendidas separadamente.

5. Limpeza de dados

Os dados no mundo real tendem a ser incompletos, fora de padrões e inconsistentes. A limpeza de dados pretende corrigir inconsistências e identificar valores atípicos. É de extrema importância que a qualidade dos dados seja assegurada devido a inúmeras caraterísticas que os dados podem apresentar [Côrtes et al., 2002]:

- Inconsistência: os valores não combinam ou contradizem valores de outros atributos do mesmo objeto. Dados inconsistentes podem resultar do processo de integração de dados de fontes ou tabelas diferentes ou da presença de ruído nos dados;
- Redundância: quando dois ou mais objetos têm os mesmos valores para dois ou mais objetos. Um atributo pode ser redundante quando o seu valor pode ser deduzido a partir do valor de um ou mais atributos. A eliminação da redundância é, normalmente, efetuada no final do processo de limpeza. Isto permite que o desempenho do algoritmo de ECD melhore significativamente;
- Incompletos: um dos problemas que é encontrado nos conjuntos de dados é a ausência de valores para determinados atributos. As causas para a ausência de valores podem ser:
    - Desconhecimento do valor de um atributo no momento do preenchimento dos valores do objeto;
    - Falta de necessidade ou obrigação de apresentar um valor para o(s) atributo(s) para determinados objetos;
    - Inexistência de um valor para alguns atributos.

Existem determinados métodos que podem ser utilizados para atribuir valores a estes atributos, como [Côrtes et al., 2002]:

- Preenchimento dos valores em falta manualmente: a utilização deste método pode não ser a mais indicada, principalmente em grandes bases de dados;
- Eliminação de objetos com valores ausentes;

- Utilização de uma constante global para preencher os valores inexistentes, atualizando todos os valores ausentes com um único valor (Exemplo: "desconhecido");
- Utilização de um atributo médio para preencher os valores ausentes: este método é usado quando o atributo é do tipo numérico e o seu significado pode ser demonstrado por um valor médio (Exemplo: média do peso para os pacientes que contêm o atributo "peso" ausente);
- Uso do valor com maior probabilidade para preencher os valores inexistentes: determinação do valor através da aplicação de uma técnica de regressão (Bayesiana ou indução por árvores de decisão).

Normalmente, os dados possuem erros ou valores que são diferentes do esperado, ou seja, não pertencem à distribuição que gerou os dados analisados. O ruído pode ser definido como uma variância ou erro aleatório no valor gerado ou medido para um atributo. Como indicador da presença de ruído são considerados os valores atípicos (*outliers*). Os valores atípicos são valores ou que estão além dos limites aceitáveis, ou que são muito diferentes dos demais valores observados para o mesmo atributo. As técnicas existentes para a redução do ruído são [Gama et al., 2012]:

- Técnicas de intervalo: após a ordenação dos valores de um atributo, estes são divididos em intervalos com o mesmo número de valores. Os valores não expectados são substituídos, por exemplo, pela média ou mediana dos valores do intervalo;
- Técnicas baseadas em agrupamento dos dados: os valores dos atributos são agrupados por uma técnica de agrupamento. Os atributos que não formem um grupo com outros valores são considerados valores atípicos;
- Técnicas baseadas em distância: o afastamento de um objeto em relação aos restantes objetos da sua classe pode determinar a presença de ruído desse mesmo objeto;
- Técnicas baseadas em regressão ou classificação: para estas técnicas é utilizada uma função de regressão que após receber um valor com ruído estima o seu valor verdadeiro. Se o valor a ser estimado for simbólico são utilizadas técnicas de classificação.

As principais razões para estas deficiências a nível dos dados são ao nível do preenchimento dos dados por seres humanos ou por equipamentos que realizam a recolha dos mesmos.

3.3 Transformação de dados

A transformação dos dados é uma etapa crucial no processo da descoberta de conhecimento devido a determinadas limitações de algumas técnicas de ECD. Esta fase permite a transformação e consolidação dos dados em formatos apropriados para a fase de mineração de dados [Dantas et al., 2008].

A existência de um número elevado de atributos pode ser um problema para as técnicas de ECD. Na maior parte dos algoritmos de ECD, para os dados serem utilizados existe a necessidade de reduzir o número de atributos. Esta redução pode ainda melhorar o desempenho do modelo induzido, reduzir o custo computacional e facilitar a interpretação dos resultados obtidos. As áreas de Reconhecimento de Padrões, Estatística e Teoria da Informação originaram técnicas que permitem reduzir o número de atributos [Gama et al., 2012] e transformar os dados em novos formatos:

1. **Agregação dos dados**
   Consiste na substituição dos atributos originais por novos atributos. Estes são formados pela combinação de grupos de atributos através da utilização de funções lineares e não lineares. A técnica Análise de Componentes Principais é um procedimento matemático que utiliza uma transformação ortogonal para converter um conjunto de observações de variáveis possivelmente relacionadas num conjunto de novos valores de variáveis não correlacionadas.

2. **Seleção de atributos**
   A seleção de atributos mantem uma parte dos atributos originais e descarta os restantes, permitindo [Bueno e Viana, 2012]:

   - Identificar os atributos mais importantes;
   - Melhorar o desempenho de várias técnicas de ECD;
   - Reduzir a necessidade de memória;
   - Melhorar o tempo de processamento;
   - Eliminar o número de atributos irrelevantes e redução do ruído;
   - Facilitar a visualização dos dados;
   - Reduzir o custo de recolha dos dados.

   Caso os atributos sejam claramente redundantes ou irrelevantes, estes podem ser manualmente eliminados. No entanto, no caso de alguns atributos torna-se difícil identificar se se pode proceder à eliminação dos mesmos. Nestes casos, são utilizadas técnicas automáticas [Gama et al., 2012]:

   - Técnicas de ordenação: os atributos são ordenados de acordo com a sua relevância para um dado critério;
   - Técnicas de Seleção de Subconjunto: a seleção de um subconjunto de atributos pode ser interpretada como um problema de procura, onde cada ponto no espaço de procura é visto como um possível subconjunto de atributos.

3. **Generalização de dados**
   A generalização de dados permite a transformação de dados primitivos (existentes nos registos das tabelas) em hierarquias de mais alto nível. Como exemplo, para o atributo idade podemos ter como valores criança, adolescente, adulto e idoso [Côrtes et al., 2002].

4. **Normalização**
   Permite a atribuição de uma nova escala a um atributo de forma a ajustar as escalas de valores para o mesmo intervalo. O propósito da normalização é minimizar os problemas originários do uso de unidades e evitar atributos com grande intervalo de valores. Os algoritmos de mineração (redes neuronais, *K-nearest neighbors,* entre outros) são beneficiados pela normalização [Alvares, 2010].

5. **Construção de atributos**
   Este método permite que novos processos sejam construídos a partir de atributos existentes para facilitar o processo de análise. Como exemplo, podemos utilizar os atributos idade, peso e altura para criar um novo atributo com o índice de massa corporal [Côrtes et al., 2002].

3.4 Mineração de dados

Normalmente, o termo mineração de dados é utilizado como sinónimo para o processo de extração de conhecimento de bases de dados. No entanto, a extração de conhecimento nos dados refere-se a todo o processo de descoberta de conhecimento e o termo mineração de dados define-se como um passo particular nesse processo com o objetivo de estabelecer relações, associações e padrões de difícil visualização [Anacleto, 2009]. Para tal, são utilizados, por exemplo, algoritmos de aprendizagem ou de classificação baseados em redes neuronais e na estatística. Os resultados são apresentados na forma de regras, hipóteses, grafos ou árvores de decisão [Bueno e Viana, 2012].

Importa salientar que a mineração de dados não elimina a necessidade de conhecimento dos dados, dos métodos analíticos e da área de negócio a que se aplicam. A mineração de dados descobre nova informação nos dados mas não menciona qual o valor dessa informação [Anacleto, 2009].

Os métodos de mineração de dados podem ser divididos em duas categorias:

- Supervisionados: os métodos supervisionados preveem um valor e necessitam da especificação de um atributo alvo (*outcome*). Estes métodos supervisionados incluem os métodos de classificação e de regressão;
- Não supervisionados: os métodos não supervisionados não necessitam da especificação de um atributo alvo e permitem descobrir a estrutura intrínseca e padrões entre os dados. Estes métodos incluem os métodos de segmentação (*clustering*) e as regras de associação.

Atendendo aos objetivos traçados para a descoberta de conhecimento, podem ser realizadas várias tarefas de mineração de dados. O conjunto de tarefas efetuadas sobre os dados define

a análise e a criação dos modelos. A Tabela 13 apresenta as principais funções de mineração de dados e algumas áreas de aplicação.

Tabela 13. Métodos da mineração de dados [Bueno e Viana, 2012].

Função	Aplicações
Classificação	– Identificar a melhor forma de tratamento de um paciente;
Regressão	– Estimar a probabilidade de um paciente morrer baseando-se nos resultados de diagnósticos médicos;
Segmentação	– Agrupar pacientes com o mesmo tipo de doença; – Agrupar clientes com comportamento de compra similar.
Associação	– Determinar os produtos que costumam ser colocados juntos num carrinho de compras.

Estas funções serão apresentadas em pormenor nos próximos subcapítulos, sendo descritas as respetivas definições, os objetivos e os algoritmos utilizados, de acordo com as diferentes técnicas.

3.4.1 Classificação

A classificação tem como objetivo a identificação de padrões nos dados que relacionam os valores das variáveis independentes e o de uma variável objetivo, permitindo assim classificar novos registos de uma forma mais precisa. O objetivo dos métodos de classificação é gerar modelos que preveem com exatidão a classe alvo. Os modelos criados podem ser aplicados a novos casos para previsão de uma classe desconhecida [Anacleto, 2009]. A construção dos modelos computacionais de classificação geralmente utiliza um dos paradigmas [Von Zuben e Attux, 2010]:

- *Top-down:* obtém o modelo de classificação a partir de informações adquiridas por especialistas;
- *Bottom-up:* adquire o modelo de classificação através da identificação dos relacionamentos entre variáveis dependentes e independentes num conjunto de dados.

Esta função de aprendizagem permite várias aplicações, como por exemplo, a identificação de clientes com crédito de risco, diagnósticos médicos, definição de segmentos alvo em campanhas de marketing, entre outras. Para tal, são utilizadas diversas técnicas de classificação [Gama et al., 2012]:

3.4.1.1 Métodos baseados em árvores de decisão

As árvores de decisão geram modelos que podem ser interpretados e aplicados a um determinado conjunto de dados para identificação da classe a que os registos pertencem [Anacleto, 2009]. Uma árvore de decisão é um grafo acíclico direcionado, em que cada nó ou é um nó de divisão, com dois ou mais sucessores, ou um nó folha. Um nó interno inclui um teste

a um atributo, um ramo representa o resultado do teste, uma folha representa uma classe e cada percurso da árvore corresponde a uma regra de classificação [Gama et al., 2012].

As árvores de decisão são construídas de forma *top-down* e usam a estratégia de "dividir para conquistar" com vista à resolução de um problema de decisão. Desta forma, a combinação das soluções dos subproblemas pode produzir uma solução do problema original. Assim sendo, é possível dividir o espaço de instâncias em subespaços e ajustar cada subespaço recorrendo a diferentes modelos.

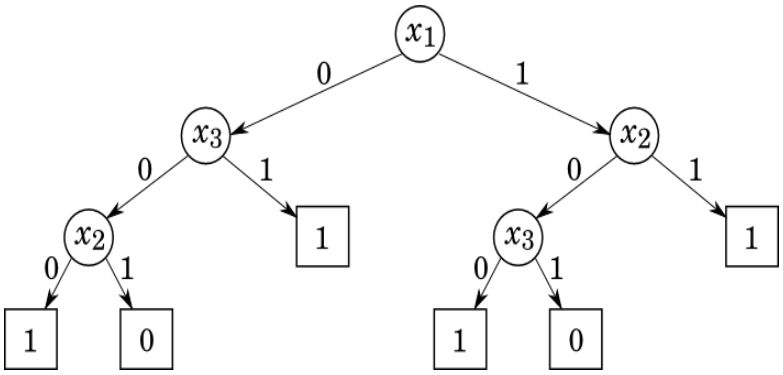

Figura 3. Exemplo de uma árvore de decisão (O'Donnell, 2015).

A Figura 3 representa uma árvore de decisão onde cada nó da árvore corresponde a uma região no espaço. Uma árvore de decisão pode efetuar predições para qualquer exemplo de entrada uma vez que abrange todo o espaço de instâncias [Gama et al., 2012].

No processo de criação de uma árvore de decisão, a poda (método conhecido por *Pruning)* é uma das fases mais importantes uma vez que conduz a erros menores de generalização [Von Zuben e Attux, 2010]. Os métodos de realização da poda são classificados como:

- Pré-poda: métodos que param a construção da árvore quando algum critério é satisfeito, evitando a perda de tempo na construção de uma estrutura que não é utilizada na árvore final;
- Pós-poda: métodos que constroem uma árvore completa e, posteriormente, realizam a respetiva poda.

Com a utilização dos métodos de poda, a árvore torna-se mais simples e, consequentemente, a interpretação por parte do utilizador torna-se facilitada.

Seguidamente, são apresentadas duas tabelas (Tabela 14 e Tabela 15) com as vantagens e desvantagens da utilização de métodos baseados em árvores de decisão.

Tabela 14. Vantagens – árvores de decisão [Gama et al., 2012].

Vantagens	
Flexibilidade	As árvores de decisão não assumem nenhuma distribuição para os dados, o espaço dos objetos é dividido em subespaços e fornece uma cobertura exaustiva do espaço de instâncias.
Robustez	As árvores são invariantes em relação a transformações monótonas de variáveis de entrada. Como tal, a sensibilidade a distribuições com grande cauda e desvios é também reduzida.
Seleção de atributos	São selecionados os atributos a usar no modelo de decisão.
Eficiência	O algoritmo para aprendizagem de uma árvore é construído de cima para baixo utilizando a estratégia de dividir para conquistar sem *backtracking*.

Tabela 15. Desvantagens – árvores de decisão [Gama et al., 2012].

Desvantagens	
Replicação	Duplicação de uma sequência de testes em diferentes ramos de uma árvore de decisão, o que resulta numa representação não concisa e com tendência a obter uma baixa precisão preditiva.
Valores ausentes	Dado que uma árvore de decisão é uma hierarquia de testes, os valores desconhecidos causam problemas na decisão do ramo que deve ser seguido.
Atributos contínuos	A presença de atributos contínuos requer uma operação de ordenação para cada atributo contínuo em cada nó de decisão.
Instabilidade	Pequenas variações no conjunto de treino podem produzir grandes variações na árvore final.

3.4.1.2 Redes Neuronais Artificiais

O cérebro humano é o responsável pelo processamento de informação e por gerar as consequentes respostas. A ação aparentemente simples de segurar um objeto é possível graças à nossa estrutura biológica. O ser humano evidencia-o quando tenta que um *robot* realize esta ação. As redes neuronais artificiais (RNAs) baseiam-se na estrutura e no funcionamento do sistema nervoso, com o intuito de simular a capacidade de aprendizagem do cérebro humano na aquisição de conhecimento [Gama et al., 2012]. O sistema nervoso "*é um conjunto complexo de células que determinam o funcionamento e o comportamento dos seres vivos*". O neurónio é a unidade funcional do sistema nervoso, sendo este constituído pelas dendrites, corpo celular e axónio. Os neurónios comunicam entre si através de sinapses, ocorrendo desta forma a propagação dos impulsos nervosos [Infoescola, 2015c].

Uma rede neuronal artificial é um sistema paralelo e distribuído, constituído por unidades de processamento simples (neurónios artificiais que processam funções matemáticas) que

permitem o armazenamento do conhecimento experimental e a disponibilização do mesmo para futura utilização [Maxwell, 2015].

As unidades são dispostas numa ou mais camadas e interligadas por um grande número de conexões que normalmente são unidirecionais. Na maioria das arquiteturas, as conexões que simulam as sinapses biológicas possuem membros que consideram a entrada recebida por cada neurónio da rede. Os pesos podem ter valores negativos ou positivos. Estes valores são ajustados num processo de aprendizagem e codificam o conhecimento adquirido pela rede [Gama et al., 2012].

A Figura 4 apresenta um exemplo da arquitetura de um neurónio artificial. Cada terminal de entrada do neurónio recebe um valor que, posteriormente, será ponderado e combinado através de uma função matemática. O resultado dessa função é a resposta do neurónio para a entrada dos valores.

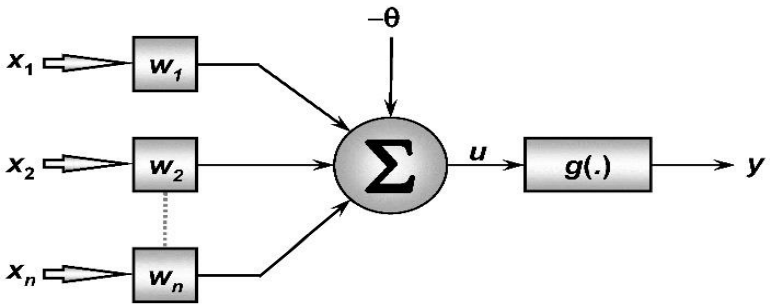

Figura 4. Neurónio artificial. Imagem retirada de [Barra, 2013].

Os neurónios podem-se apresentar numa ou mais camadas. Com uma rede multicamada, um neurónio pode receber nos seus terminais de entrada valores de saída de neurónios da camada anterior e/ou enviar o seu valor de saída para terminais de entrada de neurónios da camada seguinte. Uma rede multicamada pode ser classificada como [Gama et al., 2012]:

- Completamente conectada: quando os neurónios da rede estão conectados a todos os neurónios da camada anterior e/ou seguinte;
- Parcialmente conectada: quando os neurónios estão conectados a apenas alguns dos neurónios da camada anterior e/ou seguinte;
- Localmente conectada: são redes parcialmente conectadas, em que os neurónios conectados se encontram numa região bem definida.

Para conhecer o número de neurónios adequado na camada intermediária de uma rede de neurónios artificial é necessário ter em consideração o número de exemplos de treino, a quantidade de ruído presente nos exemplos, a complexidade da função a ser aprendida e a distribuição estatística dos dados de treino.

A ausência de um algoritmo para o treino das redes neuronais multicamadas foi ultrapassada com o algoritmo *back-propagation.* Este é um algoritmo iterativo constituído por duas fases:

- *Forward:* Cada objeto de entrada é apresentado à rede e ponderado de acordo com o peso associado. Cada neurónio numa determinada camada aplica a função de ativação à soma das suas entradas e produz um valor de saída. Este valor será utilizado como valor de entrada pelos neurónios da camada seguinte. Este processo continua até que os neurónios da camada de saída produzam um valor de saída. A diferença entre os valores de saída produzidos e desejados para cada neurónio da camada de saída indica o erro cometido pela rede para o objeto apresentado;
- *Backward:* Utiliza o valor do erro de cada neurónio para ajustar os pesos de entrada. O ajuste prossegue da camada de saída até à primeira camada intermediária. Uma vez que os valores dos erros são conhecidos apenas para os neurónios da camada de saída, o erro para os neurónios das camadas intermédias necessita de ser estimado. Este é calculado através da soma dos erros dos neurónios da camada seguinte.

As RNAs apresentam um bom desempenho quando utilizadas com um número elevado de aplicações. As redes neuronais artificiais detêm algumas propriedades e capacidades úteis, tais como [Maxwell, 2015]:

- Aprendizagem: habilidade da RNA aprender automaticamente através de um processo iterativo de ajustes aplicados a determinados parâmetros (pesos);
- Generalização: capacidade da RNA apresentar uma saída adequada para uma entrada não esperada;
- Adaptabilidade: aptidão, por parte das RNAs, de adaptação dos pesos de acordo com as modificações do meio ambiente;
- Tolerância a falhas: a falha de alguns neurónios não causa efeitos significativos na performance do sistema.

Apesar de todo o progresso efetuado no desenvolvimento das RNAs, estas ainda estão bastante aquém da capacidade do cérebro humano.

3.4.1.3 Teorema de *Bayes* e Redes *Bayesianas*

Os métodos de raciocínio probabilístico tendem a trabalhar eficazmente em situações onde existem informações incompletas ou informações não exatas. Nestes ambientes, pode ser utilizada a teoria da probabilidade com enfoque *Bayesiano.* A probabilidade "é um campo da matemática que estuda e analisa a ocorrência de fenómenos aleatórios" [Gonçalves, 2008]. Existem diferentes tipos de probabilidade:

- Probabilidade incondicional: probabilidade que não tem dependência de nenhuma condição anterior;
- Probabilidade condicional: apresentada por P(B|A) pode ser interpretada como "A probabilidade da ocorrência do evento B, dada a ocorrência do evento A" (Equação 2).

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \quad (2)$$

Teorema de Bayes

O teorema de Bayes consiste em considerar o evento A como "causa" do evento B, sendo atribuída a probabilidade do evento A atuar na ocorrência de B. Esta probabilidade é calculada antes da realização da experiência e, como tal, designa-se como a probabilidade *a priori* de A. A seguinte equação (Equação 3) apresenta o teorema de Bayes [Gonçalves, 2008]:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{3}$$

Rede *Bayesiana*

As redes *Bayesianas* foram desenvolvidas no início dos anos 80 para facilitar a tarefa de previsão e abdução em sistemas de Inteligência Artificial [Gama et al., 2012]. Uma rede *Bayesiana,* também conhecida por rede probabilística ou rede causal, é vista como "*um modelo que utiliza a teoria de grafos e distribuições de probabilidade para representação de uma situação e as respetivas variáveis e estados*" [Gonçalves, 2008]. As variáveis são os nós e os arcos identificam as relações entre as variáveis, formando um grafo acíclico dirigido (Figura 5).

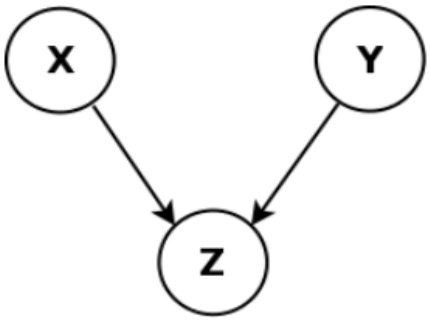

Figura 5. Grafo dirigido.

Uma rede *Bayesiana* é a representação correta de um domínio caso a condição de *Markov* seja respeitada. A condição de *Markov* declara que "*as variáveis não descendentes não geram informação adicional sobre a variável em questão*".

As redes *Bayesianas* podem ser também utilizadas para a tomada de decisões baseadas em probabilidades, explicar os resultados de uma inferência probabilística ao utilizador e decidir quais as evidências adicionais que devem ser adicionadas para que seja possível obter total conhecimento do domínio.

A aprendizagem *Bayesiana* é vista como uma maneira de adquirir a representação interna de uma rede que define um domínio de forma a facilitar a extração do conhecimento. Neste processo é essencial identificar variáveis e relações de interdependência dadas pelos arcos. A estrutura e os parâmetros da rede podem ser adquiridos:

- Por um especialista: o conhecimento é transmitido por meio de um especialista que tem como responsabilidade definir e supervisionar a construção da rede com base no seu conhecimento;
- Indutivamente: utilizando uma base de dados como exemplo para construir a rede automaticamente.

Como foi descrito anteriormente, a mineração de dados é dividida em métodos e técnicas responsáveis pela análise e extração de informação relevante contida num conjunto de dados. O próximo subcapítulo pretende demonstrar a técnica de segmentação dando a conhecer os seus objetivos e os métodos utilizados.

3.4.2 Segmentação

O processo de segmentação "*permite a partição de um conjunto de dados ou objetos em subclasses mais pequenas, com caraterísticas comuns*" [Han, 2015]. Estas subclasses denominamos segmentos. Os grupos resultantes devem ser homogéneos e bem separados. A análise de segmentos é uma aprendizagem não supervisionada, ou seja, não existem classes pré-definidas. A segmentação pode ser utilizada com o intuito de:

- Gerar um sumário compacto de dados para classificação;
- Descoberta de padrões;
- Deteção de valores atípicos*;*
- Compressão e redução de dados;
- Segmentação de clientes.

As medidas de similaridade entre duas instâncias são bastante importantes na maioria dos algoritmos de segmentação. A similaridade intra-segmento é maximizada e a similaridade inter-segmento é minimizada. Como medidas de similaridade existem [Rodrigues, 2014]:

- Distância euclidiana simples: usada na avaliação de proximidade de objetos num espaço bidimensional ou tridimensional;
- Distância euclidiana pesada: define graus de importância dos atributos;
- Distância de Mahalanobis: quando existe correlação linear entre atributos deve ser utilizada esta medida, dado que tem em consideração diferentes escalas para diferentes atributos;
- Distância Logarítmica: usada em amostras com variáveis numéricas;
- Distância Manhattan: utilizada para atributos binários.

Os algoritmos de segmentação devem ser escaláveis, deter capacidade de lidar com diferentes tipos de atributos, ter capacidade de lidar com ruído e valores isolados, lidar com elevada dimensionalidade e ter facilidade de interpretação e utilização. Os algoritmos de segmentação são divididos em três classes [Rodrigues, 2014]:

- Algoritmos de Partição: estes algoritmos constroem vários *clusters* com base no conjunto de dados inicial;

    - *K-means*: cada segmento é representado pelo valor médio de todos os objetos do segmento (centróide);
    - *K-medoids*: cada segmento é representado por um dos objetos mais próximo do centro do segmento *(*medóide).

- Algoritmos Hierárquicos: criação de uma decomposição hierárquica do conjunto de dados baseados num determinado critério;

    - Aglomerativos (Bottom-up): junção dos segmentos iterativamente, ou seja, após o cálculo da matriz de similaridade e da atribuição de cada objeto a um segmento*,* são agrupados de forma recursiva dois ou mais segmentos atómicos segundo uma medida de similaridade. Quando não existirem mais segmentos o algoritmo simplesmente deixa de efetuar a junção;
    - Divisivos (Top-down): os segmentos são particionados iterativamente, ou seja, todos os objetos são colocados no mesmo segmento sendo estes divididos em segmentos mais pequenos.

- Algoritmos baseados em Modelos: é definido um modelo para cada segmento e, em seguida, os modelos são adaptados de acordo com as instâncias atribuídas.

