UNIVERSIDADE CATÓLICA DE GOIÁS
PRÓ-REITORIA DE PÓS-GRADUAÇÃO E PESQUISA
MESTRADO EM CIÊNCIAS AMBIENTAIS E SAÚDE

APLICABILIDADE DE MEMÓRIA LÓGICA COMO FERRAMENTA COADJUVANTE NO DIAGNÓSTICO DAS DOENÇAS GENÉTICAS

Hugo Pereira Leite Filho

Goiânia – Goiás
Agosto de 2006

MESTRADO EM CIÊNCIAS
AMBIENTAIS E SAÚDE

UNIVERSIDADE CATÓLICA DE GOIÁS
PRÓ-REITORIA DE PÓS-GRADUAÇÃO E PESQUISA
MESTRADO EM CIÊNCIAS AMBIENTAIS E SAÚDE

APLICABILIDADE DE MEMÓRIA LÓGICA COMO FERRAMENTA COADJUVANTE NO DIAGNÓSTICO DAS DOENÇAS GENÉTICAS

Hugo Pereira Leite Filho

Prof. Dr. Aparecido Divino da Cruz - Orientador
Prof. Dr. Eduardo Simões de Albuquerque - Coorientador

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciências Ambientais e Saúde, da Pró-Reitoria de Pós-Graduação e Pesquisa da Universidade Católica de Goiás, como requisito parcial para obtenção do título de Mestre em Ciências Ambientais e Saúde

Goiânia – Goiás
Agosto de 2006

L533a Leite Filho, Hugo Pereira.
Aplicabilidade de memória lógica como ferramenta coadjuvante no diagnóstico das doenças genéticas / Hugo Pereira Leite Filho – 2006.
143 f.: il.

Dissertação (mestrado) – Universidade Católica de Goiás, Mestrado em Ciências Ambientais e Saúde, 2006.
"Orientador: Prof. Dr. Aparecido Divino da Cruz".
"Co-Orientador: Prof. Dr. Eduardo Simões de Albuquerque".

1. Rede bayesiana. 2. Inferência probabilística. 3. Síndrome de Turner. 4. Rede neural. 5. Citogenética. 6. Informática. 1. Título.

CDU: 575:004(043)
004:616-056.7(043)

UNIVERSIDADE
Católica
DE GOIÁS

PRÓ-REITORIA DE
PÓS-GRADUAÇÃO E PESQUISA
Av. Universitária, 1069 • Setor Universitário
Caixa Postal 86 • CEP 74605-010
Goiânia • Goiás • Brasil
Fone: (62) 227.1071 • Fax: (62) 227.1073
www.ucg.br • heck@ucg.br

DISSERTAÇÃO DO MESTRADO EM CIÊNCIAS AMBIENTAIS E SAÚDE DEFENDIDA EM 25 DE AGOSTO DE 2006 E CONSIDERADA APROVADO PELA BANCA EXAMINADORA:

1)
Dr. Aparecido Divino da Cruz (Presidente)

2)
Dr. Marcelo Ladeira (Membro Convidado)

3)
Dra. Kátia Cristina Machado Pellegrino (Membro)

4)
Dr. Eduardo Simões de Albuquerque (Co-orientador)

Dedicatória

À Márcia Beatriz Furtado da Cruz (Marcinha), um anjo em especial que, passando pela minha vida, iluminou-a com a sua presença e deixou saudades profundas com sua ausência. A cada dia sigo buscando superar e aceitar a árdua realidade da perda desta pessoa tão amada.

Márcia Beatriz Furtado da Cruz – 19/03/1976 a 26/07/2004

Agradecimentos

Ao Senhor Nosso Deus, pela sua presença constante, pelas oportunidades criadas, pelos momentos de conquistas e pelo seu consolo, que através da fé, consegue fortalecer minha rotina diária e dar-me força para enfrentar as dificuldades e alcançar os momentos felizes.

À minha filha, Giovanna Buzolo Leite, além do meu agradecimento, também o meu amor incondicional. Giovanna é meu maior e melhor presente concebido por Deus para ser a minha fonte iluminada, o alicerce da superação e de todas as minhas dificuldades e o orgulho do meu sucesso. Cada sorriso e cada olhar de Giovanna simbolizam o combustível para conquistas e realizações contínuas. Eu te amo, minha filha!

Aos meus pais, Hugo Pereira Leite e Maria Pereira Santos, meus verdadeiros heróis. A eles agradeço pelo seu amor incondicional, apoio absoluto, incentivos e lições de perseverança, valores de dignidade, caráter, hombridade, respeito ao próximo, paciência e fé. A presença de vocês em minha vida me torna um eterno aprendiz do conteúdo da vida.

À minha família, meu refúgio fortificador com abundância de carinho, paciência e compreensão. Em especial aos meus irmãos, Adriano Pereira Leite e Anderson Pereira Leite, pelos exemplos de pessoas sensatas, à minha cunhada Denise Machado Leite, pela pessoa (extrema) gentileza e, aos meus sobrinhos Ítalo Machado Leite e Beatriz Machado Leite por desempatar em mim sentimento e elo tão importantes e significativos.

À família da Marcinha, que também considero como a minha, Leonor Furtado da Cruz, Ana Carla Furtado da Cruz e José Luiz da Cruz Bastos, que sempre me apoiaram e torceram pelas minhas conquistas.

Ao querido e admirável orientador e professor Aparecido Divino da Cruz, conhecido também como Peixoto. Apesar da sua rotina extenuante, do esgotamento de expectativas,

ele enfrenta com dignidade o seu desempenho profissional. A qualidade do seu afeto incidiu sobre o trabalho de orientação dessa dissertação. Sou grato por vários motivos, pela confiança de orientar um jovem desconsolado, por mostrar o caminho do conhecimento e sabedoria por contribuir para minha maturidade profissional, sentimental e pessoal.

Meus sinceros agradecimentos ao co-orientador, professor Eduardo Simões de Albuquerque, por acreditar na minha capacidade e estar sempre presente, ajudando-me a solucionar problemas ao longo do percurso.

Ao professor Marcelo Ladeira, um exemplo de educador e pessoa. Conhecer o professor Marcelo foi sem dúvida uma das oportunidades que Deus concedeu. Seus conhecimentos foram o norte necessário para que eu aprimorasse meus estudos e consolidasse meu conhecimento.

Agradeço aos amigos, colegas e professores do mestrado, fundamentais para construção dessa dissertação, pois contribuíram com o apoio constante e conselhos fundamentais para a conclusão desse trabalho.

À família “SULEIDE/SES-GO”, que tanto apoiou e torceu pela concretização deste desafio. Sempre serei grato a cada funcionário desta brilhante Instituição, que, além de colegas de trabalho, se tornaram parte de minha família.

Aos colegas da Empresa de Correios e Telégrafos pela paciência e compreensão.

"Cada criatura, é apenas uma graduação padronizada

de um grande harmonioso".

Goethe

RESUMO

O estudo envolveu a interação entre áreas de conhecimento bastante distintas, a saber: informática, engenharia e genética, com ênfase na metodologia da construção de um sistema de apoio à tomada de decisão.

Este estudo tem como objetivo o desenvolvimento de uma ferramenta para o auxílio no diagnóstico de anomalias cromossômicas, apresentando como modelo tutorial a Síndrome de Turner. Para isso foram utilizadas técnicas de classificação baseadas em árvores de decisão, redes probabilísticas (Naïve Bayes, TAN e BAN) e rede neural MLP (do inglês, *Multi- Layer Perceptron*) com algoritmo de treinamento por retropropagação de erro.

Foi escolhido um algoritmo e uma ferramenta capaz de propagar evidências e desenvolver as técnicas de inferência eficientes capazes de gerar técnicas apropriadas para combinar o conhecimento do especialista com dados definidos em uma base de dados.

Chegamos a conclusão que a melhor solução para o domínio do problema apresentado neste estudo foi o modelo Naïve Bayes, pois este modelo apresentou maior acurácia. Os modelos árvore de decisão-ID3, TAN e BAN apresentaram soluções para o domínio do problema sugerido, mas as soluções não foram tão satisfatória quanto o Naïve Bayes. No entanto, a rede neural não promoveu solução satisfatória.

Palavras-Chave: Rede bayesiana, Inferência probabilística, Síndrome de Turner e Citogenética.

ABSTRACT

This study has involved the interaction among knowledge in very distinctive areas, or else: informatics, engineering e genetics, emphasizing the building of a taking decision backing system methodology.

The aim of this study has been the development of a tool to help in the diagnosis of chromosomal aberrations, presenting like tutorial model the Turner Syndrome. So to do that there have been used classification techniques based in decision trees, probabilistic networks (Naïve Bayes, TAN e BAN) and neural MLP network (from English, *Multi- Layer Perception*) and training algorithm by error retro propagation.

There has been chosen an algorithm and a tool able to propagate evidence and develop efficient inference techniques able to originate appropriate techniques to combine the expert knowledge with defined data in a databank.

We have come to a conclusion about the best solution to work out the shown problem in this study that was the Naïve Bayes model, because this one presented the greatest accuracy. The decision - ID3, TAN e BAN tree models presented solutions to the indicated problem, but those were not as much satisfactory as the Naïve Bayes. However, the neural network did not promote a satisfactory solution.

Key Words: Bayesian Network, Probabilists Inferences, Syndrome of Turnner and Cytogenetic.

Sumário

LISTAS DE FIGURAS

LISTAS DE QUADRO

LISTAS DE TABELAS

LISTAS DE SIGLAS

API	Application Programming Interface
BAN	BN Augmented Naïve-Bayes
BN	Bayesian Network
CBL	Cheng, Bell, and Liu
CNM	Combinal Neural Model
DW	Data Warehouse
GPL	General Public License
HGC	Heckerman, Geiger and Chickering
IA	Interligência Artificial
ID	Influency Diagram
KDD	Knowledge Discovery in Databases
LaGene	Laboratório de Citogenética Molecular
MCSA	Mestrado em Ciências Ambientais e Saúde
MD	Mineração de Dados
MDL	Minimal Description Length
MLP	Multi- Layer Perceptron
MSBN	Multiple Select Bayesian Network
SES-GO	Secretaria Estadual de Saúde do Estado de Goiás
SI	Sistema Inteligente
SRA	Srinivas, Russell and Agogino
ST	Síndrome de Turner
SULEIDE	Superintendência Leide das Neves Ferreira – Ciências e Tecnologia em Saúde do Estado de Goiás
SUS	Sistema Único de Saúde
TAN	Tree-Augmented Naïve Bayes
UCG	Universidade Católica de Goiás
WEKA	Waikato Environment for Knowledge Analysis

Capítulo 1

1. Introdução

O presente estudo contempla o processamento de conhecimento referente ao armazenamento e manipulação de dados pela máquina de forma a ser utilizada para a resolução de problemas, especificamente, como coadjuvante no diagnóstico citogenético das doenças genéticas. Além disso, tem como objetivo aplicar os conhecimentos de citogenética e informática para aumentar a qualidade do diagnóstico genético.

O presente estudo propõe uma abordagem numérica para tratar a incerteza mediante ao cálculo de probabilidades, sendo o raciocínio baseado na realização de inferências probabilísticas, isto é, no cálculo da probabilidade condicional de um evento, dada as evidências disponíveis, aplicando-se o teorema de Bayes. Em geral existem diversas evidências e a aplicação direta dessa abordagem numérica que é questionável face aos problemas de complexidade para aplicações de porte, pois requer que uma enorme matriz de probabilidades condicionais seja estimada e fornecida para o sistema, inviabilizando a aquisição de conhecimentos e implicando em elevados requisitos de tempo, armazenamento e capacidade computacional para processar informações de interesse (Flores, 2002).

É no contexto delineado acima que este estudo se insere, propondo o uso de algoritmos de aprendizagem de máquina para extrair modelos de conhecimento – redes probabilísticas - a partir de bases de dados, de modo que estas novas informações

possam auxiliar o especialista na etapa de aquisição de conhecimento que compõem o processo de construção de um sistema de apoio à tomada de decisão (Koehler e Nassar, 2002).

1.1 Objetivos

Objetivo Geral:

- Identificar uma metodologia para o auxílio na identificação de diagnóstico para anomalias cromossômicas e das doenças genéticas.

Objetivos Específicos:

- Montar um resumo sobre redes bayesianas e seus classificadores;
- Descrever os algoritmos de aprendizagem;
- Utilizar o algoritmo de aprendizagem para modelar dados na área da citogenética, propagar evidências sobre tais modelos e analisar resultados;
- Aplicar o modelo de forma tutorial à Síndrome de Turner.

1.2 Ambiente e Saúde

A relação entre a exposição individual aos agentes ambientais e o aumento no risco a saúde humana decorrente da exposição têm sido estudadas principalmente na dimensão temporal (Anto, 1989; Schwartz e Marcus, 1990). A inexistência de metodologias eficientes na análise ambiental correlacionando com saúde das populações não pode ser atribuída à carência de informações. Portanto, os estudos de sistemas especialistas, entendidos como um conjunto de técnicas de coleta, exibição e tratamento de informações especializadas (Rodrigues, 1990), permitem a análise conjunta de uma gama de variáveis sócio-ambientais, não estando, porém, livres da indução de alguns equívocos.

De acordo com Barcellos e Bastos (1996), fatores culturais, econômicos, demográficos e ambientais estão presentes em todas as escalas em que se represente o espaço, sendo talvez na escala global que as variáveis culturais apresentam maiores diferenciais. Os contrastes, porém, estão presentes na escala nacional, regional e local com menores intensidades, ou se mostram 'desbotados' em relação a outros fatores de diferenciação populacional. O espaço geográfico foi definido por Harvey (1980) como um complexo não homogêneo, talvez descontínuo e quase certamente diferente do espaço físico. No entanto, a delimitação do objeto, objetivos e hipóteses de estudo impõem uma homogeneização da unidade de análise, no interior da qual não é possível observar diferenças espaciais.

A categoria espaço tem valor intrínseco na análise das relações entre saúde e ambiente e no seu controle. Conhecer a estrutura e a dinâmica espacial permite a caracterização da situação em que ocorrem eventos que afetam a saúde. Neste sentido, o conhecimento se oferece como instrumento que clama por uma retomada da análise de situações concretas das populações e suas interações, submetidas a riscos de natureza difusa, e, por vezes, superposta (Barreto *et al.*, 1993). Além disso, o conhecimento permite o planejamento de ações de controle, alocação de recursos e a preparação de ações de intervenção, em caráter de emergência ou não.

Devido ao conjunto de elementos inter-relacionados presentes no espaço, torna-se difícil o estabelecimento de relações de causalidade entre condições ambientais e saúde. A bioinformática permite que informações ambientais e de saúde construam uma identificação de variáveis que revelem as estruturas sociais, econômicas e ambientais que oferecem riscos à saúde. Como sugere Santos (1988), a busca das causas, relacionando apenas fatores visíveis, deve ser preterida em favor do estabelecimento do contexto, no qual um evento de saúde ocorre, o que certamente não é pouco. Com isso,

a categoria espaço contribui para o entendimento dos processos envolvidos em determinado fenômeno ambiental que se deseja estudar. A bioinformática passa a ser um poderoso instrumento a serviço da pesquisa em saúde no contexto do ambiente e indo para além dele.

1.3 Estrutura do Estudo

Este trabalho é dividido em 8 capítulos e 2 anexos. No primeiro é introduzido o assunto e são relatados os objetivos, contextualização do ambiente com saúde e a estrutura do trabalho.

No capítulo 2 estão descritos os fundamentos teóricos de sistemas especialistas probabilísticos, como histórico, probabilidades bayesianas, redes bayesianas, descoberta de conhecimento na base de dados e sobre o método CRISP-*DM, do inglês, CRoss Industry Standard Process for Data Mining*.

O terceiro capítulo aborda a aprendizagem de redes bayesianas, onde são apresentados alguns métodos com suas características, algoritmos e em alguns casos, aplicações práticas já desenvolvidas. Os métodos são divididos em: método de busca e pontuação e método baseado em análise de dependência.

No capítulo 4 são comentadas as características de alguns softwares disponíveis para a propagação e aprendizado em redes bayesianas.

No quinto capítulo aborda-se o referencial teórico da Síndrome de Turner, detalhando os primeiros registros de identificação da síndrome, os sinais e sintomas que identifica a síndrome e considerações finais sobre a Síndrome de Turner.

No sexto capítulo aborda-se a metodologia da pesquisa, focando o domínio do problema, dados coletados para a pesquisa e a metodologia utilizada para o desenvolvimento da aplicação.

No capítulo 7 são apresentadas as avaliações dos resultados das amostras da pesquisa, dados estatísticos das amostras, índices de referências para identificar o melhor classificador bayesiano.

No oitavo capítulo são apresentadas as conclusões da pesquisa e os trabalhos que podem ser realizados no futuro com a proposta de melhorar os resultados obtidos.

Capítulo 2

2. Fundamentação Teórica

2.1 Histórico

Em 1763, o reverendo Bayes (Bayes, 1763) sugeriu uma regra que possibilitava estimar a probabilidade de um evento com base no conhecimento humano. A proposta passou a ser conhecida mundialmente como "Teorema de Bayes". Segundo este teorema, para os eventos cuja freqüência de ocorrência não pode ser estabelecida, a probabilidade da ocorrência pode ser dada com base no conhecimento empírico que especialistas tem sobre o mesmo.

As redes bayesiana sutilizam o Teorema de Bayes como método quantitativo para a revisão de probabilidades conhecidas, com base em uma nova informação amostral. Embora a estatística bayesiana tivesse sua origem no trabalho de Thomas Bayes, os trabalhos desenvolvidos pelo matemático francês Pierre Simon de LaPlace, em 1812, desenvolveram o teorema como é conhecido e utilizado atualmente (Herckerman, 1995).

Ao longo das décadas, a estatística tem desempenhando um papel importante na área da pesquisa científica. Foram desenvolvidos métodos usados para avaliar hipóteses e determinar as diferenças que podem ser relacionadas as chances aleatórias. A teoria da estatística formal apóia modelos de dados e métodos de previsão. Neste contexto, a

inferência bayesiana é o método estatístico mais utilizado para se estimar probabilidades em sistemas especialistas probabilísticos.

A estatística bayesiana passou a ser aplicada em sistemas de inteligência artificial no início da década de 60 (Russel e Norvig, 1995). Naquela época os formalismos das utilizações de probabilidades condicionais ainda não estavam bem definidos. Além disso, a grande quantidade de dados a serem manipulados dificultava a utilização da teoria. Assim, a partir do início da década de 70 e até a metade da década de 80, a aplicação direta da probabilidade bayesiana foi pouca explorada em pesquisas de IA (Inteligência Artificial). Com a publicação de trabalhos que definiram de forma mais concisa o método bayesiano e reduziram as quantidades de cálculos necessários, a teoria bayesiana provocou um grande impulso no campo da IA.

2.2 Probabilidade Bayesianas

As probabilidades bayesianas são usadas frequentemente nas inferências estatísticas para especificar o conhecimento *a priori* e combinar este conhecimento com os dados disponíveis através do Teorema de Bayes (Press, 1989). O conhecimento sobre o sistema antes que os dados sejam conhecidos é codificado no formulário de uma distribuição prévia da probabilidade. A fórmula de Bayes fornece então uma regra, atualizando probabilidades prévias, *a posteriori* quando os dados são conhecidos e analisados. Dado uma série de dados para a avaliação da aptidão, o melhor resultado (o mais apto) é definido como o modelo dos dados mais provável, respeitando o conhecimento prévio do domínio do problema (Zhang, 1999).

A probabilidade bayesiana é uma teoria consistente e que permite a representação de conhecimentos certos e incertos, via distribuição de probabilidade

conjunta. Tal distribuição conjunta pode ser representada pelo produto de distribuições condicionadas (Hruschka Jr, 1997).

As distribuições probabilísticas ocorrem nas relações causais incertas dentro de um domínio do problema. O raciocínio probabilístico é executado sobre estas relações com o uso do Teorema de Bayes, que pode ser expresso como segue (Rajabali *et al.*, 2004):

$$P(Xi \mid Y) = \frac{P(Xi)P(Y|Xi)}{P(X1)P(Y|X1) + P(X2)P(Y|X2) + P(X3)P(Y|X3) + ... + P(Xn)P(Y|Xn)}$$

Para $i = 1, 2, 3, ..., n$

Uma das razões para o interesse no uso de modelos probabilísticos bayesianos, quando comparados ao ponto tradicional do domínio do problema é que eles podem prever modelos estimados fornecendo ferramentas para o cálculo do risco e permitem que os responsáveis pelas tomadas de decisões combinem dados históricos com as estimativas subjetivas relatadas nas inferências probabilísticas (Pendharkar *et al.*, 2005).

A partir de observações gerais de variáveis que compõem um evento em particular, amostras em condições semelhantes podem ser inferidas como pertencente à população que contem o evento. Generalizar aspectos importantes de uma população a partir dos resultados obtidos na ocorrência de um evento de uma amostra caracteriza a inferência (Lopes, 1999). No contexto científico, a inferência probabilística é usada com dois significados:

- procedimento para se tirar conclusões a partir de valores ou de evidências; e
- procedimento necessário para tirar as conclusões conforme os resultados.

Para que as inferências probabilísticas sejam válidas, a ocorrência (amostra de um evento) deve ser representativa na população. A escolha de uma ocorrência é geralmente feita pelo princípio da conveniência, o que pode provocar um resultado de

inferência tendencioso. Para corrigir este viés, a escolha da ocorrência passou a ser aleatória o que reduz a chance de se subestimar ou superestimar um dado resultado.

Inferências probabilísticas são utilizadas em algoritmos de propagação de crenças em redes bayesianas, podendo ser do tipo **causal**, que parte das causas para os efeitos; **diagnóstico**, partindo dos efeitos para as causas; **intercausal**, quando discrimina entre causas de um efeito comum ou **misto**, caracterizado pela combinação de dois ou mais tipos previamente citados (Ladeira, 2000). Nas inferências probabilísticas, calcula-se a probabilidade de um evento, dado as evidências[1] observadas na rede.

Quando as variáveis são conhecidas, com seus respectivos valores atribuídos, a parte apreensível passa a ser a tabela de probabilidades condicionais. Estas probabilidades podem ser estimadas diretamente usando recursos de estatísticas sob o conjunto de dados, ou seja, é possível utilizar evidências estatísticas para solucionar problemas.

2.3 Redes bayesianas

As redes bayesianas foram utilizadas por muitos autores (Henrion *et al.*, 1991; Pradhan *et al.*, 1994; Laskey e Mahoney, 1997) como a técnica de modelagem de escolha para o desenvolvimento de sistemas especialistas probabilísticos (Przytula *et al.*, 2000). Uma variável é condicionada a uma ou mais outras variáveis numa relação causal. Uma distribuição pode ser representada por um grafo acíclico orientado. No grafo, cada nó representa uma variável do modelo e os arcos ligam as variáveis que estão em relação direta de causa/efeito. A esta estrutura gráfica, com a quantificação da

[1] Evidências: Observações de ocorrências de valores específicos de variáveis aleatórias, usadas nas estimativas das probabilidades de outras variáveis que assumiram certos valores (Lopes, 1999).

crença nas variáveis e seus relacionamentos, dá-se o nome de redes bayesiana ou redes causais.

Segundo Hruschka Jr, (1997) as redes bayesianas são redes de conhecimento, representada por grafos direcionados e acíclicos, nos quais os nós representam variáveis aleatórias com tabelas de probabilidades condicionais associadas e os arcos representam à interdependência entre as variáveis. A rede bayesiana permite a quantificação da força dos relacionamentos entre as variáveis e, mediante seu uso, pode-se calcular a probabilidade de um evento ocorrer condicionado à ocorrência de outro. Assim, redes de conhecimento representam a incerteza, tomando-se por base a teoria da probabilidade. Com o aparecimento de modelos de redes bayesianas, houve uma retomada do interesse em usar probabilidades para representar e manipular a incerteza nos sistemas inteligentes.

A maior dificuldade encontrada ao se trabalhar em rede bayesiana é o grande esforço computacional exigido para os cálculos probabilísticos. Pois, no cálculo de distribuições de probabilidade com a aplicação direta do Teorema de Bayes há uma explosão combinatória. No entanto, quando se explora com as redes bayesianas as independências condicionais entre as variáveis do problema, pode-se reduzir esse esforço computacional para se obter o mesmo resultado (Pearl, 1988).

Uma vantagem de uso de redes bayesianas consiste na possibilidade de potencializar o aprendizado do sistema a partir de dados inseridos. Nesse aprendizado, uma amostra é apresentada ao sistema, que, através de um algoritmo, gera a estrutura que melhor representa as relações presentes nos dados do problema (Herckerman, 1995; Buntine, 1995; Buntine, 1994b). Existem vários algoritmos e métodos para o aprendizado de redes bayesianas a partir de dados. Os algoritmos de árvores de Chow-Liu, de Lam-Bacchus e o K2, são exemplos clássicos de ferramentas baseadas no

método de busca e pontuação, sendo o K2 o algoritmo mais representativo do grupo. Existem ainda os algoritmos baseados nos métodos de análise de dependência, incluindo o Wemuth-Lauritzen (1983) e o *Constructor*.

Com as redes bayesianas pode-se representar problemas do mundo real que mantenham relações de potenciais causa e conseqüência entre as suas variáveis. A utilização de tais redes vem ganhando espaço no meio acadêmico e comercial. Existem aplicações que obtiveram resultados positivos (Doyle et al., 1994) como, por exemplo, o *Intelipath,* um sistema de diagnóstico de patologias aprovado pela Associação Americana de Medicina (Herckerman, 1991), o VISTA, um sistema de monitoramento e análise que foi utilizado pela NASA para controle de missões espaciais (Horvitz *et al.*, 1992), e os assistentes de solução de problemas da Microsoft (Herckerman *et al.*, 1995).

Segundo Nirajan (1999), atualmente vem se intensificando os métodos bayesianos dentro dos problemas para aprendizagem de máquina. O uso de metodologia bayesiana possibilita a especificação de incertezas sobre um domínio do problema de modo eficiente e ao se carregar o processo, as formas de inferências probabilísticas podem ser modeladas. Além do que foi exposto acima, uma vantagem adicional do processo de modelagem usando uma rede bayesiana consiste na habilidade de se isolar e incorporar modelos causais como probabilidades condicionais (Davis *et al.*, 1995).

De acordo com Flores (2002), representar e manipular o conhecimento em um domínio na área da citogenética, por exemplo, traz um complicador adicional. Ela é, por si, uma área aberta, onde vigora o conhecimento empírico. No entanto, raciocinar com incerteza é mais comum do que se imagina. Na realização de um exame de cariótipo, para identificação de anomalias cromossômicas, o técnico-especialista habitualmente raciocina com incerteza, pois em geral, os sinais e sintomas observados não determinam a ocorrência de uma única síndrome. Assim, os técnico-especialistas

antes de definirem qual a possível aberração cromossômica associada àqueles sinais e sintomas, analisam todo o conjunto cromossômico observando potenciais alterações no número e na forma (estrutura) de cada unidade cromossômica, geralmente recorrendo aos manuais teóricos acerca dos cromossomos e sua patologias associadas, para a conclusão e, conseqüentemente, identificação da síndrome em estudo.

Quando um estudo é tratado com incerteza, faz-se necessário utilizar recursos estratégicos para solucioná-lo, como abordagens *numérica* e *simbólica*. A abordagem simbólica é adequada para o tratamento de informações incompletas, mas não imprecisas, pois não se consegue quantificar essa incerteza. Entende-se por informação incompleta quando os sinais/sintomas não apresentam correlação com a síndrome (i.e., algumas respostas às questões relevantes não são conhecidas). Por outro lado, numa informação imprecisa as respostas são conhecidas, mas são aproximadas, face à baixa confiabilidade da fonte ou em conseqüência da imprecisão da linguagem de representação do conhecimento utilizada. Já na abordagem numérica, representa-se a incerteza como uma quantidade precisa em uma dada escala. Assim, pode-se definir um cálculo que especifica o mecanismo a ser usado para combinar e propagar a incerteza durante o processo de raciocínio (Ladeira, 2000).

2.3.1 Construindo Redes bayesianas

Conforme Zhang e colaboradores (2002), na construção da rede bayesiana descreve-se os passos para quantificar e qualificar o conhecimento de um determinado domínio do problema. Para se construir uma rede bayesiana parte-se do princípio que os dados originalmente são atuais e reflete o escopo do problema, seguindo-se as etapas:

- **Primeira etapa:** as variáveis e a interpretação da rede bayesiana são obrigatoriamente estabilizadas.

- **Segunda etapa**: as direções dos grafos acíclicos (DAG) que indicam condições independentes estão de acordo com o conhecimento do especialista;
- **Na terceira etapa:** uma ou mais redes bayesianas são construídas usando-se o conhecimento da coleta dos dados, baseando-se nas expressões de multiplicadores de probabilidades.
- **Quarta etapa:** calcula-se as distribuições de probabilidades condicionais P ($xi|$Pai).

No modelo matemático de uma rede bayesiana, os nós e arcos representam, respectivamente, as variáveis de um universo $U = (A_1, A_2,..., A_N)$ e as dependências entre elas. Na área de citogenética, a direção dos arcos, em geral, pode representar relações de causa-conseqüência entre as variáveis do domínio citogenético modelado.

Por exemplo, se houver um arco indo de um nó *A* para um nó *B*, assume-se que o nó *A* representa uma causa de *B* e adota-se como nomenclatura que *A* é um dos pais de *B*; analogamente, *B* é um dos filhos de *A*. Associado ao grafo, existe uma distribuição de probabilidades. As redes bayesianas adotam uma representação compacta onde são definidas somente as probabilidades condicionais de cada nó em relação aos seus pais. Segundo Przytula e colaboradores (2000), a topologia de uma rede bayesiana foi determinada para identificar o domínio do problema e fornecer as informações probabilísticas sobre conexões entre nós, conhecidas também como classes e atributos. São necessárias informações suficientes para se calcular as distribuições das probabilidades para as coleções dos nós conectados.

Portanto, redes bayesianas obedecem à condição de Markov que estabelece a inexistência de uma relação de dependência direta em quaisquer dois nós, a não ser que exista um arco entre eles na rede. A distribuição de probabilidade correspondente à rede é calculada a partir das probabilidades condicionais:

$$P(U) = P(A_1, A_2,..., A_N) = {}^{n}\prod_{i=1} P(A_i|pa(A_i)),$$

onde $P(U)$ é a distribuição de probabilidade conjunta para a rede, $pa(A_i)$ são os pais do nó de A_i e $P(A_i|pa(A_i))$ são as probabilidades condicionais de A_i dados os seus pais.

Cada nó possui um número finito, maior ou igual a dois, de categorias. As categorias, também comumente denominadas estados, representam os possíveis valores da variável representada pelo nó. Um nó é observado quando há conhecimento sobre o estado da variável que representa o nó. Os nós observados têm grande importância no processo de inferência realizado na rede, pois, juntamente com as probabilidades condicionais especificadas para a rede, determinam as probabilidades dos nós não observados. As probabilidades condicionais da rede após a inferência são dadas por:

$$P(A_i|E),\ A_i \in U,$$

ou seja, as probabilidades de cada nó, dado o conjunto de nós observados (E).

Em aplicações práticas, os valores da probabilidade conjunta $P(U)$ não são muito significativos na análise do problema modelado. De maior interesse são as probabilidades marginais de cada nó não observado. Utilizando-se a probabilidade conjunta, pode-se obter as probabilidades marginais somando-se, para cada estado de cada variável, todas as probabilidades em que a variável encontra-se no estado desejado. Em seguida, normaliza-se a probabilidade obtida e obtêm-se as probabilidades marginais para cada nó.

Nos quadros abaixo são representadas três variáveis, que serão determinadas na construção do gráfico acíclico direcionado como *A, B, e C*. O nó *A* é a variável que representa a causa principal, *B e C* são os efeitos de *A*. Os valores das probabilidades totalizam 100% para cada variável. Os Quadros I, II e III representam os valores de probabilidades para a rede acima:

Quadro I - Valores de P(A).

P(a1) = 0,4	P(a2) = 0,6

Quadro II - Valores de P(B|A).

P(b1/a1) = 0,2	P(b2/a2) = 0,75
P(b1/a2) = 0,25	P(b2/a1) = 0,8

Quadro III - Valores de P(C|A).

P(c1/a1) = 0,3	P(c2/a2) = 0,9
P(c1/a2) = 0,1	P(c2/a1) = 0,7

A partir dos quadros acima, obtém-se o quadro conjunto *P(A, B, C)*, contendo oito valores, dados por *P(A, B, C)= P(A) * P(B|A) * P(C|A):*

Quadro IV - Valores de P*(A, B, C*).

P(a1,b1,c1) = 0,024	P(a1,b2,c1) = 0,096
P(a1,b1,c2) = 0,056	P(a2,b1,c1) = 0,015
P(a2,b2,c1) = 0,045	P(a1,b2,c2) = 0,224
P(a2,b2,c2) = 0,405	P(a2,b1,c2) = 0,135

Para o caso em que o nó B é observado como B=b2, realiza-se um processo de inferência para os nós A e C. Para P(a1|B=b2), por exemplo, realiza-se o seguinte cálculo: P(a1|B=b2) = (P(a1,b2,c1) + P(a1,b2,c2))/ (P(a1,b2,c1) + P(a1,b2,c2) + P(a2,b2,c1) + P(a2,b2,c2)) = $P(a1 \mid B = b2) = \frac{\sum P(a, b2, c)}{\sum_c \sum_a P(a, b2, c)}$

Como é observado que *{B=b2}*, todas as probabilidades P*(A, B, C*) contendo *b1* não são utilizadas. Realizando-se o processo acima para todos os estados de *A* e *C*, obtêm-se as probabilidades condicionais para cada nó, exibidas no Quadro V.

Quadro V - Valores de P*(a1,a2,c1,c2*).

P(a1|B=b2)= 0,4156	P(c1|B=b2)= 0,1831
P(a2|B=b2)= 0,5844	P(c2|B=b2)= 0,8169

O processo de inferência realizado acima é a base da utilidade das redes bayesianas (Jensen, 2001). Em modelos causais não há relações com complexidade muito grande. A estrutura de cálculos terá uma complexidade proporcional ao número de variáveis dependentes e não dependentes ao número de variáveis do problema (Herckerman, 1995).

As redes bayesianas se destacam por representar modelos causais e por se adaptarem melhor aos modelos especialistas (Pearl, 1988). No presente estudo, a rede bayesiana representa um modelo para investigar anomalias cromossômicas em potencial, usando como parâmetros sinais e sintomas das síndromes cromossômicas.

2.4. Estado da Arte na técnica de IA e Mineração de Dados

Percebe-se um grande avanço na utilização de inferência probabilística para o desenvolvimento de sistemas inteligentes (SI) na comunidade acadêmica. Os objetivos primordiais dos pesquisadores consistem em desenvolver técnicas de inferência eficientes para o uso em SI, para os quais faz-se necessário a disponibilidade de modelos de conhecimento válidos (Koehler *et al.*, 2004). Adicionalmente outros pesquisadores se preocupam com a geração de técnicas capazes de combinar o conhecimento do especialista com informações contidas em um banco de dados para refinar os modelos de conhecimento e auxiliar na tomada de decisão *a posteriori*

(Krauser, 1998). Em conseqüência, surgiram as técnicas de mineração de dados (*do inglês, Data Mining*) e os métodos de conhecimento baseado na tecnologia de redes bayesianas para extração de modelos de conhecimento válidos a partir de base de dados.

Na tentativa de representar o conhecimento humano, a motivação para utilização de redes bayesianas é decorrente da necessidade de que os seres humanos têm de moldar os fatos e fenômenos em forma de relacionamentos causais. Sempre que um fato está sendo analisado, busca-se uma explicação que seja a causa do fato em questão, mesmo que esta causa seja imaginária e que não se possua evidências concretas (Herckerman, 1995).

Conforme Koehler e colaboradores (2004), os algoritmos de aprendizagem bayesiana, atualmente existentes, geram modelos de conhecimento com todas as variáveis existentes nas bases de dados, o que nem sempre é necessário. Assim, é solicitado ao usuário para que informe ao algoritmo quais as variáveis são mais relevantes e a sua ordenação. Por exemplo, considerando o processo de tomada de decisão na prática de exame citogenético, apenas os sintomas mais importantes são considerados necessários para a tomada de decisão sobre um determinado resultado do exame. Com isso, o técnico-especialista não elege um resultado do exame baseado em muitos sinais/sintomas, mas sim, a partir dos sinais/sintomas mais significativos para a tomada de decisão.

A construção de modelos de conhecimento bayesianos envolve três aspectos: a aprendizagem da estrutura da rede, a aprendizagem das relações de causalidade e as probabilidades associadas. O foco do presente estudo é a aprendizagem da estrutura da rede, pois a mesma é considerada a de maior relevância pela literatura científica, devido

a sua complexidade se comparada com a aprendizagem de parâmetros numéricos, que é facilmente resolvida utilizando técnicas estatísticas (Koehler *et al.*, 2004).

Para Koehler e colaboradores (2004), para gerar modelos de conhecimento bayesianos, muitos algoritmos encontrados na literatura científica necessitam das seguintes informações: a) a relação de variáveis, b) a ordenação destas variáveis, e c) as relações de dependência e independência entre as mesmas, isto é, as relações de causalidade. Com essas informações repassadas para o *software* de aprendizagem, pressupõe-se que o usuário tenha um profundo conhecimento sobre a rede a ser gerada, um pressuposto que nem sempre é verdadeiro.

2.5. Descoberta de Conhecimento em Base de Dados – KDD (Knowledge Discovery in Databases)

A terminologia descoberta de conhecimento em banco de dados (*do inglês, KDD – Knowledge Discovery in Databases*) foi proposto no primeiro *workshop* de KDD em 1989 para enfatizar que o produto final do processo de descoberta em banco de dados era o conhecimento (Fayyad *et al.*, 1996b). KDD tornou-se, então, uma área interdisciplinar específica que surgiu em resposta à necessidade de novas abordagens e soluções para viabilizar a análise de grandes e complexas bases de dados (Romão, 2002).

O processo de KDD é um método não trivial de identificação a partir de dados, de padrões que sejam válidos, novos, potencialmente úteis e compreensíveis (Fayyad, 1996b). KDD é um processo geral de descoberta de conhecimento composto por diferentes etapas, incluindo: preparação dos dados, busca de padrões, avaliação do conhecimento e refinamentos. O termo não trivial significa que envolve algum

mecanismo de busca ou inferência probabilística, e não qualquer processamento de dados direto de uma quantidade pré-definida (Romão, 2002).

A sustentabilidade da aplicação de KDD depende de aspectos práticos e técnicos. O aspecto prático inclui considerações sobre o impacto que a aplicação irá provocar, medido por critérios como rendimento, redução de custos etc. Em aplicações científicas, o impacto pode ser medido pela novidade e qualidade do conhecimento descoberto bem como pelo aumento da automação de processos de análises manuais. O aspecto técnico se refere à disponibilidade de dados suficientes, ou seja, varia de acordo com a complexidade do problema. É possível obter grande quantidade de atributos e casos, como também muitos atributos podem ser irrelevantes para o problema tratado. Então, tanto nos aspectos práticos quanto nos técnicos, o conhecimento do domínio da aplicação prepara a identificação para que ocorra a relação de dependência entre os atributos e define-se a utilidade do usuário, que poderá contribuir para a agilidade e redução da busca na tarefa de mineração de dados (MD) e, por conseqüência, reduzir o tempo necessário para a consecução das demais etapas do processo de KDD.

Muitas vezes, a aplicação de KDD se depara com situações onde prevalece base de dados grandes ou poucos dados, muitas dimensões, mudança nos dados, dados com ruído ou incompletos, interações complexas entre atributos etc. Neste contexto, é necessário organizar os dados para promover a extração de conhecimento a partir dos dados organizados.

A solução na organização dos dados pode ser na aplicação da construção de *Data Warehouse* (DW), que permite armazenar informações, anteriormente dispersas, através da identificação, compreensão, integração e agregação dos dados, de forma a posicioná-los nos locais mais apropriados, visando atender a estratégia organizacional

das empresas (Brackett,1996). Para a extração de conhecimento dos dados organizados são utilizadas ferramentas conhecidas como mineração de dados (MD), que podem incorporar técnicas estatísticas, inferências probabilísticas e/ou de IA, capazes de fornecer respostas para descobrir novos conhecimentos em grandes bases de dados.

Para inferir conhecimento que seja significante é importante estabelecer metas bem definidas. Segundo Fayyad e colaboradores (1996b), no processo KDD as metas são definidas em função dos objetivos na utilização do sistema, podendo ser de dois tipos básicos: verificação ou descoberta (Figura 1). Quando a meta é do tipo verificação, o sistema está limitado a verificar hipóteses definidas pelo usuário, enquanto que na descoberta o sistema encontra novos padrões de forma autônoma tais como a previsão e a descrição (Romão, 2002).

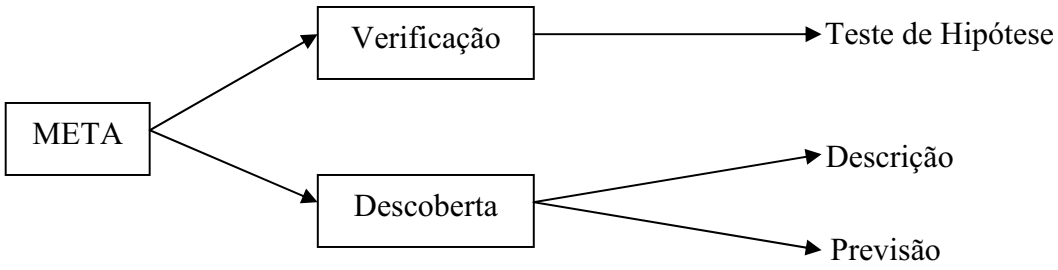


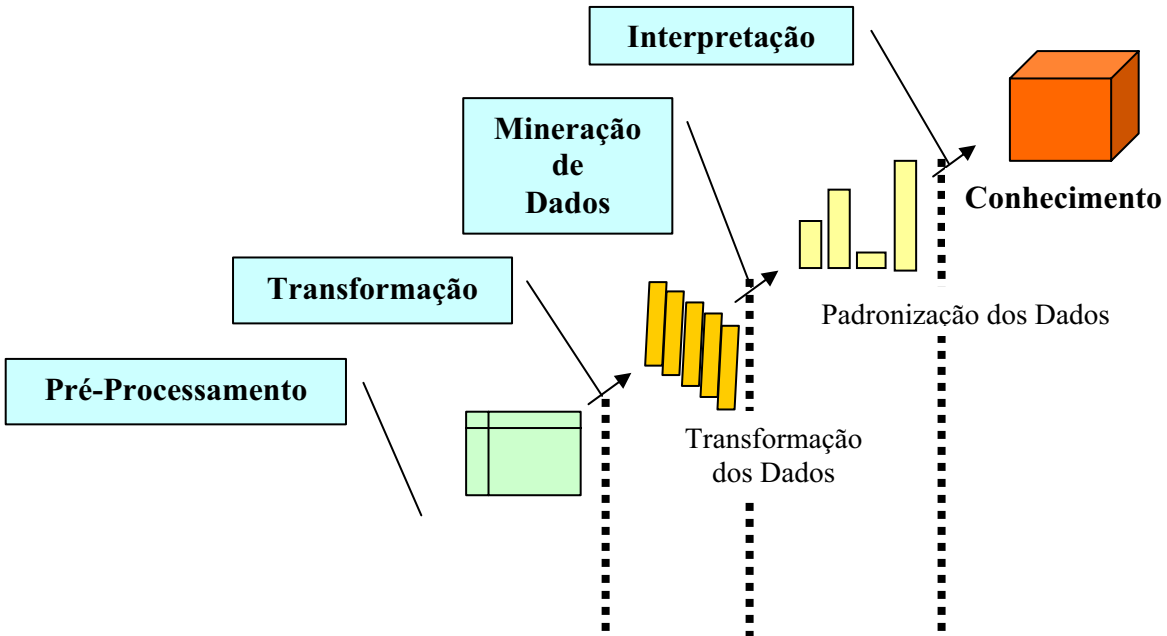
Figura 1. Escopo para definição de metas no processo KDD para a extração de conhecimento em bases de dados.

A descrição procura encontrar padrões, interpretáveis pelos usuários, que descrevem os dados de forma compreensível pelo homem. A previsão parte de diversas variáveis para se supor outras variáveis ou valores desconhecidos (Fayyad *et al.*, 1996a). As metas de previsão e descrição são alcançadas através de algumas tarefas de MD, que podem incluir a classificação, regressão, agrupamento, sumarização,

modelagem de dependência e identificação de mudanças e desvios, sendo a tarefa de classificação a mais empregada (Romão, 2002).

Segundo Fayyad e colaboradores (1996b), o processo geral de KDD é definido nas etapas que se seguem e as etapas podem ser observadas na Figura 2:

- desenvolver compreensão do domínio da aplicação, identificar o tipo de conhecimento que interessa, e identificar a meta do processo de KDD a partir do ponto de vista do usuário;
- realizar pré-processamento incluindo operações básicas, como a seleção de atributos relevantes, remoção de ruído, tratamento da ausência de valores de atributos e conversão de dados categóricos ou contínuos;
- reduzir os dados em função do objetivo da tarefa;
- escolher a tarefa de MD baseado no objetivo do processo de KDD;
- escolher o algoritmo de MD apropriado;
- realizar a MD propriamente dita;
- interpretar os padrões descobertos podendo retornar para um dos passos anteriores;
- consolidar os conhecimentos descobertos, incluindo a conferência e a solução de possíveis conflitos com conhecimentos identificados anteriormente.



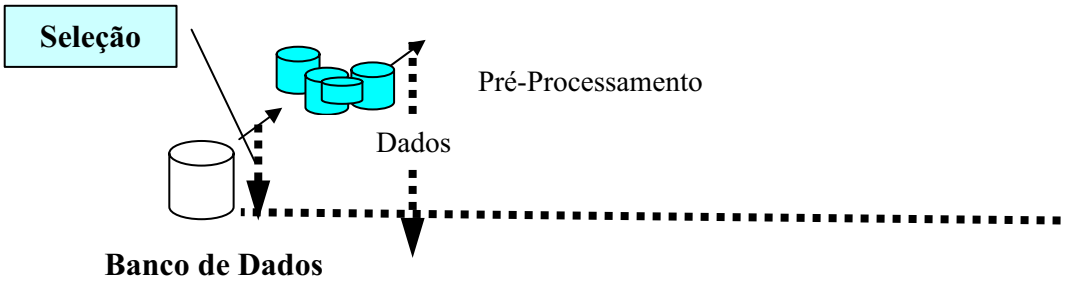


Figura 2. Esquema ilustrativo das etapas do processo de KDD, que viabilizam a análise de grandes e complexas bases de dados.

2.5.1 Características dos Dados

Características como número de registros, tipos de dados e quantidade de erros são essenciais para a aprendizagem. Mediante experimentos realizados por Rendell e Cho (1990) revelaram que algumas destas características afetam drasticamente a exatidão da aprendizagem de conceitos.

Em alguns casos, as características dos dados interagem entre si de uma forma não intuitiva. A ausência de ruído nos dados pode degradar a exatidão de classificação de forma diferente dependendo do tamanho do conceito. Comparada com os efeitos de algumas características dos dados, a escolha do algoritmo de aprendizagem se torna menos importante (Romão, 2002).

Rendell e Cho (1990) verificaram que é adequado considerar um conceito como uma função ou superfície, dentro de um limite de espaço de determinados valores possíveis de atributos, para avaliar seus efeitos na descoberta de conhecimento.

De acordo com Romão (2002), duas características podem ser enfatizadas: tamanho e concentração do conceito. O tamanho representa a proporção de exemplos positivos (por exemplo, o paciente possui anomalia cromossômica) e a concentração caracteriza a distribuição de exemplos positivos dentro do espaço de valores possíveis. A alta concentração é definida por serem poucas regiões de conceitos; baixas

concentrações são conceitos distribuídos em muitas regiões. A Figura 3 representa a variação das proporções sobre os conceitos de distribuição que caracteriza os possíveis valores que as variáveis podem receber.

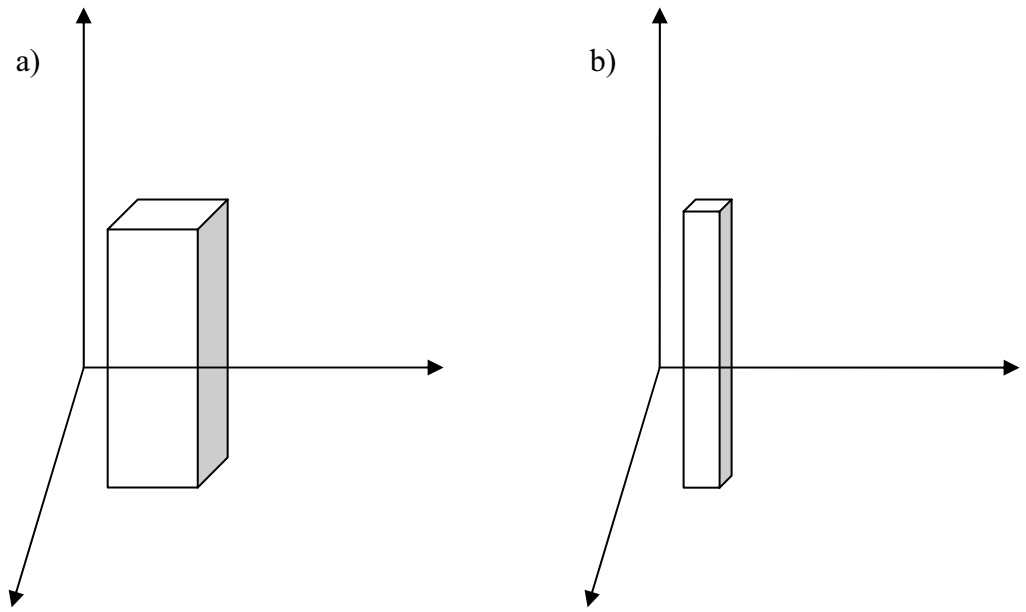


Figura 3. Variação do tamanho relativo do conceito (positivos).

Pela Figura 3 se observa que há maior probabilidade de um ruído interferir no resultado de (a) do que de (b) devido à sua maior área de abrangência. No entanto, (b) representa dados típicos encontrados em alguns problemas. Por exemplo: Casos em que foi identificado Síndrome de Turner em pacientes com sinais/sintomas de tórax em escudo, com unhas hipoplásicas e pescoço alado[2], com 100% de confiança e 18,4% de suporte de decisão[3].

[2] Para chegar a essa conclusão foi usado como parâmetro o valor de significância mínimo de suporte 7 % com um valor mínimo de confiança de 100%, e identificados 7 registros. UnBMiner foi a ferramenta usada para obter tais informações, usando o programa CNM (*Combinatorial Neural Model*).

[3] Resultados extraído do software UnBMiner usando o programa modelo combinatório neural. Os resultados partiram dos dados que constam no anexo II.

2.5.2 Pré-Processamento

Para a eficiente aplicação das técnicas de MD é necessário antes realizar uma preparação dos dados, conhecida como pré-processamento, Wang e Sundaresh, (1998) propõem que o pré-processamento inclua as seguintes etapas:

- integração de dados que consiste em remover inconsistências nos nomes ou nos valores de atributos de diferentes origens;
- limpeza de dados para detectar e corrigir eventuais erros, substituir valores perdidos etc;
- conversão de dados nominais ou em códigos, para números inteiros;
- redução do domínio (valores possíveis) para reduzir a distribuição dos valores no espaço de valores originalmente possíveis;
- construir ou derivar novos atributos;
- discretização para transformar atributos contínuos em categóricos;
- seleção de atributos: escolher atributos relevantes para a tarefa em questão.

Um dos maiores índices de insucessos dentro da MD são dados de má qualidade. Quando os dados são precários o produto de qualquer tarefa de MD também se torna precário (Romão, 2002).

Prevalecem dois tipos básicos de erros em dados: sistemáticos e não-sistemáticos (ou ruído). Os erros sistemáticos (i.e., falha na calibração de equipamento) são introduzidos de forma previsível e são potencialmente detectáveis e corrigíveis. No entanto, erros não-sistemáticos são introduzidos de forma imprevisível e são muito difíceis de detectar e corrigir. O ruído dos dados frequentemente observados incluem a informação equivocada; fornecida pelo usuário e inconsistência nos valores (Romão, 2002).

É possível também identificar situações onde prevalece a falta de alguns valores de atributos. Neste caso, a causa pode estar na coleta dos dados, remoção de dados devido à inconsistência ou no significado incompreensível do atributo. No caso de ausência de valores de atributos, o usuário possui as seguintes alternativas:

- remover registros em que faltem valores;
- prever o valor que com base nos valores de outros atributos;
- lidar com os valores ausentes dentro do algoritmo de MD;
- substituir os valores ausentes pela moda (valor mais freqüente, no caso de atributos categóricos) ou pela média ou mediana (no caso de valores contínuos).

2.5.3 Mineração de Dados

De acordo com (Wong e Leung, 2004), dado o crescimento explosivo dos dados coletados no ambiente de negócio atual, a mineração dos dados pode definir um potencial descobrimento do novo conhecimento para melhorar a tomada de decisão a nível gerencial. A mineração de dados é a tarefa de extração de informação válida, não conhecida previamente, compreensível e efetiva a partir de grandes coleções de dados e úteis para tomada de decisões (Simoudis, 1996). Segundo Fayyad e colaboradores (1996 b), prevalece uma diferença nos termos MD e KDD destacando que o componente de MD se refere apenas ao meio pelo qual padrões são extraídos e enumerados a partir dos dados. O KDD envolve a avaliação e interpretação dos padrões para decidir o que é conhecimento e o que não é. O KDD inclui a escolha do esquema de codificação, pré-processamento, amostragem e projeções realizadas antes da etapa de MD, bem como o pós–processamento naturalmente realizado depois da etapa de MD.

Para extrair conhecimentos consideráveis, existem diversas técnicas de MD disponíveis na literatura (Chen *et al.*, 1996; Cheung *et al.*, 1996). Essas técnicas podem ser aplicadas na associação, classificação e previsão em geral e também na determinação e análise de agrupamentos. As principais podem ser exemplificadas como:

- indução e/ou extração de regras;
- redes neurais;
- algoritmos evolucionários e/ou genéticos;
- técnicas estatísticas (classificadores e redes bayesianas etc.); e
- conjuntos difusos.

Segundo Fayyad e colaboradores (1996b), não há um método de mineração de dados 'universal' e a escolha de um algoritmo particular para uma aplicação particular é de certa forma uma arte.

Os algoritmos de MD diferem primariamente nos critérios utilizados para avaliar o modelo e/ou no método de busca utilizado. Não há critérios estabelecidos para se decidir quais métodos devem ser usados em dada circunstância e que muitas abordagens são aproximações heurísticas para evitar o alto custo de processamento que seria necessário para se encontrar soluções ótimas.

Três componentes primários podem ser identificados em algoritmos de MD: a) representação do modelo que é a linguagem utilizada para descrever os padrões a serem descobertos; b) critério de avaliação do modelo que corresponde à afirmação quantitativa (ou função de aptidão) da qualidade que um padrão específico possui (um modelo e seus parâmetros) em alcançar as metas do processo de KDD e, finalmente, que o método de busca é constituído por dois componentes (busca de parâmetros e busca do

modelo). Após a escolha da representação e do critério de avaliação do modelo, o problema de MD fica reduzido à tarefa de otimização, correspondendo ao encontro dos parâmetros/modelos que satisfaçam o critério de avaliação.

2.5.4 Pós-Processamento

A etapa do pós-processamento é utilizada principalmente para avaliar o processo de descoberta, melhorar a compreensão e/ou selecionar o conhecimento descoberto que seja mais relevante.

2.5.4.1 Avaliação do Processo de Descoberta

Algumas abordagens que servem para avaliar o KDD incluem a exatidão dos resultados (i.e., alguma medida da taxa de acerto), a eficiência (tempo de processamento), facilidade de compreensão do conhecimento extraído. A maior parte da literatura utiliza acurácia como o principal meio para avaliar as técnicas de KDD (Freitas, 1997), principalmente no contexto da tarefa de classificação.

Os dados utilizados para efetuar a extração de conhecimento são divididos em dois grupos exclusivos: conjunto de treinamento (equivalente a 2/3 dos dados) e conjunto de teste (1/3 dos dados) este será avaliado. O algoritmo deve descobrir regras acessando apenas os dados de treinamento. Uma vez que o processo de treinamento tenha terminado e o algoritmo tenha encontrado um conjunto de regras de classificação, a performance para estas regras é medida através da aplicação destas regras sobre os dados de teste, caracterizando uma forma de aprendizagem supervisionada.

Em casos de complementação de valores faltantes no conjunto dos dados de treinamento, a moda (média ou mediana) é calculada com base no conjunto de treinamento apenas. Para os dados de teste, eles são calculados com base em toda a

população (King *et al.*, 1995). Quando há poucos dados disponíveis, pode-se empregar todos os dados na fase de treinamento e realizar validação cruzada para calcular a taxa de acerto na fase de teste.

O método de validação cruzada (Baranauskas, 2001) possui a vantagem de utilizar todos os dados para treinamento e para teste, mantendo a separação entre os dados de treinamento e de teste. No entanto, o método de validação cruzada exige maior desempenho computacional quando comparado com a forma convencional, que utiliza apenas um conjunto de dados de teste, apresentando, portanto, um maior conjunto de recursos computacionais.

2.6 Metodologia CRISP-DM

As empresas DaimlerChrysler, a SPSS e a NCR criaram o consórcio CRISP-DM – *CRoss Industry Standard Process for Data Mining* e propuseram um modelo de referência para o processo de mineração de dados (Chapman *et al.*, 1999) não proprietário e disponível sem custos (Ladeira *et al.*, 2005).

O CRISP-DM é definido como modelo de processo que fornece uma estrutura de dados no qual são realizadas minerações de dados aplicadas aos setores de indústria, pesquisa, ciência e tecnologia (Wirth e Hipp, 2000).

O modelo CRISP-DM tem como fator positivo a não dependência da área de negócio e da tecnologia a ser utilizada na MD, além da fácil aplicação, custos mais baixos, viabilidade e facilidade da gestão dos projetos de alta ou baixa envergadura de MD.

Segundo (Hipp *et al.* 2002), a mineração dos dados é um processo complexo que requer várias ferramentas e situações diferentes. O sucesso de um projeto depende da conjectura apropriada de ferramentas auxiliares confiáveis e de especialistas hábeis que

conheçam bastante o problema. Além disso, requer uma metodologia para ser desempenhada e uma gerência de projeto eficaz. Um modelo de processo pode ajudar a compreender e controlar as interações ao longo deste processo complexo. Neste contexto, quando existem evidências de casos onde há falhas na formalização dos dados, causa-se o prejuízo nas ações para uma tomada de decisão.

Os especialistas que executam projetos baseado na mineração dos dados podem também beneficiar-se de outras maneiras. No caso de um especialista com pouca experiência, a modelagem CRISP-DM fornece a orientação, ajuda a estruturar o projeto e fornece o conceito para cada tarefa do processo. Já os especialistas experientes se beneficiam usando as listas de verificação para cada tarefa, certificando que nada de importante será esquecido. Mas o papel mais importante na aplicação do CRISP-DM é a comunicação e a documentação dos resultados. Assim, a metodologia ajuda vincular ferramentas diferentes com habilidades e fundos diversos para dar forma a um projeto eficiente e eficaz (Wirth e Hipp, 2000).

Atualmente, o CRISP-DM é considerado um padrão de facto de modelagem pela comunidade internacional em MD, permitindo que o processo de mineração possa ser realizado de forma mais rápida, confiável e com maior controle a nível gerencial (Wirth e Hipp, 2000). Figura 4 descreve as etapas da metodologia CRISP-DM que são altamente representativa e utilizados no mercado mundial.