    - Algoritmo *Kohonen*: Inicialmente cada nó possui uma posição aleatória que será ajustada na fase de aprendizagem, com base no sucesso ou insucesso de cada nó. Após os dados de entrada serem atribuídos ao nível de entrada, cada nó de saída competirá com os outros para ganhar o registo. O nó com a resposta mais forte é o vencedor. Após o término, o treino da rede dos registos similares aparecem juntos no mapa de saída;
    - Algoritmo *Two-Step*: este algoritmo é dividido em duas etapas: na primeira etapa, cada objeto é introduzido na árvore a partir do nó raiz e é guiado de forma recursiva no sentido descendente da estrutura. Após atingir o nó folha, o objeto é inserido no mesmo ou dá origem a um novo nó folha. Na segunda etapa, os subsegmentos encontrados são utilizados como objetos nesta etapa. Para tal, é aplicado um algoritmo hierárquico aglomerativo de ligação mínima.

3.4.3 Regras de Associação

A extração de conjunto de itens frequentes é um dos tópicos de investigação mais ativos na descoberta de conhecimento em bases de dados. A análise pioneira foi relativa a dados transacionais que descrevem o comportamento de consumo dos clientes com o intuito de descobrir grupos de produtos que são frequentemente comprados em conjunto [Gama et al., 2012]. Podemos afirmar que as regras de associação são "*padrões descritivos que representam a probabilidade de que um dado conjunto de itens ocorra numa transação uma vez que outro conjunto está presente*" [Vasconcelos e Carvalho, 2004]. A tarefa de associação tem como premissa básica descobrir relacionamentos ou padrões frequentes entre conjuntos de dados.

Como exemplo, se considerarmos a regra de associação {cinto, bolsa} -> {sapatos}, esta indica que o cliente que compra cinto e bolsa, de acordo com um grau de certeza, compra também sapatos. O grau de certeza para uma dada regra define-se de acordo com dois índices:

- Suporte: Considerando a regra {A1, A4} -> {A6}, o suporte consiste na percentagem de vezes em que a ocorrência {A1, A4} e {A6} ocorre. Assim sendo, o suporte é a significância estatística da regra de associação que mede a frequência dos itens {A1, A4, A6} no conjunto de dados, sendo este utilizado para eliminar regras não interessantes (Equação 4);

(4) $$Suporte = \frac{Frequência\ de\ A1, A4\ e\ A6}{Total\ de\ T}$$

- Confiança: Considerando a regra {A1, A4} -> {A6}, a confiança consiste na percentagem de casos em que a ocorrência de {A1, A4} prevê corretamente a ocorrência de {A6}, ou seja, é a probabilidade condicional que uma transação que contenha {A1 U A4} também contém A6. Com isto, indica, no conjunto de dados, o grau de correlação entre A1, A4 e A6. A confiança indica-nos a força da regra ou o grau de incerteza da regra (Equação 5).

(5) $$Suporte = \frac{Frequência\ de\ A1, A4\ e\ A6}{Frequência\ de\ A1\ e\ A4}$$

Um dos algoritmos mais utilizados para a construção de itens frequentes denomina-se algoritmo *Apriori* [Vasconcelos e Carvalho, 2004]. O algoritmo *Apriori* foi o primeiro algoritmo proposto para a extração de itens frequentes e regras de associação. Este algoritmo utiliza uma estratégia de procura em largura. Em cada nível são gerados os *itemsets* possíveis com base nos *itemsets* frequentes gerados no nível anterior [Gama et al., 2012].

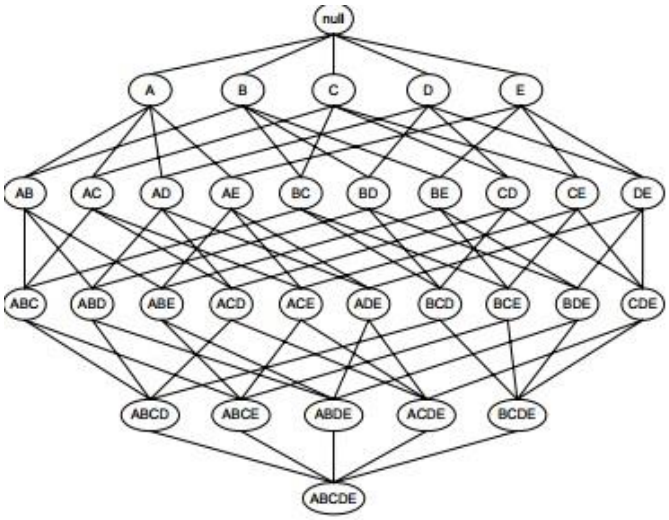

Figura 6. Itemsets frequentes – Algoritmo Apriori.

Este algoritmo inicia com os *itemsets* de tamanho 1 {A, B, C, D, E}, onde cada item é um membro do conjunto de *itemsets* candidatos. Os *itemsets* de tamanho k+1 são obtidos a partir dos *itemsets* de tamanho k. A Figura 6 representa todos os *itemsets* frequentes gerados pelo algoritmo *Apriori.* A Tabela 16 demonstra exemplos de *itemsets.*

Tabela 16. Itemsets gerados pelo algoritmo Apriori.

Tamanho	Itemsets
1	{A}, {B}, {C}, {D}, {E}
2	{A,B}, {A,C}, {A,D}, {A,E}
3	{A,B,C}, {ABD}, {ABE}
4	{A,B,C,D}, {A,B,C,E}
5	{A,B,C,D,E}

3.5 Avaliação da Técnica de Classificação

Após a criação dos modelos de classificação torna-se necessário efetuar uma avaliação. O desempenho de um modelo para além do algoritmo usado na sua construção depende, também, da distribuição das classes, do tamanho dos conjuntos de treino/teste e do custo das más classificações. Normalmente, são utilizadas duas medidas para avaliação [Rodrigues, 2014]:

- Exatidão: indica a proximidade ao valor verdadeiro;
- Precisão: mede a capacidade de reproduzir ou repetir uma medida.

O método *Holdout* é um método de amostragem que efetua a avaliação dos modelos através da separação do conjunto de dados em duas partes:

- Conjunto de treino: aqui são utilizados dois terços do conjunto inicial. Importa salientar que quanto maior o conjunto de treino, melhor o classificador;
- Conjunto de teste: utiliza um terço do conjunto inicial. Quanto maior o conjunto de teste, mais fiável a estimativa do erro.

O intervalo de confiança pode ser derivado através da equação:

(6) $$e \quad \pm z \times \sqrt{\frac{e \ \ (1 - \ e \ \ )}{nt}}$$

As variáveis representam:

- e: taxa de erro observada;
- nt: número de casos do conjunto de teste;
- z: valor da tabela da distribuição binomial para um nível de confiança de 95% => (z=1.96).

Para avaliar a performance e o interesse de cada modelo, existem várias medidas de avaliação:

- **Matriz de confusão:** tabela que permite visualizar os resultados da classificação e efetuar uma análise dos registos (Tabela 17). Cada linha da matriz representa as instâncias previstas de uma classe e cada coluna representa as instâncias reais de uma classe [Anacleto, 2009].

Tabela 17. Matriz de confusão.

		Classe prevista	
		Classe +	Classe -
Classe Real	Classe +	TP(++)	FN(+-)
	Classe -	FP(-+)	TN(--)

  - TP (Positivos Verdadeiros): número de casos positivos corretamente classificados;
  - FP (Positivos Falsos): número de casos negativos incorretamente classificados como positivos;
  - TN (Negativos Verdadeiros): número de casos negativos corretamente classificados;
  - FN (Negativos Falsos): número de casos positivos incorretamente classificados como negativos.

- **Exatidão (*Accuracy*)**: a exatidão representa a taxa de acerto de todo o classificador, ou seja, é a razão entre a soma dos acertos das classes a prever e o número total de instâncias. A fórmula utilizada é:

(7) $$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

A medida exatidão dá a mesma importância a todas as classes, tornando-se inadequada para avaliar conjuntos de dados desbalanceados e para avaliar classes com diferentes custos associados;

- **Recuperação (Recall) e Precisão (Precision):** estas medidas tendem a ser utilizadas quando a deteção de uma das classes é mais significativa. A medida precisão determina no grupo de registos, o número de registos positivos que o classificador determinou como positivos. O objetivo é minimizar os custos associados aos falsos positivos (FP). A fórmula usada para determinar a precisão é:

(8) $$Precision\ (Precisão) = \frac{TP}{TP + FP}$$

A recuperação determina o número de registos positivos que são previstos corretamente pelo classificador. O objetivo é minimizar os custos associados aos falsos negativos. A fórmula usada para calcular a medida recuperação é:

(9) $$Recall(Sensitivity) = TPR = \frac{TP}{TP + FN}$$

- **Taxa de verdadeiros positivos (TPR):** O indicador taxa de verdadeiros positivos (TPR) é um indicador que calcula a percentagem de instâncias classificadas corretamente como positivas.

(10) $$TPR = \frac{TP}{TP + FN}$$

- **Taxa de falsos positivos (False Positive Rate – FPR):** demonstra qual a percentagem de instâncias que são classificadas de forma errada como positivas. Esta medida pode ser obtida através da fórmula [Rodrigues, 2014].

(11) $$FPR = \frac{FP}{FP + TN}$$

- **Curva ROC**: é uma representação gráfica que relaciona a taxa de verdadeiros positivos (TPR), posicionada no eixo das ordenadas, com a taxa de falsos positivos (FPR),

posicionada no eixo das abcissas. Para efetuar a comparação de classificadores, a Curva ROC é reduzida a um valor escalar, representado pela área sob a curva (*Area Under Curve)* do gráfico obtido, o qual traduz a relação entre a TPR e a FPR.

A Figura 7 demonstra um exemplo de uma curva ROC. Nesta figura, pode verificar-se que um resultado excelente ocorre quando a taxa de verdadeiros positivos é elevada e, simultaneamente, a taxa de falsos positivos apresenta valores reduzidos (linha amarela). Nesta situação, o valor de área sob a curva é elevado. Quando se verifica uma proporcionalidade entre a taxa de verdadeiros positivos e a taxa de falsos positivos, o resultado deve ser rejeitado (linha azul), e neste caso, o valor determinado de área sob a curva é bastante inferior ao determinado nas situações representadas a rosa (bom resultado) e a amarelo (excelente resultado). Assim, a curva de ROC permite aferir de forma gráfica e quantitativa (área sob a curva) acerca da exatidão das assunções realizadas.

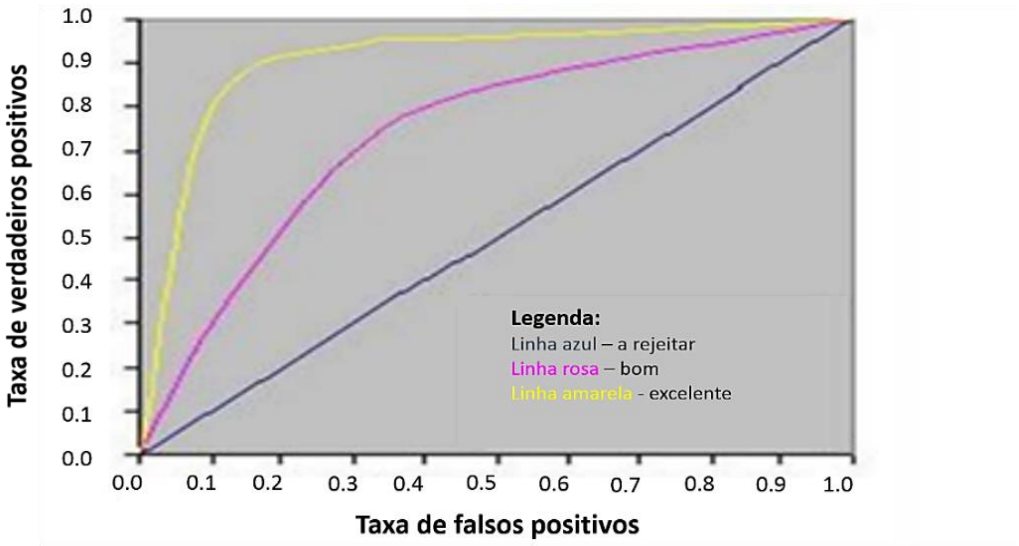

Figura 7. Comparação de curvas ROC.

3.6 Resumo

Este capítulo permitiu obter um conhecimento aprofundado acerca da mineração de dados. Após um estudo abrangente no capítulo anterior (2. Análises Sanguíneas) acerca dos atributos existentes no conjunto de dados, este capítulo (3. Descoberta de Conhecimento nos Dados) pretendeu detalhar todas as etapas de extração de conhecimento nos dados: fase de seleção dos atributos, pré-processamento de dados, transformação de dados, mineração de dados e a avaliação dos modelos. Este capítulo permite uma melhor compreensão dos próximos capítulos.

4 Exploração e Preparação dos Dados

Este capítulo divide-se em duas secções principais, que têm como objetivo apresentar as considerações obtidas após a exploração dos dados e a forma como estes foram preparados de acordo com as anomalias identificadas.

4.1 Exploração dos Dados

A compreensão dos dados é a matéria-prima para o desenvolvimento do processo de descoberta de conhecimento a partir dos dados. Para tal, o conjunto de dados disponível deve ser devidamente organizado e tratado, como abordado na secção relativa à preparação dos dados (secção 4.2).

A descoberta de conhecimento a partir dos dados disponíveis foi realizada com recurso à ferramenta Clementine 12.0, utilizando a metodologia CRISP-DM. Deste modo, a compreensão dos dados baseou-se num modelo constituído por 3 entidades: dadores, dádivas e análise. Numa primeira fase, os dados relativos aos dadores foram analisados isoladamente, enquanto os dados relativos às dádivas e à análise foram analisados em conjunto, numa proporção de 1:1 (uma análise para uma dádiva). Posteriormente, os dados relativos às 3 entidades foram agregados e analisados conjuntamente. De seguida, serão apresentadas diversas considerações relativas à compreensão dos dados referentes às entidades em análise.

4.1.1 Dadores

- Existem 41475 dadores de sangue no espaço amostral;
- Existem 21023 dadores do sexo feminino e 20452 do sexo masculino (Figura 8);

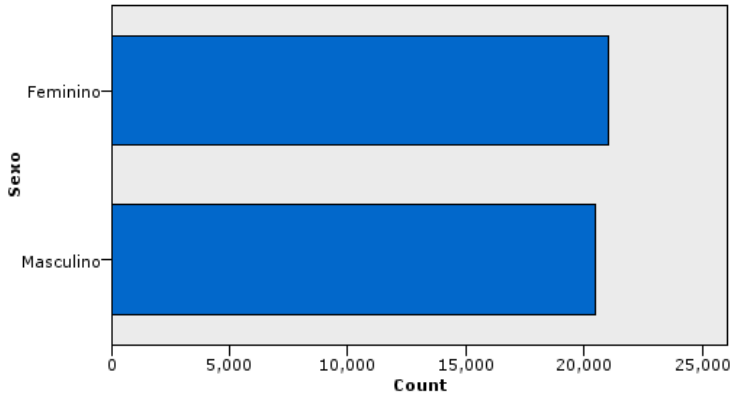

Figura 8. Dadores femininos vs. Dadores masculinos.

- Dos 41475 dadores de sangue, 27029 são casados (65.17%). Importa salientar que a maior parte dos dadores são do sexo masculino e casados (13923) (Figura 9);

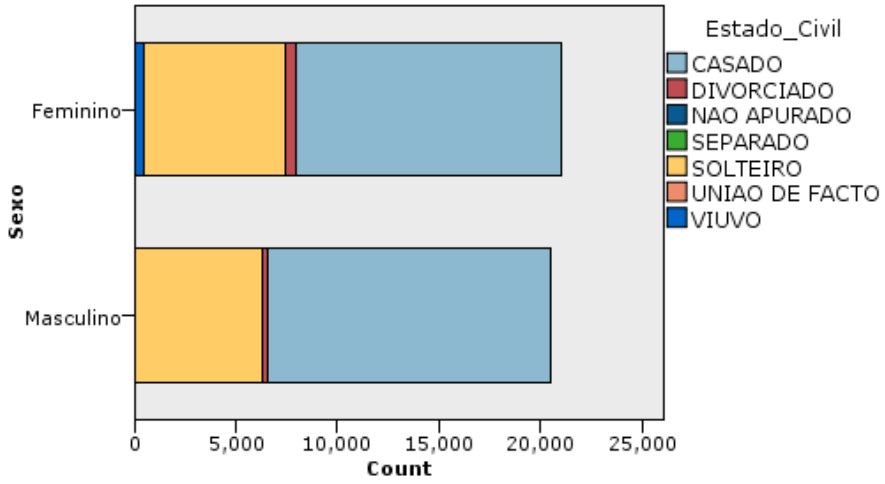

Figura 9. Sexo vs. Estado Civil.

- A idade dos dadores está compreendida entre os 30 e os 80 anos de idade. A média de idades é 51 anos;
- Apenas 1370 dadores têm grupo de sangue AB. Os grupos de sangue com maior percentagem de dadores são o A (19838) e o O (17254);
- 82.84% dos dadores têm Rh positivo;
- O tipo de sangue A com Rh positivo é encontrado pela maior parte dos dadores (16542) e, seguidamente, o sangue do tipo O com Rh positivo (14199). O tipo de sangue AB com Rh negativo é encontrado por uma pequena porção do espaço amostral (237);
- A média de peso dos dadores é 73 kg, o peso mínimo é 47 kg e o peso máximo é 170 kg (Figura 10).

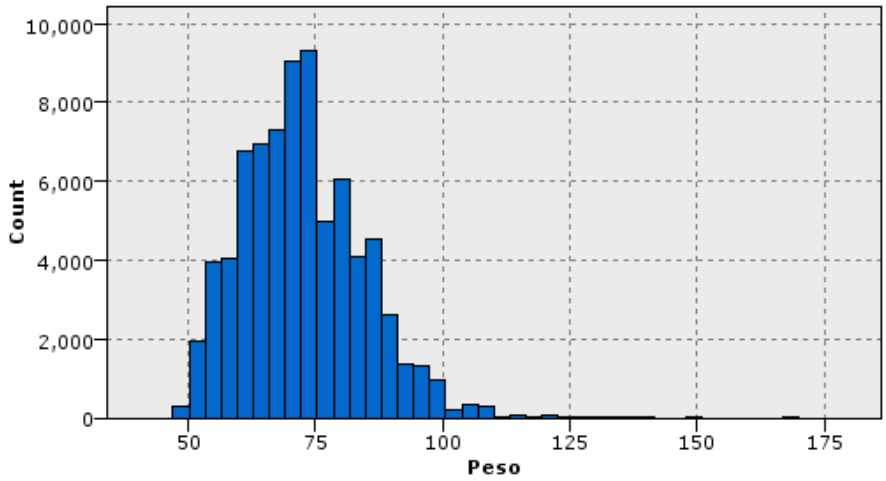

Figura 10. Dadores vs. Peso.

4.1.2 Dádivas e Análises

- Estas dádivas foram recolhidas entre 2000 e 2002;
- Neste estudo são consideradas 76582 dádivas de sangue;
- 77,39% das dádivas foram realizadas de manhã;
- Entre o ano de 2000 e 2002 o dia da semana com maior número de dádivas realizadas foi o Sábado (22451);
- O ano 2001 foi o ano onde foi obtido maior número de colheitas (31348);
- Os meses de Março e Maio foram os meses com mais dádivas efetuadas, perfazendo um total de 16496;
- O hospital onde se efetuaram mais dádivas de sangue neste período de tempo foi o Hospital distrital de Matosinhos – Pedro Hispano (8955 dádivas);
- Das 76582 dádivas apenas 33805 (44,14%) foram realizadas por dadores com a tensão arterial normal (Figura 11).

Value	Proportion	%	Count
Alta		27.01	20687
Baixa		28.84	22090
Normal		44.14	33805

Figura 11. Tensão Arterial com base nos indicadores padrão.

- 98,65% das dádivas correspondem a sangue total. Podemos aferir também que apenas existem 2 dádivas autólogas e 1029 dádivas do tipo aférese – separação do sangue à medida que é colhido;
- 99,98% das dádivas são utilizadas. Apenas 0,02% das dádivas tiveram problemas, tais como, análises de virologia, deficiência técnica, perda de validade, entre outros (Figura 12);

Value	Proportion	%	Count
Análises virologia		0.01	4
Deficiência técnica		0.0	2
Inicio sistema infor		0.0	3
Perda de validade		0.01	4
Quarentena incompleta		0.01	4
Sem amostra		0.0	1
Sem problema		99.98	76564

Figura 12. Problemas relativos às dádivas de sangue.

4.1.3 Dadores, Dádivas e Análises

- Apesar de todas as dádivas de sangue não estarem aqui representadas, é possível visualizar que o maior número de dádivas efetuadas pela mesma pessoa é 127. O dador tem como caraterísticas:
  - Tem 73 anos;
  - Casado;
  - Grupo A positivo;
  - Aposentado;
  - Vive em Gondomar;
  - Teve contato com o vírus da hepatite B. O resultado positivo vai permanecer durante toda a vida mesmo que este se encontre curado. Este dador doou sangue, em média, 4 vezes por ano dar sangue;
- Nas dádivas de sangue analisadas não foi encontrado o marcador de triagem da Hepatite C (Anti-HCV);
- Nas dádivas de sangue analisadas (76582) não foi encontrado o antigénio de superfície do vírus da Hepatite B (Ag-HBs);
- O anticorpo Anti-VIH foi encontrado apenas numa análise sanguínea. A pesquisa de anticorpos Anti-VIH é usada para efetuar uma triagem e o diagnóstico de infeção pelo VIH.
- Em todas as dádivas de sangue não foi encontrado nenhum teste VDRL positivo (identificação de pacientes com sífilis);
- Foi encontrado alanina aminotransferase (ALT) acima do valor normal em 446 dádivas. 99,42% das dádivas têm o valor entre o intervalo normal;
- Das 446 dádivas de sangue recebidas com o valor de ALT anormal pode ser salientado que 386 foram recebidas por indivíduos do sexo masculino e apenas 60 por indivíduos do sexo feminino (Figura 13);

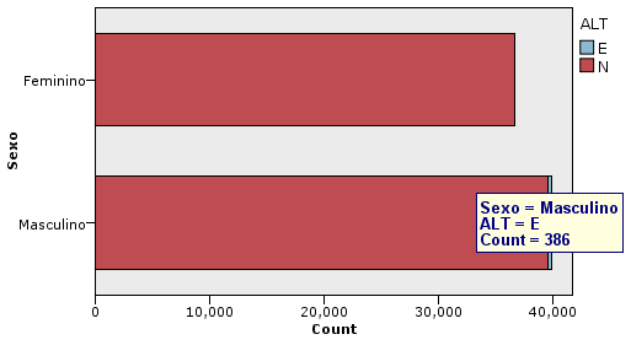

Figura 13. Sexo vs. Alanina aminotransferase (ALT).

- É no subconjunto das dádivas efetuadas por pessoas casadas que encontramos mais dádivas com um valor anormal de ALT (339 dádivas). Seguidamente, é o subconjunto das dádivas efetuadas por solteiros (98 dádivas);
- Importa salientar que valores anormais de ALT apenas se encontram em dádivas de sangue com a hemoglobina dentro dos valores normais. Também pode ser verificado que valores anormais de ALT apenas existem em dádivas cujo anticorpo da Hepatite B não se encontra. Como tal, podemos considerar que os dadores não tinham Hepatite B mas podem possuir algum problema no fígado como, por exemplo, cirrose (causada devido a excesso indevido de álcool);
- Relativamente ao anticorpo da Hepatite B (Anti-HBc):
    - 95,18% das dádivas deram resultado negativo;
    - 0.01% das dádivas deram como resultado imunidade;
    - 4,1% das dádivas reagiram (positivo);
- Das dádivas de sangue que deram como resultado imunidade à Hepatite B (11) com base no anticorpo Anti-HBc, 55% das dádivas têm o tipo O de sangue;
- Foi encontrado um número anormal de glóbulos brancos (WBC) em 27 dádivas de sangue, sendo que destas 21 dádivas pertenciam a dadores do sexo masculino. Importa também salientar que o Rhesus é positivo em 23 dádivas com um número anormal de glóbulos brancos;
- Em 36432 dádivas foram encontrados valores anormais de neutrófilos (NEU%) (Figura 14);

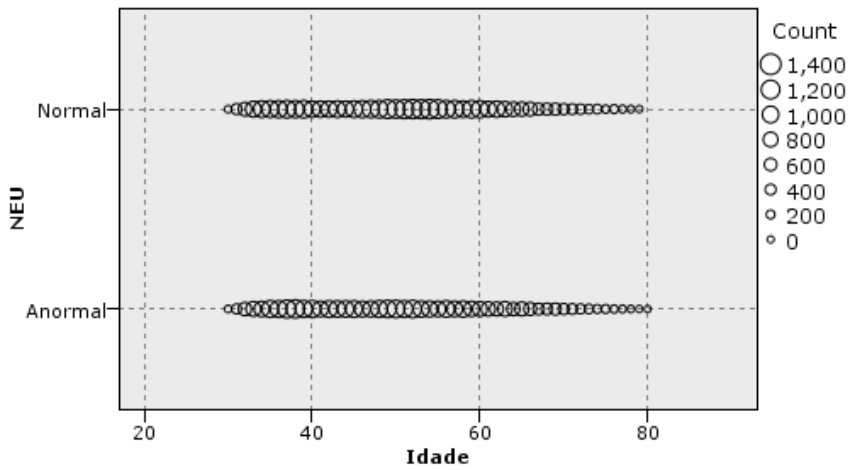

Figura 14. Idade vs. NEU%.

- Após a elaboração do exame hematócrito (HCT) foram descobertos valores anormais de glóbulos brancos no sangue em 2721 dádivas.
- Das 2721 dádivas de sangue com valores anormais de glóbulos brancos, 2748 foram realizadas por pessoas do sexo masculino;
- Após análise 7461 dádivas continham um volume corpuscular médio (MCV) acima do esperado;
- Pode ser afirmado que 20544 dádivas possuíam hemoglobina globular média (MCH) acima do pressuposto. O seguinte histograma apresenta a distribuição da

hemoglobina globular média de acordo com as idades dos indivíduos. Podemos concluir que os valores anormais de MCH encontram-se em todas as idades, mas predominam mais entre as idades de 35 e 60 anos (Figura 15);

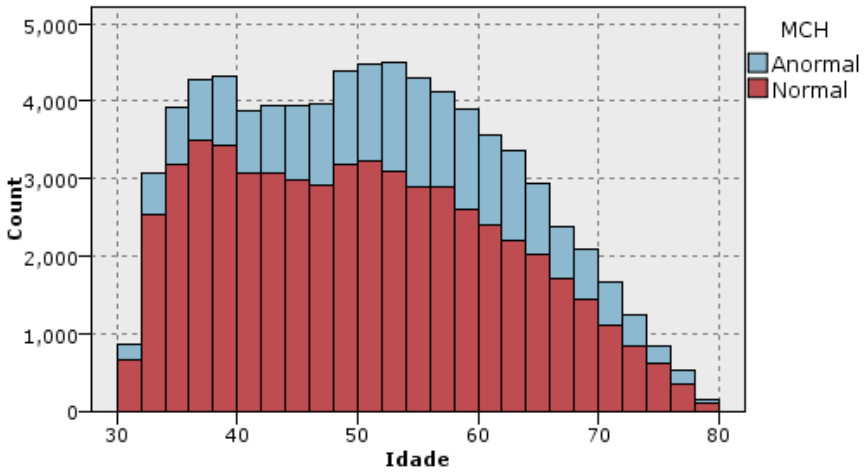

Figura 15. Idade vs. MCH.