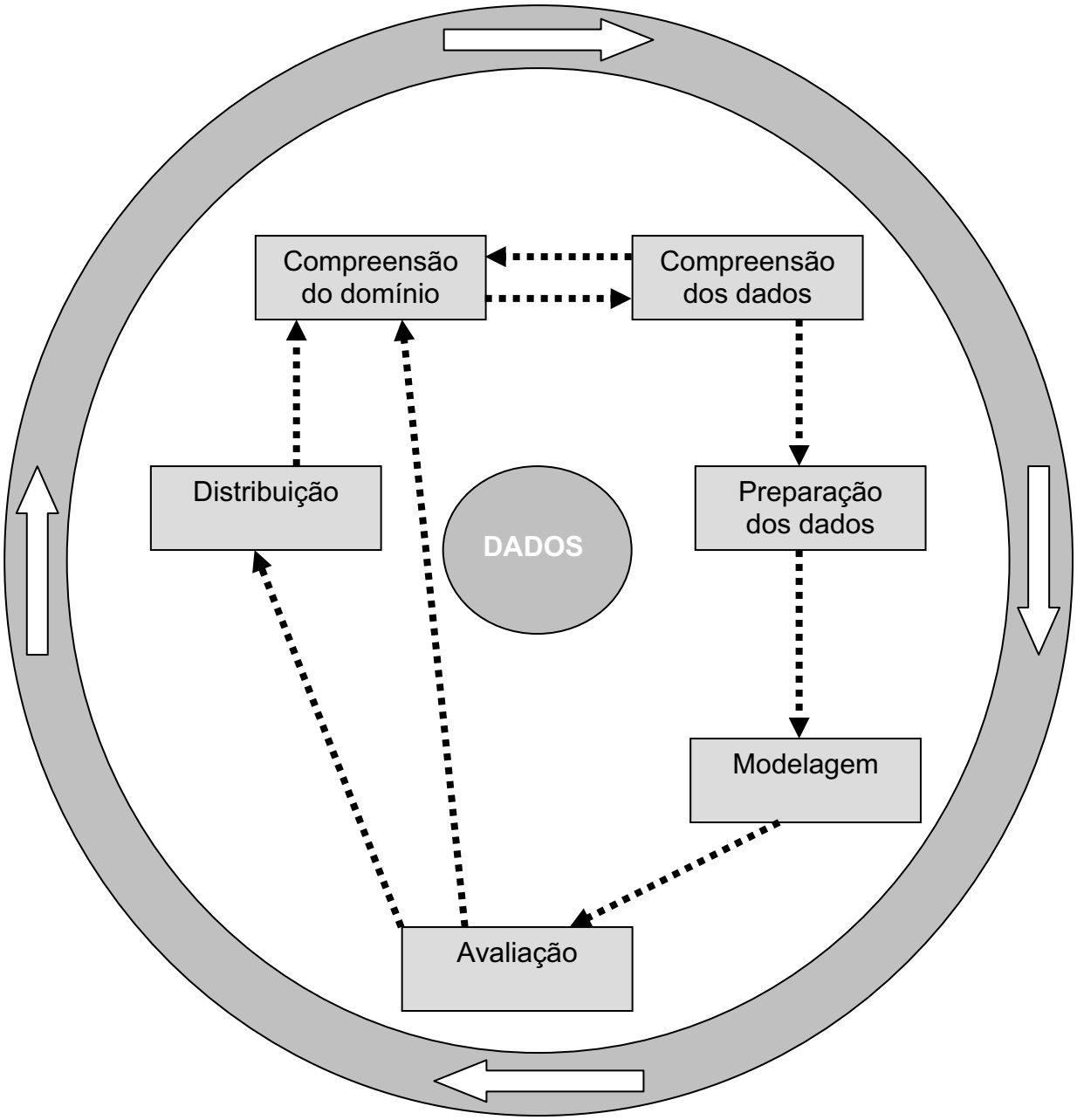


Figura 4. O modelo de processo previsto pelo consórcio CRISP-DM pode ser resumido através do ciclo de vida do processo de mineração de dados.

Quadro VI - Ciclo da metodologia do CRISP-DM

	AÇÃO	ÊNFASE	RESULTADO	TAREFAS
FASE 1	Entendimento do negócio (ou do domínio da aplicação)	Entendimento dos objetivos e requerimentos do projeto, da perspectiva do domínio, a relevância do conhecimento prévio e os objetivos do usuário final.	Plano do projeto	Avaliar recursos, criar cronogramas e listas de requisitos, identificar os riscos e eventos, planos de contingência, estabelecer um glossário da terminologia.
FASE 2	Entendimento dos dados	Selecionar o conjunto de dados (variáveis ou amostras de dados) para a MD.	Identificar problemas de qualidade dos dados, descobrir os primeiros conhecimentos ou detectar subconjuntos dos dados interessantes	Obter mais familiaridade com os dados, fazer uma descrição dos dados (formato, quantidade de registros e campos), distribuição dos atributos, relacionamentos entre pares de atributos, e identificação de agrupamentos nos dados.
FASE 3	Preparação dos dados	Cobre as atividades necessárias para a construção do banco de dados final	Seleção de atributos, limpeza, construção, integração e formatação dos dados de entrada.	Remoção de ruído ou de dados espúrios, definição de estratégias para lidar com valores faltantes, reformatação dos dados para o tipo da ferramenta a utilizar, criação de atributos derivados e de novos registros, integração de tabelas, e discretização dos dados numéricos.
FASE 4	Modelagem	Definição da melhor técnica para a preparação dos dados, identificados como fase de treinamento.	Definição de modelos e parâmetros ideais para a MD.	Permitir avaliar os modelos gerados.
FASE 5	Avaliação	Consiste em avaliar a qualidade dos modelos obtidos no treinamento	Seleção de modelos.	Verificar se objetivos do negócio foram atingidos de acordo com os critérios de sucesso adotados
FASE 6	Colocação em uso	Consolidar o conhecimento descoberto com o modelo criado	Gerar um relatório final ou implementação um processo repetitivo de mineração de dados na organização	Obter relatórios contendo informações de tempo gasto, custo envolvido, experiências que aconteceram ao longo do projeto, tarefas que não puderam ser feitas, trabalho futuro para melhorar os resultados.

Pelo Quadro VI é identificado o ciclo da metodologia CRISP-DM que especifica e detalha cada fase para planejar e realizar um projeto de MD. Usando as fases do CRISP-DM é possível dispor de instrumentos viáveis para projetar uma modelagem de dados, além de obter uma tomada de decisão mais segura, pois tem como objetivo gerir e atuar sobre a evolução da abordagem dos dados, porque permite compreender melhor as razões dos acontecimentos passados e presentes e projetar os cenários futuros mais previsíveis.

Capítulo 3

3. Algoritmos de Aprendizagem de Redes Bayesianas

Aprendizagem, no contexto das redes bayesianas, é um processo que *a priori* possui como entrada um conjunto de dados e informação e como saída *a posteriori* visualiza uma rede bayesiana. (Friedman *et al.*, 1997).

Atualmente predomina um interesse na comunidade científica em desenvolver técnicas capazes de gerar redes bayesianas com uma menor interação com um especialista, buscando, principalmente, reduzir o tempo de construção da mesma. Tais técnicas são justificadas pela massiva quantidade de dados disponível sobre um determinado domínio. Neste contexto, prevalece recente progresso das técnicas de extração de conhecimento de bases de dados, principalmente as técnicas de análise estatística (Herckerman, 1995b).

Nos quadros VII e VIII seguinte, são sugeridas algumas definições e explanações de algoritmos baseados em propagações de evidências utilizando a rede bayesiana. Os algoritmos pertencem a dois tipos de categoria. a) método de busca e pontuação e b) análise de dependência.

3.1 Método de Busca e Pontuação

Em se tratando desse método, os algoritmos visualizam o problema "aprender" como um problema de busca pela estrutura que melhor se represente os dados. O processo inicia-se com um grafo sem arcos e usam algum modelo de busca para adicionar um arco ao grafo. Depois, utiliza-se de um método de pontuação para identificar se a nova estrutura é melhor do que a anterior. Se a nova estrutura for melhor do que a anterior, então mantém o novo arco adicionado e tentam adicionar o próximo.

Esse processo se repete até que nenhuma estrutura nova seja melhor do que a estrutura atual (Cheng *et al.*, 1998).

De acordo com (Cheng *et al.*, 1997), devido a sua natureza heurística[4] os algoritmos que utilizam o método de busca e pontuação não garantem encontrar uma solução ótima. E trabalham com um espaço de modelos probabilísticos maiores do que o enfoque de análise de dependência.

A maioria das técnicas atuais para gerar uma rede bayesiana a partir de dados assume que a base de dados esteja completa, ou seja, não possuem variáveis com valores ausentes ou omissos. Quando essa suposição falha, estes métodos utilizam técnicas estatísticas para adivinhar o dado ausente, ou aproximações assintóticas para estimar a probabilidade condicional que define a rede (Ramoni e Sebastini, 1998).

Os métodos mais conhecidos que realizam a tarefa de estimar o(s) dado(s) ausente(s) na base de dados são o algoritmo EM, isto é, *Esperança e Maximização*, ou os Métodos de Monte Carlo via Cadeias de Markov (MCMC), como amostragem de *Gibbs*. A estratégica básica que permeia estes métodos está baseada no Principio da Informação Ausente (*do inglês, Missing Information Principle*), o qual preenche a informação ausente com base na informação disponível. Estes métodos de aproximação estão sujeitos a erros quando poucas informações ausentes demandam altos recursos, convergem muito lentamente, e seu tempo de execução é alto dependendo do número de dados ausentes (Buntime, 1994b; Madigan e York, 1995; Dempster *et al.*, 1977; Lauritzen, 1995; Meilisjon, 1989). O quadro VII sumariza os modelos de algoritmos utilizando métodos baseados em busca e pontuação aplicados em sistemas de apoio à tomada de decisão.

[4] Heurística é um conjunto de regras empíricas e métodos que visam à descoberta, à invenção ou à resolução de problemas (FERREIRA, 2002)

3.1.1 Métodos Baseado em Busca e Pontuação

Quadro VII - Modelos de algoritmos baseado no método busca e pontuação.

Algoritmo	Modelos	Requer ordem dos nós?	Registro dos métodos	Principais Características	Referências Bibliográficas
Chow-Liu	Árvores/ Rede de Markov	Não	Entropia	- Conjunto de dados históricos –entrada. - Estrutura de árvores – saída. - Estrutura de busca com a melhor contagem (cruzamento de dados desordenados de Kullback-Leibler). - Não utiliza a busca heurística. - Ineficiência em construir gráficos múltiplos conectados em análises de dependência. - Necessita apenas de $O(N^2)$ para calcular dependência.	Chow e Liu, 1968 Cheng *et al.*, 2000
Rebane-Pearl	Poliárvores / Rede bayesiana	Não	Entropia	- Necessita apenas de $O(N^2)$ para calcular dependência. - Extensão direta do algoritmo de Chow-Liu. - Sua estrutura não contém ciclos. - Define no máximo um caminho entre dois nodos quaisquer do grafo. - Define um método para encontrar a direcionalidade dos arcos. - Indica colisões de grafos. - A estrutura básica da poliárvore é gerada pelo método Chow-Liu. - Aplica-se uma estrutura de recuperação de poliárvore, obtendo a representação gráfica da distribuição.	Rebane e Pearl, 1987 Cheng *et al.*, 1997
K2	Rede bayesiana	Sim	Bayesiana	- É o mais representativo dos algoritmos baseados em busca e pontuação para aprendizagem em redes bayesianas. - Conjunto de dados – entrada. - Constrói uma estrutura de rede bayesiana – saída. - Encontra a estrutura da rede bayesiana mais provável. - Existem $2^{N(N1)/2}$ estruturas possíveis para representar um problema com N variáveis. - As variáveis da base de dados são discretas. - Não há casos que tenham variáveis com dados perdidos. - Todas as estruturas são igualmente prováveis no início. - Para cada nó, busca um conjunto de pais que maximiza a pontuação do nó, de acordo com a equação:	Cooper e Herskovits, 1992 Cheng *et al.*, 1998 Bouckaert, 1995 Hruschka, 1996 Larrañaga, 2002

Continuação do Quadro VII

				$g(i, Pai) = \prod_{j=1}^{qi} \frac{(ri-1)!}{(Nij+ri-1)!} \prod_{k=1}^{ri} Nijk!$ - Começa assumindo que um nó não tem pai. - A cada etapa adiciona um nó-pai se houver ganho definido por $g(i, Pai)$, maior que $g(i, Pai)$ anterior. - Algoritmo eficiente. - Utiliza métodos de busca heurística.	
HGC	Rede bayesiana	Não (requer *a priori* uma rede)	Bayesiana	- Usa *a priori* uma rede de domínio do conhecimento. - É um algoritmo baseado em pontuações bayesianas. - Utilizam os métodos chamados modularidade de parâmetro e equivalência de eventos, obtidos a partir de uma combinação do conhecimento de usuários e de dados estatísticos.	Heckerman *et al.*, 1994
Kutato	Rede bayesiana	Sim	Entropia	- Utiliza a "técnica gulosa" entre estruturas de rede. - Seleciona-se as estruturas com a menor entropia associada. - Representa a distribuição de probabilidades mais expressiva. - Utiliza uma estrutura que adota a mínima perda de informação. - O método de busca usado é similar ao usado no Algoritmo K2.	Herskovits and Cooper, 1991 Cheng *et al.*, 1998a Cooper e Herskovits, 1992
BENEDICT	Rede bayesiana	Sim	Entropia	- Emprega um método de busca heurística. - Analisa a independência condicional implícita na estrutura pela utilização do conceito de *d-separation*. - Calcula a diferença entre a independência condicional implícita e a independência condicional dos dados. - Usa uma pontuação de entropia diferente para medir a aproximidade entre uma estrutura aprendida e os dados. - Utiliza uma pontuação de entropia que é a soma dos resultados de um grupo de testes de independência condicional. - Método baseado em contagem de métrica.	Acid e Campos, 1996b Pearl, 1988

Continuação do Quadro VII

CB	Rede bayesiana	Não	Bayesiana	- Possuem habilidades para orientar arcos. - É um algoritmo híbrido que emprega ambos os métodos, análise de dependência e busca e pontuação. - Primeiro emprega uma versão modificada do Algoritmo PC (descrito no Quadro VIII) para encontrar uma ordenação entre as variáveis. - Usa uma versão modificada do Algoritmo K2 para aprender a rede bayesiana.	Singh e Valtorta, 1995 Spirtes *et al.*, 1991 Cooper e Herskovits, 1992
Suzuki	Rede bayesiana	Sim	MDL (*Minimum Description Length)*	- Não utiliza uma busca heurística. - Garante que uma estrutura ótima seja encontrada. - Utiliza a técnica "ramificar e podar" que calcula uma ramificação mínima depois da adição de um arco à estrutura e determina se uma busca adicional neste ramo é necessária. - Testes realizados com até quinhentos mil casos mostra uma eficiência do algoritmo. - É menos eficiente quando o número de casos cresce em proporções de milhares de casos.	Suzuki, 1996 Cheng *et al.*, 1998
Lam-Bacchus	Rede bayesiana	Não	MDL (*Minimum Description Length)*	- Pode orientar os arcos usando um método de busca e pontuação. - O método é bastante diferente da orientação de arcos baseada em identificação de colisão.	Lam e Bacchus, 1994
Algoritmo *Bound and Collapse* (BC)	Rede bayesiana	Não	Bayesiana	- É um método que adota a medida de pontuação bayesiana. - Baseia-se na mesma idéia do Algoritmo K2 para aprender a estrutura e as probabilidades de uma rede bayesiana a partir de dados possivelmente incompletos. - Extrai o grafo da rede a partir da informação disponível em uma base de dados através de uma versão modificada do procedimento de busca. - Extraem estimativas nas probabilidades condicionais. - Não se baseia no princípio da informação ausente. - É um método escalável, pois a complexidade do algoritmo não depende do número de dados ausentes. - É um método determinístico. -Este processo retorna intervalos probabilísticos contendo todas as possíveis estimativas consistentes com a informação disponível. Estes limites são então colapsados em um único valor via combinação convexa. - A idéia é que uma base de dados incompleta permanece habilitada para restringir as possíveis estimativas em um conjunto. - O padrão assumido para os dados ausentes, codificado com probabilidade de dados ausentes, pode ser usado para selecionar um conjunto de pontos possíveis. - É implementado pelo software *Bayesian Knowledge Discoverer (BKD)*.	Ramoni e Sebastiani, 1996a Ramoni e Sebastiani, 1996b Ramoni e Sebastiani, 1996c Ramoni e Sebastini, 1997a Ramoni e Sebastini, 1997 Ramoni e Sebastiani, 1997b Cooper e Herskovits, 1992 Ramoni e Sebastiani, 1998 Ramoni e Sebastiani, 1999

3.2 Métodos Baseados em Análise de Dependência

A independência condicional na distribuição representada por uma rede bayesiana é codificada na estrutura do grafo e pode ser encontrada usando-se o critério gráfico chamado *d-separation* (Pearl, 1988). Para a classe de algoritmo de aprendizagem baseados em análise de dependência, é assumido que a estrutura da rede representa perfeitamente as dependências e independências no domínio da aplicação, isto é, uma afirmação de independência é representada por uma estrutura se, e somente se, ela for uma independência válida para o domínio.

A validade de uma independência pode ser verificada realizando-se um teste estatístico utilizando uma base de dados sobre o domínio. A idéia básica geral dessa classe de algoritmos é a seguinte (Bouckaert, 1995): a) iniciar com um grafo não orientado sobre V (conjunto das variáveis/nodos)*;* b)remover o arco entre dois nós u e v para o qual um conjunto de variáveis S, contido em $V\backslash uv$, pode ser encontrado tal que u e v sejam condicionalmente independentes, dado S; c) selecionar o arco e nós e associar uma direção ao arco para formar uma estrutura v; d) associar direções aos arcos restantes tal que um grafo orientado acíclico seja formado. A diferença básica entre os vários algoritmos está no modo como os conjuntos de variáveis S são encontrados e nas regras de associação das direcionalidades. O Quadro VIII sumariza os modelos de algoritmos utilizando métodos baseados em análise de dependência aplicados em sistemas de apoio à tomada de decisão.

Quadro VIII - Índice dos algoritmos baseados no método de análise de dependência.

Algoritmo	Modelos	Requer ordem dos nós?	Registro dos métodos	Principais Características	Referências Bibliográficas
Wermuth-Lauritzen	Rede bayesiana	Sim	$O(N^2)$	- Pode ser usado para grandes conjuntos de dados. - Utilizam variáveis em testes. - Trata-se de um algoritmo impraticável.	Wermuth e Lauritzen, 1983 Cheng *et al.*, 1998a
SRA	Rede bayesiana	Ordenação parcial	Exponencial	- Usa o método de busca heurística.	Srinivas *et al.*, 1990
Constructor	Rede Markov	Não	Exponencial	- Utiliza a técnica de validação cruzada para evitar o sobre ajuste.	Fung e Crawford, 1990
SGS	Rede bayesiana	Não	Exponencial	- Pode orientar os arcos. - Pode apresentar todas as estruturas possíveis, baseado na teoria probabilística da causalidade. - Pode ser perfeitamente representado por uma estrutura de rede onde a direção dos arcos é interpretada como influências causais.	Spirtes *et al.*, 1990
Verma-Pearl	Rede bayesiana	Não	Exponencial	- É uma variação de SGS. - Pode orientar os arcos e o detector de conflitos na orientação dos arcos.	Verma e Pearl, 1992
PC	Rede bayesiana	Não	$O(N^{k+2})$	- Pode orientar arcos. - Aumenta a performance do algoritmo SGS. - É um algoritmo eficiente.	Spirtes e Glymour, 1991
CBL	Rede bayesiana	Não	Bayesiana	- Assume que as bases não possuem variáveis com valores ausentes. - Evita os testes de independência condicional com grandes conjuntos condicionais. - Utiliza o menor número possível dos testes de independência condicional. - A primeira versão assume ordenação de variáveis. - A segunda versão não ordenação de variáveis.	Cooper e Herskovits, 1992 Cheng *et al.*, 1997 Cheng *et al.*, 1998a

Capítulo 4

4. Softwares para redes bayesianas

Neste capítulo serão apresentados alguns modelos de softwares que manipulam as redes bayesianas. Será descrito o software no qual o domínio da presente proposta será aplicado, juntamente a metodologia utilizada na aplicação desse software.

4.1 UnBBayes

De acordo com Ladeira e colaboradores (2002), o UnBBayes é um ambiente visual e interativo para a edição e compilação de redes bayesianas (BN), diagrama de influência (ID) ou rede bayesiana múltipla selecionada (MSBN), entrada e propagação de evidências, realização de inferência probabilística e aprendizagem da topologia e/ou parâmetros de uma BN. Ele utiliza o método da árvore de junções (atual estado da arte a nível) para propagação de evidências e os algoritmos K2 (Cooper e Herskovits,1992) e B (Buntine,1991) (baseados em métodos de busca e pontuação) e CBL-A e CBL-B, baseados em análise de dependências para a aprendizagem de redes bayesianas(Cheng *et al.*,1998).

UnBBayes é desenvolvido em Java e documentado com Javadoc e JavaHelp. Através da ajuda ao usuário é possível acessar toda a documentação da API e visualizar todas as funcionalidades. O sistema é distribuído gratuitamente para uso não comercial, sob a licença GNU GPL[5]. O seu desenvolvimento envolveu o uso de técnicas de Extreme Programming[6] e refactorings (Fowler *et al.*,1999) visando obter qualidade de software.

O software UnBBayes apóia a aprendizagem da topologia ou das probabilidades de redes bayesianas com a aplicação de algoritmos de busca e pontuação ou análise de

[5] General Public License (http://www.gnu.org/licenses/gpl.html)
[6] http://www.extremeprogramming.org/rules.html

dependências em grandes bases de dados. Esse software aberto, de fácil utilização e com suporte lingüístico para internacionalização, apresenta performance comparável ao HUGIN®, líder mundial desse segmento.

Para mais informações sobre este software vejam o *home site* do *software* UnBBayes pelo endereço eletrônico: https://sourceforge.net/projects/unbbayes

4.2 UnBMiner

Conforme Ladeira e colaboradores (2005), o UnBMiner é um ambiente visual e interativo, programado em Java, independente de plataforma, para realização do processo de mineração de dados. O *framework* automatiza grande parte do processo CRISP-DM[7], do inglês *Cross-Industrial Standart Process for Data Mining*. Mais especificamente ele abrange parte da fase de preparação dos dados (módulo pré-processamento do UnBMiner), fase de modelagem (criação de modelos baseados nos formalismos Naïve Bayes, CNM, árvore de decisão – algoritmos ID3 e C4.5 – e rede neural multicamada com retropropagação) e fase de avaliação (módulo de avaliação do UnBMiner). Inicialmente o sistema desenvolvido baseou-se na ferramenta de mineração de dados WEKA[8] como inspiração para a modelagem e também para validação do programa desenvolvido, visto que é ferramenta de domínio público e bastante divulgada no meio acadêmico. No entanto, o WEKA não utiliza estruturas de dados adequadas, exigindo um alto custo de memória, e também tendo como conseqüência algoritmos não otimizados. A linguagem JAVA foi escolhida devido à sua independência de plataforma, orientação a objetos, diversos recursos implementados, facilidade para manutenção e internacionalização.

---

[7] CRISP-DM, é uma metodologia (DELMATER; HANCOCK, 2001, p.61;CHAPMAN *et al.*, 1999 *apud* COSTA SOUSA, 2003, p.47; OLIVEIRA;ALVARENGA, 2003, p.02) foi concebida em 1996, como um guia passo a passo, para a mineração de dados, e propõe um modelo gratuito de processo padrão, para mineração de dados (SPSS, 2000).

[8] http://www.cs.waikato.ac.nz/~ml/weka

Para mais informações sobre este software vejam o home site do software UnBMiner pelo endereço eletrônico: https://sourceforge.net/projec/unbbayes.

4.3 WEKA® - *Waikato Environment for Knowledge Analysis*

Conforme (Frank *et al.*, 2004), a ferramenta de MD WEKA© fornece algoritmos para diferentes tipos de problema. Na área da bioinformática, foi utilizado para a anotação automatizada da proteína (Kretschmann *et al.*, 2001; Bazzan *et al.*, 2002), na seleção para as disposições dos genes (Tobler *et al.*, 2002), experiências na automatização com o diagnóstico do câncer (Li *et al.*, 2003), descrição dos genótipos dos indivíduos (Taylor *et al.*, 2002) e classificação de perfis genéticos (Li e Wong, 2002).

A filosofia que WEKA® propaga é prover o suporte ao pesquisador definindo regras para a propagação de dados, sendo possível a geração de uma aprendizagem de máquina (Holmes *et al.*, 1994). O *software* WEKA® possui suas limitações junto a uma relação das buscas de soluções para os problemas apresentados aos usuários, devido a grande complexidade que são os domínios dos problemas. O WEKA® é definido por módulos, como: pré-processamento de dados, aprendizagem de máquina e processamento de dados para saída.

Nenhum algoritmo único é suficiente para resolver todos os problemas em se tratando em MD. O objetivo no desenvolvimento do WEKA® é permitir um máximo da flexibilidade ao dispor diversos métodos de aprendizagem de máquina para aplicação em MD. Isto inclui algoritmos de aprendizagem que definam tipos diferentes de modelos (árvores da decisão e discriminantes lineares), método de pré-processamento (discretização, transformações matemáticas arbitrárias e combinações dos atributos). Fornecendo métodos que estão disponíveis através de uma relação comum, WEKA®

facilmente compara as estratégias diferentes de soluções baseadas no mesmo método de avaliação e identifica o problema. Todas as técnicas de aprendizagem no WEKA® podem ser acessadas por linha de comando, como também por comandos de *scripts,* ou usando as API Java, pois o WEKA® é desenvolvido em linguagem de programação JAVA. WEKA® também contém uma versão alternativa usando interface gráfica, chamado de "Conhecimento de Fluxo" (Frank *et al.*, 2004).

Para mais informações sobre este software vejam o *home site* do *software* WEKA® pelo endereço eletrônico: http://www.cs.waikato.ac.nz/~ml/index.html

4.4 HUGIN®

É uma ferramenta que possibilita a construção de redes bayesianas e a propagação de evidências nessa rede (Hruschka Jr., 1997). O *software* HUGIN® (Madsen *et al.*, 2002) é uma ferramenta de uso geral para modelos gráficos probabilísticos tais como redes bayesianas, como também uma ferramenta de auxílio aos especialistas, pois trata a descoberta de conhecimento, estrutura de aprendizagem, parâmetros de estímulos e adaptação, e parâmetros na rede bayesiana.

A ferramenta HUGIN® já possui um reconhecimento significativo no mercado de *softwares* no mercado de soluções inteligentes. Estão atuando na área de soluções inteligentes há 15 anos e atualmente possui três versões de uso: comercial, acadêmico e de serviços. Todas as licenças são comerciais.

Para mais detalhes sobre este software vejam o *home site* do *software* HUGIN® pelo endereço eletrônico: http://www.hugin.com/

Capítulo 5

5.1 Síndrome de Turner: Referencial Teórico

Em 1786, o anatomista Giovanni Morgagni, durante a autopsia, descreveu uma mulher com baixa estatura, malformação renal e disgenesia gonadal. Em 1902, Funke relatou a ocorrência de disgenesia gonadal, baixa estatura, ausência de puberdade, linfedema congênito e pescoço alado em uma garota de 15 anos (Tresh e Rosenfeld, 1995). Em 1930, Ullrich (1930) descreveu algumas características, atualmente observadas na Síndrome de Turner (ST), em suas pacientes, cujo diagnóstico citogenético foi elucidado posteriormente com o advento das técnicas de cariotipagem. Porém, foi somente a partir de 1938 que a condição recebeu a denominação de ST, após a descrição apresentada por Henry H. Turner, um americano endocrinologista de Oklahoma, que estudou sete mulheres com fenótipo característico, que incluíam infantilismo sexual, pescoço alado e cúbitos valgos (Elsheikh *et al.*, 2002 *apud* Lipay *et al.*, 2005). A referida síndrome também é conhecida como Síndrome de Ullrich-Turner (Clemente-Jones, 2000).

5.2 Sinais e Sintomas da Síndrome de Turner

A ST ocorre em aproximadamente 1:2130 nascidos vivos do sexo feminino (Lipay *et al.*, 2005; Nielsen e Wohlert, 1991) e é decorrente da presença de um cromossomo X e perda total ou parcial do segundo cromossomo sexual. Apesar do vasto polimorfismo clínico, considera-se atualmente que baixa estatura, infantilismo sexual e o linfedema periférico são os achados clínicos marcantes na ST. Seu sinal clínico mais evidente e facilmente observado é a baixa estatura (Lippe, 1996), que oscila em média entre 142 e 146,8 cm (Hall e Gilchrist, 1990; Massa e Vanderschueren-Lodeweyckx,

1991). A disgenesia gonadal também é um sinal importante na ST, levando aos sinais secundários comuns como amenorréia primária, atraso no desenvolvimento puberal e esterilidade (Pasquino *et al.*, 1997). Podem também ser observadas algumas anomalias congênitas, incluindo problemas cardiovasculares e renais, deficiência auditiva, osteoporose, tendência à obesidade, e hipertelorismo de mamilos. Grandes variabilidades de sinais dismórficos também são observadas, como pescoço curto e/ou alado, tórax largo e em escudo, cúbito valgo (do latim, *cubitus valgus)*, baixa implantação dos cabelos na nuca, orelhas proeminentes e de implantação baixa, unhas hipoplásticas, estrabismo, pregas epicânticas, entre outras (Lipay *et al.*, 2005; Batch, 2002; Rieser e Underwood, 1992; Rosenfeld, 1992).