- Nas análises efetuadas foi encontrada uma concentração anormal de hemoglobina globular média (MCHC) em 421 dádivas;
- Foram encontrados valores anormais de eritrócitos (RBC) em 31 dádivas de sangue;
- Foram encontrados valores anormais e graves de hemoglobina (HGB) em 165 análises sanguíneas;
- Foram encontrados valores anormais de plaquetas (PLT) em 1821 dádivas;

4.2 Preparação dos dados

A preparação dos dados agrupa as etapas de escolha, limpeza e tratamento dos dados. Os critérios para a escolha dos dados devem ter em consideração os objetivos do problema de mineração de dados e de algumas condições técnicas (volume, tipo de dados, entre outras). A seleção dos atributos permite identificar atributos importantes, melhorar o desempenho de várias técnicas de extração de conhecimento de dados, reduzir a necessidade de memória e tempo de processamento, eliminar atributos irrelevantes e reduzir ruído, lidar com a dimensionalidade, simplificar o modelo gerado e facilitar a sua compreensão, facilitar a visualização dos dados e reduzir o custo de recolha de dados.

A limpeza e o tratamento dos dados possibilitam a correção de inconsistências, tais como a eliminação de dados incompletos, dados inconsistentes, dados redundantes e dados com ruído. É de extrema importância que a qualidade dos dados seja assegurada. É estimado que a fase de preparação dos dados demore cerca de 80% de todo o processo de mineração de dados.

Os próximos subcapítulos pretendem apresentar como foi efetuado o processo de migração dos dados para uma base de dados *SQL* e a limpeza dos dados.

4.2.1 Migração dos dados

A base de dados inicial foi desenvolvida em *Access* e contém todo o conteúdo dos indivíduos e as respetivas dádivas. A Figura 16 apresenta o modelo de dados inicial. Após a análise da normalização da mesma, foi efetuada uma reestruturação de modo a facilitar o processo de mineração dos dados. A Figura 17 apresenta o novo modelo de dados. Para tal, foi utilizada a ferramenta *SQL Server Data Tools* para migrar os dados para uma base de dados *SQL* normalizada.

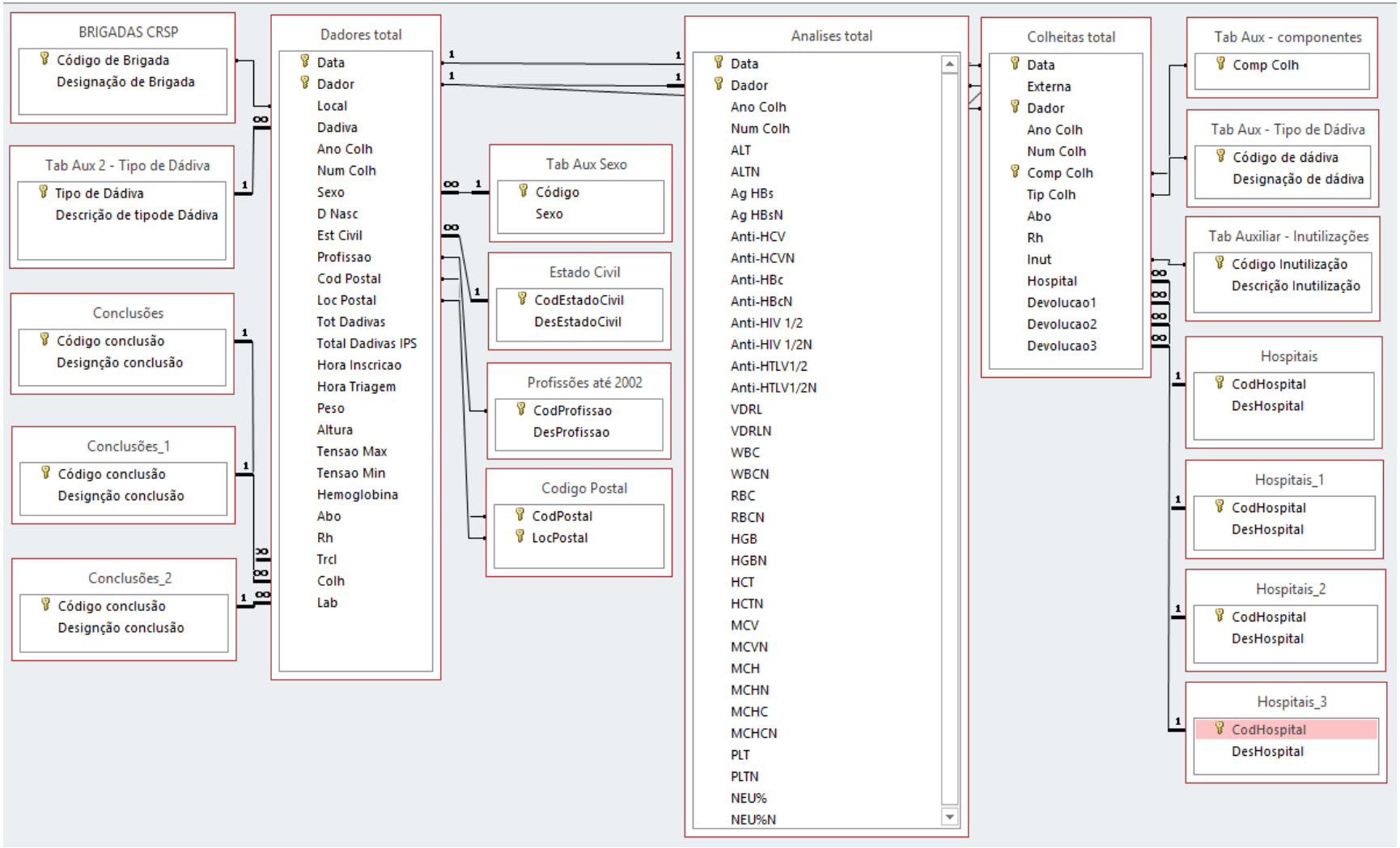

Figura 16. Modelo de dados anterior da base de dados.

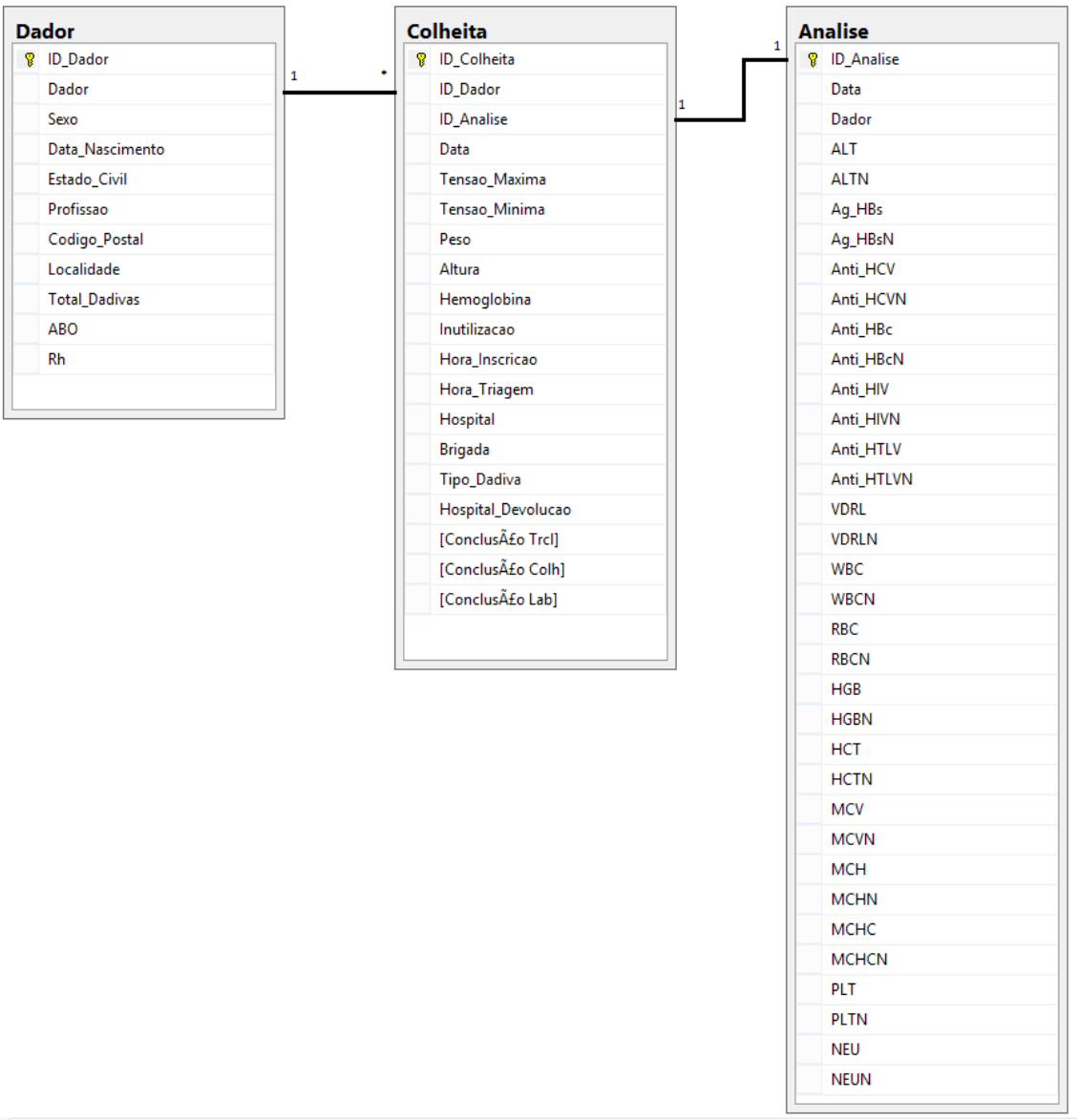

Figura 17. Novo modelo de dados da base de dados.

Foi criado um projeto do tipo *Microsoft SQL Server Integration Services* que permitiu a conceção de uma base de dados *SQL* e as respetivas tabelas. Após a criação das tabelas, todos os atributos dos registos da base de dados *Access* foram tratados e alterados para a nova tabela de acordo com a nova normalização (caso fosse justificável).

A Figura 18 apresenta o fluxo de controlo utilizado para a realização do procedimento anteriormente descrito. Todas as tarefas de fluxo de dados utilizadas podem ser visualizadas nos anexos deste documento.

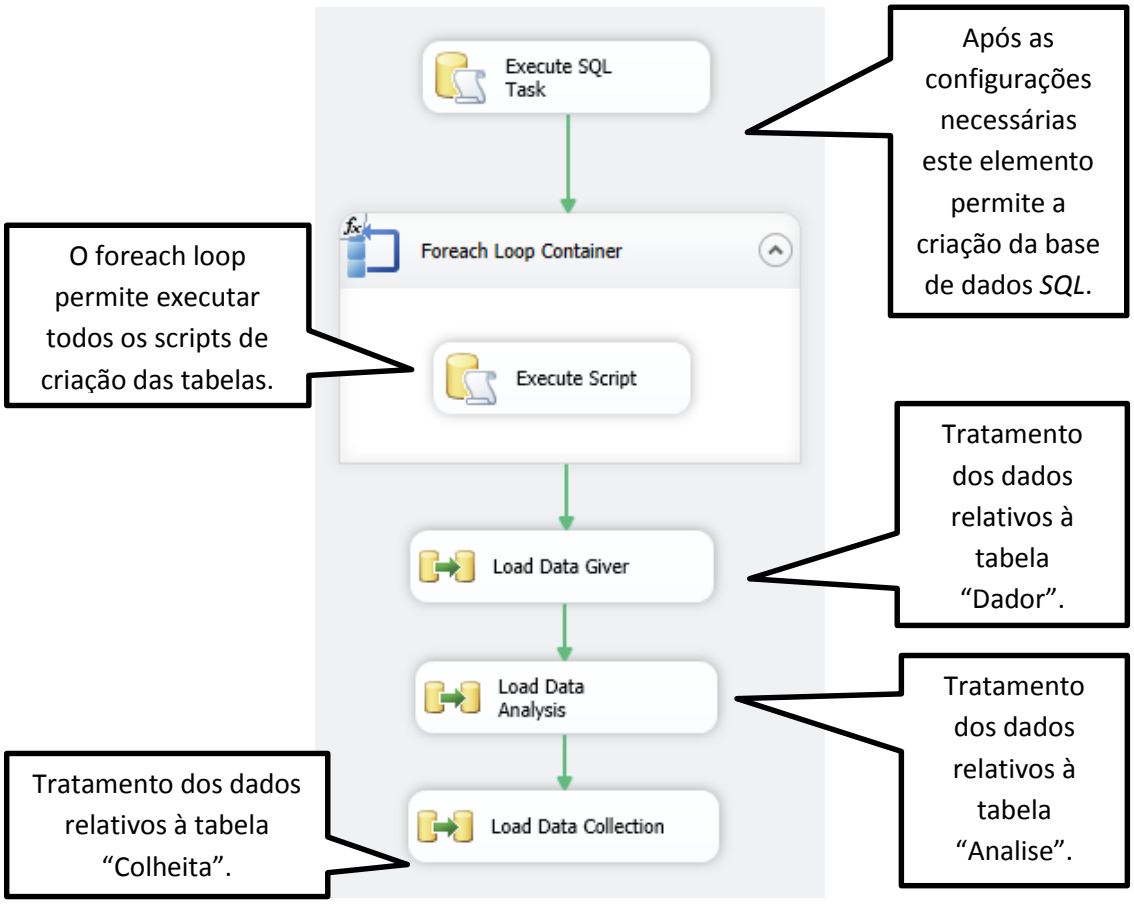

Figura 18. Controlo de Fluxo do projeto.

Após a criação da base de dados e de um projeto na ferramenta *Clementine 12.0*, foi efetuada a ligação com a base de dados *SQL*. Importa salientar que foi necessária a criação de alguns ficheiros temporários, devido a problemas na alteração das datas. Após as alterações, os dados foram guardados num ficheiro e posteriormente guardados numa nova tabela *SQL*. A Figura 19 apresenta um exemplo da descrição efetuada anteriormente.

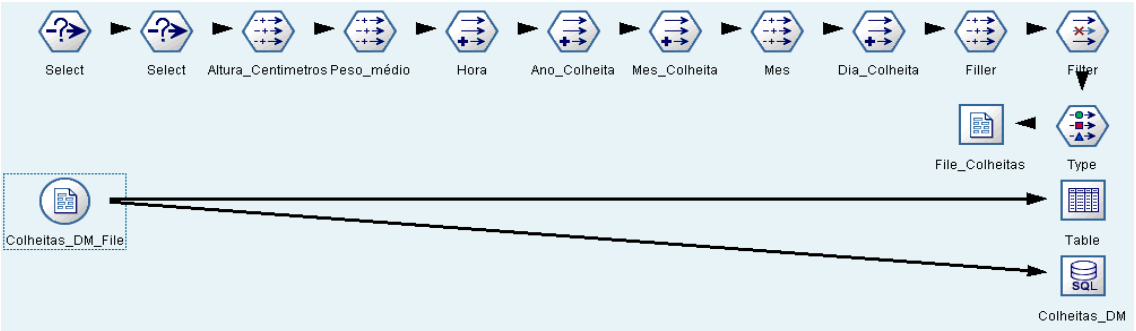

Figura 19. Tratamento dos dados (Passo intermédio de criação de um ficheiro).

O próximo subcapítulo apresenta todas as alterações efetuadas nos atributos de cada tabela para diminuir o ruído, inconsistência e redundância dos dados.

4.2.2 Limpeza dos dados

Os dados existentes no conjunto de dados continham inúmeros problemas (ruído, inconsistência e redundância). Como tal, houve a necessidade de efetuar a respetiva resolução através da adição de novos atributos, inserção manual de dados de acordo com outros atributos, utilização da média para atributos com valor a null e a exclusão de registos com inúmeros campos com valor null. A exclusão de registos diminuíram em cerca de 10% o valor total dos mesmos (8000 da tabela "Dadores", 20000 da tabela "Análises" e 12000 da tabela "Colheitas"). No final da preparação dos dados, o valor dos dadores e das dádivas com a respetiva análise é de aproximadamente 42000 e 80000, respetivamente.

4.2.2.1 Tabela Dador

- Exclusão dos registos que contêm o campo "Profissão" com o valor null;
- Exclusão dos registos que contêm o campo "Codigo_Postal" igual a 9999 com o valor null;
- Exclusão dos registos que contêm o campo "Total_Dadivas" igual a 0;
- Exclusão dos registos que contêm o campo "Estado_Civil" com o valor null;
- Exclusão dos registos que contêm o campo "Sexo" com o valor null;
- Foi efetuado o calculo a idade do dador com base na sua data de nascimento. Para os registos com o atributo data de nascimento igual a null foi introduzida a média do atributo "Idade";
- Exclusão do campo "Data_Nascimento". Isto foi efetuado devido ao cálculo da idade com base na data de nascimento;
- Exclusão dos registos que não têm grupo sanguíneo conhecido;
- Correção de erros de escrita: "A1" modificado para "A" e "A2B" modificado para "AB";
- Inserção manual da localidade com base no código postal:
    - 2780: Oeiras;
    - 4700: Braga;
    - 4430: Vila Nova de Gaia;
    - 4500: Espinho;
    - 4470: Maia.

4.2.2.2 Tabela Colheita

- Exclusão dos registos que contêm o campo "Inutilização" com o valor null;
- Exclusão do ruído existente nos dados relativos às tensões mínimas e máximas:
    - Tensão máxima maior ou igual a 20 e menor ou igual a 8;
    - Tensão mínima menor ou igual a 5 e maior ou igual a 15.
- Exclusão dos registos que contêm o campo "Hospital" com o valor null;
- Exclusão do campo "Altura" (cerca de 90% dos registos contêm este campo com valor null);
- Generalização no atributo "Hora_Inscricao":
    - Caso a hora for inferior a 13, manhã;
    - Caso a hora for maior ou igual a 13, tarde.

- Divisão do atributo “Data” em 3 atributos:
    - “Ano”;
    - “Mês”;
    - “Dia”.
- Exclusão dos atributos “Hora_Inscricao” e “Hora_Triagem”;
- Exclusão do atributo “Hemoglobina” (cerca de 90% dos registos contêm este campo com o valor null);
- Para o atributo “Peso” com o valor igual a null, foi introduzida a média do peso dos dadores.

4.2.2.3 Tabela Análise

- Exclusão dos registos que contêm o campo “Inutilização” com o valor null;
- Exclusão dos registos que contêm o campo “ALT” com o valor null;
- Exclusão dos registos que contêm o campo “ALTN” com o valor null;
- Exclusão dos registos que contêm o campo “ALTN” inferior a 2 (eliminar ruído e inconsistência dos dados);
- Exclusão dos registos que contêm ‘?’ no campo “Ag_HBs”;
- Exclusão dos registos que contêm o campo “Anti_HCV” com o valor null;
- Exclusão dos registos que contêm ‘?’ no campo “Ag_HBs”;
- Exclusão dos registos que contêm ‘?’ no campo “Anti_HCV”;
- Exclusão dos registos que contêm o campo “Anti_HBc” com o valor null;
- Exclusão dos registos que contêm ‘?’ no campo “Anti_HBc”;
- Exclusão dos registos que contêm ‘P’ no campo “Anti_HBc”;
- Exclusão dos registos que contêm o campo “Anti_HIV” com o valor null;
- Exclusão dos registos que contêm ‘?’ no campo “Anti_HIV”;
- Exclusão dos registos que contêm o campo “Anti_HTLV” com o valor null;
- Exclusão dos registos que contêm o campo “Ag_HBsN” com o valor null;
- Exclusão dos registos que contêm o campo “Anti_HTLVN” com o valor null;
- Exclusão dos registos que contêm o campo “Anti_HCVN” com o valor null;
- Exclusão dos registos que contêm o campo “Anti_HBcN” com o valor null;
- Exclusão dos registos que contêm o campo “Anti_HIVN” com o valor null;
- Exclusão dos registos que contêm o campo “VDRL” com o valor null;
- Exclusão dos registos que contêm o campo “WBC” com o valor null;
- Exclusão dos registos que contêm ‘?’ no campo “WBC”;
- Exclusão dos registos que contêm o campo “RBCN” com o valor null;
- Adição de um novo atributo relativo ao “HGBN” denominado “HGB” onde:
    - Caso “HGBN” seja inferior a 8 então “HGBN” é considerado grave;
    - Caso “HGBN” seja superior ou igual a 12 e “HGBN” inferior ou igual a 18 é considerado normal;
    - Outro valor é considerada anormal.
- Adição de um novo atributo relativo ao “HCTN” denominado “HTC” onde:

    - Caso “HCTN” seja superior a 35 e “HCTN” inferior ou igual a 50 é considerado normal;
    - Outro valor é considerada anormal.
- Adição de um novo atributo relativo ao “MCVN” denominado “MCV” onde:
    - Caso “MCVN” seja superior ou igual a 80 e “MCVN” inferior ou igual a 95 é considerado normal;
    - Outro valor é considerada anormal.
- Adição de um novo atributo relativo ao “MCHN” denominado “MCH” onde:
    - Caso “MCHN” seja superior ou igual a 27 e “MCHN” inferior ou igual a 31 é considerado normal;
    - Outra possibilidade é considerada anormal.
- Adição de um novo atributo relativo ao “MCHCN” denominado “MCHC” onde:
    - Caso “MCHCN” seja superior ou igual a 32 e “MCHCN” inferior ou igual a 36 é considerado normal;
    - Outro valor é considerada anormal.
- Adição de um novo atributo relativo ao “PLTN” denominado “PLT” onde:
    - Caso “PLTN” seja superior ou igual a 150 e “PLTN” inferior ou igual a 450 é considerado normal;
    - Outro valor é considerada anormal.
- Adição de um novo atributo relativo ao “NEUN” denominado “NEU” onde:
    - Caso “NEUN” seja superior ou igual a 55 e “NEUN” inferior ou igual a 70 é considerado normal;
    - Outro valor é considerada anormal.
- Adição de um novo atributo relativo ao “RBCN” denominado “RBC” onde:
    - Caso “RBCN” seja superior ou igual a 4 é considerado normal;
    - Outro valor é considerada anormal.
- Para os atributos “Ag_HBs”, “Anti_HCV”, “Anti_HBc”, “Anti-HIV” e “WBC” os registos com o valor “NE” são modificados para “N”;
- Para os registos que contêm os atributos “HGBN”, “HCTN”, “WBCN”, “MCVN”, “MCHN” e “MCHCN” com o valor null, estes são modificados pela média dos mesmos.

5 Modelação e Avaliação

Este capítulo divide-se em três subcapítulos – Classificação, Segmentação e Regras de Associação – e tem como objetivo demonstrar quais os algoritmos e parâmetros utilizados para encontrar os resultados pretendidos e responder aos objetivos e cenários delineados. Em cada subcapítulo é apresentada a criação dos modelos necessários com base nas técnicas utilizadas e é efetuada a avaliação dos mesmos.

5.1 Classificação

A técnica de classificação tem como objetivo a identificação de padrões nos dados que relacionem os valores das variáveis independentes e o de uma variável objetivo. Os métodos de classificação permitem gerar modelos que preveem com exatidão a classe alvo. Como foi mencionado anteriormente, foram criados modelos que permitissem responder a determinados cenários.

5.1.1 Quais os perfis dos dadores que podem vir a contrair possíveis problemas no fígado, tais como, cirrose, hepatite ou colestase?

Após a escolha dos atributos e da seleção do atributo objetivo (ALT), foi efetuado um balanceamento dos dados com o fator 0.01 para o atributo ALT (alanina aminotransferase) com o valor igual a “N” (Figura 20).

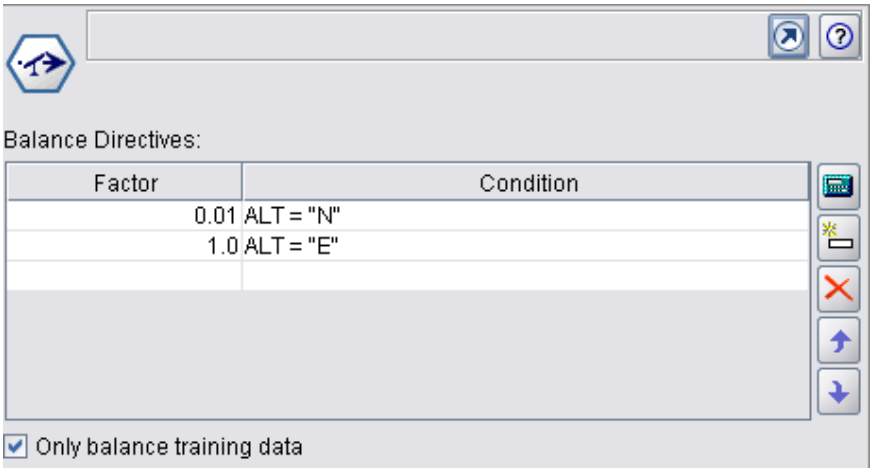

Figura 20. Balanceamento dos dados.

O método utilizado para avaliação designa-se de método *Holdout* que divide o conjunto inicial de dados aleatoriamente em dois conjuntos disjuntos: conjunto treino que equivale a 2/3 do conjunto inicial e o conjunto teste que equivale a 1/3 do conjunto inicial. O modelo foi criado com base no algoritmo *C5.0*. A Figura 21 demonstra o fluxo do que foi referido anteriormente.