5.3 Achados Citogenéticos na Síndrome de Turner

A ST é um resultado da monossomia completa ou parcial do cromossomo X, que resulta em indivíduos fenotipicamente femininos. A ST pode ser identificada ao nascimento, favorecendo o acompanhamento clínico durante a infância. No entanto, a maioria dos casos é identificada pelos especialistas depois do início da puberdade. Dados empíricos estimam que a ST afeta aproximadamente 3% do todos os fetos femininos concebidos (Cockwell *et al.*, 1991). No entanto, a maioria (99%) destas concepções evolui para o aborto espontâneo, geralmente durante o primeiro trimestre de gravidez (Lipay *et al.*, 2005; Frías *et al.*, 2003). Assim, a incidência da perda total ou parcial do cromossomo X varia de 1:2000 para 1:5000 nascidos-vivos, fenotipicamente femininos (Cockwell *et al.*, 1991; Nielsen e Wohlert, 1991). Cerca de 80% das meninas com 45 cromossomos, que apresenta monossomia de X e ST, herdaram maternalmente o único cromossomo X, enquanto que em 20% delas, o cromossomo X é de origem paterna (Hassold *et al.*, 1988). A ocorrência da ST é geralmente um evento esporádico

dentro de uma família e seu risco de recorrência na irmandade é muito baixo (Lipay *et al.*, 2005).

Quando há suspeita clínica de ST, o teste apropriado para a confirmação diagnóstica e o cariótipo com bandas G, que possibilita a identificação e análise do cromossomo. Para a cariotipagem executa-se a cultura de linfócitos, em curto prazo (72 h), obtida do sangue periférico da paciente. Aproximadamente em 50% dos casos, a notação cariotípica clássica é de 45,X, que inclui a monossomia do X, conforme Figura 5.

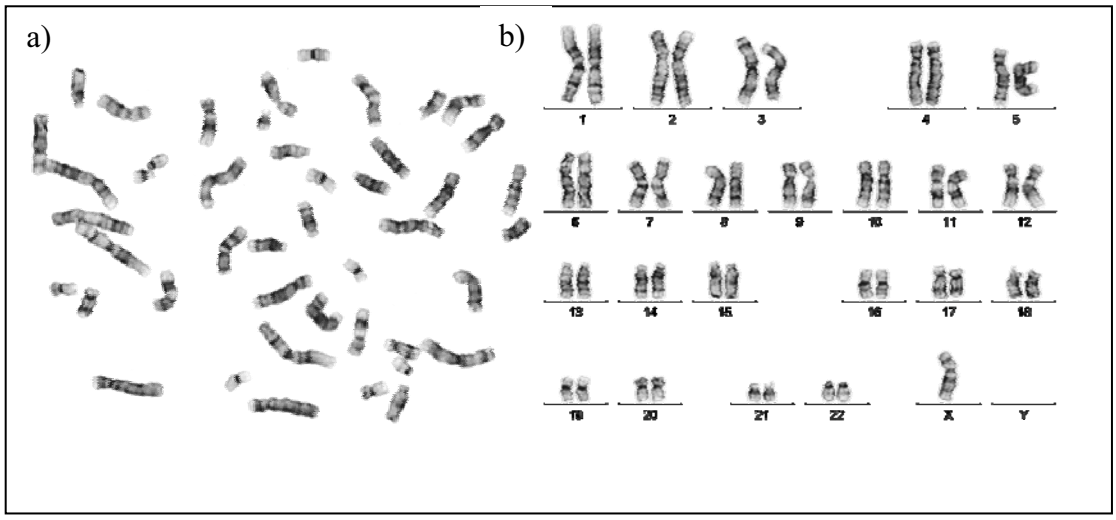


Figura 5. Cariótipo humano exibindo monossomia de X, a alteração cromossômica mais freqüentemente observada na Síndrome de Turner. **A:** Metáfase contendo os cromossomos espalhados, que foram obtidos após a cultura de linfócitos por 72h e corados pelo método de GTG (Tripsina + Giemsa) **B:** Pareamento cromossômico evidenciando a ausência de um cromossomo sexual, resultando na notação cariotípica 45,X, correspondente ao diagnóstico citogenético da Síndrome de Turner.
Fonte. Laboratório de Citogenética Humana de Genética Molecular (LaGene), fotografia gentilmente cedido por Dr. Cláudio C. Silva.

Adicionalmente ao cariótipo clássico, outras aberrações cromossômicas podem ser encontradas nos pacientes com ST, totalizando cerca de 50% dos casos de ST com outros cariótipos (Lipay *et al.*, 2005; Frias *et al.*, 2003), a saber:

- mosaicos: 45,X/46,XX e 45,X/47,XXX, etc;
- aberrações estruturais do cromossomo X: mais freqüentemente 46,X,i(Xq), 46,X,i(Xp), (46,X,r(X), 46,X,del(Xq) ou 46,X,del(Xp), respectivamente

isocromossomo de braço longo, isocromossomo de braço curto, cromossomo X em anel, deleção de braço longo e deleção de braço curto. Podem ocorrer ainda translocações herdadas ou *de novo*;

- cromossomos marcadores, que são estruturalmente anômalos e de origem indefinida.

5.4 Considerações Finais

Embora a inteligência de pacientes com ST possa ser normal, alguns casos podem apresentar riscos nos domínios cognitivos, comportamental e sociais. Estes incluem inabilidades de aprendizagem, particularmente no que diz respeito à percepção espacial, integração visual-motora, memória, habilidade de formularem objetivos, seqüências nas ações e déficit de atenção. Os desequilíbrios comportamentais diferem pela idade. Meninas mais jovens podem apresentar comportamento com hiperatividade, imaturidade e ansiedade. A ansiedade, depressão e insatisfação nos relacionamentos são mais comuns em meninas mais velhas (McCauley *et al.*, 1994).

Mulheres com Síndrome de Turner apresentam um risco relativo elevado para algumas intercorrências médicas que requerem cuidados durante toda a vida. Neste contexto, o acompanhamento das pacientes durante a infância, a puberdade e o envelhecimento devem ser seguidas por uma equipe multidisciplinar que atenda aos problemas de saúde específicos associados com essa síndrome. O acompanhamento por endocrinologista, cardiologista, fonoaudiólogo, psicólogo, ginecologista e um conselheiro geneticista devem estar à frente da assistência promovida para os casos de ST, com o objetivo de atender as necessidades e as demandas específicas de cada casos (Elsheikh, 2002).

Capítulo 6

6 Metodologia

6.1 Abordagem Adotada:

A proposta do Mestrado em Ciências Ambientais e Saúde (MCAS) da Universidade Católica de Goiás (UCG) têm como meta a articulação das diferentes áreas da ciência em torno da temática da saúde humana e do meio ambiente, através de uma abordagem e metodologia interdisciplinares. Essa visão se concretiza na definição de um conhecimento emergente, de caráter amplo, que considera todo um conjunto de disciplinas das ciências naturais, da saúde e sociais, para construir um conhecimento e uma racionalidade social orientados a um desenvolvimento sustentável, eqüitativo e duradouro, com ênfase nos aspectos da saúde e qualidade de vida da população humana.

Buscando a proposta do MCAS, identificamos a multidisciplinaridade/ interdisciplinaridade entre os campos da genética humana e a informática.

A presente proposta de estudo consiste no desenvolvimento de uma ferramenta computacional que facilite o controle e agilize o processo de exames citogenéticos.

6.2 Característica do Estudo

Através dessa dissertação propõe-se o desenvolvimento de um sistema probabilístico de apoio à tomada de decisão, ferramenta dotada de conhecimentos gerados a partir de estudos multidisciplinares e multiinstitucionais, o que lhe confere a capacidade não apenas de avaliar a qualidade dos exames citogenéticos, mas também de recomendar o seu uso de maneira adequada e segura na avaliação de resultados dos exames, afinal possui uma base de dados que cruzam informações gerando resultados para uma tomada de decisão. Para uma melhor definição, utilizamos também o conceito

de que sistema de apoio à tomada de decisão que consiste em fatos e heurísticas. Os fatos constituem um corpo de informação que são largamente compartilhados, publicamente disponível e geralmente aceito pelos especialistas em um campo do conhecimento. As heurísticas são em sua maioria privadas, regras pouco discutidas de bom discernimento (regras do raciocínio plausível, regras de boa conjectura), que caracterizam a tomada de decisão no nível de especialista na área. O nível de desempenho de um sistema de apoio à tomada de decisão é função principalmente do tamanho e da qualidade do banco de conhecimento que possui (Harmon, 1988).

Pretende-se que através dos dados determinísticos, ou seja, dados apresentados como constantes, dados probabilísticos sejam geradas, com base em repetições de eventos ou em dados experimentais (Kenarangui e Seifi, 1994).

6.3 Descrição dos Métodos do Processo de Exame Citogenético

Os métodos atuais para a análise são baseados na suspeita clínica informada pelo médico-assistente. Segue uma análise minuciosa e criteriosa do técnico–especialista utilizando manualmente uma pesquisa em livros didáticos para localizar qual a doença cromossômica que está relacionada com a suspeita clínica. Assim, há uma possível identificação prévia dos cromossomos para deduzir a presença de alguma síndrome. O diagnóstico citogenético se inicia na coleta de sangue do paciente, referenciado ao laboratório pelo clínico que o acompanha. Após a análise, o resultado é concluído e um laudo contendo o cariograma é emitido.

A proposta objetiva do estudo foi seguir as etapas da metodologia CRISP-DM, assim sendo, são necessários uma atenção aos atributos referente ao nosso escopo do estudo. Com base nos dados coletados pelo técnico-especialista, foram identificados os atributos (sinais/sintomas) e a classe (síndrome) do problema.

Tanto a classe quanto os atributos receberam valores “SIM” ou “NAO” e convertido para “1” e “0” respectivamente, conforme ANEXO III. Foi utilizada uma classe: “TURNER”. A classe TURNER foi diagnosticada por 9 atributos independentes, representados por: “sexo feminino; baixa estatura; tórax em escudo; disgenesia gonadal; unhas hipoplásicas; cúbito valgo; pescoço alado; hipertelorismo de mamilos e tendência à obesidade”. Foram 84 registros de pacientes para os casos de ST.

Foi necessária a limpeza dos dados, dessa forma também foi utilizada uma regressão usando o método *step-wise* para executar a seleção das variáveis e identificar os atributos que são relevantes para a classe. Como o número de casos foi pequeno, então foi necessário aplicar métodos que validem os dados com resultados coerentes, neste caso o método de escolha foi o *4-fold-cross-validation.*

De acordo com Bouckaert e Frank (2004), antes do *cross-validation* ser executada, os dados são randomizados de modo que cada um dos conjuntos de aprendizado e teste resultantes exiba a mesma distribuição. A forma ideal é que o resultado do teste fosse independente da partição particular resultante do processo de randomização, porque isso tornaria muito mais fácil replicar resultados experimentais publicados na literatura. Contudo, na prática há sempre uma certa sensibilidade da partição usada. Para medir a replicabilidade, são necessários repetir o mesmo teste várias vezes com os mesmos dados e com diferentes partições aleatórias – neste estudo foram quatro repetições – e contou com a freqüência de coincidência dos resultados que foi o mesmo.

Consideramos testes baseados em número r de vezes dos 4-*fold cross-validations* onde r e k podem ter qualquer valor. Observamos diferenças $x_{ij} = a_{ij} - b_{ij}$ para a repartição i e série j. Alguém poderia simplesmente usar $m = \frac{1}{k.r} \sum_{i=1}^{k} \sum_{j=1}^{r} x_{i,j}$ como uma estimativa para a média é $\sigma^2 = \frac{1}{k.r-1} \sum_{i=1}^{k} \sum_{j=1}^{r} (x_{i,j} - m)^2$ como uma

estimativa para a variância. Então, supondo que os vários valores de x_{ij} são independentes, o teste estatístico $t = m \Big/ \sqrt{(1/k.r)\sigma^2}$ é distribuído de acordo com uma distribuição-*t* com df = k.r – 1 graus de liberdade. O *cross-validation* é um caso especial de sub-amostragem aleatória onde asseguramos que os conjuntos de teste em uma série não se sobrepõem. Naturalmente, os conjuntos de teste de séries diferentes se sobreporão. O modelo apresentado, resultou na seguinte estatística:

$$t = \frac{\frac{1}{k.r}\sum_{i=1}^{k}\sum_{j=1}^{r} x_{i,j}}{\sqrt{\left(\frac{1}{k.r} + \frac{n_2}{n_1}\right)\sigma^2}},$$

onde n_1 é o número de exemplos usados para aprendizado (treinamento), e n_2 o número de exemplos usados para teste (avaliação). Assim, o teste acima descrito é chamado de "*teste k-fold cv corrigido e repetido*" (Bouckaert e Frank, 2004).

6.4 Coleta de dados

A forma de coleta de dados consiste no uso de casos atendidos no LaGene/SULEIDE/SES-GO e caracterizaram uma medição de parte de uma população. As estatísticas calculadas a partir da amostras foram usadas para predizer vários parâmetros populacionais (Larson e Farber, 2004).

Os dados foram coletados e analisados pelo técnico-especialista mediante pedido de exames laboratoriais. Apesar do número intenso de pedidos laboratoriais foram identificadas duas realidades que dificultam o levantamento dos dados consistentes. A primeira dificuldade são os casos em que foram indicados como “caso a esclarecer”, pois os atributos não foram identificados o que determina a exclusão desses casos. A segunda dificuldade encontrada foram os casos onde o técnico-especialista identificou parte dos atributos, assim não formou uma correlação entre os atributos e as respectivas classes, nestes casos também houveram a exclusão dos dados. Portanto, foram identificados 84 casos onde a classe é TURNER.

6.4.1 Aplicação dos Algoritmos de Aprendizagem

O primeiro algoritmo de aprendizado implementado foi o algoritmo k2, onde determinou a topologia do Naïve Bayes; pelo classificador TAN foi utilizado uma versão modificada do algoritmo Chow-Liu; já pelo classificador BAN utilizou uma versão mais atualizada do algoritmo CBL-B. Ressalta-se que a aplicação do classificador TAN e BAN teve como objetivo identificar as redes montadas por seus algoritmos, e não foram utilizadas para comparação com os demais classificadores, haja vista, que a ferramenta que descreve os classificadores TAN e BAN está em fase experimental. Foram aplicadas, também, as metodologias usando os conceitos de rede neural MLP (do inglês, *Multi- Layer Perceptron*) com algoritmo de treinamento por retropropagação de erro e árvore de decisão com o algoritmo ID3.

Usando os dados coletados pelo especialista foi treinada a rede bayesiana, usando os métodos KDD e CRISP-DM.

6.4.3 Definição do Problema e Justificativa do Estudo

O diagnóstico citogenético se inicia na coleta de sangue do paciente, referenciado ao laboratório pelo clínico que o acompanha. A partir da suspeita clínica, o técnico especialista procede com uma análise minuciosa e criteriosa. Após a análise, o resultado é concluído e um laudo, contendo o cariograma, é emitido.

Atualmente, leva-se em torno de 1 mês para se obter o resultado, um prazo considerável para a emissão do laudo, que afeta a rotina laboratorial, causando um acúmulo de exames e um retardo nos procedimentos clínicos e na assistência ao paciente.

Com base no exposto acima, a proposta deste estudo foi aplicar as técnicas de mineração de dados ao aprendizado de classificadores baseados em redes probabilísticas, árvores de decisão ou redes neurais, visando obter um modelo acurado,

eficaz e efetivo para facilitar e agilizar o processo de exames citogenéticos. Conseqüentemente, os usuários do Sistema Único de Saúde (SUS) obterão os resultados de seus testes em um menor lapso de tempo, uma vez que a ferramenta proposta poderá ser usada para organizar e priorizar o atendimento para realização de exames. A Figura 6 ilustra o esboço do procedimento para a realização do exame citogenético, identificando o domínio do problema contido neste estudo.

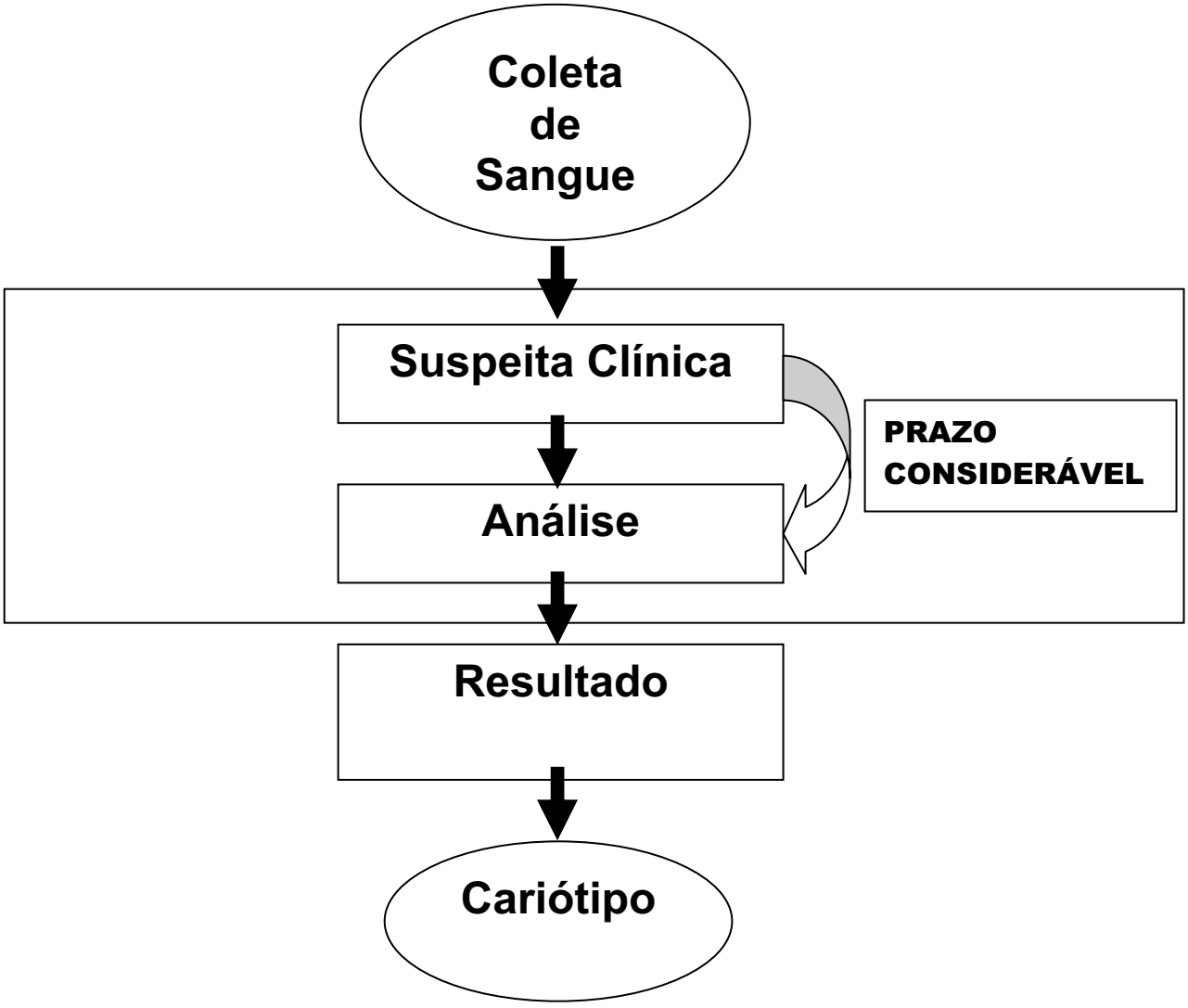


Figura 6. Ciclo do processo para obter um resultado de cariótipo e a identificação do problema (quadro em destaque) a ferramenta desenvolvida mediante a elaboração do presente estudo propõe solução para o problema e, consequentemente, melhora no atendimento do paciente.

6.4.4 Montagem dos Dados da Pesquisa para Efetivar a Mineração de Dados – MD.

As principais tarefas de mineração de dados incluem classificação[9], regressão[10], agrupamentos[11],sumarização[12], modelagem de dependências[13], detecção de desvios[14] e associação[15]. Neste contexto, o presente estudo se concentrou na tarefa de classificação. Para efeito de construção de classificadores, um conjunto de n variáveis foi dividido em n-1 variável atributos e uma variável classe. Assim, o classificador é um modelo construído a partir de um conjunto de casos de atributos-classe que utiliza os valores associados aos estados das variáveis atributos para inferir o estado corrente da variável classe.

Em geral, existe apenas uma variável classe. Portanto é usual se referir aos estados dessa variável como a classe possível. Neste estudo foram utilizados os classificadores probabilísticos – Naïve Bayes, BAN e TAN, o classificador árvore de decisão usando o algoritmo ID3, e neural – rede neural MLP (do inglês, *Multi- Layer Perceptron*) com algoritmo de treinamento por retropropagação de erro.

6.4.5 Classificadores baseados em Rede Bayesianas

Naïve Bayes, BAN e TAN são redes bayesianas específicas construídas com base em hipóteses de restrições de independência entre as variáveis, que serão ilustradas nas Figuras 7, 8 e 9 abaixo.

---

[9] Aprendizagem de função, definida sobre variáveis de atributos, que classifica a variável de interesse em um conjunto de classes pré-definidas

[10] Aprendizagem de função, definida sobre variáveis de atributos, que estima o valor real da variável de interesse.

[11] Identificação de conjunto finito de categorias que descrevem os dados

[12] Descrição compacta para um subconjunto de dados ou descoberta de relações funcionais entre variáveis

[13] Obter modelo para descrever dependências significativas entre as variáveis de interesse

[14] Encontrar as alterações mais significantes nos dados em relação aos valores históricos médios

[15] Identificação de padrões descritivos intrínsicos entre subconjunto de itens em uma mesma transação, representados, em geral, através de regras de associação entre variáveis que ocorrem na transação

6.4.5.1 Classificador da Topologia Naïve Bayes

Por Cheng e Greiner (1999) a topologia da rede Naïve Bayes é representada por uma árvore de altura unitária onde a raiz é a variável de classe e as folhas são os variáveis atributos.

Naïve Bayes foi utilizado como um classificador eficaz por muitos anos. Possui algumas vantagens em relação aos demais classificadores. Primeiramente, é fácil de construir porque a estrutura é dada *a priori* (i.e. onde não é requerido nenhum procedimento de aprendizagem da estrutura). Além dessa facilidade o processo da classificação é muito eficiente. Ambas as vantagens são devido a suposição de que todas as variáveis atributos sejam condicionalmente independentes, dado a variável de classe. Embora esta suposição da independência seja obviamente problemática (Cheng e Greiner, 1999).

Surpreendentemente o classificador Naïve Bayes supera muitos classificadores sofisticados, especialmente onde os atributos não são fortemente correlacionadas.

Segundo (Cheng e Greiner, 1999) em anos recentes, muitos dos esforços para em melhorar o classificador Naïve Bayes seguiram duas aproximações gerais: selecionando o subconjunto da característica e relaxando suposições da independência.

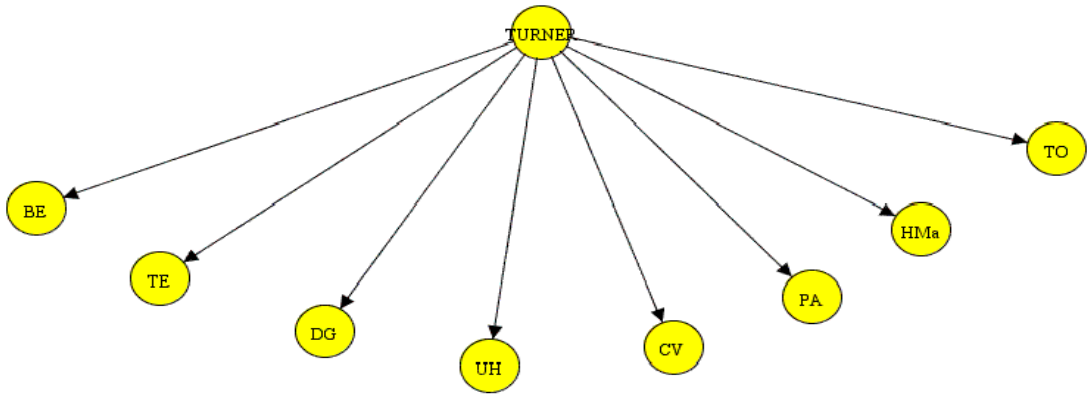


Figura 7. Classificador Naïve Bayes para Síndrome de Turner. Legenda: SF - Sexo feminino, BE - Baixa estatura, TE - Tórax em escudo, DG - Disgenesia gonadal, UH - Unhas hipoplásicas, CV - Cúbito valgo, PA - Pescoço alado, HMa - Hipertelorismo de mamilos, TO - Tendência à obesidade, TURNER - Síndrome de TURNER.

6.4.5.2 Classificador TAN (*Tree-Augmented Naïve Bayes*)

O algoritmo de aprendizagem do classificador TAN (Friedman *et al.*, 1997) primeiro constrói uma estrutura de uma árvore com as variáveis em X \ {"TURNER"}, executando teste onde são propagadas informações mutuamente condicionadas de "TURNER". Após se adiciona uma ligação da classe-nó "TURNER" a cada atributo-nó, similar a uma estrutura de Naïve Bayes (i.e., a nó-classe é um pai para todos os nós-atributos) – veja a Figura 8 (Cheng e Greiner, 1999).

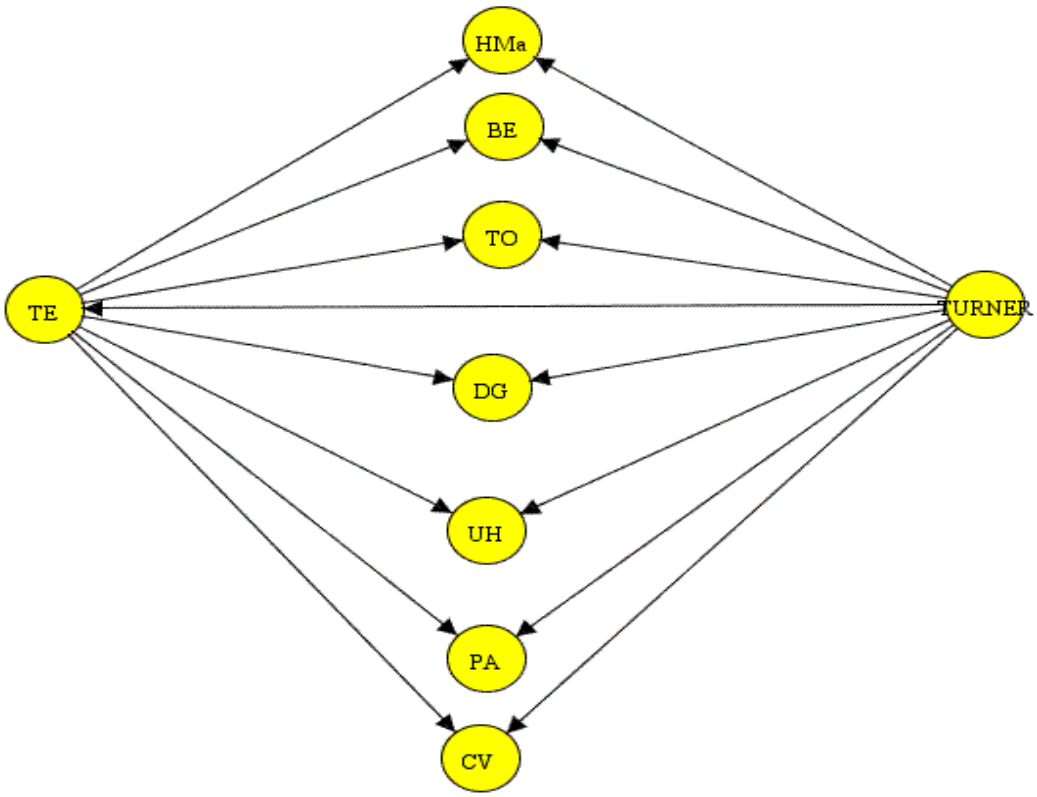


Figura 8. Classificador TAN para Síndrome de Turner. Legenda: SF - Sexo feminino, BE - Baixa estatura, TE - Tórax em escudo, DG - Disgenesia gonadal, UH - Unhas hipoplásicas, CV - Cúbito valgo, PA - Pescoço alado, HMa - Hipertelorismo de mamilos, TO - Tendência à obesidade, TURNER - Síndrome de TURNER

6.4.5.3 Classificador BAN (do inglês, BN Augmented Naïve-Bayes)

O classificador BAN é uma extensão do classificador TAN permitindo que os atributos dêem forma a um gráfico arbitrário acíclico e orientado (Friedman *et al.*, 1997). conforme Figura 9.

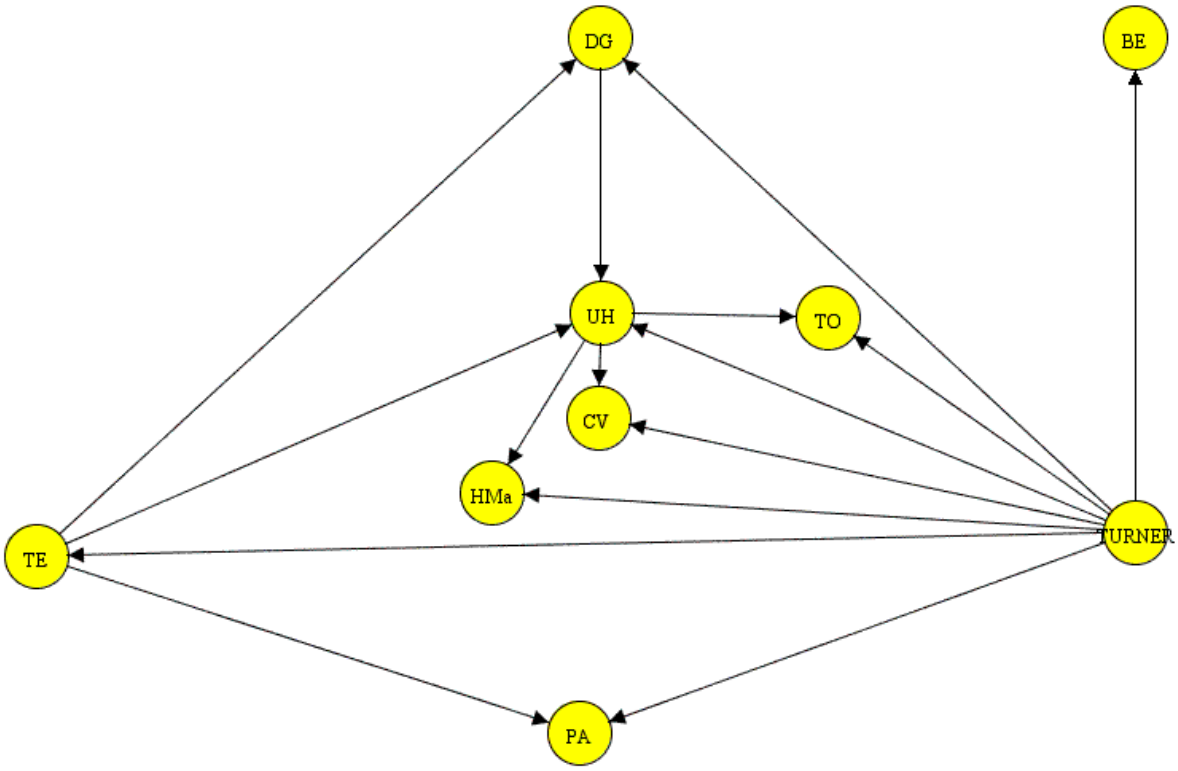


Figura 9. Classificador BAN para Síndrome de Turner. Legenda: SF - Sexo feminino, BE - Baixa estatura, TE - Tórax em escudo, DG - Disgenesia gonadal, UH - Unhas hipoplásicas, CV - Cúbito valgo, PA - Pescoço alado, HMa - Hipertelorismo de mamilos, TO - Tendência à obesidade, TURNER - Síndrome de TURNER.

6.4.5.4 Comparação dos classificadores: Naïve Bayes, TAN e BAN

No Quadro IX foram descritos comparativamente os três classificadores das redes probabilísticas.

Quadro IX - Comparação entre os classificadores Naïve Bayes, BAN e TAN

Naïve Bayes	TAN	BAN
Permite que o nó-classe seja pai dos todos os nós-atributos	Monta uma fase de treinamento usando X \ {"TURNER"} como entrada	Monta uma fase de treinamento usando X \ {"TURNER"} (juntamente com os nós ordenados) como entrada
Aprendizagem usando os parâmetros (grava os valores estimados empiricamente por freqüência) e imprima as probabilidades	Executa o algoritmo Chow-Liu com uma versão modificada.	Executa o algoritmo CBLA com uma versão modificada.
-	Adiciona a classe "TURNER" como pai de todos os atributos.	Adiciona a classe "TURNER" como pai de todos os atributos.
-	Informa os parâmetros e imprime uma rede TAN.	Informa os parâmetros e imprime uma rede BAN.

Capítulo 7

7. Avaliação dos Resultados e Conclusões

Para a geração dos resultados da fase de teste e treinamento desse estudo, foram utilizadas as ferramentas UnBBayes e UnBMiner.

Na UnBBayes foram propostas a topologia da rede bayesiana e as probabilidades condicionais estimadas para cada variável. As topologias das redes bayesianas foram avaliadas criteriosamente e aprovado pelo técnico-especialista. Foram avaliadas 3 redes pelos técnico-especialistas; a topologia do Naïve Bayes, Figura 7; a topologia do TAN, Figura 8 e a topologia do BAN, Figura 9. Portanto, de acordo com o técnico-especialista a topologia Naïve Bayes representada, pode ser utilizada como modelo qualitativo casual de fácil entendimento do relacionamento das variáveis analisadas para uma definição de ocorrências da Síndrome de Turner.

Usando a UnBMiner foi possível realizar as etapas definidas pelo CRISP-DM. A primeira etapa foi constituída pelo pré-processamento, onde foram identificados os atributos. A segunda etapa foi as seleções dos algoritmos que propagou as inferências probabilísticas. A terceira etapa foi a evolução dos dados, que gerou resultados para a identificação da melhor solução para o domínio do problema. Os resultados foram apresentados no ANEXO I. O UnBMiner foi ainda usado para gerar os resultados a partir da aplicaçào de Árvore de Decisão e Rede Neural.

7.1 Entendimento dos Dados

Seleção dos dados: Foram selecionados 84 casos de pacientes para realização de exames citogenéticos com o objetivo de identificar anomalias cromossômicas, neste estudo foi refinada a pesquisa para uma síndrome em especial, a Síndrome de Turner. Todos os exames foram realizados no Laboratório de Citogenética Molecular do Estado

de Goiás – LaGene/SULEIDE/SES/GO. Os casos selecionados foram utilizados no treinamento dos modelos de classificadores da rede bayesiana que avaliou os resultados e identificou qual modelo que mais interpretou a realidade do técnico especialista. A tabela I contém a análise estatística descritiva dos 84 casos.

Tabela 7.1 - Descrição Estatísticas dos Dados em relação aos sinais e sintomas dos pacientes com ST incluídos neste estudo.

Atributos / Classe	Legenda	Tipo	Média	Desvio Padrão	Mínimo	Máximo
Sexo feminino	SF	Numérico	1,00	0,000	1	1
Baixa estatura	BE	Numérico	0,93	0,259	0	1
Tórax em escudo	TE	Numérico	0,21	0,413	0	1
Disgenesia gonadal	DG	Numérico	0,24	0,428	0	1
Unhas hipoplásicas	UH	Numérico	0,12	0,326	0	1
Cúbito valgo	CV	Numérico	0,13	0,339	0	1
Pescoço alado	PA	Numérico	0,24	0,428	0	1
Hipertelorismo de mamilos	Hma	Numérico	0,15	0,364	0	1
Tendência à obesidade	TO	Numérico	0,13	0,339	0	1
TURNER – Classe	TURNER	Numérico	0,33	0,474	0	1

7.2 Preparação dos Dados

Foram identificados 9 sinais e sintomas, caracterizados como atributos e uma classe, conforme Tabela I. Para efetivar a limpeza dos dados foi conduzida uma avaliação de regressão utilizando-se o método *step-wise*. Na aplicação da regressão concluiu-se que para os modelos com as variáveis dependentes da classe TURNER, a variável "Sexo Feminino - SF" foi constante e não apresentou correlação, assim, a variável foi retirada das análises.

7.3 Modelagem

Os classificadores probabilísticos Naïve Bayes, TAN e BAN, para a classe TURNER, foram apresentados anteriormente nas Figuras 7, 8 e 9, respectivamente.

O classificador baseado em rede neural foi treinado e avaliado utilizando os parâmetros apresentados na Tabela II.

Tabela 7.2 - Parâmetro de configuração da Rede Neural.

Parâmetro	Valor
Neurônios na camada de entrada	9
Neurônios na camada escondida	5
Neurônios na camada de saída	1
Épocas	400
Taxa de aprendizagem	0,3
Momento	0,2
Função de ativação	Sigmóide
Normalização	Linear

Os classificadores probabilísiticos Naïve Bayes, TAN, BAN foram gerados pela ferramenta UnBBayes, enquanto que os classificadores Rede Neural e Árvore de Decisão foram gerados pela ferramenta UnBMiner. A base de dados foram treinados a partir dos 84 casos registrados (ANEXO II).

7.4 Avaliação

Para discriminar os modelos gerados foram utilizados índices de referências através da matriz de confusão, que representa uma matriz bidimensional com uma linha e uma coluna para cada classe. Cada elemento da matriz de confusão apresenta um número de casos avaliados na qual a classe real (R) é a linha e a classe predita (P) pelo classificador é a coluna. Bons resultados são identificados quando a diagonal principal apresenta valores altos e a diagonal secundária apresenta valores nulos, pois a diagonal principal caracteriza os números de acertos e a diagonal secundária os números de não-acertos (Ladeira, 2006).

Quadro X - Matriz de confusão com duas classes.

R\P	SIM	NÃO
SIM	Verdadeiro Positivo (VP)	Falso Negativo (FN)
NÃO	Falso Positivo (FP)	Verdadeiro Negativo (VN)

7.5 Seleção do Modelo

Os modelos treinados foram avaliados considerando-se uma base de dados com 84 casos. Como o número de casos consistiu em uma amostra discreta, um modelo expressou 9 variáveis, ou seja, seria possível no mínimo 512 condições possíveis de resultados. Neste caso, foi aplicado o método de validação cruzada dividido em 4 partes, chamadas de *4-fold cross-validation,* conforme Quadro XI.

Quadro XI - Classificação dos grupos para a aplicação do método *four-fold cross-validation* objetivando-se a redução dos viéis associados aos dados.

Grupos	Números de Registros	Subconjuntos	Conjunto	Procedimentos
I	21	Sc1	- Sc1…….. - Sc2 a Sc4…….	- Avaliar - Treinamento
II	21	Sc2	- Sc2…….. - Sc1, Sc3 e Sc4…	- Avaliar - Treinamento
III	21	Sc3	- Sc3…….. - Sc1, Sc2 e Sc4…	- Avaliar - Treinamento
IV	21	Sc4	- Sc4…….. - Sc1 a Sc3…….	- Avaliar - Treinamento

Fonte: Prof. Marcelo Ladeira, 2006.

Nesta análise foram utilizados os índices apresentados na Quadro XII. Todos assumem valores no intervalo [0,1] e puderam ser calculados a partir da matriz de confusão.

Quadro XII - Índice para discriminação entre os classificadores dicotômicos

Sensibilidade (S)	Especificidade (E)	Acurácia (Ac)	SE
S = VP / (VP+FN)	E = VN / (VN+FP)	Ac = (VP+VN) / (VP+FP+VN+FN)	SE = S * E

A sensibilidade é a taxa do verdadeiro positivo (Tvp), já a especificidade retrata a taxa de verdadeiro negativo (Tvn), a acurácia é a probabilidade de acerto do

classificador. O produto da sensibilidade pela especificidade, implica no produto da taxa de verdadeiro positivo pelo complemento da taxa do falso positivo.

O melhor classificador é aquele que maximiza a taxa de verdadeiro positivo e, simultaneamente, minimiza a taxa de falso positivo. Portanto, o melhor classificador é aquele que apresenta maior valor para a acurácia e o maior valor para o produto de sensibilidade vezes especificidade. Os resultados obtidos no presente estudo para cada modelo foram apresentados na Tabela III.

Tabela 7.3 - Parâmetros de resultados obtidos por cada classificador.

Modelo	Sensibilidade	Especificidade	Acurácia	S*E
Árvore de Decisão - Id3	0,806	0,771	0,786	0,621
Naïve Bayes	0,761	0,958	0,893	0,729
TAN	0,656	0,938	0,845	0,615
BAN	0,656	0,958	0,857	0,628
Rede Neural	0,761	0,536	0,607	0,408

Tanto a acurácia quanto o produto sensibilidade/especificidade foram critérios utéis para a avaliação geral do modelo. Estes parâmetros mostraram que o Naïve Bayes foi o melhor modelo para prever a variável classe TURNER.

7.6 Análise dos Dados Estatísticos: Correlação

Após a análise da correlação entre os atributos em relação à classe TURNER, a única variável que apresentou valor não-significativo foi a variável BE-Baixa Estatura. Todas as demais variáveis apresentaram valores significativos para a correlação. É importante ressaltar que, na prática médica, o sinal/sintoma BE-Baixa Estatura reflete uma correlação positiva para a Síndrome de Turner, o que pode ser confirmado no Anexo II, pois todos os casos de Turner apresentaram baixa estatura. No entanto, o confundimento ocorreu devido ao discreto tamanho amostral, que incluem os casos de baixa estatura que não foram diagnosticados com a ST. O teste do χ^2 mostrou que todos os sinais e sintomas analisados, exceto sexo feminino e baixa estatura, foram significativos para a predizerem a ocorrência da ST (Tabela IV)

Tabela 7.4 - Dados estatísticos para Qui-quadrada com índice de confiança a 95% e grau de significância menor que 0,001

Atributos	Classe: TURNER			
	Ausente	**Presente**	**Total**	χ^2
TE	56	10	66	*45,818*
DG	53	11	64	*31,533*
UH	56	18	74	*22,703*
CV	56	17	73	*25,315*
PA	55	9	64	*44,920*
HMa	55	16	71	*24,072*
TO	56	17	73	*25,315*

IC=95%; $p < 0,001$

Capítulo 8 – Conclusões

O presente estudo demonstrou que a teoria que envolve a rede bayesiana proveu resultados consistentes que possibilitam a construção de sistemas baseados em conhecimentos. Assim, foi possível integrar o aprendizado e a propagação de evidências, obtendo-se bons resultados.

O problema encontrado na construção da rede bayesiana e da formação de um sistema especialista probabilísticos foi a obtenção do conhecimento. Pois, a maioria dos dados que serviu de parâmetro para a obtenção dos resultados continha informações incompletas, fazendo com que o número de casos úteis para a experiência fosse reduzido.

A base de dados foi dividida em 4 partes igual de forma separada e definida aleatoriamente, ou seja, foram geradas fases de testes e fases de treinamento para a classe (TURNER) em especifico. Foram avaliados os classificadores rede neural e árvore de decisão, além dos classificadores probabilísticos Naïve Bayes, BAN e TAN.

De acordo com os resultados apresentados na Tabela III, concluímos que a melhor solução foi o modelo Naïve Bayes, pois este modelo apresentou maior acurácia. Os modelos Árvore de Decisão, TAN e BAN apresentaram soluções para o domínio do problema sugerido, mas as soluções não foram tão satisfatória quanto o Naïve Bayes. No entanto, a Rede Neural não promoveu solução satisfatória.

8.1 Contribuições

O presente estudo contribuiu para a área de conhecimento mediante a condensação dos assuntos abordados em um referencial teórico que descreveu o estado da arte na teoria da propagação de crenças em redes bayesianas, como a implementação dos softwares livres UnbBayes e UnbMiner. Adicionalmente, os softwares livres foram usados para se calcular os parâmetros numéricos para definir a confiabilidade dos relacionamentos entre as variáveis de uma rede bayesiana com a utilização de resultados probabilísticos, e os resultados de uma aplicação da teoria bayesiana (aprendizagem e propagação de evidências) no domínio do diagnóstico da Síndrome de Turner.

8.2 Trabalhos Futuros

O presente estudo possibilitou desenvolver um procedimento que possibilite, durante a entrevista dos técnico-especialistas com o paciente, o levantamento exato dos sinais/sintomas que possibilitarão maior definição e exatidão no diagnóstico da síndrome.

A estrutura da rede e a definição do melhor método podem ser aplicadas de acordo com novos sinais/sintomas e permitem a inclusão de novas síndromes genéticas no modelo apresentado.

A área de aprendizado em redes bayesianas ainda encontra-se em crescimento. As descobertas e propostas nesta área são de grande importância na construção das bases do conhecimento causal.

Referência Bibliográfica

1. Acid, S., Campos, BENEDICT An Algorithm for Learning Probabilistic Belief Networks, Univ. Granada, Spain, Proceedings of the sixth International Conference IPMU'96, 1996.
2. Acid, S., Campos, LM. Algorithm for Finding Minimum d-Separating Sets in Belief Networks, Proceedings of UAI96, 003010, E. Horvitz, F. Jensen, MorganKaufmann; 1996.
3. Anto, JM. Community outbreaks of asthma associated with inhalation of soybean dust. The New England Journal of Medicine, 320:1097-1102; 1989.
4. Baranauskas, JÁ. Extração Automática de Conhecimento por Múltiplos Indutores. [Tese] Univ. São Paulo, São Carlos–SP; 2001.
5. Barcellos, C., Bastos, FI. Are geoprocessing, environment, and health a possible combination?. Cad. Saúde Pública, vol.12, no.3, p.389-397. ISSN 0102-311X; jul./set. 1996
6. Barreto, ML., Carmo, EH, Noronha, CV. Mudança dos padrões de morbi-mortalidade: uma revisão crítica das abordagens epidemiológicas. Physis, 31:127-146; 1993.
7. Bastos, FI.,Barcellos, C. A geografia social da AIDS no Brasil. Revista de Saúde Pública, 29:52-62; 1995.
8. Batch J. Turner syndrome in childhood and adolescence. Best Pract Res Clin Endocrinol Metab. p. 465-82, 2002.
9. Bayes, T. An essay to towards solving a problem in the doctrine of charges. Philosophic Transactions of the Royal Society of London, 53:570-418; 1763.
10. Bazzan, AL., Engel,PM., Schroeder,L.F., Da Silva,SC. Automated annotation of keywords for proteins related to mycoplasmataceae using machine learning techniques. Bioinformatics, 18 (Suppl. 2), 35S–43S; 2002.
11. Bösze P, Eiben OG, Gaal M, Laszlo J. Body measurements of patients with streak gonads and their bearing upon the karyotype. Hum Genet. p. 355-60, 1980.
12. Bouckaert RR, Frank E. Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms. Ed. Springer Berlin/Heidelberg. vol.3056/2004, p.3-12, 2004.
13. Bouckaert, RR., Bayesian Belief Networks. [Tese]. Univ. Utrech, June; 1995.
14. Brackett, M. H. The data warehouse challenge: taming data chaos. New York: John Wiley & Sons, 1996.
15. Buntime, W. Learning with Graphical Models. Technical Report FIA-94-02 from the Artificial Intelligence Branch AT Nasa Ames; 1994.
16. Buntime, W. Learning in networks. Invited paper for 50th Session of the International Statistical Institute, Bejing, China; 1995.
17. Buntine, W. Theory Refinement on Bayesian Networks. In: Conference on Uncertainty in Artificial Intelligence, 7. Proceedings. Morgan Kaufmann,

p.52-60; 1991.

18. Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C, Wirth R. CRISP-DM 1.0: Step-by-step data mining guide. SPSS, 1999.

19. Chen, M-S., Han, J., Yu, P. S. Data mining: an overview from a database perspective. IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No. 6, p.886-883; 1996.

20. Cheng J, Bell D, Liu W. Learning Bayesian Networks from Data: An Efficient Approach Based on Information Theory, 2000. Procurar mais referencia.

21. Cheng, J, Greiner, R. Comparing Bayesian Network Classifiers, Proceedings of the fifteenth international conference on uncertainty in artificial intelligence, 1999.

22. Cheng, J., Bell, DA., Liu, W. An algorithm for Bayesian belief networks construction from data. In Proceedings of AI. STAT97, p. 83-90, Florida; 1997.

23. Cheng, J., Bell, DA., Liu, W. Documento disponível no site: http://www.cs.ualberta.ca/~jcheng/bnpc.htm

24. Cheng, J., Bell, DA., Liu, W. Learning belief networks from data: An efficient approach based on information theory. Technical Reports; 1998.

25. Cheung, DW., Ng, VT., Fu, AW. Efficient mining of association rules in distributed databases. IEEE Transactions on Knowledge and Data Engineering, v.8, n. 6, p. 911-922, 1996.

26. Chow C, Liu C. Approximating discrete probability distributions with dependence trees. Information Theory, IEEE Transactions on. Vol. 14. nº 3. p.462-467, 1968.

27. Chu CE, Connor JM, Donaldson MD, Kelnar CJ, Smail PJ, Greene SA. Detection of Y mosaicism in patients with Turner's syndrome. J Med Genet. p. 578–580, 1995.

28. Clement-Jones M, Schiller S, Rao E, Blaschke RJ, Zuninga A, Zeller R, et al. The short stature homeobox gene SHOX is involved in skeletal abnormalities in Turner syndrome. Oxford University Press. Human Molecular Genetics. vol. 9, Nº5. p. 695-702, 2000.

29. Cockwell AE, MacKenzie M, Youings S, et al. A cytogenetic and molecular study of a series of 45,X fetuses and their parents. J Med Genet. p. 152–155, 1991.

30. Cooper, G., Herskovits, E. A Bayesian method for the induction of probabilistic networks from data: Machine Learning, No.9. p. 309-347; 1992.

31. Davis, DT., Chen, Z., Hwang, J-N., Tsang, L., Njoku, E.. Solving Inverse Problems by Bayesian Iterative Inversion of a Forward Model with Applications to Parameter Mapping Using SMMR Remote Sensing Data. IEEE Transactions On Geoscience And Remote Sensing, Vol 33, No 5. September; 1995

32. Dempster, A., Laird, N., Rubin, D. Maximum Likelihood from Incomplete Data Via the EM Algorithm. J. Royal Statist. Soc. ,B., 39: 138; 1977.

33. Doyle, J., Dean, T. *et al.*, Strategics directions in artificial intelligent. ACM Computing Survey, 28, Nº4, December; 1994.

34. Elsheikh M, Dunger DB, Conway GS, Wass JAH. Turner's Syndrome in Adulthood. Printed in U.S.A. Endocrine Reviews. vol. 23, nº1. p. 120-140, 2002.

35. Fayyad, UM., Piatetsky-Shapiro, G., Smyth, P. Advances in knowledge discovery & data mining. Chapter 1: From data mining to knowledge discovery: an overview. AAAI/MIT, 1996a.

36. Fayyad, UM., Piatetsky-Shapiro, G., Smyth, P. Knowledge Discovery and Data Mining: Towards a Unifying Framework. Second International Conference on KD & DM. Portland, Oregon; 1996b.

37. Flores CD, Fundamentos dos Sistemas Especialistas. Porto Alegre, Rio Grande do Sul, 2002.

38. Flores, CD. Fundamentos dos Sistemas Especialistas. Univ. Federal do Rio Grande do Sul Porto Alegre: PPGC/UFRGS; 2000.

39. Ford CE, Jones KW, Polani PE, Almeida JC, Briggs JH A. Sex chromosome anomaly in a case of gonadal dysgenesis (Turner's syndrome). Lancet. P.711-3, 1959.

40. Fowler, M., Beck, K., Brant, J., Opdyke, W., and Roberts, D. Refactoring:

41. Improving the Design of Existing Code. Addison Wesley, p.431; 1999.

42. Frank, E. Hall, M., Trigg, L., Holmes, G., Witten, IH. Data mining in bioinformatics using Weka. Bioinformatics Applications Note. Vol. 20, No. 15, p 2479–2481; 2004.

43. Freitas, AA. Generic, Set-Oriented Primitives to Support Data-Parallel Knowledge Discovery in Relational Database Systems. [Ph.D. Thesis], University of Essex, UK; July 1997.

44. Frías JL, Davenport ML, Committee on Genetics and Section on Endocrinology. Health Supervision for Children With Turner Syndrome. Pediatrics. p. 692-702. DOI: 10.1542/peds.111.3.692, 2003.

45. Friedman, N., Geiger, D., Goldszmidt, M. Bayesian Networks Classifier. Machine Learning, 29, p. 131-161; 1997.

46. Fung, R.M e Crawford, S. L. Constructor: a System for Induction of Probabilistic Models. Proceedings of AAAI p. 762769, Boston, MA: MIT Press; 1990.

47. Gravholt CH, Juul S, Naeraa RW. Morbity in Turner Syndrome. J Epidemiol. p. 147-58, 1998.

48. Gravholt CH, Naeraa RW. Reference values for body proportions and body composition in adult women with Ullrich- Turner's syndrome. Am J Med Genet. p. 403-8, 1997.

49. Guimarães MM, Guerra CTG, Alves STF, Cunha MCSA, Marins LA, Barreto LFM, et al. Intercorrências Clínicas na Síndrome de Turner. Arq Bras Endocrinol Metab. Vol. 45 nº 4, 2001.

50. Hall JG, Gilchrist DM. Turner syndrome and its variants. Pediatr Clin North Am. P.1421-44, 1990.

51. Harmon P, King D. Sistemas Especialistas. A inteligência artificial chega ao mercado. Editora Campus, Rio de Janeiro, 1988.

52. Harvey, D.. A Justiça Social e a Cidade. São Paulo: Hucitec; 1980.

53. Hassold T, Benham F, Leppert M. Cytogenetic and molecular analysis of sex-chromosome monosomy. Am J Hum Genet. p. 534–541, 1988.

54. Held KR, Kerber S, Kaminsky E, et al. Mosaicism in 45, X Turner syndrome: does survival in early pregnancy depend on the presence of two sex chromosomes? Hum Genet. P. 288–294, 1992.

55. Henrion, M., Breese J., Horvitz, E. Decision Analysis and Expert Systems, AI Magazine: Winter 1991.

56. Herckerman, D. A bayesian approach to learning causal network. Technical Report: MSR-TR-95-04. Microsoft Research, March; 1995b.

57. Herckerman, D. A tutorial on learning bayesian networks. Technical Reports MSR-TR-95-06. Microsoft Research, Advanced Technology Division, Microsoft Corporation; 1995

58. Herckerman, D. A tutorial on learning bayesian networks. Technical Report: MSR-TR-95-06. Microsoft Research, Advanced Technology Division, Microsoft Corporation; 1995a.

59. Herckerman, D. Probabilistic similarity networks. MIT Press, Cambrige, MA.; 1991

60. Herckerman, D., Breese, J., Rommelse, K. Decision Theoretic troubleshooting. Communication of ACM 38, Nº 3, pp. 49-57, March; 1995.

61. Herckerman, D., Geiger D., Chickering, DM. Learning Bayesian Networks. The Combinational of Knowledge and Statistic data. Technical Report MSR-TR-94-09, Microsoft Research. Advanced Technology Division, July; 1994.

62. Hipp, J., Güntzer U., Nakhaeizadeh, G. Data Mining of Association Rules and the Process of Knowledge Discovery in Databases. In Data Mining in E-Commerce, Medicine, and Knowledge Management, editors: Springer. P. 15-36, 2002.

63. Holmes, G., Donkin, A., Witten, IH. WEKA: A Machine Learning Workbench. Proceedings of the 1994 Second Australian and New Zealand; 1994.

64. Hook EB, Warburton D. The distribution of chromosomal genotypes associated with Turner's syndrome: livebirth prevalence rates and evidence for diminished fetal mortality and severity in genotypes associated with structural X abnormalities or mosaicism. Hum Genet. p. 24–27, 1983.

65. Horvitz, EJ., Srinivas, S., Rouokangas,C., Barry, M. A decision-theoretic approach to the display of information for time-critical decisions: The VISTA project. Proceedings of SOAR'92, Conference on Space Operations Automation and Research, National Aeronautics and Space Administration; 1992.

66. Hruschka Jr, ER., Propagação de Evidências em Redes bayesianas: Diagnóstico sobre Doenças Pulmonares [dissertação]. Brasília(DF): Universidade de Brasília; 1997.

67. Hruschka Jr., ER., da Silva WT. Propagação de Crença e Aprendizagem em Redes bayesianas. Relatório Técnico – TR. 96-03. Univ. de Brasília. Agosto; 1996.

68. Hughes PC, Ribeiro J, Hughes IA. Body proportions in Turner's syndrome. Arch Dis Child. p. 506-7, 1986.

69. Jensen, FV. Bayesianas Networks and Decision Graphs New York: Springer; 2001.

70. Jordan, MI. Learning in Graphical Models. MIT Press, 1999.

71. Kenarangui R, Seifi A. Fuzzy power flow analysis. Univ. Tabriz, Tabriz-IRAN, vol. 29, nº2, p. 105-109, 1994.

72. King, RD., Feng, C., Sutherland, A. Statlog: Comparison of classification algorithms on large real-world problems. Applied Artificial Intelligence Vol.3, pp. 289-333, May/June 1995.

73. Koehler C, Nassar SM, editors. Modelagem de Redes bayesianasa partir de Dados Médicas. Symposio de Informática y Salud; 2002.

74. Koehler, C; Vicari MR; Flores, CD; Nassar, MS. Mineração de Rede bayesianas a partir de Base de Dados Médicos: Proposta de Algoritmo. Univ. de Caxias do Sul (UCS), Caxias do Sul. Brasil; 2004.

75. Krause, P.J. Learning Probabilistic Networks. Philips Research Laboratories Tech. Report., 1998.

76. Kretschmann,E., Fleischmann,W., Apweiler,R. Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT. Bioinformatics, No.17, p.920–926; 2001.

77. Ladeira, M. Aprendizagem de Modelos para Diagnóstico da Síndrome da Apnéia Obstrutiva do Sono. Comunicação pessoal, 2006.

78. Ladeira, M. Diagrama de Influências Múltiplo Secionado. [Tese]. Porto Alegre:[s.n], 2000.

79. Ladeira, M., da Silva, DC., Lima Jr., FJF., Onishi, MS., Carvalho, RN., da Silva, WT. Ferramenta Aberta e Independente de Plataforma para Redes Probabilísticas; 2002.

80. Ladeira, M., Vieira, MHP., Prado, H.A, Noivo, RM., Castanheira, DBS. UnBMiner® – Ferramenta Aberta para Mineração de Dados, Univ. de Brasília; 2005.

81. Larrañaga P. Algoritmos de Estimación de Distribuciones iguales Computación Evolutiva + Modelos Gráficos Probabilísticos. Dpto. Lenguajes y Ciencias de la Computación - Universidad de Málaga. 30 Septiembre 2002.

82. Larson R, Farber B. Estatística aplicada; trad. Cyro de Carvalho Patarra – São Paulo: Prentice Hall, 2004.

83. Laskey, K., Mahoney S. Network Fragments: Representing Knowledge for Constructing Probabilistic Models, Uncertainty in Artificial Intelligence: Proceedings of the Thirteenth Conference; 1997.

84. Lauritzen, S. L. The EM algorithm for Graphical Association Models with Missing Data. Computational Statistics and Data Analysis, 19, Vol.2, 191-

201; 1995.

85. Li,J., Liu,H., Downing,J.R., Yeoh,A.E., Wong,L. Simple rules underlying gene expression profiles of more than six subtypes of acute lymphoblastic leukemia (ALL) patients. Bioinformatics, 19, 71–78; 2003b.

86. Li,J., Liu,H., Ng,S.K., Wong,L.. Discovery of significant rules for classifying cancer diagnosis data. Bioinformatics, 19 (Suppl. 2), II93–II102; 2003a.