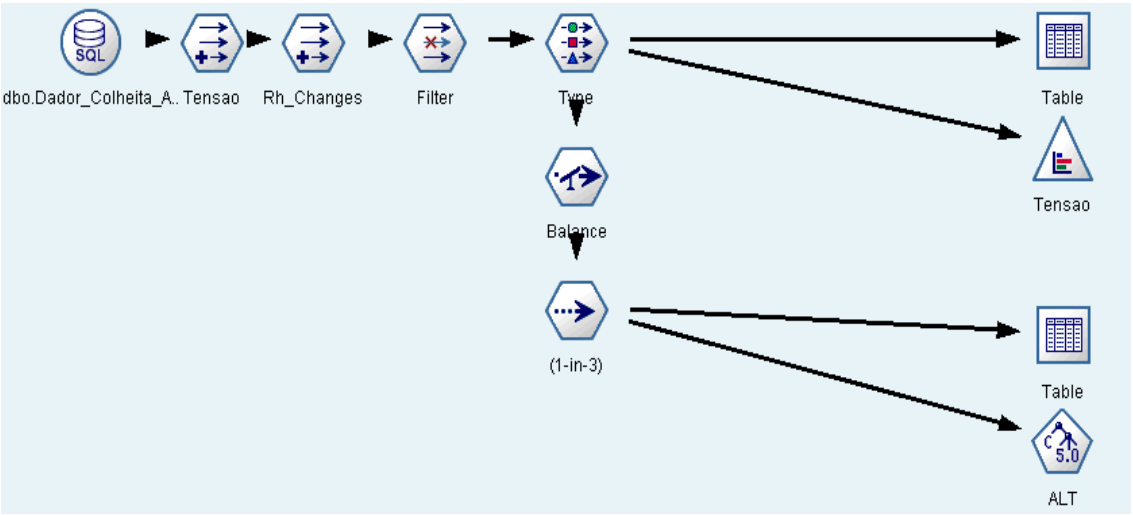

Figura 21. Criação do modelo com base no algoritmo C5.0 – atributo objetivo ALT.

Após o modelo criado (Figura 22) com base no algoritmo *C5.0*, verificou-se que a taxa de acerto é de 81%.

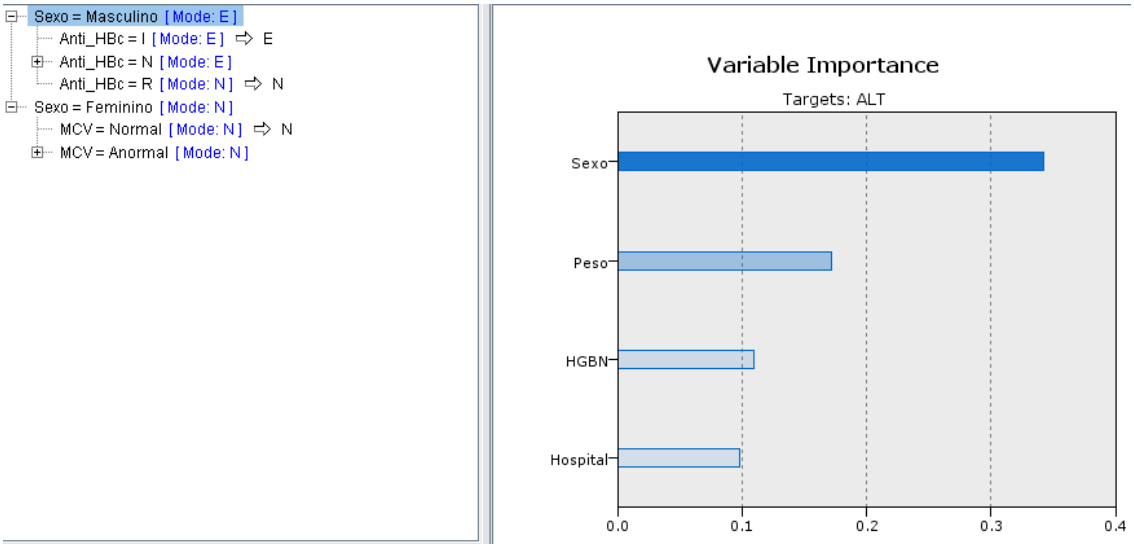

Figura 22. Apresentação inicial do modelo criado com base no algoritmo C5.0 e com o atributo objetivo ALT.

O modelo criado pode ser visualizado mais pormenorizadamente na secção dos anexos (Anexo C – Figura 56). A Tabela 18 representa a matriz de confusão relativa ao modelo criado.

Tabela 18. Matriz de confusão – atributo objetivo ALT.

ALT	**E**	**N**
E	76	73
N	4789	20589

Esta matriz apresenta os valores resultantes da previsão do modelo criado onde:

- 76 dádivas de sangue com valores anormais de ALT foram previstas corretamente, enquanto que 73 dádivas de sangue foram previstas como tendo valores normais de ALT que na realidade têm valores anormais de ALT;
- 20589 dádivas de sangue com valores normais de ALT foram previstas corretamente, enquanto que 4789 dádivas foram previstas como tendo valores anormais de ALT que na realidade detêm valores normais de ALT.

Utilizando o modelo, podem ser descritos alguns perfis de indivíduos que possuem ou podem vir a apresentar valores anormais de ALT e, consequentemente, problemas de fígado (Tabela 19):

Tabela 19. Perfis de indivíduos com valores anormais de ALT – Árvore de decisão.

Perfil 1	Sexo Masculino
	Imune à Hepatite B (Anti-HBc)
Perfil 2	Sexo Masculino
	Anti-HBc negativo
	Peso superior a 70 Kg
	Valor de hemoglobina inferior ou igual a 16
Perfil 3	Sexo Masculino
	Anti-HBc negativo
	Peso superior a 70 Kg
	Entrega da dádiva efetuada no Centro Regional do Sangue de Lisboa
	Valores de Neutrófilos acima do valor normal
	Valor de hemoglobina superior ou igual a 14
Perfil 4	Sexo Masculino
	Anti-HBc negativo
	Peso superior a 70 Kg
	Entrega da dádiva realizada no Centro Regional Oncologia do Porto
	Valor anormal da hemoglobina globular média (MCH)

Importa salientar que foi utilizado a técnica de segmentação (algoritmos *TwoSteps*, *Kohonen* e *K-means*), mas não foram conseguidos melhores resultados.

5.1.2 Quais os perfis dos dadores que podem apresentar um défice de fatores essenciais (ferro, vitamina B12 ou ácido fólico) para uma quantidade de hemoglobina adequada por glóbulo vermelho?

Após a escolha dos atributos, da seleção do atributo objetivo (MCH) e da utilização do método de avaliação *Holdout,* foram criados modelos com base nos algoritmos *C5.0*, *CHAID* e *C&R tree*. Após a criação dos modelos, foi feita a avaliação dos mesmos (Figura 23).

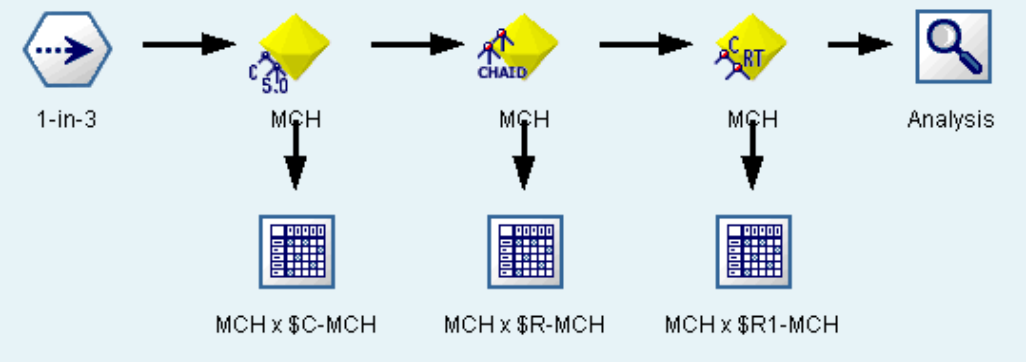

Figura 23. Avaliação dos modelos – atributo objetivo MCH.

Após análise dos modelos concebidos, foi concluído que o modelo com maior taxa de acerto foi aquele que foi desenvolvido com base no algoritmo *C5.0* (95%) (Figura 24).

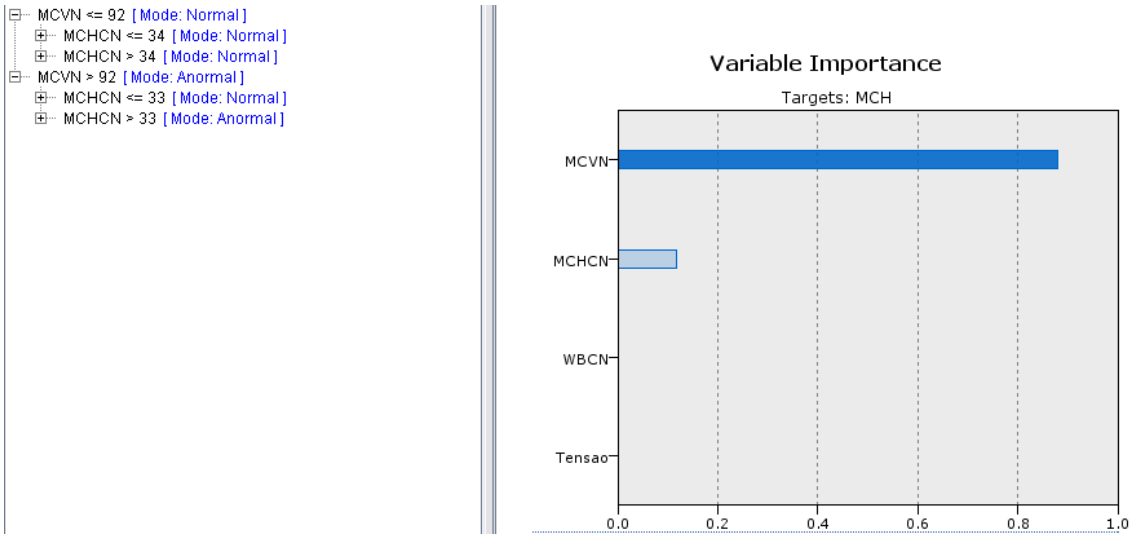

Figura 24. Apresentação inicial do modelo criado com base no algoritmo C5.0 e com o atributo objetivo MCH.

O modelo criado pode ser visualizado mais pormenorizadamente na secção dos anexos (Anexo C – Figura 57). A Tabela 20 apresenta a matriz de confusão relativa ao modelo.

Tabela 20. Matriz de confusão – atributo objetivo MCH.

MCH	Anormal	Normal
Anormal	6148	763
Normal	616	18000

Esta matriz apresenta os valores resultantes da previsão do modelo criado onde:

- 6148 dádivas de sangue com valores anormais de MCH foram previstas corretamente, enquanto que 763 dádivas de sangue foram previstas como tendo valores normais de MCH que na realidade têm valores anormais de MCH;
- 18000 dádivas de sangue com valores normais de MCH foram previstas corretamente, enquanto que 616 dádivas foram previstas como tendo valores anormais de MCH que na realidade detêm valores normais de MCH.

Utilizando o modelo com maior taxa de acerto, serão apresentados perfis de indivíduos que possuem ou podem vir a apresentar uma quantidade de hemoglobina inadequada por glóbulo vermelho (Tabela 21):

Tabela 21. Perfis de indivíduos com valores anormais de MCH.

Perfil 1	Volume médio dos glóbulos vermelhos do sangue (MCV) superior a 92 fentolitros
	Concentração média de hemoglobina por glóbulo vermelho (MCHC) superior a 33 gramas por decilitro
Perfil 2	Volume médio dos glóbulos vermelhos do sangue (MCV) superior a 92 fentolitros
	Concentração média de hemoglobina por glóbulo vermelho (MCHC) superior a 33 gramas por decilitro
	Quantidade de glóbulos vermelhos no volume total de sangue inferior a 41%
	Dádiva realizada no ano de 2000 ou 2001
	Valor da hemoglobina (HGB) inferior a 14
Perfil 3	Volume médio dos glóbulos vermelhos do sangue (MCV) superior a 92 fentolitros
	Concentração média de hemoglobina por glóbulo vermelho (MCHC) superior a 33 gramas por decilitro
	Quantidade de glóbulos vermelhos no volume total de sangue superior a 40%
	Dádiva efetuada no mês de Abril
	O grupo sanguíneo do indivíduo é o A
	Quantidade de eritrócitos inferior a 5 unidades
	Tensão baixa

5.1.3 Quais os perfis dos indivíduos dadores que podem vir a contrair doenças autoimunes, úlceras gástricas, doenças renais ou bloqueios nos vasos sanguíneos?

Para a criação do modelo utilizado foram escolhidos os atributos pretendidos, efetuou-se a seleção do atributo objetivo (PLT) e foi selecionado o método de avaliação (*Holdout).* Foram criados dois modelos, com base nos algoritmos *C5.0* e *CHAID*.

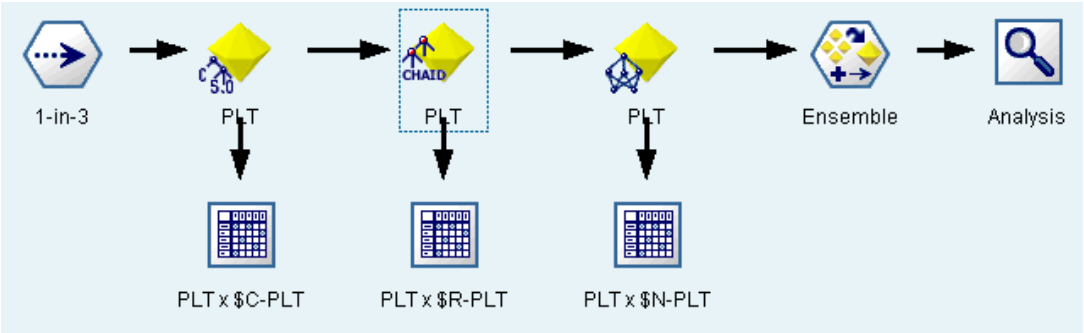

Figura 25. Avaliação dos modelos – atributo objetivo PLT.

Após a avaliação dos modelos (Figura 25), foi concluído que o modelo com maior taxa de acerto foi criado com o algoritmo *C5.0*. Este modelo, tem como taxa de acerto 86.18%. No

entanto, o modelo criado com o algoritmo *CHAID* também resultou numa taxa de acerto de 84.81%. As Figuras 26 e 27 apresentam os modelos iniciais criados.

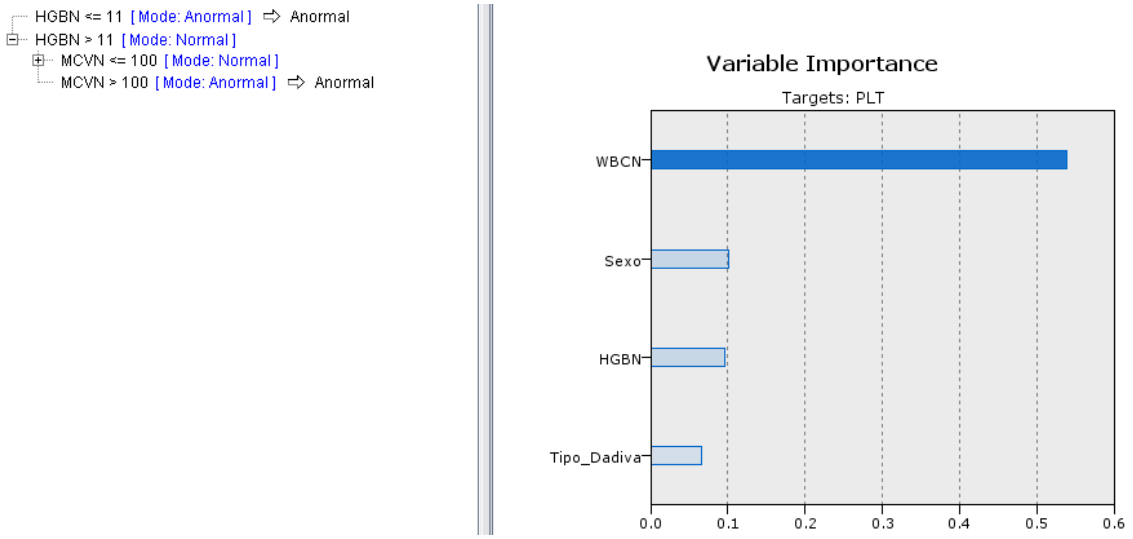

Figura 26. Apresentação inicial do modelo criado com base no algoritmo C5.0 e com o atributo objetivo PLT.

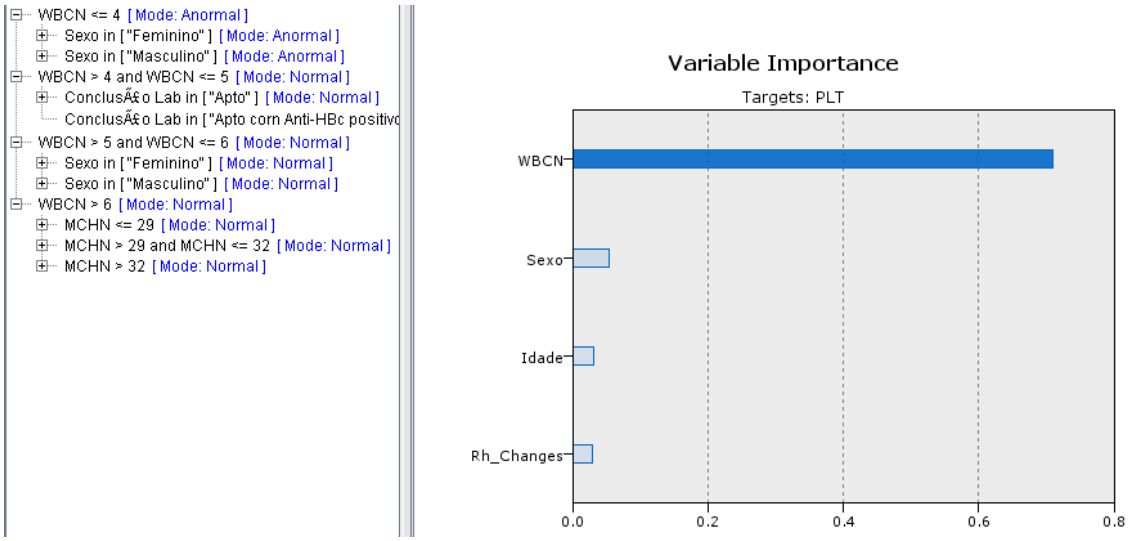

Figura 27. Apresentação inicial do modelo criado com base no algoritmo CHAID e com o atributo objetivo PLT.

Os modelos criados podem ser visualizados mais pormenorizadamente na secção dos anexos (Anexo C – Figuras 58 e 59).

A Tabela 22 e a Tabela 23 apresentam a matriz de confusão para ambos os modelos.

Tabela 22. Matriz de confusão de acordo com o algoritmo C5.0 – atributo objetivo PLT.

PLT	Anormal	Normal
Anormal	249	375
Normal	3154	21749

Esta matriz apresenta os valores resultantes da previsão do modelo criado onde:

- 249 dádivas de sangue com valores anormais de PLT foram previstas corretamente, enquanto que 375 dádivas de sangue foram previstas como tendo valores normais de PLT que na realidade têm valores anormais de PLT;
- 21749 dádivas de sangue com valores normais de PLT foram previstas corretamente, enquanto que 3154 dádivas foram previstas como tendo valores anormais de PLT que na realidade detêm valores normais de PLT.

Podem ser retiradas as mesmas conclusões da Tabela 22 para a Tabela 23, considerando apenas que o número de dádivas é diferente para cada célula da tabela.

Tabela 23. Matriz de confusão de acordo com o algoritmo CHAID – atributo objetivo PLT.

PLT	Anormal	Normal
Anormal	221	403
Normal	3474	21429

De acordo com os modelos criados com uma taxa de acerto superior a 80%, podem ser apresentados diferentes perfis de indivíduos com possibilidade de virem a contrair doenças renais, úlceras gástricas, entre outras (Tabela 24 e Tabela 25):

Tabela 24. Perfis de indivíduos com valores anormais de PLT – algoritmo C5.0.

Perfil 1	Valor da hemoglobina inferior a 12 gramas por decilitro
Perfil 2	Valor da hemoglobina superior a 11 gramas por decilitro
	Volume médio dos glóbulos vermelhos no sangue superior a 100 fentolitros

Tabela 25. Perfis de indivíduos com valores anormais de PLT – algoritmo CHAID.

Perfil 1	Valor de leucócitos inferior a 4 mm^3
	Indivíduos do sexo masculino
Perfil 2	Valor de leucócitos inferior a 4 mm^3
	Indivíduos do sexo feminino
	Colheita efetuada no ano 2000

5.1.4 Quais os perfis dos dadores que contraem ou podem vir a contrair doenças hepáticas (diferentes tipos de anemia)?

Para a criação do modelo utilizado foram escolhidos os atributos pretendidos, efetuou-se a seleção do atributo objetivo (MCV) e foi selecionado o método de avaliação (*Holdout).* Após a criação dos modelos com base nos algoritmos *C&R tree, C5.0* e *CHAID* foi efetuada a avaliação dos mesmos e verifica-se que o modelo com maior taxa de acerto foi criado com o algoritmo *C5.0* (Figura 28).

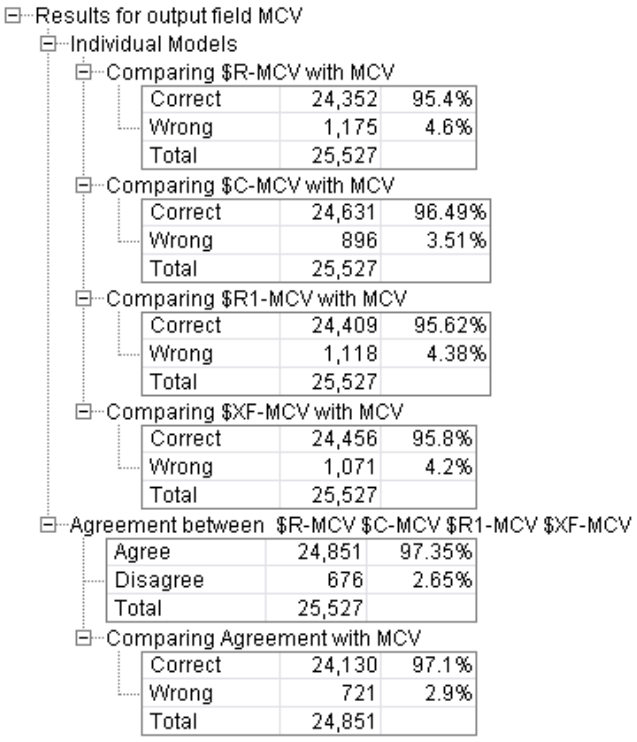

Figura 28. Taxas de acerto dos diferentes modelos.

Seguidamente, é apresentada a matriz de confusão do modelo criado (Tabela 26).

Tabela 26. Matriz de confusão – atributo objetivo MCV.

MCV	Anormal	Normal
Anormal	1940	569
Normal	327	22691

Utilizando o modelo com maior taxa de acerto (Figura 27), serão apresentados perfis de indivíduos que possuem ou podem vir a contrair doenças hepáticas, como por exemplo um tipo de anemia (Tabela 27).

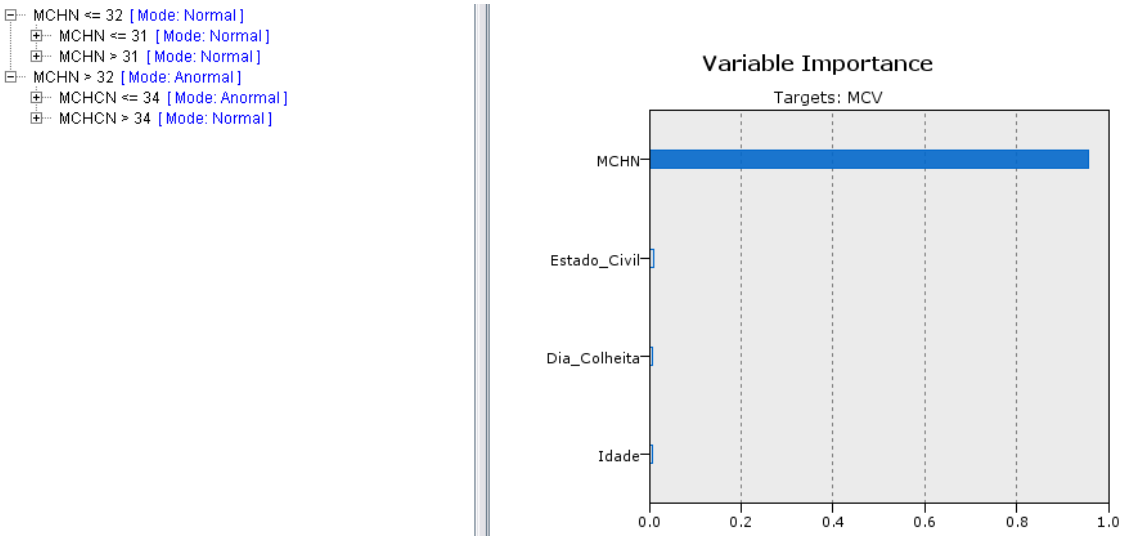

Figura 29. Apresentação inicial do modelo criado com base no algoritmo C5.0 e com o atributo objetivo MCV.

O modelo criado pode ser visualizado mais pormenorizadamente na secção dos anexos (Anexo C – Figura 60).

Tabela 27. Perfis de indivíduos com valores anormais de MCV.

Perfil 1	Valor de hemoglobina contida por glóbulo vermelho superior a 32 picogramas
	Concentração média de hemoglobina por glóbulo vermelho inferior a 35 gramas por decilitro
	Valor Eritrócitos inferior a 5 unidades
Perfil 2	Valor de hemoglobina contida por glóbulo vermelho superior a 32 picogramas
	Concentração média de hemoglobina por glóbulo vermelho inferior a 35 gramas por decilitro
	Valor Eritrócitos superior a 4 unidades
	Quantidade de glóbulos vermelhos no volume total de sangue superior a 43%
Perfil 3	Valor de hemoglobina contida por glóbulo vermelho superior a 32 picogramas
	Concentração média de hemoglobina por glóbulo vermelho inferior a 35 gramas por decilitro
	Valor Eritrócitos inferior a 5 unidades
	Quantidade de glóbulos vermelhos no volume total de sangue inferior a 43%
	Sexo masculino
	Idade inferior a 53 anos
	Número de plaquetas superior a 187

5.1.5 Quais os perfis dos dadores com uma concentração de hemoglobina globular média por glóbulo vermelho fora dos valores normais?

Para a criação do modelo utilizado foram escolhidos os atributos pretendidos, efetuou-se a seleção do atributo objetivo (MCHC) e foi selecionado o método de avaliação (*Holdout).* A Figura 30 apresenta todo o fluxo utilizado para a criação dos modelos.