87. Li,J., Wong,L. Identifying good diagnostic gene groups from gene expression profiles using the concept of emerging patterns. Bioinformatics, 18, 725–734; 2002.

88. Lipay MVN, Bianco B, Verreschi ITN. Gonadal dysgenesis and tumors: genetic and clinical features. Arq Bras Endocrinol Metab. vol. 49, nº 1. p. 60-70, 2005.

89. Lippe B. Turner syndrome. In: Sperling MA, editor. Pediatric endocrinology. Phyladelphia: W.B. Saunders Co. p.387-422, 1996.

90. Lippe BM. Turner Syndrome. In: Sperling MA, ed. Pediatric Endocrinology. Philadephia:WB Saunders Company, 1996:387-421.

91. Lopes, Paulo Afonso. – Probabilidade e estatística, Rio de Janeiro: Reichmann & Affonso Editores; 1999.

92. Madigan, D. York, J. Bayesian Graphical Models for Discrete Data. International Statistic Review, 63:215-232; 1995.

93. Madsen AL, Lang M, Kjarulff UB, Jensen F. The Hugin Tool for Learning Bayesian Networks.First European Workshop on Probabilistic Graphical Models, 2002

94. Massa GG, Vanderschueren-Lodeweyckx M. Age and height at diagnosis in Turner syndrome: influence of paternal height. Pediatrics, p. 1148-52, 1991.

95. Mathur A, Stekol L, Schatz D, MacLaren NK, Scott ML, Lippe B. The parental origin of the single X chromosome in Turner syndrome: lack of correlation with parental age or clinical phenotype. Am J Hum Genet. p. 682–686, 1991.

96. McCauley E, Ross J, Sybert V. Self-concept and behavioral profiles in females with Turner syndrome. In: Stabler B, Underwood LE, eds. Growth, Stature, and Adaptation: Behavioral, Social, and Cognitive Aspects of Growth Delay. Chapel Hill, NC: University of North Carolina. p. 181–194, 1994.

97. Meilisjon, I. A fast Improvement to the EM Algorithm on its Own Terms. J. Royal Statist. Soc., B,. 51 Vol.01, 127138; 1989.

98. Naeraa RW, Nielsen J. Standards for growth and final height in Turner's Syndrome. Acta Paediatr Scand. p. 182-90, 1990.

99. Nielsen J, Wohlert M. Chromosome abnormalities found among 34,910 newborn children: results from a 13-year incidence study in Arhus, Denmark. Hum Genet. p. 81–83, 1991.

100. Niranjan, M. Why I am Not a Non-Bayesian? [abstract].The Institution of Electrical Engineers.Printed and published by the IEE. Savoy Place, London WC2ROBL, UK; 1999

101. Pasquino AM, Passeri F, Pucarelli I, Segni M, Municchi G. Spontaneous pubertal

development in Turner's syndrome. Italian Study Group for Turner's Syndrome. J Clin Endocrinol Metab. p. 1810–1813, 1997.

102. Patsalis PC, Hadjimarkou MI, Velissariou V, et al. Supernumerary marker chromosomes (SMCs) in Turner syndrome are mostly derived from the Y chromosome. Clin Genet. p. 184–190, 1997.

103. Pearl, J. Probabilistic Reasoning in Intelligent Systems Networks of Plausible Inference. San Mateo. Morgan Kaufmann, 1988

104. Pendharkar, PC., Subramanian, GH., Rodger, JA. A Probabilistic Model for Predicting Software Development Effort. IEEE Transactions on Software Engineering, Vol. 31, No. 7; July 2005.

105. Pradhan, M., Provan G., Middleton, B., Henrion M., Knowledge Engineering for Large Belief Networks, Uncertainty in Artificial Intelligence: Proceedings of the Tenth Conference, 1994.

106. Press, SJ., Bayesian Statistics: Principles, Models, and Applications, Wiley; 1989.

107. Przytula, K., Lu T., Thompson D., Bayesian Network Probabilities for Diagnostic Problems, p. 193-200; 2000.

108. Rajabally, E., Sen, P., Whittle, S., Dalton, J., Aids to Bayesian Belief Network Construction Second IEEE Internattonal Conference On Intelligent Systems; June 2004. p. 7803-8278.

109. Ramoni, M. e Sebastiani, P., Discovering Bayesian Networks in Incomplete Databases, KMi. Technical Report KMiTR46, Knowledge Media Institute. The Open University, United Kingdom; 1997b

110. Ramoni, M. e Sebastiani, P., Efficient Learning Bayesian Networks from Incomplete Databases, KMi. Technical Report KMiTR41, Knowledge Media Institute. The Open University, United Kingdom, 1996c.

111. Ramoni, M. e Sebastiani, P., Parameter Estimation in Bayesian Networks from Incomplete Databases. Intelligent Data Analysis Journal, 2; 1998

112. Ramoni, M. e Sebastiani, P., Robust Learning with Missing Data KMi. Technical Report KMiTR28, Knowledge Media Institute. The Open University, United Kingdom; 1997a.

113. Ramoni, M. e Sebastiani, P., The Use of Exogenous Knowledge to Learn Bayesian Networks from Incomplete Databases, KMi. Technical Report KMiTR44, Knowledge Media Institute. The Open University, United Kingdom; 1996b.

114. Ramoni, M., Sebastiani, P., Learning Bayesian Networks from Incomplete Databases. KMi. Technical Report KMiTR43, Knowledge Media Institute. The Open University, United Kingdom; 1996a.

115. Ramoni, M., Sebastiani, P., Learning Conditional Probabilities from Incomplete Data an Experimental Comparison. In proceedings of the Seventh International Workshop on Artificial Intelligent and Statistics, Morgan Kaufmam, San Mateo, CA; 1999.

116. Ranke MB, Grauer ML. Adult height in Turner's syndrome: Results of a

maturational survey 1993. Horm Res. p. 90-4, 1994.

117. Ranke MB. Turner and Noonan Syndrome S: Disease-Specific Growth and Growth—Promoting Therapies. In: Kelnar CJH, Savage MO, Stirling HF, Saenger P, ed.Growth Disorders. London:Chapman & Hall. P. 623-39, 1998.

118. Rebane, G., Pearl, J. The Recovery of Causal Polytrees from Statistical Dat. Proceedings of UAI'87. p.222-228, Seattle; 1987.

119. Rendell, L., Cho, H. Empirical Learning as a Function of Concept Character, Machine Learning, Volume 5, Issue 3, p.267-298; Aug 1990.

120. Rieser RN, Underwood LE. Turner Syndrome: a guide for families. California:The Turner Syndrome Society; 1992.

121. Rochiccioli P, David M, Malpuech G, Colle M, Limal JM, Battin J, et al. Study of final height in Turner's Syndrome: Ethnic and genetic influences. Acta Paediatr Scand. p. 305-8, 1994.

122. Rodrigues, M. Introdução ao geoprocessamento.In: Simpósio Brasileiro de Geoprocessamento.São Paulo: Sagres Editora; 1990.

123. Romão, W. Descoberta de Conhecimento Relevante em Banco e Dados sobre Ciência e Tecnologia. [Tese]. Univ. Federal de Santa Catarina, Santa Catarina-Brasil; 2002.

124. Rongen Westerlaken C, Rikken B, Vastrick P, Jeuken AH, de Lange MY, Wit JM, et al. Body proportions in individuals with Turner's Syndrome. The Dutch Growth Hormone Working Group. Eur J Pediatr. p. 813-7, 1993.

125. Rosenfeld, RG. Turner syndrome: a guide for physicians. California:The Turner Syndrome Society; 1992.

126. Ross J, Feuillan P, Long L. Lipid abnormalities in Turner's syndrome. J Pediatr. p. 242-5, 1995.

127. Russel, S., Norvig, P. Artificial intelligent: a modern approach. ed. Prentice Hall: New Jersey;1995.

128. Santos, M. Espaço e Método. ed. São Paulo: Nobel; 1988.

129. Schwartz, J., Marcus, A. Mortality and air pollution in London: a time series analysis. American Journal of Epidemiology, 131:185-193; 1990.

130. Simoudis, E. Reality Check for Data Mining. IEEE Expert, Los Alamitos, Vol.11, No.5, p. 26-33; Out. 1996.

131. Singh, M., Valtorta, M. Construction of Bayesian Networks Structures from Data: Belief Survey and an Efficient Algorithm. International Journal of Approximate reasoning, 12, p. 11-131; 1995.

132. Skuse DH, James RS, Bishop DV, et al. Evidence from Turner's syndrome of an imprinted X-linked locus affecting cognitive function. Nature. p. 705–708, 1997.

133. Spirtes, P., Glymour, C., Scheines, R. An Algorithm for Fast Recovery of Sparse Causal Graphs, Social Science Computer Review, 9, 6272; 1991.

134. Suzuki, J. Learning Bayesian Belief Networks Based on the MDL Principle: An Efficient Algorithm Using the Branch and Bound Technique, Proceedings of

the International Conference on Machine Learning. Bally, Italy; 1996.

135. Taylor,J., King,R.D., Altmann,T., Fiehn,O. Application of metabolomics to plant genotype discrimination using statistics and machine learning. Bioinformatics, 18 (Suppl. 2), 241S–248S; 2002.

136. Tesch LG, Rosenfeld RG. Morgagni, Ullrich and Turner: the discovery of gonadal dysgenesis. Endocrinologist 5:327–328, 1995.

137. Tobler, JB., Molla, MN., Nuwaysir, EF., Green, RD., Shavlik, JW. Evaluating machine learning approaches for aiding probe selection for gene-expression arrays. Bioinformatics, 18 (Suppl. 1), 164S–171S; 2002.

138. Turner HH. A syndrome of infantilism, congenital webbed neck and cubitus valgus. Endocrinology 23:566–574, 1938.

139. Ullrich O. Uber typische Kombinationsbilder multipler Abartungen. Z Kinderheilk 49:271–276, 1930.

140. Varrela J, Vinkka H, Alvesalo L. The phenotype of 45,X females: An anthropometric quantification. Ann Hum Biol. p. 53-66, 1994.

141. Wang, K., Sundaresh, S. Selecting Features By Vertical Compactness of Data. In: Liu. J. & Motoda, H. (Eds) Feature Extraction, Construction and Selection: a data mining perspective. Kluwer, 1998.

142. Weiss I. Additional evidence of gradual loss of germ cells in the pathogenesis of streak ovaries in Turner's syndrome. J Med Genet. P. 540 –544, 1971.

143. Wermuth, N., Lauritzen, S. Graphical and Recursive Models for Contingency Tables. Biometrika, 72, 537552; 1983.

144. Wirth, R., Hipp, J. CRISP-DM: Towards a Standard Process Model for Data Mining, DaimlerChrysler Research & Technology FT3/KL e Wilhelm-Schickard-Institute, Univ. of Tübingen; 2000.

145. Wong, ML., Leung, WS. An Efficient Data Mining Method for Learning Bayesian Networks Using an Evolutionary Algorithm-Based Hybrid Approach. IEEE Transactions On Evolutionary Computation, Vol. 8, No. 4, p. 378-404; August 2004.

146. Zhang S-Z., Yang N-H., Wang X-K., Construction and application of bayesian networks in flood decision supporting system. Proceedings of the Fist International Conference on Machine Learning and Cybernetics, Beijing, 4-5 November; 2002.

147. Zhang, B-T,. A Bayesian framework for evolutionary computation. Proceedings of the 1999 Congress on Evolutionary Computation (CEC99), 1:722-728; 1999.

ANEXO I

Acurácia dos casos de Síndromes de Turner aplicando modelo de Rede Neural – Grupo I – Conteúdo para avaliar.

Quadro XIII - Informações sobre variáveis executado pelo UnBMiner aplicando o modelo de Rede Neural

=== Informações Gerais ===		
Modelo Aplicado:		**Rede Neural**
Relacionamentos:		**Null**
Instancias:		**21 casos**
Número de Atributos:		**9**
Tipo	**Sigla**	**Descrição**
Atributos:	**SF**	**- Sexo feminino**
Atributos:	**BE**	**- Baixa estatura**
Atributos:	**TE**	**- Tórax em escudo**
Atributos:	**DG**	**- Disgenesia gonadal**
Atributos:	**UH**	**- Unhas hipoplásicas**
Atributos:	**CV**	**- Cúbito valgo**
Atributos:	**PA**	**- Pescoço alado**
Atributos:	**HMa**	**- Hipertelorismo de mamilos**
Atributos:	**TO**	**- Tendência à obesidade**
Classe:	**TURNER**	**- Síndrome de TURNER**

Tabela I.5 - Informações sobre o modelo de teste executado pelo UnBMiner aplicando o modelo de Rede Neural

Teste do Modelo: === Classificação do Modelo ===	
Taxa de Aprendizagem:	**0.3**
Momento:	**0.2**
Hidden Layer Size:	**5**
Training Time:	**400**
Activation Function:	**Sigmoid**
Decaimento da Taxa de Aprendizagem:	**False**
Entrada numérica de Normalização:	**No**
Função Íngreme de Ativação:	**1.0**

Tabela I.6 - Informações do resumo executado pelo UnBMiner aplicando o modelo de Rede Neural

=== Resumo ===		
Descrição	**Qtde**	**Porcentagem**
Instâncias Corretamente Classificadas:	**21**	**100 %**
Instâncias Incorretamente Classificadas:	**0**	**0 %**
Quadratic loss function:	**-**	**-**
Número Total de Instâncias:	**21**	**-**

Tabela I.7 - Informações sobre o detalhamento da classe executado pelo UnBMiner aplicando o modelo de Rede Neural

=== Acurácia Detalhada por Classe ===

VP Rate –Taxa Verdadeiro Positivo	FP Rate – Taxa Falso Positivo	VN Rate-Taxa Verdadeiro Negativo	FN Rate– Taxa Falso Negativo	Classe
1	0	1	0	NÃO
1	0	1	0	SIM

Tabela I.8 - Informações da matriz de confusão executado pelo UnBMiner aplicando o modelo de Rede Neural.

=== Matriz de Confusão ===

a b <-- classificado como:

a	b		
13	0	**a =**	**NÃO**
0	8	**b =**	**SIM**

Interpretação dos dados:
1 - Através da base dos dados foram identificados que a instância ‘SINDROME DE TURNER ’ assumiu valores ‘NAO’ em 13 casos.

2 – Propagando as evidências usando a classificação por Rede Neural identifica-se que foram gerados 13 casos com os valores ‘NÃO’ para a instância ‘SINDROME DE TURNER ’, ou seja, valores verdadeiros positivos. Assim, obtém-se um resultado de 1 ou 100 % de acertos.

3 – Em geral as evidências probabilísticas geradas pelo modelo rede neural usando como referências dados para ocorrências em Síndromes de Turner foram identificadas com 21 casos, 100 % dos casos como instâncias corretamente classificadas, e com 0 casos com instâncias incorretamente classificados.

4 – Devido ao valor significativo que a rede neural identificou verifica-se que o resultado é satisfatório.

5 – Para uma confirmação mais exata dos resultados devemos ressaltar a importância do número de amostra. Neste caso, considera-se um número pequeno de 84 casos.

Acurácia dos casos de Síndromes de Turner aplicando modelo de Rede Neural – Grupo I – Conteúdo para treinamento.

Quadro XIV - Informações sobre variáveis executado pelo UnBMiner aplicando o modelo de Rede Neural

=== Informações Gerais ===		
Modelo Aplicado:		**Rede Neural**
Relacionamentos:		**Null**
Instancias:		**65 casos**
Número de Atributos:		**9**
Tipo	**Sigla**	**Descrição**
Atributos:	**SF**	**- Sexo feminino**
Atributos:	**BE**	**- Baixa estatura**
Atributos:	**TE**	**- Tórax em escudo**
Atributos:	**DG**	**- Disgenesia gonadal**
Atributos:	**UH**	**- Unhas hipoplásicas**
Atributos:	**CV**	**- Cúbito valgo**
Atributos:	**PA**	**- Pescoço alado**
Atributos:	**HMa**	**- Hipertelorismo de mamilos**
Atributos:	**TO**	**- Tendência à obesidade**
Classe:	**TURNER**	**- Síndrome de TURNER**

Tabela I.9 - Informações sobre o modelo de teste executado pelo UnBMiner aplicando o modelo de Rede Neural

Teste do Modelo: === Classificação do Modelo ===	
Taxa de Aprendizagem:	**0.3**
Momento:	**0.2**
Hidden Layer Size:	**5**
Training Time:	**400**
Activation Function:	**Sigmoid**
Decaimento da Taxa de Aprendizagem:	**false**
Entrada numérica de Normalização:	**No**
Função Íngreme de Ativação:	**1.0**

Tabela I.10 - Informações do resumo executado pelo UnBMiner aplicando o modelo de Rede Neural

=== Resumo ===		
Descrição	**Qtde**	**Porcentagem**
Instâncias Corretamente Classificadas:	**65**	**100 %**
Instâncias Incorretamente Classificadas:	**0**	**0 %**
Quadratic loss function:	**-**	**-**
Número Total de Instâncias:	**65**	**-**

Tabela I.11 - Informações sobre o detalhamento da classe executado pelo UnBMiner aplicando o modelo de Rede Neural

=== Acurácia Detalhada por Classe ===

VP Rate –Taxa Verdadeiro Positivo	FP Rate – Taxa Falso Positivo	VN Rate-Taxa Verdadeiro Negativo	FN Rate– Taxa Falso Negativo	Classe
1	0	1	0	NÃO
1	0	1	0	SIM

Tabela I.12 - Informações da matriz de confusão executado pelo UnBMiner aplicando o modelo de Rede Neural.

=== Matriz de Confusão ===

a	b	<-- classificado como:	
45	0	**A =**	**NÃO**
0	20	**B =**	**SIM**

Interpretação dos dados:
1 - Através da base dos dados foram identificados que a instância ‘SINDROME DE TURNER ’ assumiu valores ‘NAO’ em 45 casos.

2 – Propagando as evidências usando a classificação por Rede Neural identifica-se que foram gerados 45 casos com os valores ‘NÃO’ para a instância ‘SINDROME DE TURNER ’, ou seja, valores verdadeiros positivos. Assim, obtém-se um resultado de 1 ou 100 % de acertos.

3 – Em geral as evidências probabilísticas geradas pelo modelo rede neural usando como referências dados para ocorrências em Síndromes de Turner foram identificadas com 65 casos, 100 % dos casos como instâncias corretamente classificadas, e com 0 casos com instâncias incorretamente classificados.

4 – Devido ao valor significativo que a rede neural identificou verifica-se que o resultado é satisfatório.

5 – Para uma confirmação mais exata dos resultados devemos ressaltar a importância do número de amostra. Neste caso, considera-se um número pequeno de 84 casos.

Acurácia dos casos de Síndromes de Turner aplicando modelo de Rede Neural – Grupo II – Conteúdo para avaliar.

Quadro XV - Informações sobre variáveis executado pelo UnBMiner aplicando o modelo de Rede Neural

=== Informações Gerais ===		
Modelo Aplicado:		**Rede Neural**
Relacionamentos:		**Null**
Instancias:		**21 casos**
Número de Atributos:		**9**
Tipo	**Sigla**	**Descrição**
Atributos:	**SF**	**- Sexo feminino**
Atributos:	**BE**	**- Baixa estatura**
Atributos:	**TE**	**- Tórax em escudo**
Atributos:	**DG**	**- Disgenesia gonadal**
Atributos:	**UH**	**- Unhas hipoplásicas**
Atributos:	**CV**	**- Cúbito valgo**
Atributos:	**PA**	**- Pescoço alado**
Atributos:	**HMa**	**- Hipertelorismo de mamilos**
Atributos:	**TO**	**- Tendência à obesidade**
Classe:	**TURNER**	**- Síndrome de TURNER**

Tabela I. 13 - Informações sobre o modelo de teste executado pelo UnBMiner aplicando o modelo de Rede Neural

Teste do Modelo: === Classificação do Modelo ===	
Taxa de Aprendizagem:	**0.3**
Momento:	**0.2**
Hidden Layer Size:	**5**
Training Time:	**400**
Activation Function:	**Sigmoid**
Decaimento da Taxa de Aprendizagem:	**false**
Entrada numérica de Normalização:	**No**
Função Íngreme de Ativação:	**1.0**

Tabela I.14 - Informações do resumo executado pelo UnBMiner aplicando o modelo de Rede Neural

=== Resumo ===		
Descrição	**Qtde**	**Porcentagem**
Instâncias Corretamente Classificadas:	**21**	**100 %**
Instâncias Incorretamente Classificadas:	**0**	**0 %**
Quadratic loss function:	**-**	**-**
Número Total de Instâncias:	**21**	**-**

Tabela I.15 - Informações sobre o detalhamento da classe executado pelo UnBMiner aplicando o modelo de Rede Neural

=== Acurácia Detalhada por Classe ===

VP Rate –Taxa Verdadeiro Positivo	FP Rate – Taxa Falso Positivo	VN Rate-Taxa Verdadeiro Negativo	FN Rate– Taxa Falso Negativo	Classe
1	0	1	0	NÃO
1	0	1	0	SIM

Tabela I.16 - Informações da matriz de confusão executado pelo UnBMiner aplicando o modelo de Rede Neural.

=== Matriz de Confusão ===

a	**b**	**<-- classificado como:**	
17	0	**a =**	**NÃO**
0	4	**b =**	**SIM**

Interpretação dos dados:
1 - Através da base dos dados foram identificados que a instância ‘SINDROME DE TURNER ’ assumiu valores ‘NAO’ em 17 casos.

2 – Propagando as evidências usando a classificação por Rede Neural identifica-se que foram gerados 17 casos com os valores ‘NÃO’ para a instância ‘SINDROME DE TURNER ’, ou seja, valores verdadeiros positivos. Assim, obtém-se um resultado de 1 ou 100 % de acertos.

3 – Em geral as evidências probabilísticas geradas pelo modelo rede neural usando como referências dados para ocorrências em Síndromes de Turner foram identificadas com 21 casos, 100 % dos casos como instâncias corretamente classificadas, e com 0 casos com instâncias incorretamente classificados.

4 – Devido ao valor significativo que a rede neural identificou verifica-se que o resultado é satisfatório.

5 – Para uma confirmação mais exata dos resultados devemos ressaltar a importância do número de amostra. Neste caso, considera-se um número pequeno de 84 casos.

Acurácia dos casos de Síndromes de Turner aplicando modelo de Rede Neural – Grupo II – Conteúdo para treinamento.

Quadro XVI - Informações sobre variáveis executado pelo UnBMiner aplicando o modelo de Rede Neural

=== Informações Gerais ===		
Modelo Aplicado:		**Rede Neural**
Relacionamentos:		**Null**
Instancias:		**64 casos**
Número de Atributos:		**9**
Tipo	**Sigla**	**Descrição**
Atributos:	**SF**	**- Sexo feminino**
Atributos:	**BE**	**- Baixa estatura**
Atributos:	**TE**	**- Tórax em escudo**
Atributos:	**DG**	**- Disgenesia gonadal**
Atributos:	**UH**	**- Unhas hipoplásicas**
Atributos:	**CV**	**- Cúbito valgo**
Atributos:	**PA**	**- Pescoço alado**
Atributos:	**HMa**	**- Hipertelorismo de mamilos**
Atributos:	**TO**	**- Tendência à obesidade**
Classe:	**TURNER**	**- Síndrome de TURNER**

Tabela I.17 - Informações sobre o modelo de teste executado pelo UnBMiner aplicando o modelo de Rede Neural

Teste do Modelo: === Classificação do Modelo ===	
Taxa de Aprendizagem:	**0.3**
Momento:	**0.2**
Hidden Layer Size:	**5**
Training Time:	**400**
Activation Function:	**Sigmoid**
Decaimento da Taxa de Aprendizagem:	**false**
Entrada numérica de Normalização:	**No**
Função Íngreme de Ativação:	**1.0**

Tabela I.18 - Informações do resumo executado pelo UnBMiner aplicando o modelo de Rede Neural

=== Resumo ===		
Descrição	**Qtde**	**Porcentagem**
Instâncias Corretamente Classificadas:	**64**	**100 %**
Instâncias Incorretamente Classificadas:	**0**	**0 %**
Quadratic loss function:	**-**	**-**
Número Total de Instâncias:	**64**	**-**

Tabela I.19 - Informações sobre o detalhamento da classe executado pelo UnBMiner aplicando o modelo de Rede Neural

=== Acurácia Detalhada por Classe ===

VP Rate –Taxa Verdadeiro Positivo	FP Rate – Taxa Falso Positivo	VN Rate-Taxa Verdadeiro Negativo	FN Rate– Taxa Falso Negativo	Classe
1	0	1	0	NÃO
1	0	1	0	SIM

Tabela I.20 - Informações da matriz de confusão executado pelo UnBMiner aplicando o modelo de Rede Neural.

=== Matriz de Confusão ===

a	**b**	**<-- classificado como:**	
40	0	**a =**	**NÃO**
0	24	**b =**	**SIM**

Interpretação dos dados:
1 - Através da base dos dados foram identificados que a instância ‘SINDROME DE TURNER ’ assumiu valores ‘NAO’ em 40 casos.

2 – Propagando as evidências usando a classificação por Rede Neural identifica-se que foram gerados 40 casos com os valores ‘NÃO’ para a instância ‘SINDROME DE TURNER ’, ou seja, valores verdadeiros positivos. Assim, obtém-se um resultado de 1 ou 100 % de acertos.

3 – Em geral as evidências probabilísticas geradas pelo modelo rede neural usando como referências dados para ocorrências em Síndromes de Turner foram identificadas com 64 casos, 100 % dos casos como instâncias corretamente classificadas, e com 0 casos com instâncias incorretamente classificados.

4 – Devido ao valor significativo que a rede neural identificou verifica-se que o resultado é satisfatório.

5 – Para uma confirmação mais exata dos resultados devemos ressaltar a importância do número de amostra. Neste caso, considera-se um número pequeno de 84 casos.

Acurácia dos casos de Síndromes de Turner aplicando modelo de Rede Neural – Grupo III – Conteúdo para avaliar.

Quadro XVII - Informações sobre variáveis executado pelo UnBMiner aplicando o modelo de Rede Neural

=== Informações Gerais ===		
Modelo Aplicado:		**Rede Neural**
Relacionamentos:		**Null**
Instancias:		**21 casos**
Número de Atributos:		**9**
Tipo	**Sigla**	**Descrição**
Atributos:	**SF**	**- Sexo feminino**
Atributos:	**BE**	**- Baixa estatura**
Atributos:	**TE**	**- Tórax em escudo**
Atributos:	**DG**	**- Disgenesia gonadal**
Atributos:	**UH**	**- Unhas hipoplásicas**
Atributos:	**CV**	**- Cúbito valgo**
Atributos:	**PA**	**- Pescoço alado**
Atributos:	**HMa**	**- Hipertelorismo de mamilos**
Atributos:	**TO**	**- Tendência à obesidade**
Classe:	**TURNER**	**- Síndrome de TURNER**

Tabela I.21 - Informações sobre o modelo de teste executado pelo UnBMiner aplicando o modelo de Rede Neural

Teste do Modelo: === Classificação do Modelo ===	
Taxa de Aprendizagem:	**0.3**
Momento:	**0.2**
Hidden Layer Size:	**5**
Training Time:	**400**
Activation Function:	**Sigmoid**
Decaimento da Taxa de Aprendizagem:	**false**
Entrada numérica de Normalização:	**No**
Função Íngreme de Ativação:	**1.0**

Tabela I.22 Informações do resumo executado pelo UnBMiner aplicando o modelo de Rede Neural

=== Resumo ===		
Descrição	**Qtde**	**Porcentagem**
Instâncias Corretamente Classificadas:	**21**	**100 %**
Instâncias Incorretamente Classificadas:	**0**	**0 %**
Quadratic loss function:	**-**	**-**
Número Total de Instâncias:	**21**	**-**

Tabela I.23 Informações sobre o detalhamento da classe executado pelo UnBMiner aplicando o modelo de Rede Neural

=== Acurácia Detalhada por Classe ===

VP Rate –Taxa Verdadeiro Positivo	FP Rate – Taxa Falso Positivo	VN Rate-Taxa Verdadeiro Negativo	FN Rate– Taxa Falso Negativo	Classe
1	0	1	0	NÃO
1	0	1	0	SIM

Tabela I.24 Informações da matriz de confusão executado pelo UnBMiner aplicando o modelo de Rede Neural.

=== Matriz de Confusão ===

a	**b**	**<-- classificado como:**
14	0	**a = NÃO**
0	7	**b = SIM**

Interpretação dos dados:
1 - Através da base dos dados foram identificados que a instância ‘SINDROME DE TURNER ’ assumiu valores ‘NAO’ em 14 casos.

2 – Propagando as evidências usando a classificação por Rede Neural identifica-se que foram gerados 14 casos com os valores ‘NÃO’ para a instância ‘SINDROME DE TURNER ’, ou seja, valores verdadeiros positivos. Assim, obtém-se um resultado de 1 ou 100 % de acertos.

3 – Em geral as evidências probabilísticas geradas pelo modelo rede neural usando como referências dados para ocorrências em Síndromes de Turner foram identificadas com 21 casos, 100 % dos casos como instâncias corretamente classificadas, e com 0 casos com instâncias incorretamente classificados.

4 – Devido ao valor significativo que a rede neural identificou verifica-se que o resultado é satisfatório.

5 – Para uma confirmação mais exata dos resultados devemos ressaltar a importância do número de amostra. Neste caso, considera-se um número pequeno de 84 casos.

Acurácia dos casos de Síndromes de Turner aplicando modelo de Rede Neural – Grupo III – Conteúdo para treinamento.