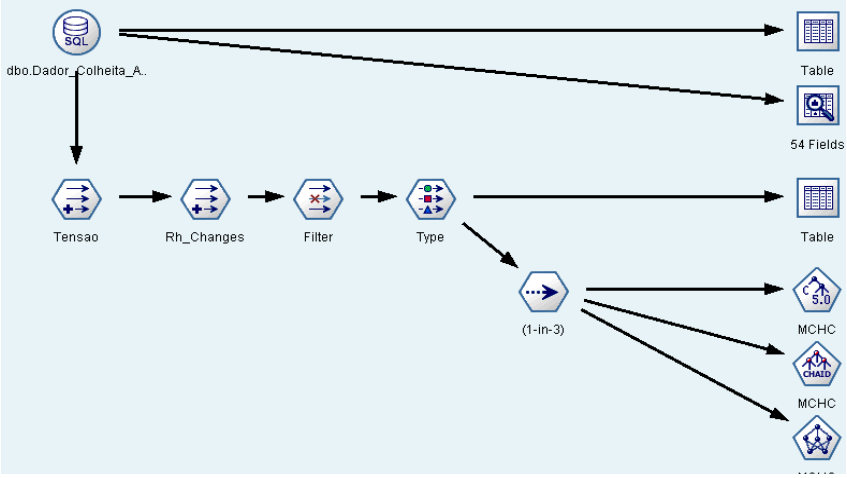

Figura 30. Criação dos modelos – atributo objetivo MCHC.

Após análise dos modelos obtidos, foi concluído que o modelo com maior taxa de acerto foi aquele que foi desenvolvido com base no algoritmo *C5.0* (99.76%) (Figura 31).

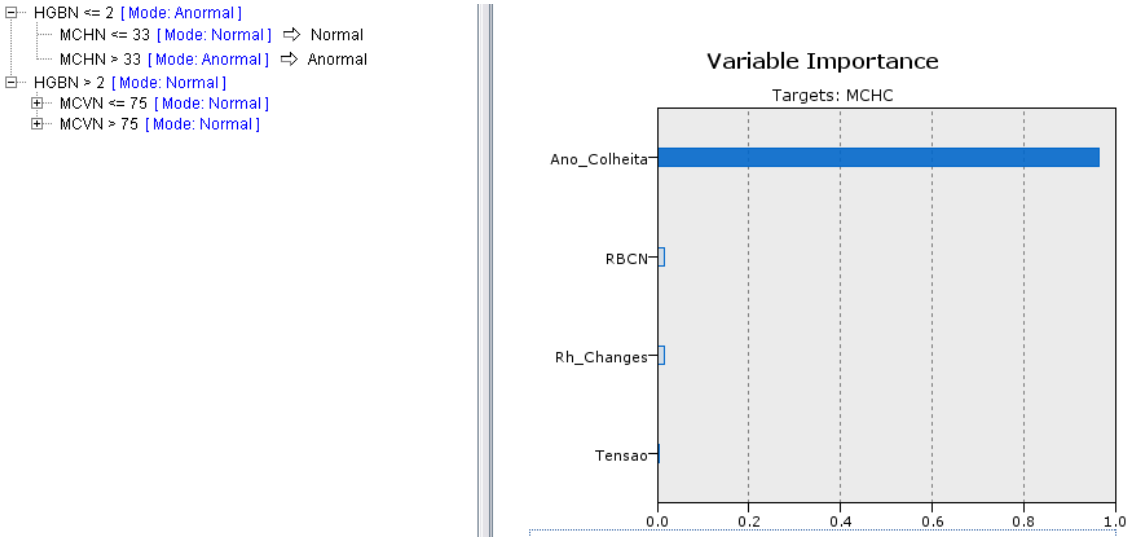

Figura 31. Apresentação inicial do modelo criado com base no algoritmo C5.0 e com o atributo objetivo MCHC.

O modelo criado pode ser visualizado mais pormenorizadamente na secção dos anexos (Anexo C – Figura 61). A Tabela 28 apresenta a matriz de confusão relativa ao modelo.

Tabela 28. Matriz de confusão – atributo objetivo MCHC.

MCHC	Anormal	Normal
Anormal	89	49
Normal	12	25377

Utilizando o modelo com maior taxa de acerto, são apresentados perfis de indivíduos que possuem ou podem vir a apresentar uma quantidade de hemoglobina inadequada por glóbulo vermelho (Tabela 29):

Tabela 29. Perfis de indivíduos com valores anormais de MCHC.

Perfil 1	Hemoglobina inferior a 3 gramas por decilitro
	Valor de hemoglobina contida por glóbulo vermelho superior a 33 picogramas
Perfil 2	Hemoglobina superior a 2 gramas por decilitro
	Volume médio dos glóbulos vermelhos do sangue superior a 76 fentolitros
	Valor de hemoglobina contida por glóbulo vermelho inferior a 24 picogramas
	Valor dos neutrófilos inferior a 54%

5.1.6 Quais os perfis dos indivíduos dadores que têm ou podem vir a contrair policitemia, hemólise ou leucemia?

Após a escolha dos atributos e da seleção do atributo objetivo (HCT), foi efetuado um balanceamento dos dados com o fator 0.01 para o atributo HCT com o valor igual a "Normal". A Figura 32 apresenta todo o fluxo da criação dos modelos.

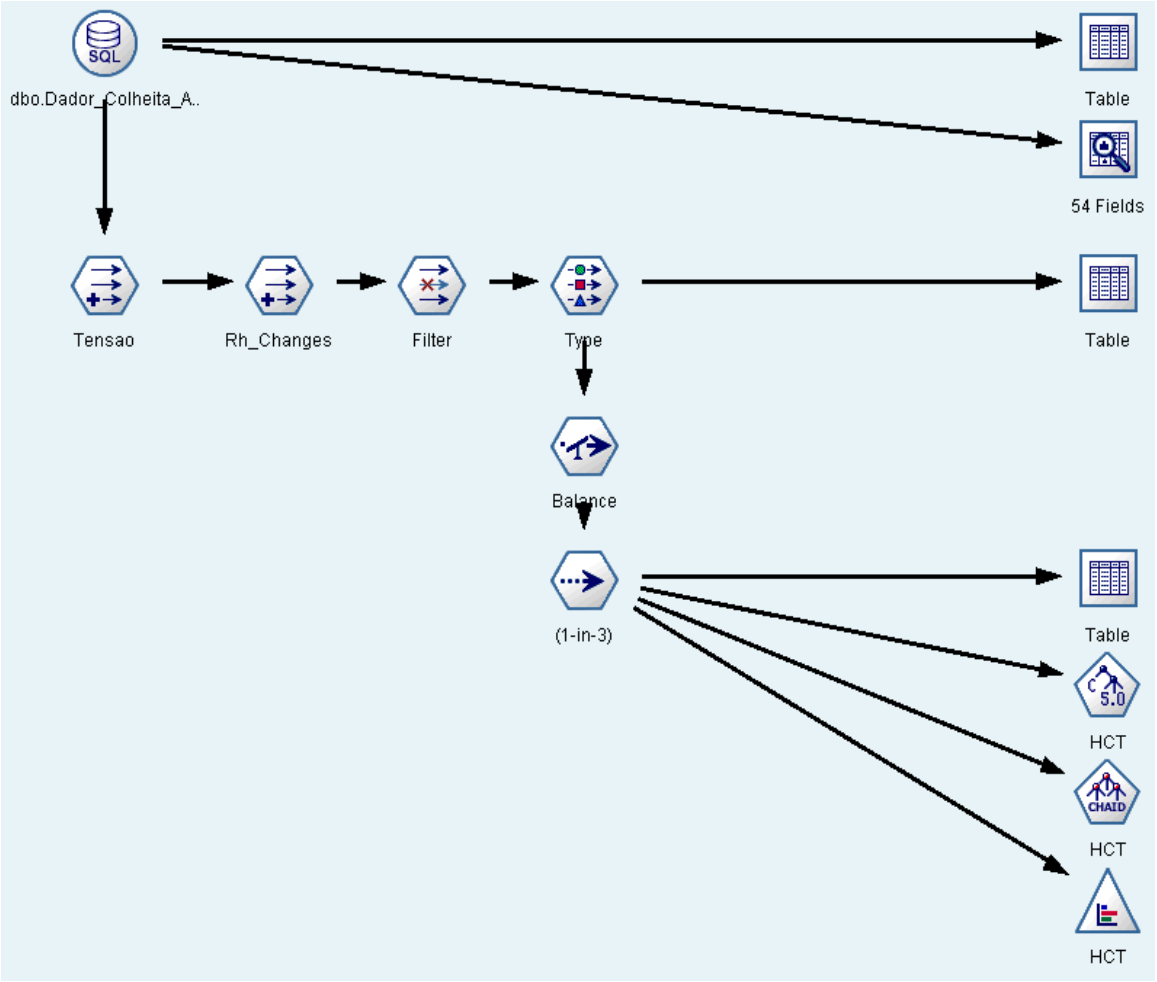

Figura 32. Criação dos modelos – atributo objetivo HCT.

Após a criação dos modelos (algoritmo *CHAID* e *C5.0*), foi efetuada a avaliação dos mesmos para validar os resultados obtidos. A taxa de acerto do modelo criado pelo algoritmo *C5.0* é de 96.56% e a taxa de acerto do modelo criado pelo algoritmo *CHAID* é de 95.44% (Figuras 33 e 34).

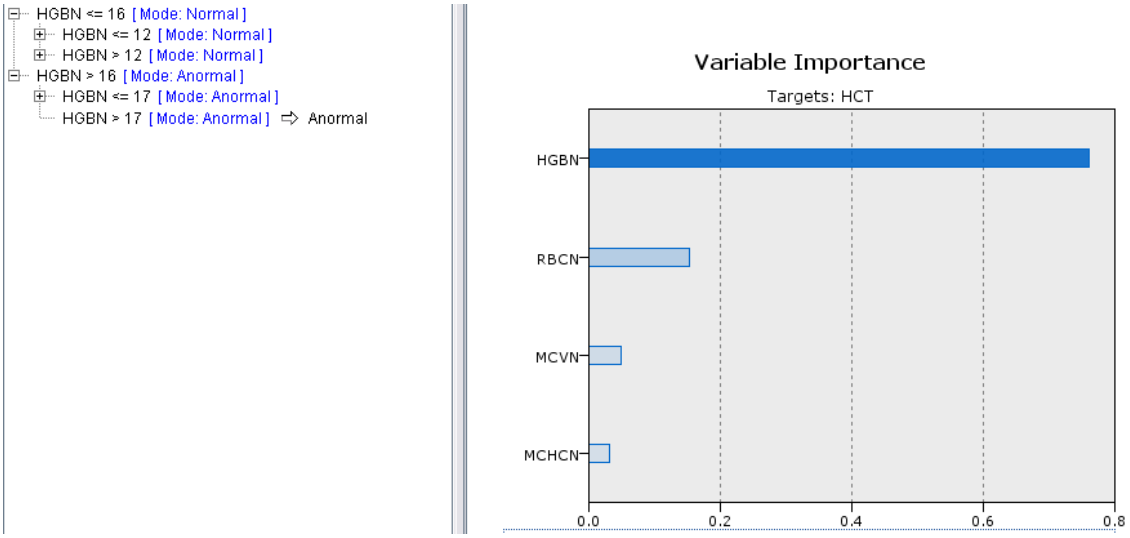

Figura 33. Apresentação inicial do modelo criado com base no algoritmo C5.0 e com o atributo objetivo HCT.

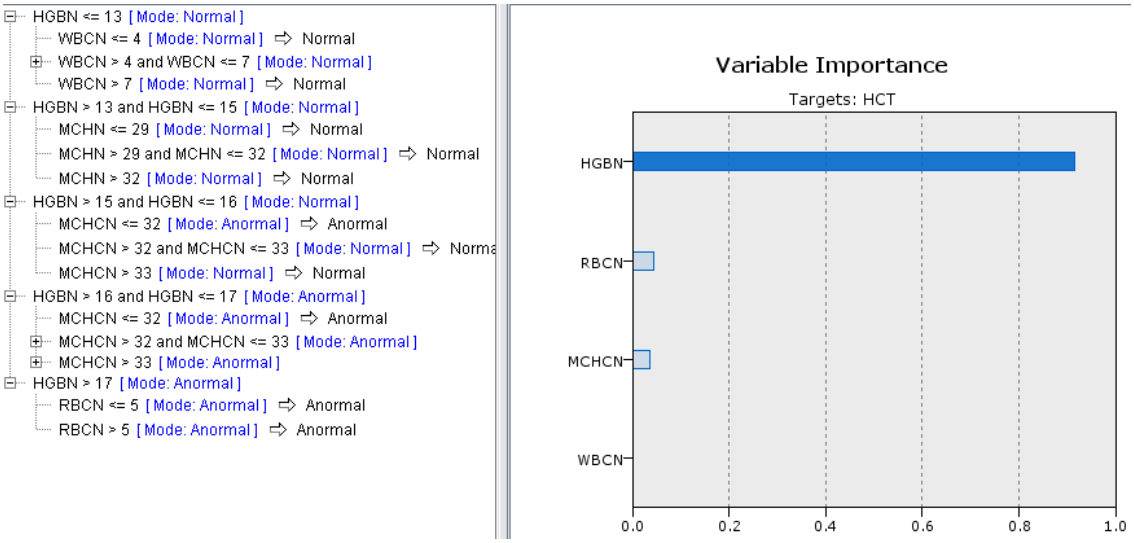

Figura 34. Apresentação inicial do modelo criado com base no algoritmo CHAID e com o atributo objetivo HCT.

Os modelos criados podem ser visualizados mais pormenorizadamente na secção dos anexos (Anexo C – Figuras 62 e 63). As Tabela 30 e 31 apresentam a matriz de confusão para cada modelo.

Tabela 30. Matriz de confusão de acordo com o algoritmo C5.0 – atributo objetivo HCT.

HCT	Anormal	Normal
Anormal	863	28
Normal	851	23785

Podem ser retiradas as mesmas conclusões da Tabela 30 para a Tabela 31, considerando apenas que o número de dádivas é diferente para cada célula da tabela.

Tabela 31. Matriz de confusão de acordo com o algoritmo CHAID – atributo objetivo HCT.

HCT	Anormal	Normal
Anormal	836	55
Normal	1110	23526

De acordo com os modelos criados com uma taxa de acerto superior a 80%, podem ser apresentados diferentes perfis de indivíduos com possibilidade de virem a contrair doenças como a policitemia, hemólise ou leucemia. As seguintes tabelas (Tabela 32 e 33) apresentam, respetivamente, os perfis criados pelo modelo criado pelo algoritmo *C5.0* e pelo modelo criado pelo algoritmo *CHAID*:

Tabela 32. Perfis de indivíduos com valores anormais de HCT – algoritmo C5.0.

Perfil 1	Valor da hemoglobina superior a 16 gramas por decilitro
Perfil 2	Valor de hemoglobina igual a 17 gramas por decilitro
	Concentração média de hemoglobina por glóbulo vermelho superior a 33 gramas por decilitro
	Quantidade de eritrócitos inferior a 6 unidades
	Volume médio dos glóbulos vermelhos do sangue superior a 92 fentolitros

Tabela 33. Perfis de indivíduos com valores anormais de HCT – algoritmo CHAID.

Perfil 1	Valor de hemoglobina igual a 17 gramas por decilitro
	Concentração média de hemoglobina por glóbulo vermelho superior a 33 gramas por decilitro
	Quantidade de eritrócitos inferior a 6 unidades
	Volume médio dos glóbulos vermelhos do sangue superior a 92 fentolitros
Perfil 2	Valor de hemoglobina superior a 17 gramas por decilitro
	Quantidade de eritrócitos inferior a 6 unidades

5.1.7 Quais os perfis dos indivíduos dadores que têm ou podem vir a contrair problemas com o funcionamento da medula óssea, falência renal ou úlcera gástrica?

Para responder ao objetivo delineado, foi elaborado todo o processo de criação de modelos para obter padrões inerentes nos dados. Inicialmente, foi selecionado o atributo HGB como atributo objetivo, foram escolhidos os atributos e foi feito o balanceamento do atributo HGB (Figura 35). Seguidamente, foram utilizados os algoritmos *CHAID* e *C5.0* para a criação dos modelos.

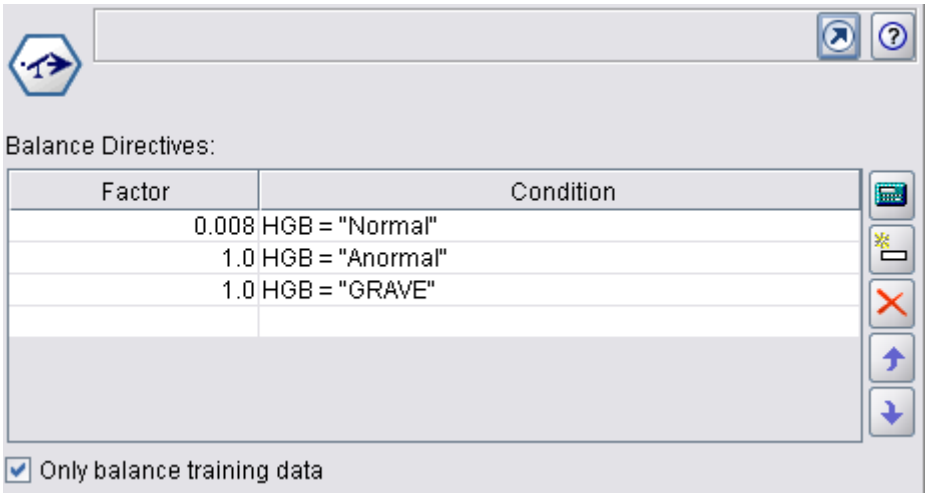

Figura 35. Balanceamento do atributo HGB.

Seguidamente, os modelos foram analisados (Figuras 36 e 37) e as taxas de acerto para ambos os modelos criados (algoritmo *C5.0* e *CHAID*) foram, respetivamente, de 99.08% e 97.21%.

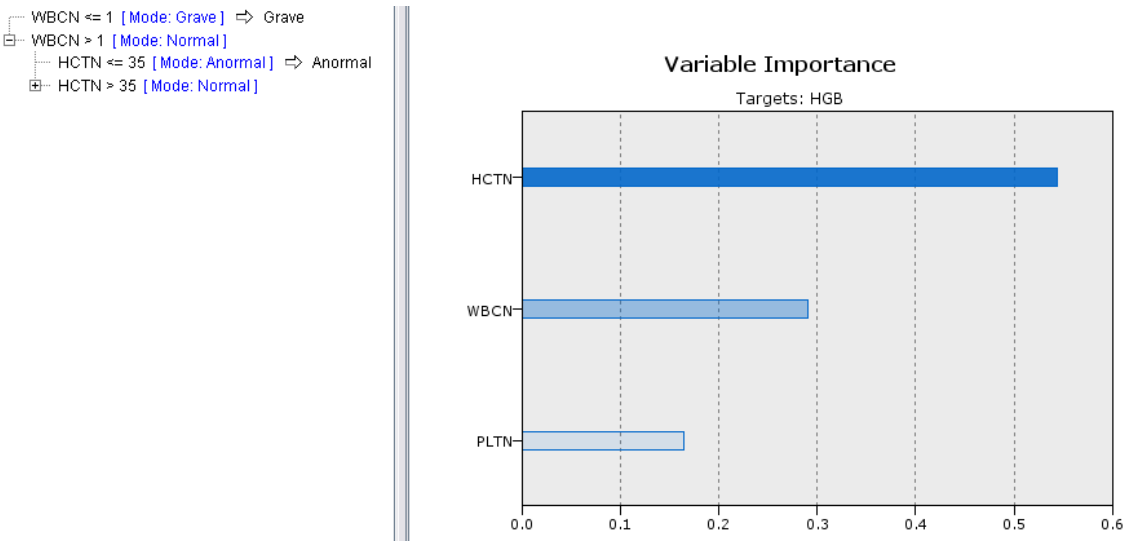

Figura 36. Apresentação inicial do modelo criado com base no algoritmo C5.0 e com o atributo objetivo HGB.

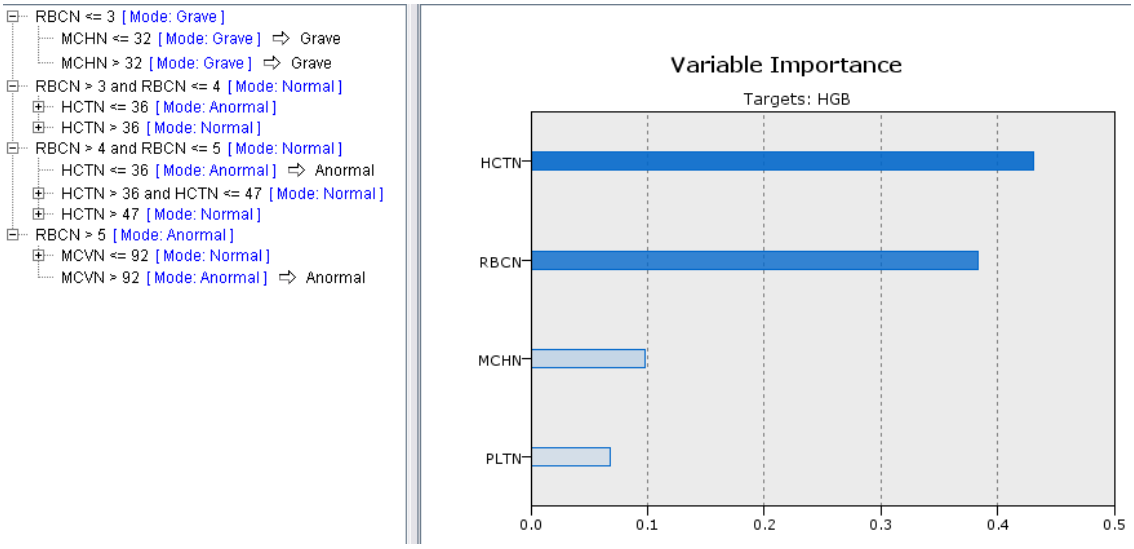

Figura 37. Apresentação inicial do modelo criado com base no algoritmo CHAID e com o atributo objetivo HGB.

Os modelos criados podem ser visualizados mais pormenorizadamente na secção dos anexos (Anexo C – Figuras 64 e 65). As Tabelas 34 e 35 apresentam a matriz de confusão para cada um dos modelos.

Tabela 34. Matriz de confusão de acordo com o algoritmo CHAID – atributo objetivo HGB.

HGB	Anormal	Grave	Normal
Anormal	38	1	1
Grave	0	10	1
Normal	193	39	25244

Podem ser retiradas as mesmas conclusões da Tabela 34 para a Tabela 35, considerando apenas que o número de dádivas é diferente para cada célula da tabela.

Tabela 35. Matriz de confusão de acordo com o algoritmo C5.0 – atributo objetivo HGB.

HGB	Anormal	Grave	Normal
Anormal	30	3	7
Grave	0	8	3
Normal	280	418	24778

Após analisar as árvores criadas, foram selecionados perfis de indivíduos que possuem ou poderão contrair determinadas doenças, tais como, a medula óssea, falência renal ou úlcera gástrica (Tabela 36 e Tabela 37):

Tabela 36. Perfis de indivíduos com valores anormais de HGB – algoritmo C5.0.

Perfil 1	Valor dos leucócitos inferior a 2 unidades
Perfil 2	Valor dos leucócitos superior a 1 unidade
	Quantidade de glóbulos vermelhos inferior a 37%

Tabela 37. Perfis de indivíduos com valores anormais de HGB – algoritmo CHAID.

Perfil 1	Valor de Eritrócitos inferior a 4 unidades
Perfil 2	Valor de Eritrócitos superior a 5 unidades
	Volume corpuscular médio dos glóbulos vermelhos do sangue superior a 92 fentolitros

5.2 Segmentação

O método de segmentação permite efetuar uma partição de um conjunto de dados em subclasses mais pequenas com caraterísticas comuns [Han, 2015]. A análise de segmentos é uma aprendizagem não supervisionada, ou seja, não existem classes pré-definidas. Este método foi utilizado quando a técnica de classificação, por si só, não conseguiu obter bons resultados e para descobrir grupos com caraterísticas comuns. Seguidamente, são apresentados os cenários onde foi necessário criar modelos com base na técnica de segmentação.

5.2.1 Quais os perfis dos indivíduos dadores que têm ou têm maior probabilidade de virem a contrair leucemia mieloide crónica, leucemia aplástica ou cirrose?

Para responder a este objetivo foi necessária a utilização da técnica de segmentação. Após a escolha dos atributos e da seleção do atributo objetivo (NEU) foram criados os modelos de acordo com o algoritmo *C5.0* e *CHAID*. Após a avaliação dos mesmos, as taxas de acerto eram, respetivamente, de 58.69% e 57.88%. Como tal, foi utilizada a técnica de segmentação. A Figura 38 apresenta a criação dos modelos com base nos algoritmos *TwoStep, K-means* e *Kohonen*.

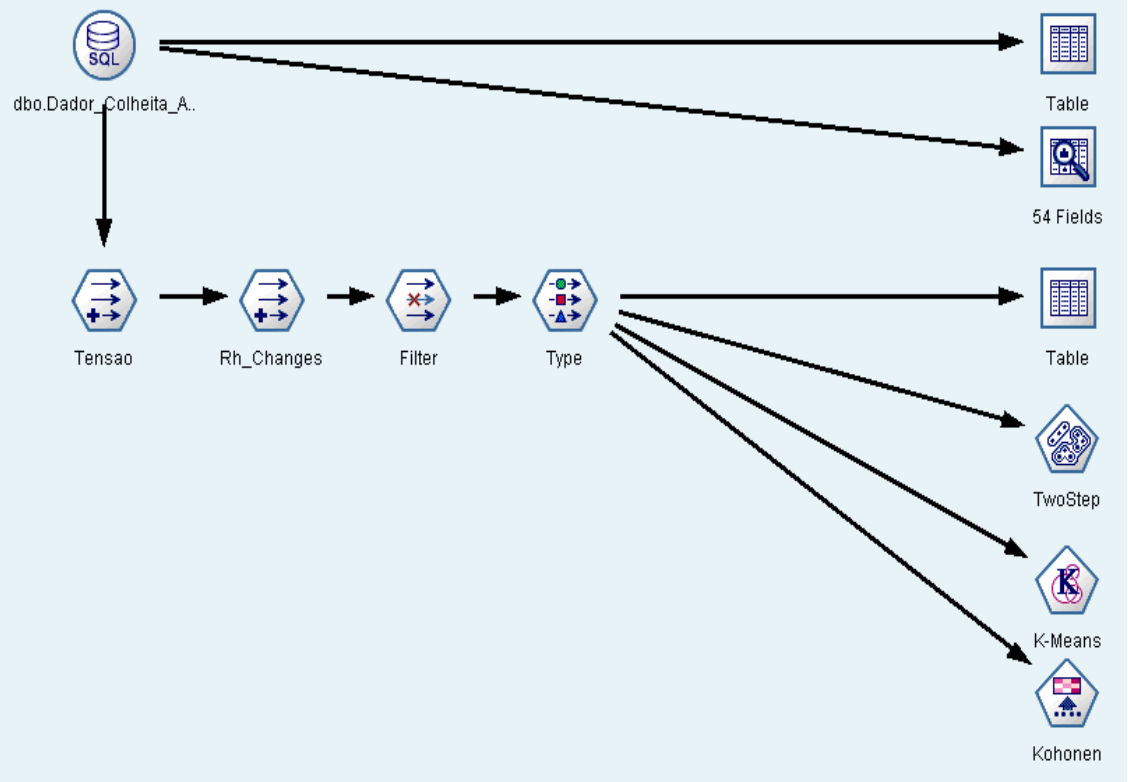

Figura 38. Segmentação.

Seguidamente, foi criada uma nova *stream* para criar os modelos de acordo com os segmentos criados. Para tal, foi utilizado o algoritmo *C5.0* e o método de avaliação para cada um dos modelos denomina-se *Cross-validate* (Figura 39). Importa salientar que o algoritmo que deu melhores resultados foi o *TwoStep.*

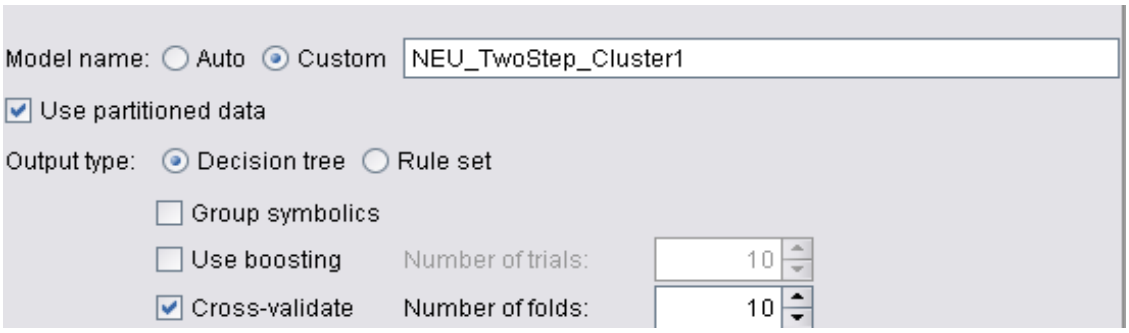

Figura 39. Método de avaliação utilizado.

A Figura 40 ilustra todo o processo de criação dos modelos de todos os segmentos criados pelo algoritmo *TwoStep*.

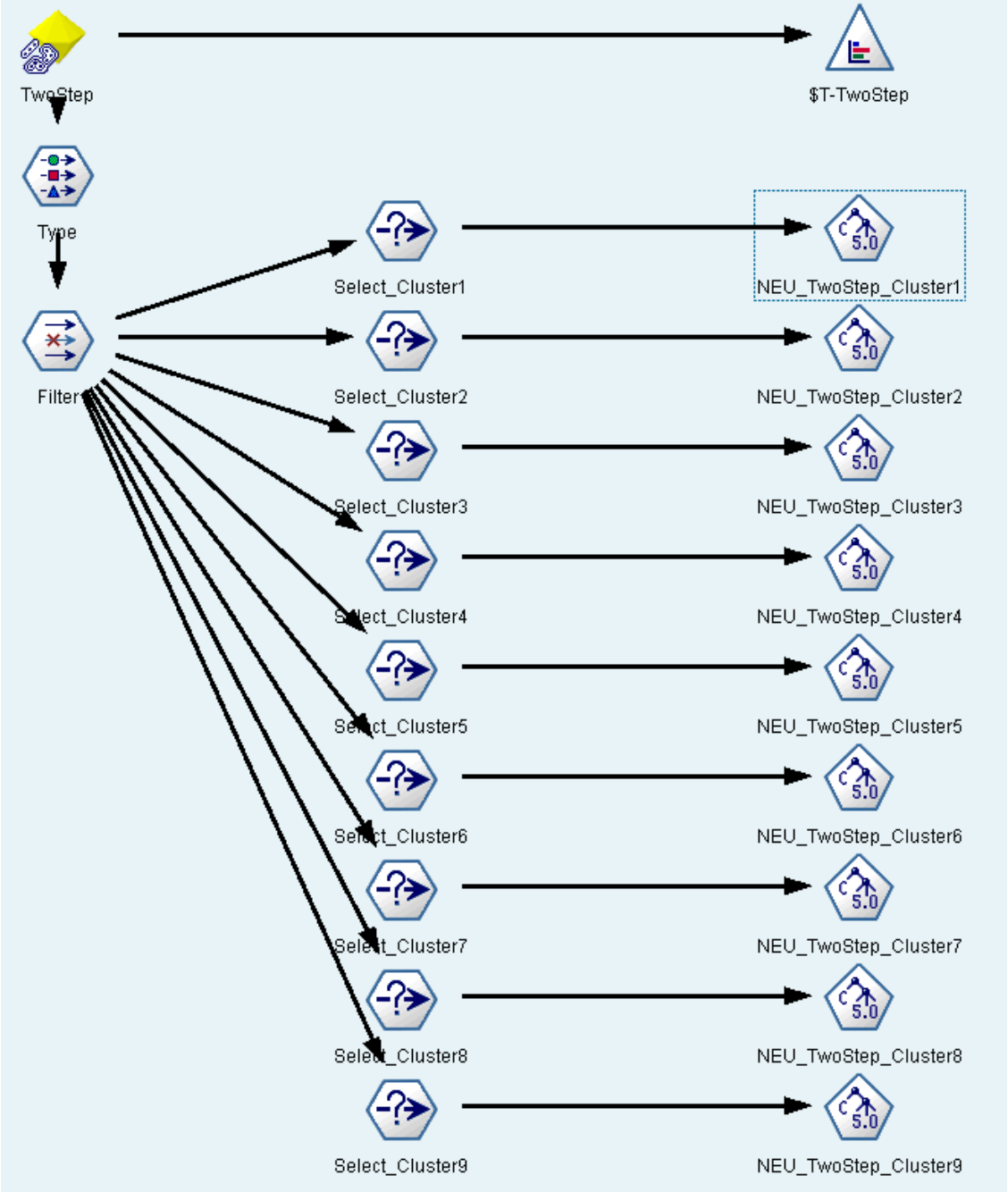

Figura 40. Criação dos modelos – atributo objetivo NEU.

A Figura 41 apresenta todos os nós utilizados para efetuar a avaliação dos modelos criados.

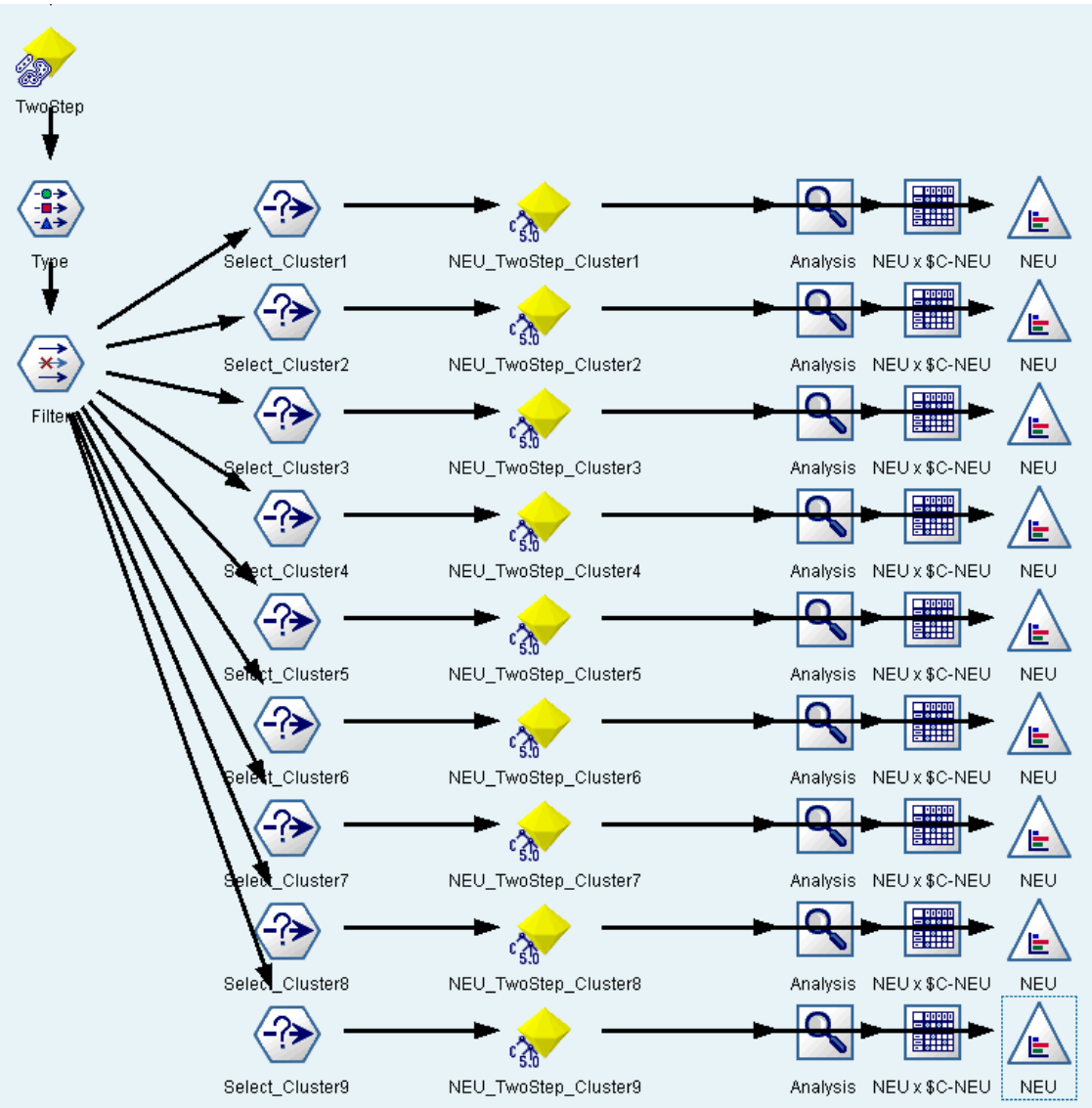

Figura 41. Avaliação dos modelos – atributo objetivo NEU.

Após a utilização da técnica de segmentação, verifica-se um aumento significativo nas taxas de acerto dos modelos (Tabela 38).

Tabela 38. Taxa de acerto para cada modelo.

Segmento	Taxa de acerto
1	77.51%
2	77.65%
3	78.14%
4	75.89%
5	100%
6	76.73%
7	100%
8	100%
9	99.92%

A figura seguinte (Figura 42) ilustra todos os segmentos e a quantidade de neutrófilos (normal ou anormal). Verifica-se que os segmentos 7 e 8 apenas contêm valores anormais de eritrócitos.

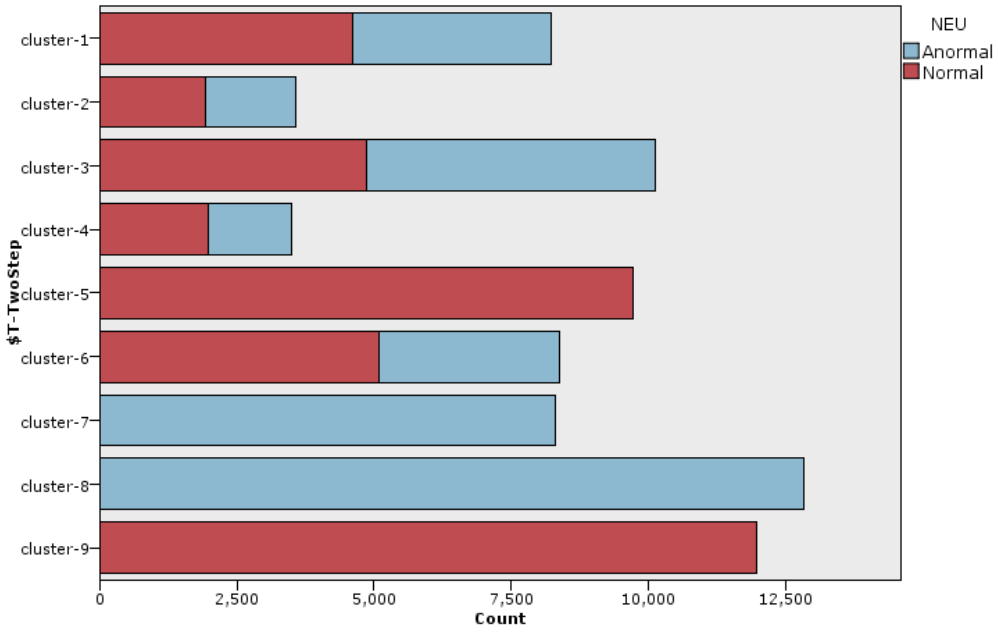

Figura 42. Distribuição do valor dos eritrócitos pelos segmentos.

Após analisar as árvores criadas para cada segmento, foram selecionados perfis de indivíduos que possuem ou poderão contrair certas doenças, tais como, a leucemia mieloide crónica, leucemia aplástica ou cirrose):

Tabela 39. Perfis de indivíduos com valores anormais de percentagem de neutrófilos - Segmento 1.

Perfil 1	Valor dos leucócitos inferior a 6 unidades por litro
	Volume médio dos glóbulos vermelhos no sangue inferior a 92 fentolitros
	Colheita efetuada no mês de Abril
	Valor da alanina aminotransferase inferior a 21 unidades por litro
Perfil 2	Valor dos leucócitos inferior a 6 unidades por litro
	Volume médio dos glóbulos vermelhos no sangue superior a 92 fentolitros
	Estado civil do dador divorciado ou solteiro

Tabela 40. Perfis de indivíduos com valores anormais de NEU% - Segmento 2.

Perfil 1	Dádiva do tipo sangue total
	Valor inferior a 8 leucócitos por centímetro cúbico de sangue
	Valor de hemoglobina superior a 12 gramas por decilitro
	Dadores do sexo masculino
	Rhesus positivo

Tabela 41. Perfis de indivíduos com valores anormais de NEU% - Segmento 3.

Perfil 1	Valor inferior a 7 leucócitos por centímetro cúbico de sangue
	Valor da alanina aminotransferase inferior a 36 unidades por litro
	Dádiva efetuada na parte da manhã
	Idade superior a 60 anos
	Tensão baixa ou alta

Tabela 42. Perfis de indivíduos com valores anormais de NEU% - Segmento 4.

Perfil 1	Valor dos glóbulos vermelhos superior a 4 milhões por milímetro cúbico

Tabela 43. Perfis de indivíduos com valores anormais de NEU% - Segmento 6.

Perfil 1	Dadores do sexo masculino
	Grupo sanguíneo O
	Valor dos glóbulos vermelhos inferior a 4 milhões por milímetro cúbico

5.2.2 É possível efetuar uma divisão do conjunto de dados de acordo com as caraterísticas comuns dos indivíduos dadores?

Para encontrar resposta a este cenário foram utilizadas técnicas de segmentação de dados. Inicialmente, foi criado um modelo com base no algoritmo *TwoStep* com todos os atributos do conjunto de dados. Este modelo originou três segmentos. Seguidamente, foi colocado o atributo originado pelo algoritmo *TwoStep* (*$T-Clusters-Group*) como atributo objetivo para a criação de um modelo com base no algoritmo C5.0 (Figura 43).

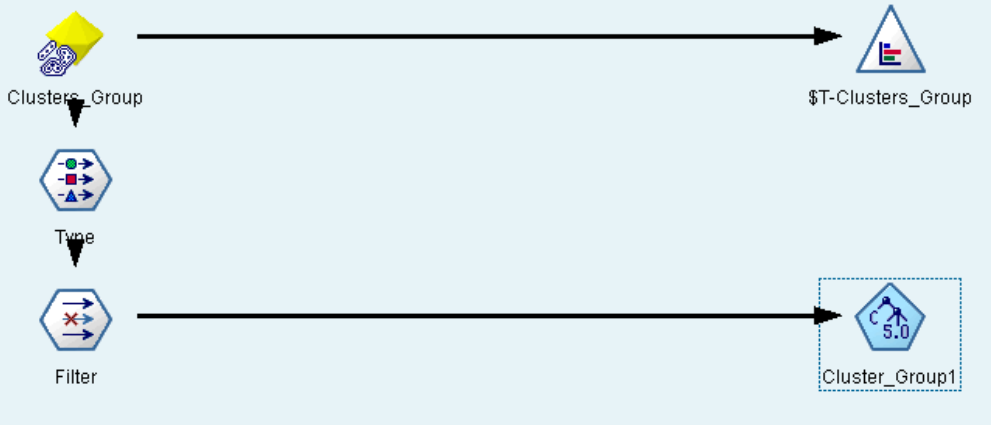

Figura 43. Criação do modelo com o atributo objetivo *$T-Clusters-Group*.

Após a criação do modelo (algoritmo C5.0), foi efetuada a avaliação do mesmo (Figura 44) para validar os resultados obtidos. A taxa de acerto do modelo criado foi de 98.39%.

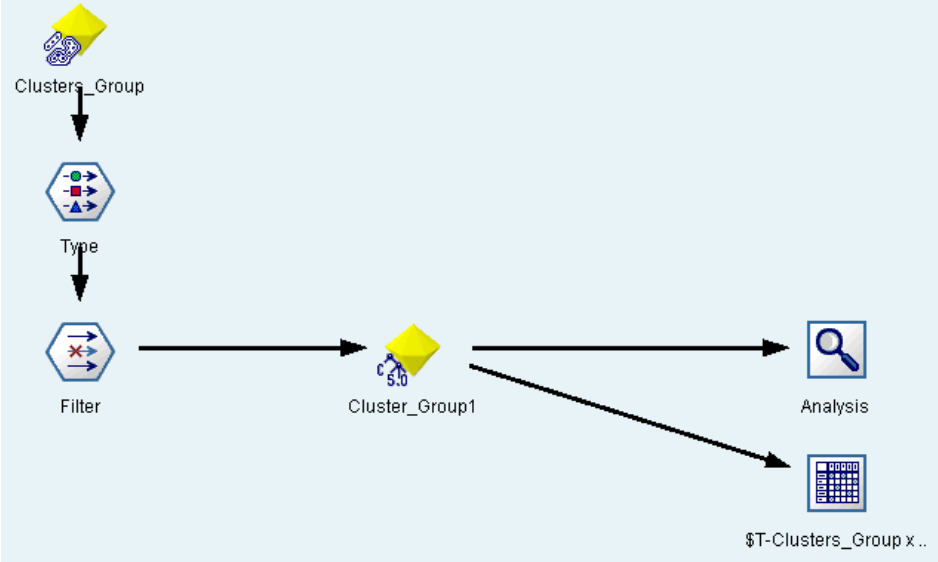

Figura 44. Avaliação do modelo – atributo objetivo *$T-Clusters-Group*.

O conjunto de regras criado pelo modelo (Figura 45) permitiu efetuar uma divisão do conjunto de dados em dois grupos distintos: indivíduos saudáveis e indivíduos não saudáveis. O terceiro grupo (segmento 2) não permite tirar qualquer tipo de conclusões.

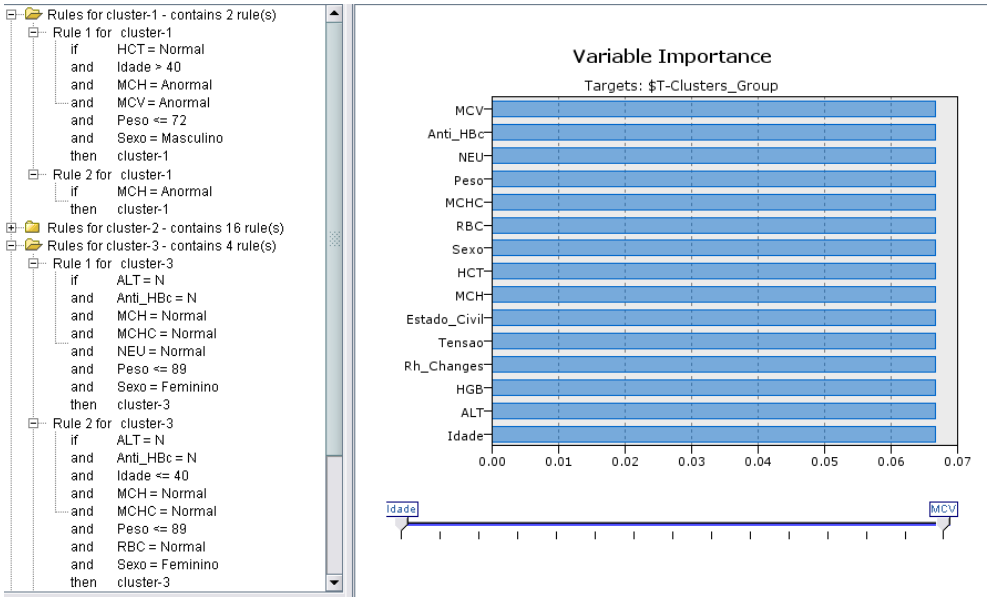

Figura 45. Modelo criado (C5.0) com o *$T-Clusters-Group* como atributo objetivo.

De acordo com a Figura 45, pode ser verificado que o segmento 1 criou regras que demonstram um possível grupo não saudável e o segmento 3 criou regras que apresentam um possível grupo saudável (Tabela 44 e 45).

Tabela 44. Conjunto de regras para o grupo dos indivíduos saudáveis.

Grupo Saudável	**Regra 1**	ALT com valores normais
		MCH com valores normais
		MCHC com valores normais
		Peso inferior a 90 Kg
		Sexo feminino
	Regra 2	ALT com valores normais
		Idade inferior a 41 anos
		MCH com valores normais
		MCHC com valores normais
		Peso inferior a 90 Kg
		RBC normal
		Sexo feminino

Tabela 45. Conjunto de regras para o grupo dos indivíduos não saudáveis.

Grupo Não Saudável	**Regra 1**	HCT normal
		Idade superior a 40 anos
		MCH com valores anormais
		MCV anormal
		Peso inferior a 73 Kg
		Sexo feminino
	Regra 2	MCH anormal

5.3 Regras de Associação

As regras de associação pretendem representar a probabilidade de um conjunto de itens ocorra numa transação da presença de outro conjunto [Vasconcelos e Carvalho, 2004]. A tarefa de associação tem como premissa básica descobrir relacionamentos ou padrões frequentes entre conjuntos de dados. Seguidamente, através de técnicas de associação, serão demonstrados os modelos criados que permitiram responder aos cenários traçados.

5.3.1 Quais os valores anormais que geralmente se encontram associados num boletim analítico?

Para a solução deste problema foi necessária a utilização de regras de associação. Inicialmente, foi criado um ficheiro com todos os valores anormais por dador. O conteúdo deste ficheiro foi utilizado pelo nó "SetToFlag" para criar outro ficheiro que será utilizado pelos algoritmos de associação. O conteúdo do ficheiro é demonstrado na Figura 46.

	Dador	ALT	Anti_HBc	HCT	HGB	MCH	MCHC	MCV	NEU	PLT	RBC	WBC
1	80338	F	F	F	F	F	F	F	T	F	F	F
2	20381	F	F	F	F	F	F	F	T	F	F	F
3	6570	F	F	F	F	F	F	F	T	F	F	F
4	58678	F	F	F	F	F	F	F	T	F	F	F
5	11075	F	F	F	F	F	F	F	T	F	F	F
6	52412	F	F	F	F	F	F	F	T	F	F	F
7	83882	F	F	F	F	F	F	F	T	F	F	F
8	79820	F	F	F	F	F	F	F	T	F	F	F
9	72100	F	F	F	F	F	F	F	T	F	F	F
10	80063	F	F	F	F	F	F	F	T	F	F	F
11	81217	F	F	F	F	F	F	F	T	F	F	F
12	49904	F	F	F	F	F	F	F	T	F	F	F
13	79302	F	F	F	F	F	F	F	T	F	F	F
14	50783	F	F	F	F	F	F	F	T	F	F	F
15	49043	F	F	F	F	F	F	F	T	F	F	F
16	14058	F	F	F	F	F	F	F	T	F	F	F
17	59207	F	F	F	F	F	F	F	T	F	F	F
18	9303	F	F	F	F	F	F	F	T	F	F	F
19	39276	F	F	F	F	F	F	F	T	F	F	F
20	46982	F	F	F	F	F	F	F	T	F	F	F
21	79084	F	F	F	F	F	F	F	T	F	F	F

Figura 46. Ficheiro criado pelo nó "SetToFlag".

Este ficheiro demonstra quem tem valores normais "F" e quem tem valores anormais "T" para todos os parâmetros acima demonstrados. A Figura 47 ilustra a força dos valores anormais que geralmente se encontram associados.

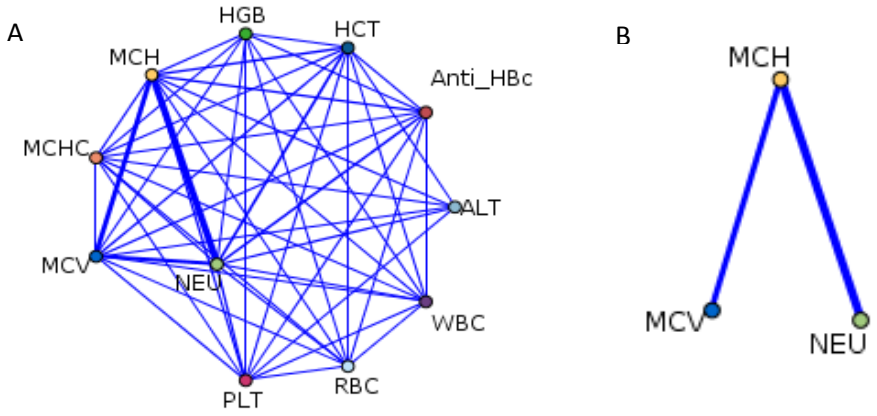

Figura 47. Nó "Web". A) Vista geral; B) Vista ao pormenor.

Após a criação dos modelos com os algoritmos *Apriori, Carma e GRI* foram analisadas as regras de acordo com o suporte e a confiança (Tabela 46):

Tabela 46. Regras de Associação.