Quadro XVIII - Informações sobre variáveis executado pelo UnBMiner aplicando o modelo de Rede Neural

=== Informações Gerais ===		
Modelo Aplicado:		**Rede Neural**
Relacionamentos:		**Null**
Instancias:		**64 casos**
Número de Atributos:		**9**
Tipo	**Sigla**	**Descrição**
Atributos:	**SF**	**- Sexo feminino**
Atributos:	**BE**	**- Baixa estatura**
Atributos:	**TE**	**- Tórax em escudo**
Atributos:	**DG**	**- Disgenesia gonadal**
Atributos:	**UH**	**- Unhas hipoplásicas**
Atributos:	**CV**	**- Cúbito valgo**
Atributos:	**PA**	**- Pescoço alado**
Atributos:	**HMa**	**- Hipertelorismo de mamilos**
Atributos:	**TO**	**- Tendência à obesidade**
Classe:	**TURNER**	**- Síndrome de TURNER**

Tabela I.25 - Informações sobre o modelo de teste executado pelo UnBMiner aplicando o modelo de Rede Neural

Teste do Modelo: === Classificação do Modelo ===	
Taxa de Aprendizagem:	**0.3**
Momento:	**0.2**
Hidden Layer Size:	**5**
Training Time:	**400**
Activation Function:	**Sigmoid**
Decaimento da Taxa de Aprendizagem:	**false**
Entrada numérica de Normalização:	**No**
Função Íngreme de Ativação:	**1.0**

Tabela I.26 - Informações do resumo executado pelo UnBMiner aplicando o modelo de Rede Neural

=== Resumo ===

Descrição	**Qtde**	**Porcentagem**
Instâncias Corretamente Classificadas:	**64**	**100 %**
Instâncias Incorretamente Classificadas:	**0**	**0 %**
Quadratic loss function:	**-**	**-**
Número Total de Instâncias:	**64**	**-**

Tabela I.27 Informações sobre o detalhamento da classe executado pelo UnBMiner aplicando o modelo de Rede Neural

=== Acurácia Detalhada por Classe ===

VP Rate –Taxa Verdadeiro Positivo	FP Rate – Taxa Falso Positivo	VN Rate-Taxa Verdadeiro Negativo	FN Rate– Taxa Falso Negativo	Classe
1	0	1	0	NÃO
1	0	1	0	SIM

Tabela I.28 Informações da matriz de confusão executado pelo UnBMiner aplicando o modelo de Rede Neural.

=== Matriz de Confusão ===

a b <-- classificado como:

a	b		
43	0	**a =**	**NÃO**
0	21	**b =**	**SIM**

Interpretação dos dados:
1 - Através da base dos dados foram identificados que a instância ‘SINDROME DE TURNER ’ assumiu valores ‘NAO’ em 43 casos.

2 – Propagando as evidências usando a classificação por Rede Neural identifica-se que foram gerados 43 casos com os valores ‘NÃO’ para a instância ‘SINDROME DE TURNER ’, ou seja, valores verdadeiros positivos. Assim, obtém-se um resultado de 1 ou 100 % de acertos.

3 – Em geral as evidências probabilísticas geradas pelo modelo rede neural usando como referências dados para ocorrências em Síndromes de Turner foram identificadas com 64 casos, 100 % dos casos como instâncias corretamente classificadas, e com 0 casos com instâncias incorretamente classificados.

4 – Devido ao valor significativo que a rede neural identificou verifica-se que o resultado é satisfatório.

5 – Para uma confirmação mais exata dos resultados devemos ressaltar a importância do número de amostra. Neste caso, considera-se um número pequeno de 84 casos.

Acurácia dos casos de Síndromes de Turner aplicando modelo de Rede Neural – Grupo IV – Conteúdo para avaliar.

Quadro XIX - Informações sobre variáveis executado pelo UnBMiner aplicando o modelo de Rede Neural

=== Informações Gerais ===		
Modelo Aplicado:		**Rede Neural**
Relacionamentos:		**Null**
Instancias:		**21 casos**
Número de Atributos:		**9**
Tipo	**Sigla**	**Descrição**
Atributos:	**SF**	**- Sexo feminino**
Atributos:	**BE**	**- Baixa estatura**
Atributos:	**TE**	**- Tórax em escudo**
Atributos:	**DG**	**- Disgenesia gonadal**
Atributos:	**UH**	**- Unhas hipoplásicas**
Atributos:	**CV**	**- Cúbito valgo**
Atributos:	**PA**	**- Pescoço alado**
Atributos:	**HMa**	**- Hipertelorismo de mamilos**
Atributos:	**TO**	**- Tendência à obesidade**
Classe:	**TURNER**	**- Síndrome de TURNER**

Tabela I.29 - Informações sobre o modelo de teste executado pelo UnBMiner aplicando o modelo de Rede Neural

Teste do Modelo: === Classificação do Modelo ===	
Taxa de Aprendizagem:	**0.3**
Momento:	**0.2**
Hidden Layer Size:	**5**
Training Time:	**400**
Activation Function:	**Sigmoid**
Decaimento da Taxa de Aprendizagem:	**false**
Entrada numérica de Normalização:	**No**
Função Íngreme de Ativação:	**1.0**

Tabela I.30 - Informações do resumo executado pelo UnBMiner aplicando o modelo de Rede Neural

=== Resumo ===		
Descrição	**Qtde**	**Porcentagem**
Instâncias Corretamente Classificadas:	**21**	**100 %**
Instâncias Incorretamente Classificadas:	**0**	**0 %**
Quadratic loss function:	**-**	**-**
Número Total de Instâncias:	**21**	**-**

Tabela I.31 - Informações sobre o detalhamento da classe executado pelo UnBMiner aplicando o modelo de Rede Neural

=== Acurácia Detalhada por Classe ===

VP Rate –Taxa Verdadeiro Positivo	FP Rate – Taxa Falso Positivo	VN Rate-Taxa Verdadeiro Negativo	FN Rate– Taxa Falso Negativo	Classe
1	0	1	0	NÃO
1	0	1	0	SIM

Tabela I.32 - Informações da matriz de confusão executado pelo UnBMiner aplicando o modelo de Rede Neural.

=== Matriz de Confusão ===

a	**b**	**<-- classificado como:**	
12	0	**A =**	**NÃO**
0	9	**B =**	**SIM**

Interpretação dos dados:
1 - Através da base dos dados foram identificados que a instância ‘SINDROME DE TURNER ’ assumiu valores ‘NAO’ em 12 casos.

2 – Propagando as evidências usando a classificação por Rede Neural identifica-se que foram gerados 12 casos com os valores ‘NÃO’ para a instância ‘SINDROME DE TURNER ’, ou seja, valores verdadeiros positivos. Assim, obtém-se um resultado de 1 ou 100 % de acertos.

3 – Em geral as evidências probabilísticas geradas pelo modelo rede neural usando como referências dados para ocorrências em Síndromes de Turner foram identificadas com 21 casos, 100 % dos casos como instâncias corretamente classificadas, e com 0 casos com instâncias incorretamente classificados.

4 – Devido ao valor significativo que a rede neural identificou verifica-se que o resultado é satisfatório.

5 – Para uma confirmação mais exata dos resultados devemos ressaltar a importância do número de amostra. Neste caso, considera-se um número pequeno de 84 casos.

Acurácia dos casos de Síndromes de Turner aplicando modelo de Rede Neural – Grupo IV – Conteúdo para treinamento.

Quadro XX - Informações sobre variáveis executado pelo UnBMiner aplicando o modelo de Rede Neural

=== Informações Gerais ===		
Modelo Aplicado:		**Rede Neural**
Relacionamentos:		**Null**
Instancias:		**63 casos**
Número de Atributos:		**9**
Tipo	**Sigla**	**Descrição**
Atributos:	**SF**	**- Sexo feminino**
Atributos:	**BE**	**- Baixa estatura**
Atributos:	**TE**	**- Tórax em escudo**
Atributos:	**DG**	**- Disgenesia gonadal**
Atributos:	**UH**	**- Unhas hipoplásicas**
Atributos:	**CV**	**- Cúbito valgo**
Atributos:	**PA**	**- Pescoço alado**
Atributos:	**HMa**	**- Hipertelorismo de mamilos**
Atributos:	**TO**	**- Tendência à obesidade**
Classe:	**TURNER**	**- Síndrome de TURNER**

Tabela I.33 - Informações sobre o modelo de teste executado pelo UnBMiner aplicando o modelo de Rede Neural

Teste do Modelo: === Classificação do Modelo ===	
Taxa de Aprendizagem:	**0.3**
Momento:	**0.2**
Hidden Layer Size:	**5**
Training Time:	**400**
Activation Function:	**Sigmoid**
Decaimento da Taxa de Aprendizagem:	**false**
Entrada numérica de Normalização:	**No**
Função Íngreme de Ativação:	**1.0**

Tabela I.34 - Informações do resumo executado pelo UnBMiner aplicando o modelo de Rede Neural

=== Resumo ===		
Descrição	**Qtde**	**Porcentagem**
Instâncias Corretamente Classificadas:	**63**	**100 %**
Instâncias Incorretamente Classificadas:	**0**	**0 %**
Quadratic loss function:	**-**	**-**
Número Total de Instâncias:	**63**	**-**

Tabela I.35 - Informações sobre o detalhamento da classe executado pelo UnBMiner aplicando o modelo de Rede Neural

=== Acurácia Detalhada por Classe ===

VP Rate –Taxa Verdadeiro Positivo	FP Rate – Taxa Falso Positivo	VN Rate-Taxa Verdadeiro Negativo	FN Rate– Taxa Falso Negativo	Classe
1	0	1	0	NÃO
1	0	1	0	SIM

Tabela I.36 - Informações da matriz de confusão executado pelo UnBMiner aplicando o modelo de Rede Neural.

=== Matriz de Confusão ===

a	b	<-- classificado como:	
44	0	a =	NÃO
0	19	b =	SIM

Interpretação dos dados:
1 - Através da base dos dados foram identificados que a instância ‘SINDROME DE TURNER ’ assumiu valores ‘NAO’ em 44 casos.

2 – Propagando as evidências usando a classificação por Rede Neural identifica-se que foram gerados 44 casos com os valores ‘NÃO’ para a instância ‘SINDROME DE TURNER ’, ou seja, valores verdadeiros positivos. Assim, obtém-se um resultado de 1 ou 100 % de acertos.

3 – Em geral as evidências probabilísticas geradas pelo modelo rede neural usando como referências dados para ocorrências em Síndromes de Turner foram identificadas com 63 casos, 100 % dos casos como instâncias corretamente classificadas, e com 0 casos com instâncias incorretamente classificados.

4 – Devido ao valor significativo que a rede neural identificou verifica-se que o resultado é satisfatório.

5 – Para uma confirmação mais exata dos resultados devemos ressaltar a importância do número de amostra. Neste caso, considera-se um número pequeno de 84 casos.

Acurácia dos casos de Síndromes de Turner aplicando modelo de Naïve Bayes – Grupo I – Conteúdo para avaliar.

Quadro XXI - Informações sobre variáveis executado pelo UnBMiner aplicando o modelo de Naïve Bayes

=== Informações Gerais ===		
Modelo Aplicado:		**Naïve Bayes**
Relacionamentos:		**Null**
Instancias:		**21 casos**
Número de Atributos:		**9**
Tipo	**Sigla**	**Descrição**
Atributos:	**SF**	**- Sexo feminino**
Atributos:	**BE**	**- Baixa estatura**
Atributos:	**TE**	**- Tórax em escudo**
Atributos:	**DG**	**- Disgenesia gonadal**
Atributos:	**UH**	**- Unhas hipoplásicas**
Atributos:	**CV**	**- Cúbito valgo**
Atributos:	**PA**	**- Pescoço alado**
Atributos:	**HMa**	**- Hipertelorismo de mamilos**
Atributos:	**TO**	**- Tendência à obesidade**
Classe:	**TURNER**	**- Síndrome de TURNER**
Modelo:	**Rede bayesiana**	

Tabela I.37 - Informações do resumo executado pelo UnBMiner aplicando o modelo de Naïve Bayes

=== Resumo ===

Descrição	**Qtde**	**Porcentagem**
Instâncias Corretamente Classificadas:	**19**	**90,47 %**
Instâncias Incorretamente Classificadas:	**2**	**9,52 %**
Quadratic loss function:	**-**	**-**
Número Total de Instâncias:	**21**	**-**

Tabela I.38 - Informações sobre o detalhamento da classe executado pelo UnBMiner aplicando o modelo de Naïve Bayes

=== Acurácia Detalhada por Classe ===

VP Rate –Taxa Verdadeiro Positivo	**FP Rate – Taxa Falso Positivo**	**VN Rate-Taxa Verdadeiro Negativo**	**FN Rate– Taxa Falso Negativo**	**Classe**
1	0,25	0,75	0	NÃO
0,75	0	1	0,25	SIM

Tabela I.39 - Informações da matriz de confusão executado pelo UnBMiner aplicando o modelo de Naïve Bayes.

=== Matriz de Confusão ===

a	**b**	**<-- classificado como:**	
a	**b**		
13	0	**a =**	**NÃO**
2	6	**b =**	**SIM**

Interpretação dos dados:
1 - Através da base dos dados foram identificados que a instância ‘SINDROME DE TURNER ’ assumiu valores ‘NAO’ em 13 casos.

2 – Propagando as evidências usando a classificação por Naïve Bayes identifica-se que foram gerados 13 casos com os valores ‘NÃO’ para a instância ‘SINDROME DE TURNER ’, ou seja, valores verdadeiros positivos. Assim, obtém-se um resultado de 0,9047 ou aproximadamente 91% de acerto.

3 – Em geral as evidências probabilísticas geradas pelo modelo Naïve bayes usando como referências dados para ocorrências em Síndromes de Turner foram identificadas com 21 casos, aproximadamente 91%, dos casos como instâncias corretamente classificadas, e com aproximadamente 9%, com instâncias incorretamente classificadas.

4 – Para uma confirmação mais exata dos resultados devemos ressaltar a importância do número de amostra. Neste caso, considera-se um número pequeno de 84 casos.

Acurácia dos casos de Síndromes de Turner aplicando modelo de Naïve Bayes – Grupo I – Conteúdo para treinamento.

Quadro XXII - Informações sobre variáveis executado pelo UnBMiner aplicando o modelo de Naïve Bayes

=== Informações Gerais ===		
Modelo Aplicado:		**Naïve Bayes**
Relacionamentos:		**Null**
Instancias:		**65 casos**
Número de Atributos:		**9**
Tipo	**Sigla**	**Descrição**
Atributos:	**SF**	**- Sexo feminino**
Atributos:	**BE**	**- Baixa estatura**
Atributos:	**TE**	**- Tórax em escudo**
Atributos:	**DG**	**- Disgenesia gonadal**
Atributos:	**UH**	**- Unhas hipoplásicas**
Atributos:	**CV**	**- Cúbito valgo**
Atributos:	**PA**	**- Pescoço alado**
Atributos:	**HMa**	**- Hipertelorismo de mamilos**
Atributos:	**TO**	**- Tendência à obesidade**
Classe:	**TURNER**	**- Síndrome de TURNER**
Modelo:	**Rede bayesiana**	

Tabela I.40 - Informações do resumo executado pelo UnBMiner aplicando o modelo de Naïve Bayes

=== Resumo ===

Descrição	**Qtde**	**Porcentagem**
Instâncias Corretamente Classificadas:	**58**	**89,23 %**
Instâncias Incorretamente Classificadas:	**7**	**10,76 %**
Quadratic loss function:	**-**	**-**
Número Total de Instâncias:	**65**	**-**

Tabela I.41 - Informações sobre o detalhamento da classe executado pelo UnBMiner aplicando o modelo de Naïve Bayes

=== Acurácia Detalhada por Classe ===

VP Rate –Taxa Verdadeiro Positivo	**FP Rate – Taxa Falso Positivo**	**VN Rate-Taxa Verdadeiro Negativo**	**FN Rate– Taxa Falso Negativo**	**Classe**
1	0,35	0,65	0	NÃO
0,65	0	1	0,35	SIM

Tabela I.42 - Informações da matriz de confusão executado pelo UnBMiner aplicando o modelo de Naïve Bayes.

=== Matriz de Confusão ===

a b <-- classificado como:

a	**b**		
45	0	**a =**	**NÃO**
7	13	**b =**	**SIM**

Interpretação dos dados:
1 - Através da base dos dados foram identificados que a instância ‘SINDROME DE TURNER ’ assumiu valores ‘NAO’ em 45 casos.

2 – Propagando as evidências usando a classificação por Naïve Bayes identifica-se que foram gerados 45 casos com os valores ‘NÃO’ para a instância ‘SINDROME DE TURNER ’, ou seja, valores verdadeiros positivos. Assim, obtém-se um resultado de 0,8923 ou aproximadamente 89% de acerto.

3 – Em geral as evidências probabilísticas geradas pelo modelo Naïve bayes usando como referências dados para ocorrências em Síndromes de Turner foram identificadas com 65 casos, aproximadamente 89%, dos casos como instâncias corretamente classificadas, e com aproximadamente 11%, com instâncias incorretamente classificadas.

4 – Para uma confirmação mais exata dos resultados devemos ressaltar a importância do número de amostra. Neste caso, considera-se um número pequeno de 84 casos.

Acurácia dos casos de Síndromes de Turner aplicando modelo de Naïve Bayes – Grupo II – Conteúdo para avaliar.

Quadro XXIII - Informações sobre variáveis executado pelo UnBMiner aplicando o modelo de Naïve Bayes

=== Informações Gerais ===		
Modelo Aplicado:		**Naïve Bayes**
Relacionamentos:		**Null**
Instancias:		**21 casos**
Número de Atributos:		**9**
Tipo	**Sigla**	**Descrição**
Atributos:	**SF**	**- Sexo feminino**
Atributos:	**BE**	**- Baixa estatura**
Atributos:	**TE**	**- Tórax em escudo**
Atributos:	**DG**	**- Disgenesia gonadal**
Atributos:	**UH**	**- Unhas hipoplásicas**
Atributos:	**CV**	**- Cúbito valgo**
Atributos:	**PA**	**- Pescoço alado**
Atributos:	**HMa**	**- Hipertelorismo de mamilos**
Atributos:	**TO**	**- Tendência à obesidade**
Classe:	**TURNER**	**- Síndrome de TURNER**
Modelo:	**Rede bayesiana**	

Tabela I.43 - Informações do resumo executado pelo UnBMiner aplicando o modelo de Naïve Bayes

=== Resumo ===

Descrição	**Qtde**	**Porcentagem**
Instâncias Corretamente Classificadas:	**19**	**90,47 %**
Instâncias Incorretamente Classificadas:	**2**	**9,52 %**
Quadratic loss function:	**0.1905**	**-**
Número Total de Instâncias:	**21**	**-**

Tabela I.44 - Informações sobre o detalhamento da classe executado pelo UnBMiner aplicando o modelo de Naïve Bayes

=== Acurácia Detalhada por Classe ===

VP Rate –Taxa Verdadeiro Positivo	**FP Rate – Taxa Falso Positivo**	**VN Rate-Taxa Verdadeiro Negativo**	**FN Rate– Taxa Falso Negativo**	**Classe**
1	0,5	0,5	0	NÃO
0,5	0	1	0,5	SIM

Tabela I.45 - Informações da matriz de confusão executado pelo UnBMiner aplicando o modelo de Naïve Bayes.

=== Matriz de Confusão ===

a b <-- classificado como:

a	**b**		
17	0	**a =**	**NÃO**
2	2	**b =**	**SIM**

Interpretação dos dados:
1 - Através da base dos dados foram identificados que a instância ‘SINDROME DE TURNER ’ assumiu valores ‘NAO’ em 17 casos.

2 – Propagando as evidências usando a classificação por Naïve Bayes identifica-se que foram gerados 17 casos com os valores ‘NÃO’ para a instância ‘SINDROME DE TURNER ’, ou seja, valores verdadeiros positivos. Assim, obtém-se um resultado de 0,9047 ou aproximadamente 91% de acerto.

3 – Em geral as evidências probabilísticas geradas pelo modelo Naïve bayes usando como referências dados para ocorrências em Síndromes de Turner foram identificadas com 21 casos, aproximadamente 91%, dos casos como instâncias corretamente classificadas, e com aproximadamente 9%, com instâncias incorretamente classificadas.

4 – Para uma confirmação mais exata dos resultados devemos ressaltar a importância do número de amostra. Neste caso, considera-se um número pequeno de 84 casos.

Acurácia dos casos de Síndromes de Turner aplicando modelo de Naïve Bayes – Grupo II – Conteúdo para treinamento.

Quadro XXIV - Informações sobre variáveis executado pelo UnBMiner aplicando o modelo de Naïve Bayes

=== Informações Gerais ===		
Modelo Aplicado:		**Naïve Bayes**
Relacionamentos:		**Null**
Instancias:		**64 casos**
Número de Atributos:		**9**
Tipo	**Sigla**	**Descrição**
Atributos:	**SF**	**- Sexo feminino**
Atributos:	**BE**	**- Baixa estatura**
Atributos:	**TE**	**- Tórax em escudo**
Atributos:	**DG**	**- Disgenesia gonadal**
Atributos:	**UH**	**- Unhas hipoplásicas**
Atributos:	**CV**	**- Cúbito valgo**
Atributos:	**PA**	**- Pescoço alado**
Atributos:	**HMa**	**- Hipertelorismo de mamilos**
Atributos:	**TO**	**- Tendência à obesidade**
Classe:	**TURNER**	**- Síndrome de TURNER**
Modelo:	**Rede bayesiana**	

Tabela I.46 - Informações do resumo executado pelo UnBMiner aplicando o modelo de Naïve Bayes

=== Resumo ===

Descrição	**Qtde**	**Porcentagem**
Instâncias Corretamente Classificadas:	**58**	**90,62 %**
Instâncias Incorretamente Classificadas:	**6**	**9,37 %**
Quadratic loss function:	**0.1875**	**-**
Número Total de Instâncias:	**64**	**-**

Tabela I.47 - Informações sobre o detalhamento da classe executado pelo UnBMiner aplicando o modelo de Naïve Bayes

=== Acurácia Detalhada por Classe ===

VP Rate –Taxa Verdadeiro Positivo	**FP Rate – Taxa Falso Positivo**	**VN Rate-Taxa Verdadeiro Negativo**	**FN Rate– Taxa Falso Negativo**	**Classe**
1	0,25	0,75	0	NÃO
0,75	0	1	0,25	SIM

Tabela I.48 - Informações da matriz de confusão executado pelo UnBMiner aplicando o modelo de Naïve Bayes.

=== Matriz de Confusão ===

a b <-- classificado como:

a	**b**		
40	0	**a =**	**NÃO**
6	18	**b =**	**SIM**

Interpretação dos dados:
1 - Através da base dos dados foram identificados que a instância ‘SINDROME DE TURNER ’ assumiu valores ‘NAO’ em 40 casos.

2 – Propagando as evidências usando a classificação por Naïve Bayes identifica-se que foram gerados 40 casos com os valores ‘NÃO’ para a instância ‘SINDROME DE TURNER ’, ou seja, valores verdadeiros positivos. Assim, obtém-se um resultado de 0,906 ou aproximadamente 91% de acerto.

3 – Em geral as evidências probabilísticas geradas pelo modelo Naïve bayes usando como referências dados para ocorrências em Síndromes de Turner foram identificadas com 65 casos, aproximadamente 91%, dos casos como instâncias corretamente classificadas, e com aproximadamente 9%, com instâncias incorretamente classificadas.

4 – Para uma confirmação mais exata dos resultados devemos ressaltar a importância do número de amostra. Neste caso, considera-se um número pequeno de 84 casos.

Acurácia dos casos de Síndromes de Turner aplicando modelo de Naïve Bayes – Grupo III – Conteúdo para avaliar.

Quadro XXV - Informações sobre variáveis executado pelo UnBMiner aplicando o modelo de Naïve Bayes

=== Informações Gerais ===		
Modelo Aplicado:		**Naïve Bayes**
Relacionamentos:		**Null**
Instancias:		**21 casos**
Número de Atributos:		**9**
Tipo	**Sigla**	**Descrição**
Atributos:	**SF**	**- Sexo feminino**
Atributos:	**BE**	**- Baixa estatura**
Atributos:	**TE**	**- Tórax em escudo**
Atributos:	**DG**	**- Disgenesia gonadal**
Atributos:	**UH**	**- Unhas hipoplásicas**
Atributos:	**CV**	**- Cúbito valgo**
Atributos:	**PA**	**- Pescoço alado**
Atributos:	**HMa**	**- Hipertelorismo de mamilos**
Atributos:	**TO**	**- Tendência à obesidade**
Classe:	**TURNER**	**- Síndrome de TURNER**
Modelo:	**Rede bayesiana**	

Tabela I.49 - Informações do resumo executado pelo UnBMiner aplicando o modelo de Naïve Bayes

=== Resumo ===

Descrição	**Qtde**	**Porcentagem**
Instâncias Corretamente Classificadas:	**17**	**80,95 %**
Instâncias Incorretamente Classificadas:	**4**	**19,04 %**
Quadratic loss function:	**0,381**	**-**
Número Total de Instâncias:	**21**	**-**

Tabela I.50 - Informações sobre o detalhamento da classe executado pelo UnBMiner aplicando o modelo de Naïve Bayes

=== Acurácia Detalhada por Classe ===

VP Rate –Taxa Verdadeiro Positivo	**FP Rate – Taxa Falso Positivo**	**VN Rate-Taxa Verdadeiro Negativo**	**FN Rate– Taxa Falso Negativo**	**Classe**
1	0,571	0,429	0	NÃO
0,429	0	1	0,571	SIM

Tabela I.51 Informações da matriz de confusão executado pelo UnBMiner aplicando o modelo de Naïve Bayes.

=== Matriz de Confusão ===

a b <-- classificado como:

a	**b**		
14	0	**a =**	**NÃO**
4	3	**b =**	**SIM**

Interpretação dos dados:
1 - Através da base dos dados foram identificados que a instância ‘SINDROME DE TURNER ’ assumiu valores ‘NAO’ em 14 casos.

2 – Propagando as evidências usando a classificação por Naïve Bayes identifica-se que foram gerados 14 casos com os valores ‘NÃO’ para a instância ‘SINDROME DE TURNER ’, ou seja, valores verdadeiros positivos. Assim, obtém-se um resultado de 0,8045 ou aproximadamente 81% de acerto.

3 – Em geral as evidências probabilísticas geradas pelo modelo Naïve bayes usando como referências dados para ocorrências em Síndromes de Turner foram identificadas com 21 casos, aproximadamente 81%, dos casos como instâncias corretamente classificadas, e com aproximadamente 19%, com instâncias incorretamente classificadas.

4 – Para uma confirmação mais exata dos resultados devemos ressaltar a importância do número de amostra. Neste caso, considera-se um número pequeno de 84 casos.

Acurácia dos casos de Síndromes de Turner aplicando modelo de Naïve Bayes – Grupo III – Conteúdo para treinamento.

Quadro XXVI - Informações sobre variáveis executado pelo UnBMiner aplicando o modelo de Naïve Bayes

=== Informações Gerais ===		
Modelo Aplicado:		**Naïve Bayes**
Relacionamentos:		**Null**
Instancias:		**64 casos**
Número de Atributos:		**9**
Tipo	**Sigla**	**Descrição**
Atributos:	**SF**	**- Sexo feminino**
Atributos:	**BE**	**- Baixa estatura**
Atributos:	**TE**	**- Tórax em escudo**
Atributos:	**DG**	**- Disgenesia gonadal**
Atributos:	**UH**	**- Unhas hipoplásicas**
Atributos:	**CV**	**- Cúbito valgo**
Atributos:	**PA**	**- Pescoço alado**
Atributos:	**HMa**	**- Hipertelorismo de mamilos**
Atributos:	**TO**	**- Tendência à obesidade**
Classe:	**TURNER**	**- Síndrome de TURNER**
Modelo:	**Rede bayesiana**	

Tabela I.52 - Informações do resumo executado pelo UnBMiner aplicando o modelo de Naïve Bayes

=== Resumo ===

Descrição	**Qtde**	**Porcentagem**
Instâncias Corretamente Classificadas:	**59**	**92,18 %**
Instâncias Incorretamente Classificadas:	**5**	**7,81 %**
Quadratic loss function:	**0.1563**	**-**
Número Total de Instâncias:	**64**	**-**

Tabela I.53 - Informações sobre o detalhamento da classe executado pelo UnBMiner aplicando o modelo de Naïve Bayes

=== Acurácia Detalhada por Classe ===

VP Rate –Taxa Verdadeiro Positivo	**FP Rate – Taxa Falso Positivo**	**VN Rate-Taxa Verdadeiro Negativo**	**FN Rate– Taxa Falso Negativo**	**Classe**
1	0,238	0,762	0	NÃO
0,762	0	1	0,238	SIM

Tabela I.54 - Informações da matriz de confusão executado pelo UnBMiner aplicando o modelo de Naïve Bayes.

=== Matriz de Confusão ===

a b <-- classificado como:

a	**b**		
43	0	**A =**	**NÃO**
5	16	**B =**	**SIM**

Interpretação dos dados:
1 - Através da base dos dados foram identificados que a instância ‘SINDROME DE TURNER ’ assumiu valores ‘NAO’ em 43 casos.

2 – Propagando as evidências usando a classificação por Naïve Bayes identifica-se que foram gerados 43 casos com os valores ‘NÃO’ para a instância ‘SINDROME DE TURNER ’, ou seja, valores verdadeiros positivos. Assim, obtém-se um resultado de 0,9218 ou aproximadamente 92% de acerto.

3 – Em geral as evidências probabilísticas geradas pelo modelo Naïve bayes usando como referências dados para ocorrências em Síndromes de Turner foram identificadas com 64 casos, aproximadamente 92%, dos casos como instâncias corretamente classificadas, e com aproximadamente 8%, com instâncias incorretamente classificadas.

4 – Para uma confirmação mais exata dos resultados devemos ressaltar a importância do número de amostra. Neste caso, considera-se um número pequeno de 84 casos.

Acurácia dos casos de Síndromes de Turner aplicando modelo de Naïve Bayes – Grupo IV – Conteúdo para avaliar.