Consequente	Antecedente	Confiança (%)	Suporte (%)
MCH	MCV, NEU	97.735	10.721
Com base nesta informação pode ser concluído que 10.721% das dádivas têm valores anormais no volume médio dos glóbulos vermelhos no sangue (MCV), neutrófilos (NEU) e hemoglobina contida por glóbulo vermelho (MCH). Pode ser também concluído que em 97.735% dos casos que tiverem valores anormais de MCV e NEU terão, consequentemente, valores anormais de MCH.			
MCH	MCV	97.619	16.317
Com base nesta informação pode ser concluído que 16.317% das dádivas têm valores anormais no volume médio dos glóbulos vermelhos no sangue (MCV) e valores anormais de hemoglobina contida por glóbulo vermelho (MCH). Pode ser também concluído que em 97.619% dos casos que tiverem valores anormais de MCV terão, consequentemente, valores anormais de MCH.			
NEU	MCH	63.14	43.47
Com base nesta informação pode ser concluído que 43.47% das dádivas têm valores anormais no volume médio dos glóbulos vermelhos no sangue (MCH) e neutrófilos (NEU). Pode ser também concluído que em 63.14% dos casos que tiverem valores anormais de MCH terão, consequentemente, valores anormais de NEU.			

5.3.2 Quais os parâmetros de um boletim analítico que podem levar a que outros parâmetros se tornem anormais?

Para responder a este cenário, foi utilizado um algoritmo de criação de regras de associação sequencial. Inicialmente, foi criado um ficheiro com a identificação do dador, os valores anormais de cada dador e a data da realização da dádiva de sangue. A Figura 48 apresenta um excerto do ficheiro.

	Dador	ALT	Data
1	17	NEU	2000-12-01
2	17	NEU	2001-05-01
3	17	NEU	2001-11-01
4	20	MCV	2002-05-01
5	20	MCH	2002-05-01
6	20	NEU	2002-05-01
7	21	MCH	2001-11-01
8	23	MCH	2001-11-01
9	24	MCH	2000-12-01
10	24	NEU	2000-12-01
11	24	MCH	2001-11-01
12	24	NEU	2001-11-01
13	25	MCH	2000-12-01
14	25	MCV	2000-12-01
15	25	NEU	2000-12-01
16	25	NEU	2001-05-01
17	25	MCH	2001-05-01
18	25	MCV	2002-05-01
19	25	MCH	2002-05-01
20	25	NEU	2002-05-01
21	27	NEU	2001-05-01

Figura 48. Ficheiro com a identificação do dador, doença e data da dádiva efetuada.

Seguidamente, foi utilizado este ficheiro para criar o modelo. A Figura 49 apresenta as configurações efetuadas para criação do mesmo.

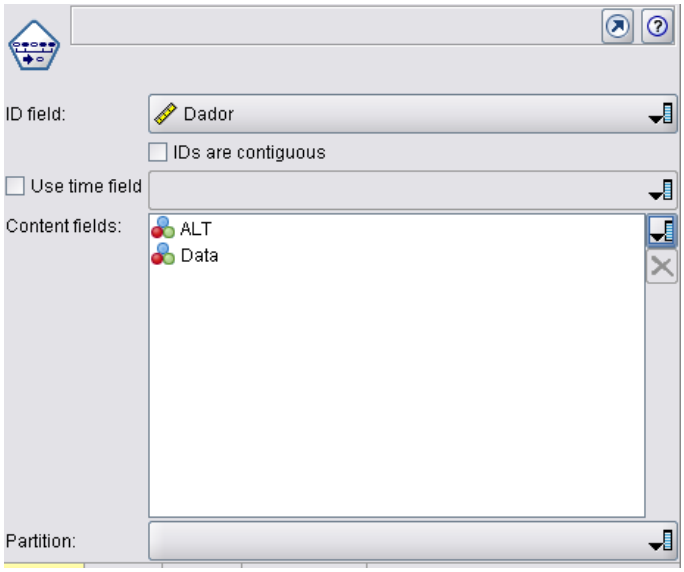

Figura 49. Algoritmo de sequência.

Tabela 47. Resultado do algoritmo de sequência.

Antecedente	Consequente	Suporte (%)	Confidência (%)
NEU	MCH	79.447	26.368

A Tabela 47 apresenta o resultado do modelo. Com base nestes resultados pode ser concluído que 79.447% das dádivas têm valores anormais neutrófilos (NEU) e valores anormais de hemoglobina contida por glóbulo vermelho (MCH). Pode ser também concluído que em 26.368% dos casos que tiverem valores anormais de NEU terão, futuramente, valores anormais de MCH.

6 Conclusões

Após todo o trabalho de análise realizado no âmbito desta dissertação e do estudo realizado na área de mineração de dados, surgiram algumas conclusões de carácter objetivo.

O volume de dados produzido duplica a cada dois anos, sendo que 90% dos mesmos são dados não estruturados. Dado que mais informação não implica mais conhecimento, foram criadas novas técnicas capazes de trabalhar com uma elevada quantidade de dados com o objetivo de procurar extrair o que de relevante se encontra implícito nos dados.

O processo de mineração de dados é constituído por um conjunto de técnicas e ferramentas que através de algoritmos, permite explorar grandes quantidades de dados com o intuito de descobrir padrões e correlações, que poderão ser úteis para inúmeras situações.

6.1 Objetivos Alcançados

Após a simplificação da base de dados, foi efetuada a preparação dos dados. Esta fase demorou mais de 80% do processo de mineração dos dados. Existiam elevados problemas nos dados (inconsistência, desbalanceamento, ruído) que tiveram de ser contornados. Seguidamente, analisaram-se os dados e foram criados os modelos que permitiram responder aos objetivos/cenários traçados com as técnicas e os algoritmos necessários. Após vários testes, concluiu-se que era necessária a utilização de quase todos os atributos existentes. Depois da avaliação dos modelos criados, foi extraído o conhecimento necessário para responder aos objetivos/cenários pretendidos.

As regras criadas pelos modelos envolveram muitos atributos de acordo as árvores de decisão produzidas pelos modelos. Os modelos criados tiveram taxas de acerto superiores a 80%, ou seja, foi obtido conhecimento que poderá ser utilizado e analisado futuramente. Quando o valor de 80% não foi conseguido na técnica de classificação, utilizou-se a técnica de segmentação que aumentou consideravelmente a taxa de acerto. Podemos verificar que os

atributos mais importantes para a construção das árvores de decisão foram, essencialmente, o HGB (hemoglobina), MCH (quantidade de hemoglobina nos glóbulos vermelhos) e ALT (alanina aminotransferase).

As regras de associação permitiram saber quais os elementos comuns num determinado conjunto de dados, ou seja, quais os elementos/doenças comuns nos dadores. Os valores de suporte e confiança foram bastante aceitáveis. Por exemplo, verificou-se que 10.721% das dádivas têm valores anormais no volume médio dos glóbulos vermelhos no sangue (MCV), neutrófilos (NEU) e valores anormais de hemoglobina contida por glóbulo vermelho (MCH). Verificou-se também que em 97.735% dos casos que tiverem valores anormais de MCV e NEU terão, consequentemente, valores anormais de MCH.

Foi também utilizado o algoritmo de sequência para apresentar quais os parâmetros com valores anormais que podem desencadear com que outros parâmetros se tornem também anormais. Contudo, os resultados do suporte e da confiança foram relativamente baixos.

Para concluir, podemos dizer que os modelos criados responderam com boas taxas de acerto aos cenários propostos e permitiram retirar conclusões que poderão ser ou não importantes na área da saúde.

6.2 Trabalho Futuro

No âmbito desta dissertação, este processo pode tornar-se útil na melhoria dos serviços de saúde, como por exemplo, evitar determinadas doenças ou problemas de saúde. Para tal, será necessário apresentar os resultados a um profissional de saúde para que seja efetuada uma avaliação aos mesmos para concluir da respetiva importância.

Como trabalho futuro, seria proveitosa a junção de novas bases de dados com informação de novos parâmetros/testes sanguíneos ou com informação hospitalar dos pacientes para analisar e obter padrões mais consolidados e potencialmente úteis.

Em síntese, o presente trabalho foi uma boa experiência para desenvolver novas capacidades de análise e espírito crítico, acerca das técnicas e algoritmos que devem ser utilizados para que os processos aqui abordados consigam obter conhecimento útil e que possa ser utilizado para ajudar a melhorar a ciência e, consequentemente, a qualidade de vida da sociedade.

Referências

ALBANESI.IT. 2015. *Emocromo* [Online]. albanesi.it La voce degli italiani moderni. Disponível em: http://www.albanesi.it/salute/esami/emocromo.htm [Último acesso: 25 de maio de 2015].

ALVARES, L. O. 2010. Mineração de Dados - Análise Exploratória de Dados. Universidade Federal de Ciências da Saúde de Porto Alegre.

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

BARRA, F. 2013. *Redes neurais artificiais* [Online]. Disponível em: http://blogdopetcivil.com/2013/07/05/redes-neurais-artificiais/ [Ùltimo acesso: 13 de Junho de 2015].

BIAL. 2009. *Hipertensão Arterial - Informação para o utente* [Online]. Disponível em: https://www.bial.com/imagem/Caderno%20saude_Hipertensao%20arterial_V2.pdf [Ùltimo acesso: 5 de Setembro de 2015].

BIOCAMPELLO. 2015. *Interpretação de Análises Clínicas* [Online]. Disponível em: http://www.biocampello.com/Interpretacao-das-suas-Analises-Clinicas [Último acesso: 25 de Maio de 2015].

BUENO, M. F. & VIANA, M. R. 2012. Mineração de dados: Aplicações, Eficiência e Usabilidade. *Anais do Congresso de iniciação científica do INATEL - INCITEL 2012.* Brasil.

FUNDAÇÃO PORTUGUESA DE CARDIOLOGIA. 2014. *Tensão e Hipertensão arterial* [Online]. Disponível em: http://www.fpcardiologia.pt/saude-do-coracao/factores-de-risco/hipertensao/ [Ùltimo acesso: 5 de Setembro de 2015].

CARVALHO, D. R., ESCOBAR, L. F. A. & TSUNODA, D. 2014. Pontos de atenção para o uso da mineração de dados na saúde. *Informação & Informação,* 19**,** 249-273.

CDC. 2015. *Interpretation of Hepatitis B Serologic Test Results* [Online]. Disponível em: http://www.cdc.gov/hepatitis/HBV/PDFs/SerologicChartv8.pdf [Ùltimo acesso: 23 de Setembro de 2015].

CENTRO DE CIÊNCIA JÚNIOR. 2013. *Sangue, presente da vida* [Online]. Disponível em: http://www.centrocienciajunior.com/novidades/novidade.asp?id=1270 [Último acesso: 24 de Maio de 2015].

CÔRTES, S., PORCARO, R. M. & LIFSCHITZ, S. 2002. *Mineração de dados - funcionalidades, técnicas e abordagens*, PUC.

DANTAS, E. R. G., JÚNIOR, J. C. A. P., LIMA, D. S. & AZEVEDO, R. R. 2008. O Uso da Descoberta de Conhecimento em Base de Dados para Apoiar a Tomada de Decisões. *SEGeT – Simpósio de Excelência em Gestão e Tecnologia.* Rio de Janeiro.

DE VASCONCELOS, L. M. R. & DE CARVALHO, C. L. 2004. Aplicação de Regras de Associação para Mineração de Dados na Web. Instituto de Informática da Universidade Federal de Goiás.

DEVMEDIA. 2015. *Mineração e Análise de Dados em SQL* [Online]. Disponível em: http://www.devmedia.com.br/mineracao-e-analise-de-dados-em-sql/29337 [Último acesso: 3 de Junho de 2015].

DIREÇÃO-GERAL DE SAÚDE. 2011. Norma nº 020/2011 de 28/09/2011.

SISTEMA GALILEU DE EDUCAÇÃO ESTATÍSTICA. 2015. *Escalas de medida* [Online]. Disponível em: http://www.galileu.esalq.usp.br/mostra_topico.php?cod=92 [Último acesso: 3 de junho de 2015].

FAYYAD, U., PIATETSKY-SHAPIRO, G. & SMYTH, P. 1996. From Data Mining to Knowledge Discovery in Databases. *American Association for Artificial Intelligence***,** 37-54.

FISIOTERAPIA PARA TODOS. 2014. *Hemograma completo em exames de sangue* [Online]. Disponível em: http://www.fisioterapiaparatodos.com/p/exame/hemograma-completo-em-exames-de-sangue/ [Último acesso: 25 de maio de 2015].

FREITAS, J. A. S. 2006. *Uso de Técnicas de Data Mining para Análise de Bases de Dados Hospitalares com Finalidades de Gestão.* Faculdade de Economia da Universidade do Porto.

GAMA, J., CARVALHO, A. P. L., FACELI, K., LORENA, A. C. & OLIVEIRA, M. 2012. *Extração de Conhecimento de Dados - Data mining*, Edições Sílabo.

GONÇALVES, A. R. 2008. *Redes Bayesianas* [Online]. Disponível em: http://www-users.cs.umn.edu/~andre/arquivos/pdfs/bayesianas.pdf [Último acesso: 10 de junho de 2015].

HAN, J. 2015. Cluster Analysis in Data Mining. Coursera Course. Disponível em: https://www.coursera.org/course/clusteranalysis. [Último acesso: 25 de Maio de 2015].

HEMOMINAS, F. 2015. *Critérios gerais de doação - doação de plaquetas* [Online]. Disponível em: http://www.hemominas.mg.gov.br/en/home-2/64-duvidas/doacao-de-sangue/662-criterios-gerais-de-doacao#pode-se-doar-sangue-com-a-pressão-alta-ou-baixa [Ùltimo acesso: 5 de Setembro de 2015].

IBM, 2011. *IBM SPSS Modeler CRISP-DM Guide* [Online]. Disponível em: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/14.2/en/CRISP_DM.pdf/ [Ùltimo acesso: 20 de Junho de 2015].

INFOESCOLA. 2015a. *Plaquetas* [Online]. Disponível em: http://www.infoescola.com/sangue/plaquetas/ [Último acesso: 25 de Maio de 2015]

INFOESCOLA. 2015b. *Plasma Sanguíneo* [Online]. Disponível em: http://www.infoescola.com/sangue/plasma-sanguineo/ [Ùltimo acesso: 25 de Maio de 2015].

INFOESCOLA. 2015c. *Sistema Nervoso* [Online]. Disponível em: http://www.infoescola.com/biologia/sistema-nervoso/ [Último acesso: 23 de Junho de 2015].

INSTITUTO PORTUGUÊS DO SANGUE E DA TRANSPLANTAÇÃO. 2015. *Dê Sangue* [Online]. Disponível em: http://darsangue.pt/ [Último acesso: 24 de Maio de 2015].

LABLUXOR. 2012. *Teste HIV/SIDA* [Online]. Disponível em: http://www.labluxor.com/teste-hiv-sida [Ùltimo acesso: 23 de Setembro de 2015].

LABORATÓRIO PIOLEDO. 2015. *Análises Clínicas - Hemograma* [Online]. Disponível em: http://www.laboratoriopioledo.pt/servicos/index.php?action=getDetalhe&id=9 [Último acesso: 25 de Maio de 2015].

MACHADO, E. L. & LADEIRA, M. 2007. Um estudo de limpeza em base de dados desbalanceada com sobreposição de classes.

MAXWELL. 2015. *Redes Neurais Artificiais* [Online]. Disponível em: http://www.maxwell.vrac.puc-rio.br/13380/13380_4.PDF [Último acesso: 15 de Junho de 2015].

MD.SAÚDE. 2015a. *Hemograma - Entenda os seus resultados* [Online]. Disponível em: http://www.mdsaude.com/2009/11/hemograma.html [Último acesso: 25 de Maio de 2015].

MD.SAÚDE. 2015b. *Sífilis | sintomas e tratamento* [Online]. Disponível em: http://www.mdsaude.com/2009/01/dst-sifilis.html [Ùltimo acesso: 24 de Setembro de 2015].

MEDICAMENTOS E SAÚDE. 2014. *Exames de Hepatologia – Alanina Aminotransferase – ALT* [Online]. Disponível em: http://www.medicamentosesaude.com/exames-de-hepatologia-alanina-aminotransferase-alt/ [Último acesso: 25 de Maio de 2015].

MEDICINENET. 2014. *Hemoglobin* [Online]. Disponível em: http://www.medicinenet.com/hemoglobin/article.htm [Último acesso: 30 de Maio de 2015].

O'DONNELL, R. 2015. *Subspaces and decision trees* [Online]. Disponível em: http://www.contrib.andrew.cmu.edu/~ryanod/?p=547 [Último acesso: 16 de Junho de 2015].

OLIVEIRA, M. 2011. *Leucócitos ou glóbulos brancos* [Online]. Disponível em: http://www.conhecersaude.com/adultos/3375-leucocitos-ou-globulos-brancos.html [Ùltimo acesso: 24 de Setembro de 2015].

OMS. 2015a. *Hepatitis B* [Online]. Disponível em: http://www.who.int/mediacentre/factsheets/fs204/en/ [Ùltimo acesso: 23 de Setembro de 2015].

OMS. 2015b. *Hepatitis C* [Online]. Disponível em: http://www.who.int/mediacentre/factsheets/fs164/en/ [Ùltimo acesso: 23 de Setembro de 2015].

HERMES PARDINI. 2015. *Marcadores das Hepatites Virais* [Online]. Disponível em: http://www.labhpardini.com.br/lab/imunologia/hepatite.htm [Ùltimo acesso: 23 de Setembro de 2015].

PERRY, A. G. & POTTER, P. 2013. *Fundamentos de Enfermagem*. Elsevier, 8ª edição.

RIBEIRO, L. S. 2010. *Uma abordagem semântica para seleção de atributos no processo de KDD.*

ROCHE. 2015a. *Tipos de Hepatites: A,B,C,D,E,G. Marcadores.* [Online]. Disponível em: http://www.roche.pt/hepatites/marcadores.cfm [Ùltimo acesso: 23 de Setembro de 2015].

ROCHE. 2015b. *O que é a SIDA* [Online]. Disponível em: http://www.roche.pt/sida/o_que_e_a_sida/ [Ùltimo acesso: 23 de Junho de 2015].

RODRIGUES, F. 2014. *Descoberta de Conhecimento - Clustering* [Online]. Disponível em: https://moodle.isep.ipp.pt/ [Ùltimo acesso: 20 de Junho de 2015].

SIMSCIENCE. 2015. *Blood: composition and functions* [Online]. Disponível em: http://simscience.org/membranes/advanced/essay/blood_comp_and_func1.html [Ùltimo acesso: 21 de Setembro de 2015].

TODABIOLOGIA. 2015. *Hemácias* [Online]. Disponível em: http://www.todabiologia.com/anatomia/hemacias.htm [Último acesso: 24 de Maio de 2015].

TUASAÚDE. 2012. *Neutrófilos* [Online]. Disponível em: http://www.tuasaude.com/neutrofilos/ [Ùltimo acesso: 25 de Maio de 2015].

TUASAÚDE. 2014a. *Hematócrito* [Online]. Disponível em: http://www.tuasaude.com/hematocrito-hct/ [Ùltimo acesso: 25 de Maio de 2015].

TUASAÚDE. 2014b. *Plaquetas* [Online]. Disponível em: http://www.tuasaude.com/plaquetas/ [Ùltimo acesso: 25 de Maio de 2015].

VON ZUBEN, F. J. & ATTUX, R. R. F. 2010. *Árvores de Decisão* [Online]. Disponível em: ftp://ftp.dca.fee.unicamp.br/pub/docs/vonzuben/ia004_1s10/notas_de_aula/topico7_IA004_1s10.pdf [Ùltimo acesso: 12 de Junho de 2015].

Anexos

Anexo A: Fluxo de Dados – Projeto Microsoft SQL Server Integration Services

Com o objetivo de facilitar o processo de mineração de dados procedeu-se à criação de um projeto *Microsoft SQL Server Integration Services.* De seguida, serão apresentados os fluxos de dados das três entidades criadas.

Na Figura 50 encontra-se representado o fluxo de dados relativo à entidade "Dador", a partir do qual os registos são migrados para a tabela da nova base de dados *SQL*.

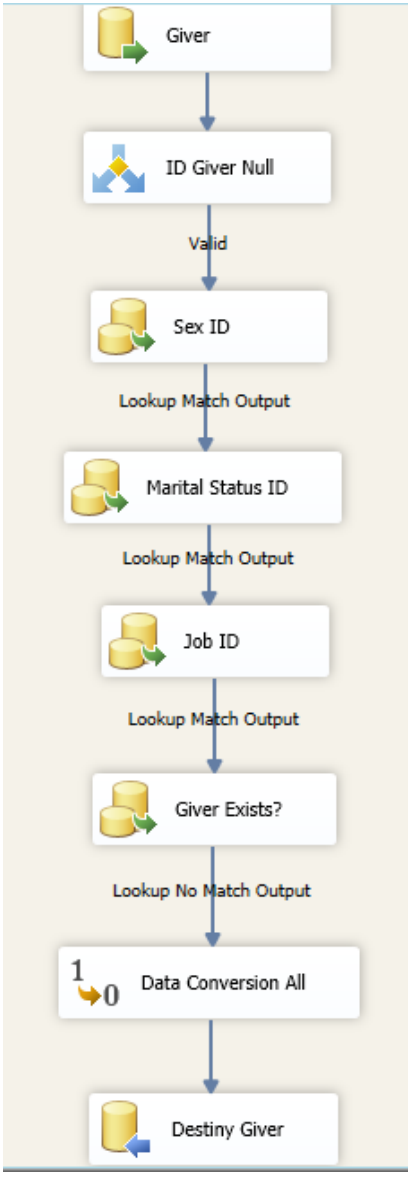

Figura 50. Fluxo de Dados (Load Data Giver).

Um procedimento semelhante ao descrito anteriormente foi aplicado para a entidade "Análise", tal como demonstrado na Figura 51.

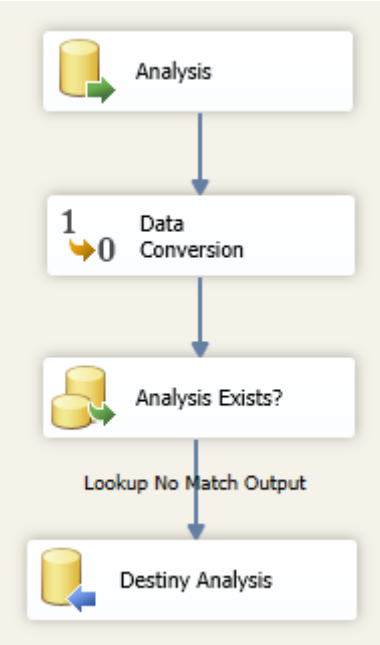

Figura 51. Fluxo de Dados (Load Data Analysis).

O fluxo de dados relativo à entidade “Colheita” encontra-se representado nas Figuras 52, 53 e 54.

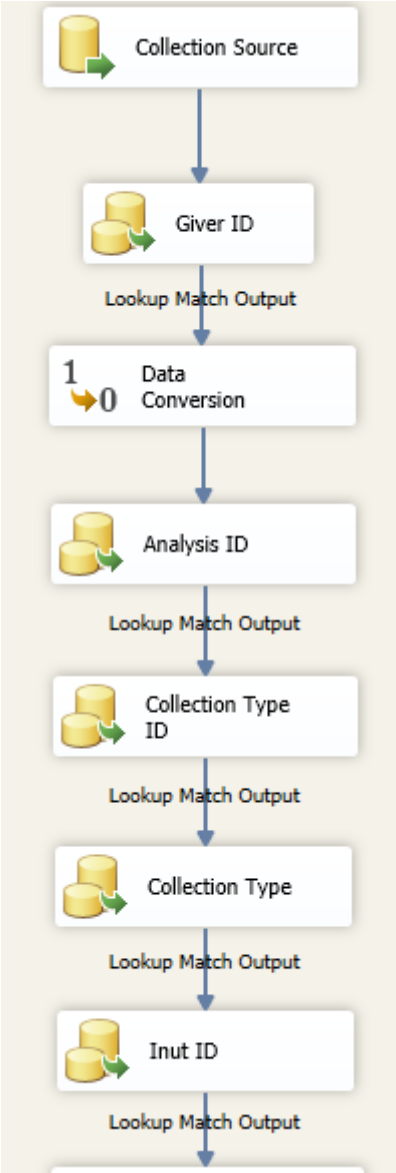

Figura 52. Fluxo de Dados (1 - Load Data Collection).

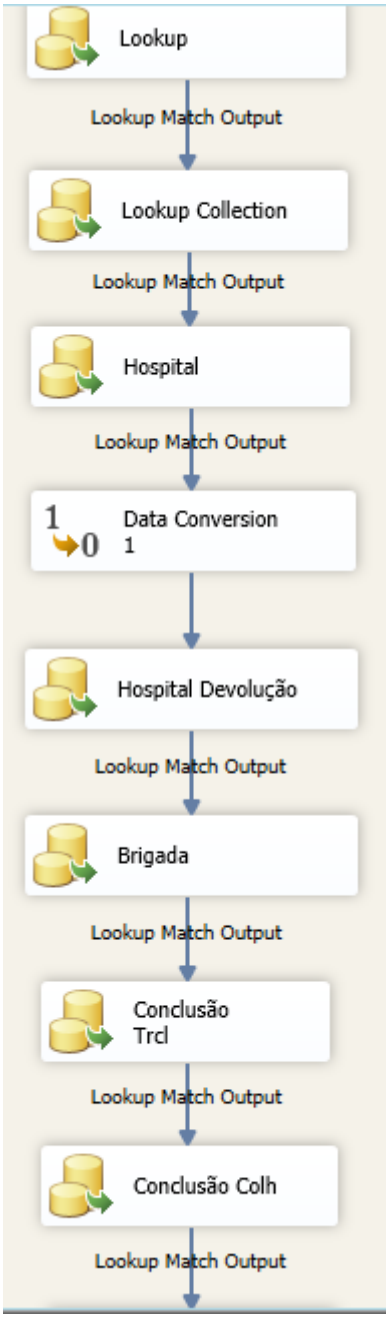

Figura 53. Fluxo de Dados (2 - Load Data Collection).

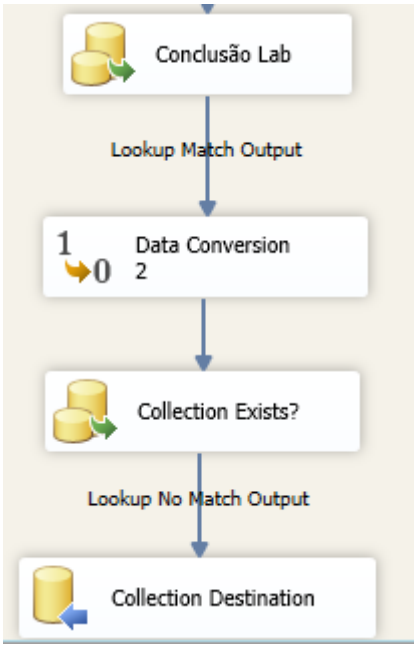

Figura 54. Fluxo de Dados (3 - Load Data Collection).

Anexo B: Ferramenta Clementine 12.0 – Modelos

A seguinte figura (Figura 55) ilustra todos os modelos criados para efetuar a análise estudada ao longo do documento.

Figura 55. Palete dos modelos criados.

Anexo C: Ferramenta Clementine 12.0 – Modelos criados

Ao longo da elaboração deste projeto foram desenvolvidos diferentes modelos com base em algoritmos distintos, que se encontram ilustrados de forma parcial no capítulo “Modelação e Avaliação” (Capítulo 5). De seguida, serão apresentados com maior detalhe (até ao 5º nível) alguns dos modelos gerados.

A Figura 56 ilustra a árvore gerada pelo algoritmo C5.0 (atributo objetivo ALT) e o gráfico com a importância das variáveis utilizadas.

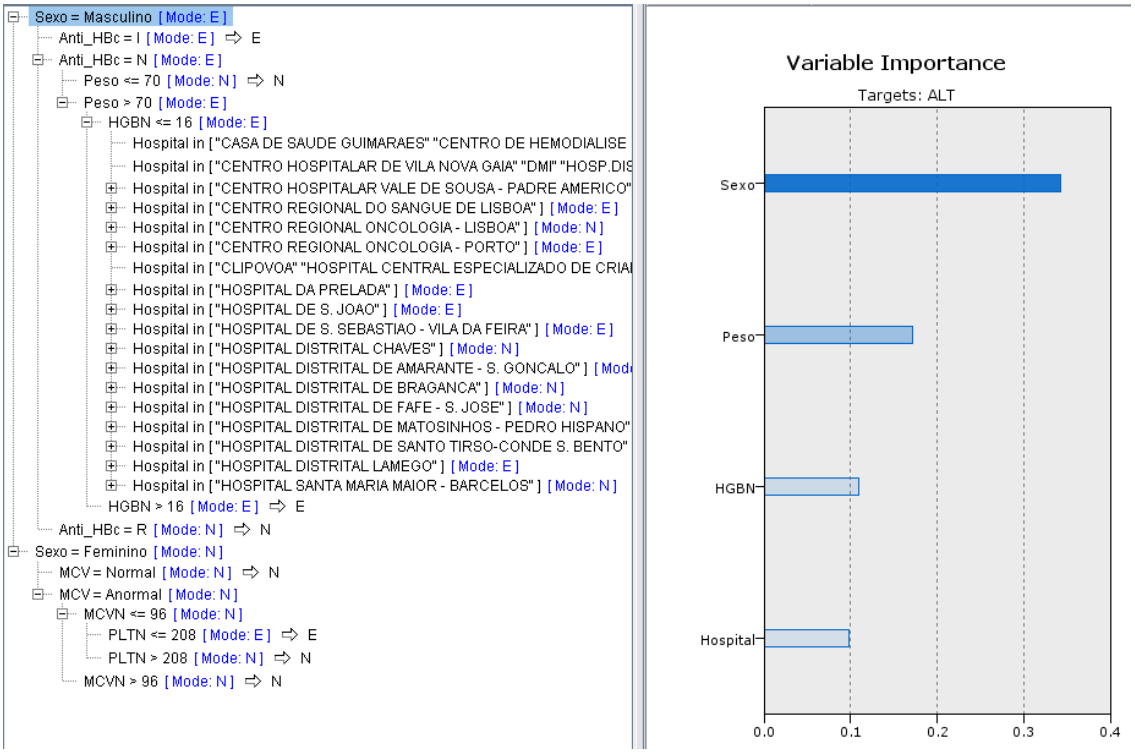

Figura 56. Modelo criado com base no algoritmo C5.0 – atributo objetivo ALT.

A árvore gerada com o algoritmo C5.0 (atributo objetivo MCH) e o gráfico com a importância das variáveis utilizadas podem ser visualizados na Figura 57.

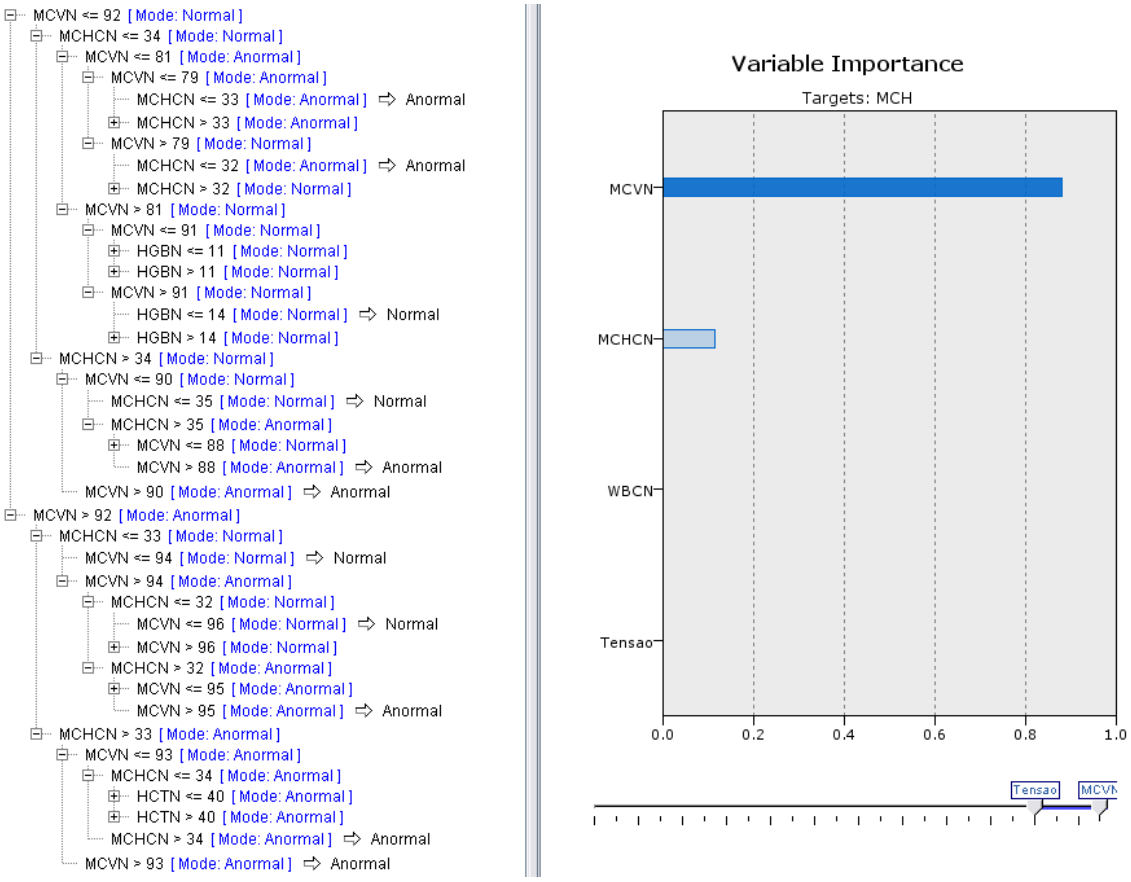

Figura 57. Modelo criado com base no algoritmo C5.0 – atributo objetivo MCH.

Nas Figuras 58 e 59 são apresentadas as árvores geradas pelos algoritmos C5.0 e CHAID, respetivamente, para o atributo objetivo ALT, assim como o gráfico com a importância das variáveis utilizadas.

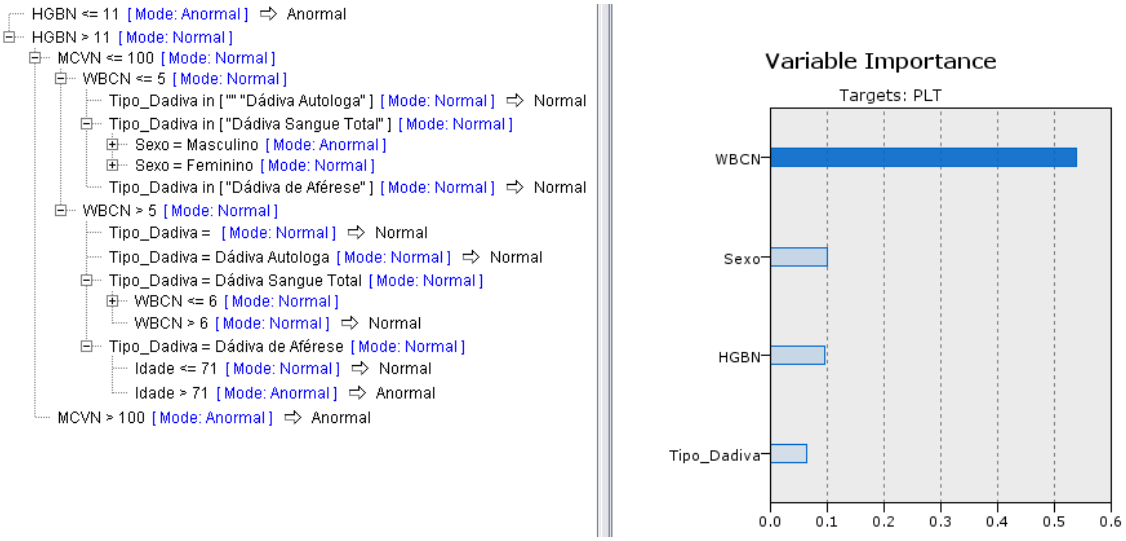

Figura 58. Modelo criado com base no algoritmo C5.0 – atributo objetivo PLT.

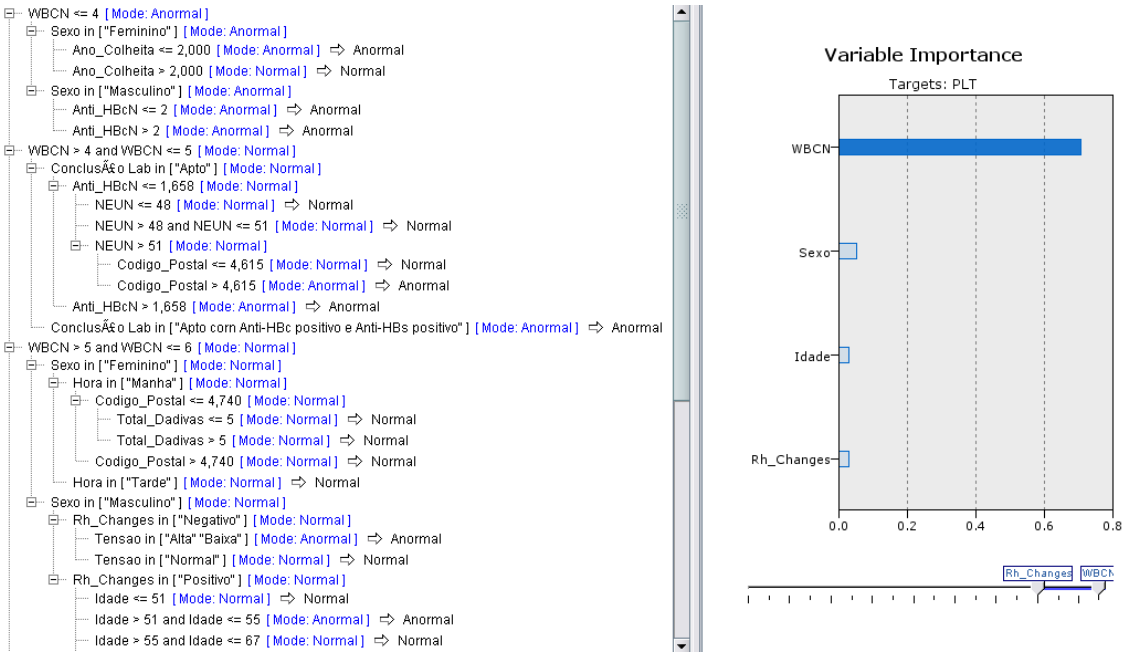

Figura 59. Modelo criado com base no algoritmo CHAID – atributo objetivo PLT.

A Figura 60 apresenta a árvore gerada pelo algoritmo C5.0 (atributo objetivo MCV) e o gráfico com a importância das variáveis utilizadas.

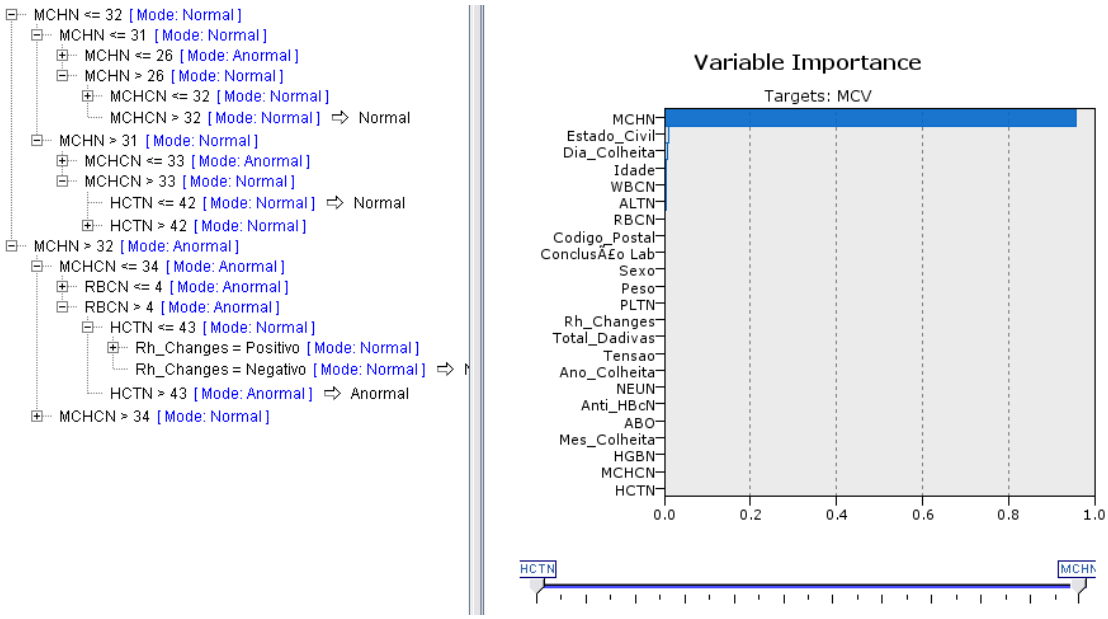

Figura 60. Modelo criado com base no algoritmo C5.0 – atributo objetivo MCV.

A árvore gerada com o algoritmo C5.0 (atributo objetivo MCHC) e o gráfico com a importância das variáveis utilizadas podem ser visualizados na Figura 61.

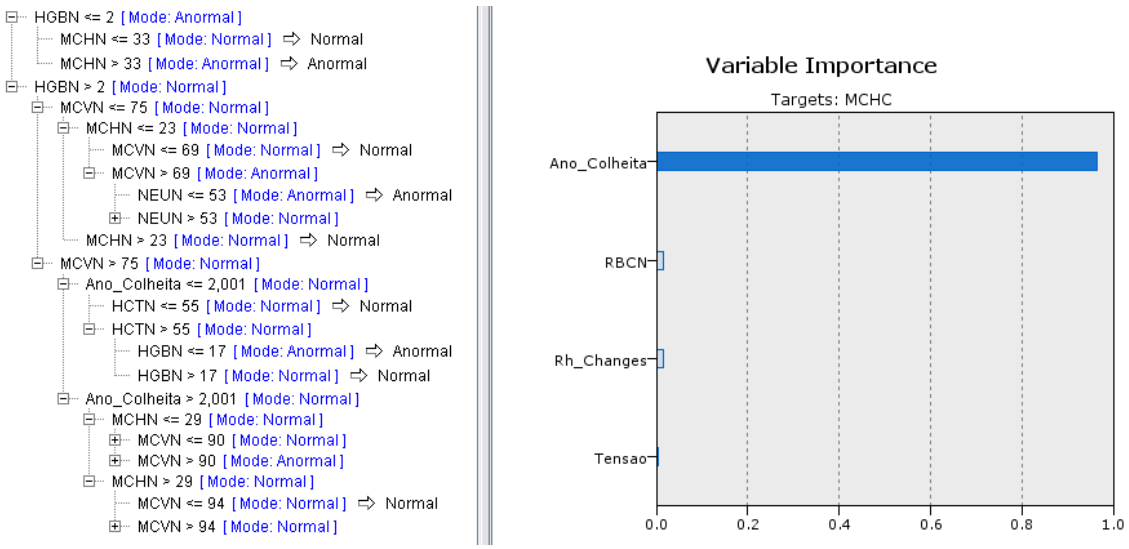

Figura 61. Modelo criado com base no algoritmo C5.0 – atributo objetivo MCHC.

Nas Figuras 62 e 63 são apresentadas as árvores geradas pelos algoritmos C5.0 e CHAID, respetivamente, para o atributo objetivo HCT, assim como o gráfico com a importância das variáveis utilizadas.

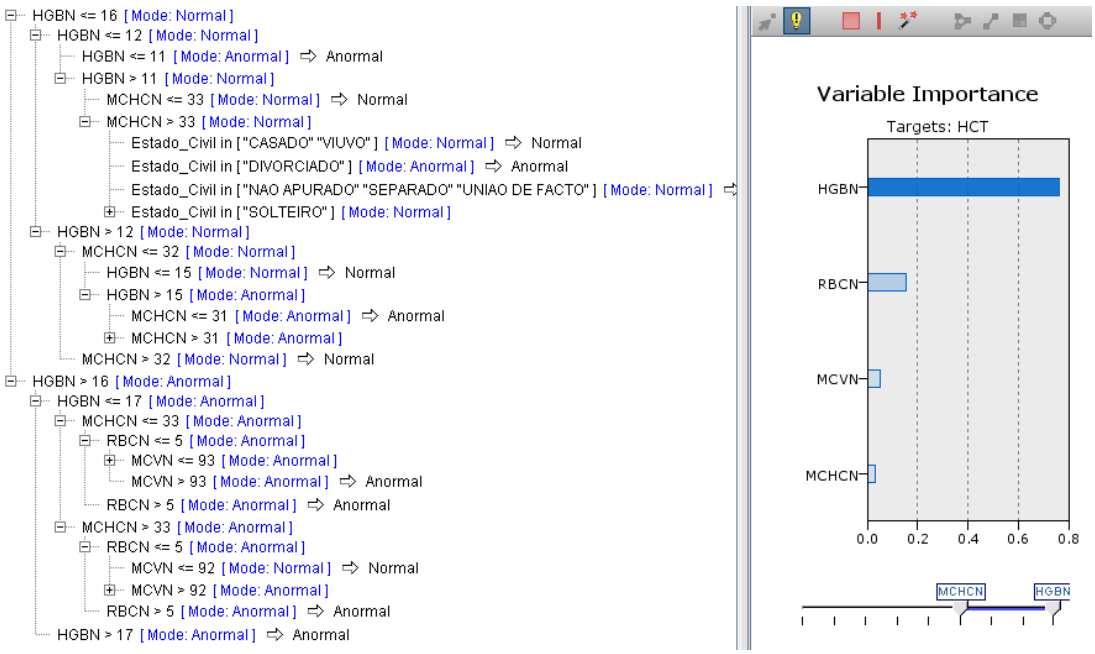

Figura 62. Modelo criado com base no algoritmo C5.0 – atributo objetivo HCT.

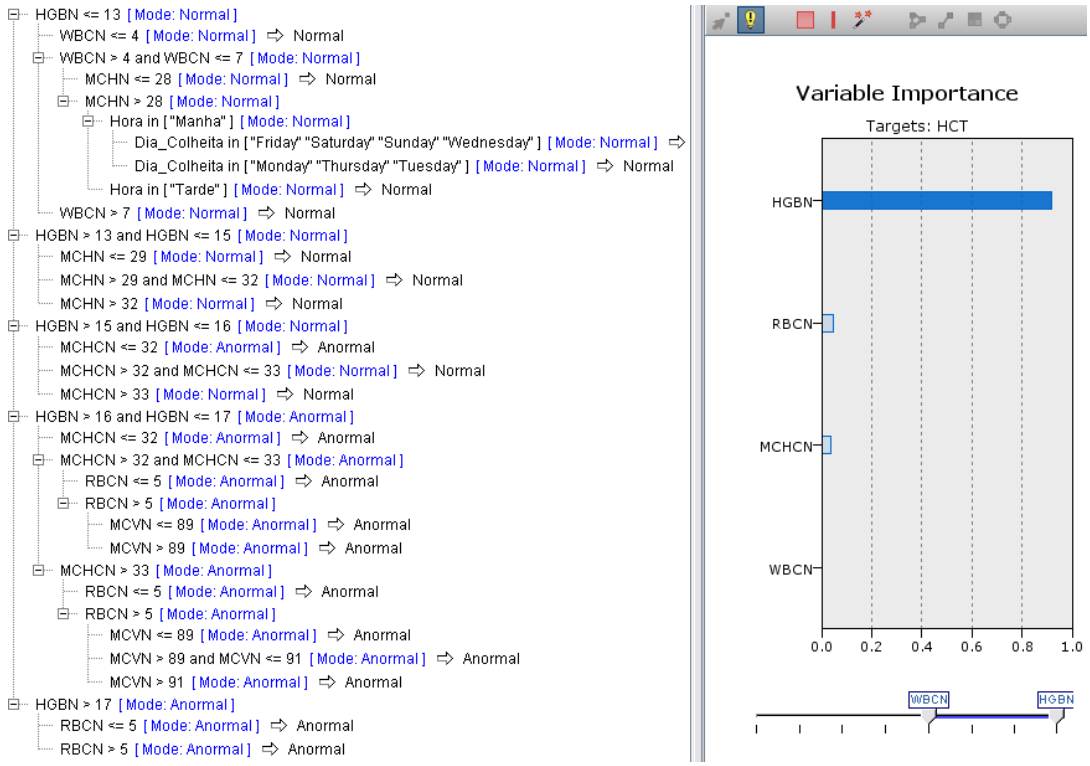

Figura 63. Modelo criado com base no algoritmo CHAID – atributo objetivo HCT.

O mesmo sucede para as Figuras 64 e 65 para o atributo objetivo HGB.

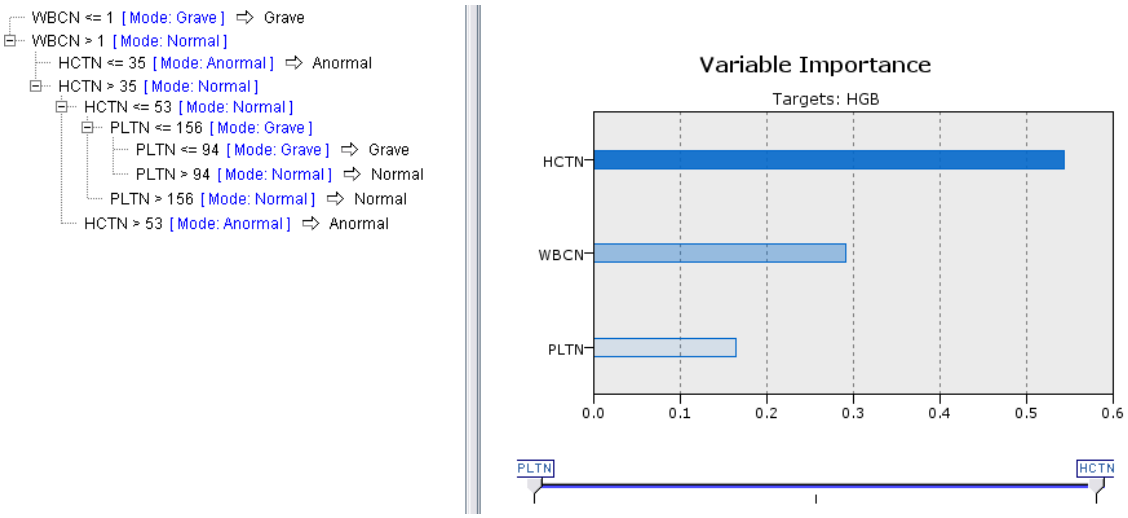

Figura 64. Modelo criado com base no algoritmo C5.0 – atributo objetivo HGB.

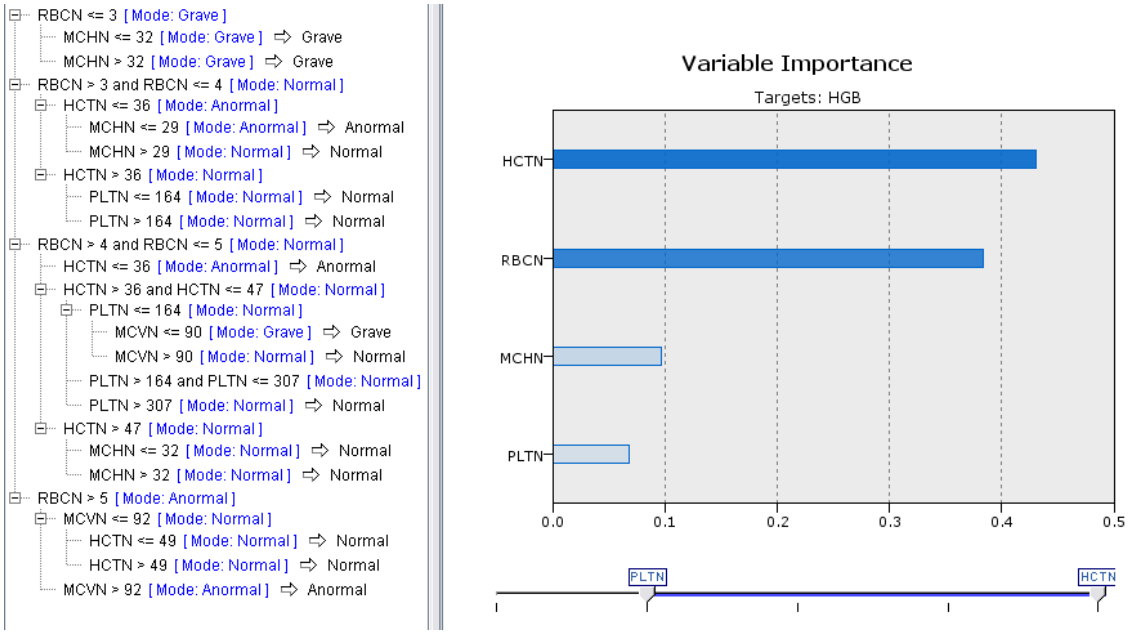

Figura 65. Modelo criado com base no algoritmo CHAID – atributo objetivo HGB.