Quadro XXVII - Informações sobre variáveis executado pelo UnBMiner aplicando o modelo de Naïve Bayes

=== Informações Gerais ===		
Modelo Aplicado:		**Naïve Bayes**
Relacionamentos:		**Null**
Instancias:		**21 casos**
Número de Atributos:		**9**
Tipo	**Sigla**	**Descrição**
Atributos:	**SF**	**- Sexo feminino**
Atributos:	**BE**	**- Baixa estatura**
Atributos:	**TE**	**- Tórax em escudo**
Atributos:	**DG**	**- Disgenesia gonadal**
Atributos:	**UH**	**- Unhas hipoplásicas**
Atributos:	**CV**	**- Cúbito valgo**
Atributos:	**PA**	**- Pescoço alado**
Atributos:	**HMa**	**- Hipertelorismo de mamilos**
Atributos:	**TO**	**- Tendência à obesidade**
Classe:	**TURNER**	**- Síndrome de TURNER**
Modelo:	**Rede bayesiana**	

Tabela I.55 - Informações do resumo executado pelo UnBMiner aplicando o modelo de Naïve Bayes

=== Resumo ===		
Descrição	**Qtde**	**Porcentagem**
Instâncias Corretamente Classificadas:	**18**	**85,71 %**
Instâncias Incorretamente Classificadas:	**3**	**14,28 %**
Quadratic loss function:	**0.2857**	**-**
Número Total de Instâncias:	**21**	**-**

Tabela I.56 - Informações sobre o detalhamento da classe executado pelo UnBMiner aplicando o modelo de Naïve Bayes

=== Acurácia Detalhada por Classe ===				
VP Rate –Taxa Verdadeiro Positivo	**FP Rate – Taxa Falso Positivo**	**VN Rate-Taxa Verdadeiro Negativo**	**FN Rate– Taxa Falso Negativo**	**Classe**
1	0,333	0,667	0	NÃO
0,667	0	1	0,333	SIM

Tabela I.57 - Informações da matriz de confusão executado pelo UnBMiner aplicando o modelo de Naïve Bayes.

=== Matriz de Confusão ===			
a b <-- classificado como:			
a	**b**		
12	0	**a =**	**NÃO**
3	6	**b =**	**SIM**

Interpretação dos dados:
1 - Através da base dos dados foram identificados que a instância ‘SINDROME DE TURNER ’ assumiu valores ‘NAO’ em 12 casos.

2 – Propagando as evidências usando a classificação por Naïve Bayes identifica-se que foram gerados 12 casos com os valores ‘NÃO’ para a instância ‘SINDROME DE TURNER ’, ou seja, valores verdadeiros positivos. Assim, obtém-se um resultado de 0,8571 ou aproximadamente 86% de acerto.

3 – Em geral as evidências probabilísticas geradas pelo modelo Naïve bayes usando como referências dados para ocorrências em Síndromes de Turner foram identificadas com 21 casos, aproximadamente 86%, dos casos como instâncias corretamente classificadas, e com aproximadamente 14%, com instâncias incorretamente classificadas.

4 – Para uma confirmação mais exata dos resultados devemos ressaltar a importância do número de amostra. Neste caso, considera-se um número pequeno de 84 casos.

Acurácia dos casos de Síndromes de Turner aplicando modelo de Naïve Bayes – Grupo IV – Conteúdo para treinamento.

Quadro XXVIII - Informações sobre variáveis executado pelo UnBMiner aplicando o modelo de Naïve Bayes

=== Informações Gerais ===		
Modelo Aplicado:		**Naïve Bayes**
Relacionamentos:		**Null**
Instancias:		**63 casos**
Número de Atributos:		**9**
Tipo	**Sigla**	**Descrição**
Atributos:	**SF**	**- Sexo feminino**
Atributos:	**BE**	**- Baixa estatura**
Atributos:	**TE**	**- Tórax em escudo**
Atributos:	**DG**	**- Disgenesia gonadal**
Atributos:	**UH**	**- Unhas hipoplásicas**
Atributos:	**CV**	**- Cúbito valgo**
Atributos:	**PA**	**- Pescoço alado**
Atributos:	**HMa**	**- Hipertelorismo de mamilos**
Atributos:	**TO**	**- Tendência à obesidade**
Classe:	**TURNER**	**- Síndrome de TURNER**
Modelo:	**Rede bayesiana**	

Tabela I.58 - Informações do resumo executado pelo UnBMiner aplicando o modelo de Naïve Bayes

=== Resumo ===

Descrição	**Qtde**	**Porcentagem**
Instâncias Corretamente Classificadas:	**56**	**88,88 %**
Instâncias Incorretamente Classificadas:	**7**	**11,11 %**
Quadratic loss function:	**0.2222**	**-**
Número Total de Instâncias:	**63**	**-**

Tabela I.59 - Informações sobre o detalhamento da classe executado pelo UnBMiner aplicando o modelo de Naïve Bayes

=== Acurácia Detalhada por Classe ===

VP Rate –Taxa Verdadeiro Positivo	**FP Rate – Taxa Falso Positivo**	**VN Rate-Taxa Verdadeiro Negativo**	**FN Rate– Taxa Falso Negativo**	**Classe**
1	0,368	0,632	0	NÃO
0,632	0	1	0,368	SIM

Tabela I.60 - Informações da matriz de confusão executado pelo UnBMiner aplicando o modelo de Naïve Bayes.

=== Matriz de Confusão ===

a b <-- classificado como:

a	**b**		
44	0	**a =**	**NÃO**
7	12	**b =**	**SIM**

Interpretação dos dados:
1 - Através da base dos dados foram identificados que a instância ‘SINDROME DE TURNER ’ assumiu valores ‘NAO’ em 44 casos.

2 – Propagando as evidências usando a classificação por Naïve Bayes identifica-se que foram gerados 44 casos com os valores ‘NÃO’ para a instância ‘SINDROME DE TURNER ’, ou seja, valores verdadeiros positivos. Assim, obtém-se um resultado de 0,8888 ou aproximadamente 89% de acerto.

3 – Em geral as evidências probabilísticas geradas pelo modelo Naïve bayes usando como referências dados para ocorrências em Síndromes de Turner foram identificadas com 63 casos, aproximadamente 89%, dos casos como instâncias corretamente classificadas, e com aproximadamente 11%, com instâncias incorretamente classificadas.

4 – Para uma confirmação mais exata dos resultados devemos ressaltar a importância do número de amostra. Neste caso, considera-se um número pequeno de 84 casos.

Acurácia dos casos de Síndromes de Turner aplicando modelo de Classificador ID3, do inglês (decision tree classifier) – rede bayesiana– Grupo I – Conteúdo para avaliar.

Quadro XXIX - Informações sobre variáveis executado pelo UnBMiner aplicando o modelo de Árvore de Decisão

=== Informações Gerais ===		
Modelo Aplicado:		**Árvore de Decisão – ID3**
Relacionamentos:		**Null**
Instancias:		**21 casos**
Número de Atributos:		**9**
Tipo	**Sigla**	**Descrição**
Atributos:	**SF**	**- Sexo feminino**
Atributos:	**BE**	**- Baixa estatura**
Atributos:	**TE**	**- Tórax em escudo**
Atributos:	**DG**	**- Disgenesia gonadal**
Atributos:	**UH**	**- Unhas hipoplásicas**
Atributos:	**CV**	**- Cúbito valgo**
Atributos:	**PA**	**- Pescoço alado**
Atributos:	**HMa**	**- Hipertelorismo de mamilos**
Atributos:	**TO**	**- Tendência à obesidade**
Classe:	**TURNER**	**- Síndrome de TURNER**

Quadro XXX - Informações modelo de teste executado pelo UnBMiner aplicando o modelo de Árvore de Decisão

Teste do Modelo: === Classificação do Modelo ===
TE = 1: TURNER = 1 (6.0) TE = 0 | DG = 1 | | BE = 1: TURNER = 1 (2.0) | | BE = 0: TURNER = 0 (1.0) | DG = 0: TURNER = 0 (12.0)

Tabela I.61 - Informações do resumo executado pelo UnBMiner aplicando o modelo de Árvore de Decisão

=== Resumo ===

Descrição	**Qtde**	**Porcentagem**
Instâncias Corretamente Classificadas:	**21**	**100 %**
Instâncias Incorretamente Classificadas:	**0**	**0 %**
Quadratic loss function:	**0**	**-**
Número Total de Instâncias:	**21**	**-**

Tabela I.62 - Informações sobre o detalhamento da classe executado pelo UnBMiner aplicando o modelo de Árvore de Decisão

=== Acurácia Detalhada por Classe ===

VP Rate –Taxa Verdadeiro Positivo	FP Rate – Taxa Falso Positivo	VN Rate-Taxa Verdadeiro Negativo	FN Rate– Taxa Falso Negativo	Classe
1	0	1	0	NÃO
1	0	1	0	SIM

Tabela I.63 - Informações da matriz de confusão executado pelo UnBMiner aplicando o modelo de Árvore de Decisão.

=== Matriz de Confusão ===

a	**b**	**<-- classificado como:**	
13	0	**a =**	**NÃO**
0	8	**b =**	**SIM**

Interpretação dos dados:
1 - Através da base dos dados foram identificados que a instância ‘SINDROME DE TURNER ’ assumiu valores ‘NAO’ em 21 casos.

2 – Propagando as evidências usando a classificação por árvores de decisão (ID3) identifica-se que foram gerados 13 casos com os valores ‘NÃO’ para a instância ‘SINDROME DE TURNER ’, ou seja, valores verdadeiros positivos. Assim, obtém-se um resultado de 1 ou 100 %.

3 – Em geral as evidências probabilísticas geradas pelo modelo árvores de decisão (ID3) usando como referências dados para ocorrências em Síndromes de Turner foram identificadas com 21 casos, 100 %, dos casos como instâncias corretamente classificadas, e com nenhum caso com instâncias incorretamente classificadas.

4 – Para uma confirmação mais exata dos resultados devemos ressaltar a importância do número de amostra. Neste caso, considera-se um número pequeno de 84 casos.

Acurácia dos casos de Síndromes de Turner aplicando modelo de Classificador ID3, do inglês (decision tree classifier) – Rede bayesiana– Grupo I – Conteúdo para treinamento.

Quadro XXXI - Informações sobre variáveis executado pelo UnBMiner aplicando o modelo de Árvore de Decisão

=== Informações Gerais ===		
Modelo Aplicado:		**Árvore de Decisão – ID3**
Relacionamentos:		**Null**
Instancias:		**65 casos**
Número de Atributos:		**9**
Tipo	**Sigla**	**Descrição**
Atributos:	**SF**	**- Sexo feminino**
Atributos:	**BE**	**- Baixa estatura**
Atributos:	**TE**	**- Tórax em escudo**
Atributos:	**DG**	**- Disgenesia gonadal**
Atributos:	**UH**	**- Unhas hipoplásicas**
Atributos:	**CV**	**- Cúbito valgo**
Atributos:	**PA**	**- Pescoço alado**
Atributos:	**HMa**	**- Hipertelorismo de mamilos**
Atributos:	**TO**	**- Tendência à obesidade**
Classe:	**TURNER**	**- Síndrome de TURNER**

Quadro XXXII - Informações modelo de teste executado pelo UnBMiner aplicando o modelo de Árvore de Decisão

Teste do Modelo: === Classificação do Modelo ===
PA = 1: TURNER = 1 (13.0) PA = 0 | DG = 1 | | BE = 1: TURNER = 1 (5.0) | | BE = 0: TURNER = 0 (2.0) | DG = 0 | | CV = 1: TURNER = 1 (1.0) | | CV = 0 | | | HMa = 1 | | | | BE = 1: TURNER = 1 (1.0) | | | | BE = 0: TURNER = 0 (1.0) | | | HMa = 0: TURNER = 0 (42.0)

Tabela I.64 - Informações do resumo executado pelo UnBMiner aplicando o modelo de Árvore de Decisão

=== Resumo ===		
Descrição	**Qtde**	**Porcentagem**
Instâncias Corretamente Classificadas:	**65**	**100 %**
Instâncias Incorretamente Classificadas:	**0**	**0 %**
Quadratic loss function:	**0**	**-**
Número Total de Instâncias:	**65**	**-**

Tabela I.65 - Informações sobre o detalhamento da classe executado pelo UnBMiner aplicando o modelo de Árvore de Decisão

=== Acurácia Detalhada por Classe ===

VP Rate –Taxa Verdadeiro Positivo	FP Rate – Taxa Falso Positivo	VN Rate-Taxa Verdadeiro Negativo	FN Rate– Taxa Falso Negativo	Classe
1	0	1	0	NÃO
1	0	1	0	SIM

Tabela I.66 - Informações da matriz de confusão executado pelo UnBMiner aplicando o modelo de Árvore de Decisão.

=== Matriz de Confusão ===

a	**b**	**<-- classificado como:**	
45	0	**a =**	**NÃO**
0	20	**b =**	**SIM**

Interpretação dos dados:
1 - Através da base dos dados foram identificados que a instância ‘SINDROME DE TURNER ’ assumiu valores ‘NAO’ em 45 casos.

2 – Propagando as evidências usando a classificação por árvores de decisão (ID3) identifica-se que foram gerados 45 casos com os valores ‘NÃO’ para a instância ‘SINDROME DE TURNER’, ou seja, valores verdadeiros positivos. Assim, obtém-se um resultado de 1 ou 100 %.

3 – Em geral as evidências probabilísticas geradas pelo modelo árvores de decisão (ID3) usando como referências dados para ocorrências em Síndromes de Turner foram identificadas com 65 casos, 100 %, dos casos como instâncias corretamente classificadas, e com nenhum caso com instâncias incorretamente classificadas.

4 – Para uma confirmação mais exata dos resultados devemos ressaltar a importância do número de amostra. Neste caso, considera-se um número pequeno de 84 casos.

Acurácia dos casos de Síndromes de Turner aplicando modelo de Classificador ID3, do inglês (decision tree classifier) – Rede bayesiana– Grupo II – Conteúdo para avaliar.

Quadro XXXIII - Informações sobre variáveis executado pelo UnBMiner aplicando o modelo de Árvore de Decisão

=== Informações Gerais ===		
Modelo Aplicado:		**Árvore de Decisão – ID3**
Relacionamentos:		**Null**
Instancias:		**21 casos**
Número de Atributos:		**9**
Tipo	**Sigla**	**Descrição**
Atributos:	**SF**	**- Sexo feminino**
Atributos:	**BE**	**- Baixa estatura**
Atributos:	**TE**	**- Tórax em escudo**
Atributos:	**DG**	**- Disgenesia gonadal**
Atributos:	**UH**	**- Unhas hipoplásicas**
Atributos:	**CV**	**- Cúbito valgo**
Atributos:	**PA**	**- Pescoço alado**
Atributos:	**HMa**	**- Hipertelorismo de mamilos**
Atributos:	**TO**	**- Tendência à obesidade**
Classe:	**TURNER**	**- Síndrome de TURNER**

Quadro XXXIV - Informações modelo de teste executado pelo UnBMiner aplicando o modelo de Árvore de Decisão

Teste do Modelo: === Classificação do Modelo ===
TE = 1: TURNER = 1 (6.0) TE = 0 | DG = 1 | | BE = 1: TURNER = 1 (2.0) | | BE = 0: TURNER = 0 (1.0) | DG = 0: TURNER = 0 (12.0)

Tabela I.67 - Informações do resumo executado pelo UnBMiner aplicando o modelo de Árvore de Decisão

=== Resumo ===

Descrição	**Qtde**	**Porcentagem**
Instâncias Corretamente Classificadas:	**21**	**100 %**
Instâncias Incorretamente Classificadas:	**0**	**0 %**
Quadratic loss function:	**0**	**-**
Número Total de Instâncias:	**21**	**-**

Tabela I.68 - Informações sobre o detalhamento da classe executado pelo UnBMiner aplicando o modelo de Árvore de Decisão

=== Acurácia Detalhada por Classe ===

VP Rate –Taxa Verdadeiro Positivo	FP Rate – Taxa Falso Positivo	VN Rate-Taxa Verdadeiro Negativo	FN Rate– Taxa Falso Negativo	Classe
1	0	1	0	NÃO
1	0	1	0	SIM

Tabela I.69 Informações da matriz de confusão executado pelo UnBMiner aplicando o modelo de Árvore de Decisão.

=== Matriz de Confusão ===

a	b	<-- classificado como:	
a	b		
17	0	a =	NÃO
0	4	b =	SIM

Interpretação dos dados:
1 - Através da base dos dados foram identificados que a instância ‘SINDROME DE TURNER ’ assumiu valores ‘NAO’ em 17 casos.

2 – Propagando as evidências usando a classificação por árvores de decisão (ID3) identifica-se que foram gerados 17 casos com os valores ‘NÃO’ para a instância ‘SINDROME DE TURNER ’, ou seja, valores verdadeiros positivos. Assim, obtém-se um resultado de 1 ou 100 %.

3 – Em geral as evidências probabilísticas geradas pelo modelo árvores de decisão (ID3) usando como referências dados para ocorrências em Síndromes de Turner foram identificadas com 21 casos, 100 %, dos casos como instâncias corretamente classificadas, e com nenhum caso com instâncias incorretamente classificadas.

4 – Para uma confirmação mais exata dos resultados devemos ressaltar a importância do número de amostra. Neste caso, considera-se um número pequeno de 84 casos.

Acurácia dos casos de Síndromes de Turner aplicando modelo de Classificador ID3, do inglês (decision tree classifier) – Rede bayesiana– Grupo II – Conteúdo para treinamento.

Quadro XXXV - Informações sobre variáveis executado pelo UnBMiner aplicando o modelo de Árvore de Decisão

=== Informações Gerais ===		
Modelo Aplicado:		**Árvore de Decisão – ID3**
Relacionamentos:		**Null**
Instancias:		**64 casos**
Número de Atributos:		**9**
Tipo	**Sigla**	**Descrição**
Atributos:	**SF**	**- Sexo feminino**
Atributos:	**BE**	**- Baixa estatura**
Atributos:	**TE**	**- Tórax em escudo**
Atributos:	**DG**	**- Disgenesia gonadal**
Atributos:	**UH**	**- Unhas hipoplásicas**
Atributos:	**CV**	**- Cúbito valgo**
Atributos:	**PA**	**- Pescoço alado**
Atributos:	**HMa**	**- Hipertelorismo de mamilos**
Atributos:	**TO**	**- Tendência à obesidade**
Classe:	**TURNER**	**- Síndrome de TURNER**

Quadro XXXVI - Informações modelo de teste executado pelo UnBMiner aplicando o modelo de Árvore de Decisão

Teste do Modelo: === Classificação do Modelo ===
PA = 1: TURNER = 1 (13.0) PA = 0 | DG = 1 | | BE = 1: TURNER = 1 (5.0) | | BE = 0: TURNER = 0 (2.0) | DG = 0 | | CV = 1: TURNER = 1 (1.0) | | CV = 0 | | | HMa = 1 | | | | BE = 1: TURNER = 1 (1.0) | | | | BE = 0: TURNER = 0 (1.0) | | | HMa = 0: TURNER = 0 (42.0)

Tabela I.70 - Informações do resumo executado pelo UnBMiner aplicando o modelo de Árvore de Decisão

=== Resumo ===		
Descrição	**Qtde**	**Porcentagem**
Instâncias Corretamente Classificadas:	**61**	**95,31 %**
Instâncias Incorretamente Classificadas:	**3**	**4,68 %**
Quadratic loss function:	**0.0938**	**-**
Número Total de Instâncias:	**64**	**-**

Tabela I.71 - Informações sobre o detalhamento da classe executado pelo UnBMiner aplicando o modelo de Árvore de Decisão

=== Acurácia Detalhada por Classe ===

VP Rate –Taxa Verdadeiro Positivo	FP Rate – Taxa Falso Positivo	VN Rate-Taxa Verdadeiro Negativo	FN Rate– Taxa Falso Negativo	Classe
0.975	0.083	0.917	0.025	NÃO
0.917	0.025	0.975	0.083	SIM

Tabela I.72 - Informações da matriz de confusão executado pelo UnBMiner aplicando o modelo de Árvore de Decisão.

=== Matriz de Confusão ===

a	**b**	**<-- classificado como:**
39	1	**a = NÃO**
2	22	**b = SIM**

Interpretação dos dados:

1 - Através da base dos dados foram identificados que a instância ‘SINDROME DE TURNER ’ assumiu valores ‘NAO’ em 39 casos.

2 – Propagando as evidências usando a classificação por árvores de decisão (ID3) identifica-se que foram gerados 39 casos com os valores ‘NÃO’ para a instância ‘SINDROME DE TURNER ’, ou seja, valores verdadeiros positivos. Assim, obtém-se um resultado de 0,9535 ou aproximadamente 95 %.

3 – Em geral as evidências probabilísticas geradas pelo modelo árvores de decisão (ID3) usando como referências dados para ocorrências em Síndromes de Turner foram identificadas com 64 casos, 95 %, dos casos como instâncias corretamente classificadas, e com 5% caso com instâncias incorretamente classificadas.

4 – Para uma confirmação mais exata dos resultados devemos ressaltar a importância do número de amostra. Neste caso, considera-se um número pequeno de 84 casos.

Acurácia dos casos de Síndromes de Turner aplicando modelo de Classificador ID3, do inglês (decision tree classifier) – Rede bayesiana– Grupo III – Conteúdo para avaliar.

Quadro XXXVII - Informações sobre variáveis executado pelo UnBMiner aplicando o modelo de Árvore de Decisão

=== Informações Gerais ===		
Modelo Aplicado:		**Árvore de Decisão – ID3**
Relacionamentos:		**Null**
Instancias:		**21 casos**
Número de Atributos:		**9**
Tipo	**Sigla**	**Descrição**
Atributos:	**SF**	**- Sexo feminino**
Atributos:	**BE**	**- Baixa estatura**
Atributos:	**TE**	**- Tórax em escudo**
Atributos:	**DG**	**- Disgenesia gonadal**
Atributos:	**UH**	**- Unhas hipoplásicas**
Atributos:	**CV**	**- Cúbito valgo**
Atributos:	**PA**	**- Pescoço alado**
Atributos:	**HMa**	**- Hipertelorismo de mamilos**
Atributos:	**TO**	**- Tendência à obesidade**
Classe:	**TURNER**	**- Síndrome de TURNER**

Quadro XXXVIII - Informações modelo de teste executado pelo UnBMiner aplicando o modelo de Árvore de Decisão

Teste do Modelo: === Classificação do Modelo ===
TE = 1: TURNER = 1 (6.0) TE = 0 | DG = 1 | | BE = 1: TURNER = 1 (2.0) | | BE = 0: TURNER = 0 (1.0) | DG = 0: TURNER = 0 (12.0)

Tabela I.73 - Informações do resumo executado pelo UnBMiner aplicando o modelo de Árvore de Decisão

=== Resumo ===

Descrição	**Qtde**	**Porcentagem**
Instâncias Corretamente Classificadas:	**21**	**100 %**
Instâncias Incorretamente Classificadas:	**0**	**0 %**
Quadratic loss function:	**0**	**-**
Número Total de Instâncias:	**21**	**-**

Tabela I.74 - Informações sobre o detalhamento da classe executado pelo UnBMiner aplicando o modelo de Árvore de Decisão

=== Acurácia Detalhada por Classe ===

VP Rate –Taxa Verdadeiro Positivo	FP Rate – Taxa Falso Positivo	VN Rate-Taxa Verdadeiro Negativo	FN Rate– Taxa Falso Negativo	Classe
1	0	1	0	NÃO
1	0	1	0	SIM

Tabela I.75 - Informações da matriz de confusão executado pelo UnBMiner aplicando o modelo de Árvore de Decisão.

=== Matriz de Confusão ===

a	**b**	**<-- classificado como:**	
14	0	**a =**	**NÃO**
0	7	**b =**	**SIM**

Interpretação dos dados:
1 - Através da base dos dados foram identificados que a instância ‘SINDROME DE TURNER ’ assumiu valores ‘NAO’ em 14 casos.

2 – Propagando as evidências usando a classificação por árvores de decisão (ID3) identifica-se que foram gerados 14 casos com os valores ‘NÃO’ para a instância ‘SINDROME DE TURNER ’, ou seja, valores verdadeiros positivos. Assim, obtém-se um resultado de 1 ou 100 %.

3 – Em geral as evidências probabilísticas geradas pelo modelo árvores de decisão (ID3) usando como referências dados para ocorrências em Síndromes de Turner foram identificadas com 21 casos, 100 %, dos casos como instâncias corretamente classificadas, e com nenhum caso com instâncias incorretamente classificadas.

4 – Para uma confirmação mais exata dos resultados devemos ressaltar a importância do número de amostra. Neste caso, considera-se um número pequeno de 84 casos.

Acurácia dos casos de Síndromes de Turner aplicando modelo de Classificador ID3, do inglês (decision tree classifier) – Rede bayesiana– Grupo III – Conteúdo para treinamento.

Quadro XXXIX - Informações sobre variáveis executado pelo UnBMiner aplicando o modelo de Árvore de Decisão

=== Informações Gerais ===		
Modelo Aplicado:		**Árvore de Decisão – ID3**
Relacionamentos:		**Null**
Instancias:		**64 casos**
Número de Atributos:		**9**
Tipo	**Sigla**	**Descrição**
Atributos:	**SF**	**- Sexo feminino**
Atributos:	**BE**	**- Baixa estatura**
Atributos:	**TE**	**- Tórax em escudo**
Atributos:	**DG**	**- Disgenesia gonadal**
Atributos:	**UH**	**- Unhas hipoplásicas**
Atributos:	**CV**	**- Cúbito valgo**
Atributos:	**PA**	**- Pescoço alado**
Atributos:	**HMa**	**- Hipertelorismo de mamilos**
Atributos:	**TO**	**- Tendência à obesidade**
Classe:	**TURNER**	**- Síndrome de TURNER**

Quadro XL - Informações modelo de teste executado pelo UnBMiner aplicando o modelo de Árvore de Decisão

Teste do Modelo: === Classificação do Modelo ===
PA = 1: TURNER = 1 (13.0) PA = 0 | DG = 1 | | BE = 1: TURNER = 1 (5.0) | | BE = 0: TURNER = 0 (2.0) | DG = 0 | | CV = 1: TURNER = 1 (1.0) | | CV = 0 | | | HMa = 1 | | | | BE = 1: TURNER = 1 (1.0) | | | | BE = 0: TURNER = 0 (1.0) | | | HMa = 0: TURNER = 0 (42.0)

Tabela I.76 - Informações do resumo executado pelo UnBMiner aplicando o modelo de Árvore de Decisão

=== Resumo ===		
Descrição	**Qtde**	**Porcentagem**
Instâncias Corretamente Classificadas:	**64**	**100 %**
Instâncias Incorretamente Classificadas:	**0**	**0 %**
Quadratic loss function:	**0**	**-**
Número Total de Instâncias:	**64**	**-**

Tabela I.77 - Informações sobre o detalhamento da classe executado pelo UnBMiner aplicando o modelo de Árvore de Decisão

=== Acurácia Detalhada por Classe ===

VP Rate –Taxa Verdadeiro Positivo	FP Rate – Taxa Falso Positivo	VN Rate-Taxa Verdadeiro Negativo	FN Rate– Taxa Falso Negativo	Classe
1	0	1	0	NÃO
1	0	1	0	SIM

Tabela I.78 - Informações da matriz de confusão executado pelo UnBMiner aplicando o modelo de Árvore de Decisão.

=== Matriz de Confusão ===

a	**b**	**<-- classificado como:**	
43	0	**a =**	**NÃO**
0	21	**b =**	**SIM**

Interpretação dos dados:
1 - Através da base dos dados foram identificados que a instância ‘SINDROME DE TURNER ’ assumiu valores ‘NAO’ em 43 casos.

2 – Propagando as evidências usando a classificação por árvores de decisão (ID3) identifica-se que foram gerados 43 casos com os valores ‘NÃO’ para a instância ‘SINDROME DE TURNER ’, ou seja, valores verdadeiros positivos. Assim, obtém-se um resultado de 1 ou 100 %.

3 – Em geral as evidências probabilísticas geradas pelo modelo árvores de decisão (ID3) usando como referências dados para ocorrências em Síndromes de Turner foram identificadas com 64 casos, 100 %, dos casos como instâncias corretamente classificadas, e com nenhum caso com instâncias incorretamente classificadas.

4 – Para uma confirmação mais exata dos resultados devemos ressaltar a importância do número de amostra. Neste caso, considera-se um número pequeno de 84 casos.

Acurácia dos casos de Síndromes de Turner aplicando modelo de Classificador ID3, do inglês (decision tree classifier) – Rede bayesiana– Grupo IV – Conteúdo para avaliar.

Quadro XLI - Informações sobre variáveis executado pelo UnBMiner aplicando o modelo de Árvore de Decisão

=== Informações Gerais ===		
Modelo Aplicado:		**Árvore de Decisão – ID3**
Relacionamentos:		**Null**
Instancias:		**21 casos**
Número de Atributos:		**9**
Tipo	**Sigla**	**Descrição**
Atributos:	**SF**	**- Sexo feminino**
Atributos:	**BE**	**- Baixa estatura**
Atributos:	**TE**	**- Tórax em escudo**
Atributos:	**DG**	**- Disgenesia gonadal**
Atributos:	**UH**	**- Unhas hipoplásicas**
Atributos:	**CV**	**- Cúbito valgo**
Atributos:	**PA**	**- Pescoço alado**
Atributos:	**HMa**	**- Hipertelorismo de mamilos**
Atributos:	**TO**	**- Tendência à obesidade**
Classe:	**TURNER**	**- Síndrome de TURNER**

Quadro XLII - Informações modelo de teste executado pelo UnBMiner aplicando o modelo de Árvore de Decisão

Teste do Modelo: === Classificação do Modelo ===
TE = 1: TURNER = 1 (6.0) TE = 0 | DG = 1 | | BE = 1: TURNER = 1 (2.0) | | BE = 0: TURNER = 0 (1.0) | DG = 0: TURNER = 0 (12.0)

Tabela I.79 - Informações do resumo executado pelo UnBMiner aplicando o modelo de Árvore de Decisão

=== Resumo ===

Descrição	**Qtde**	**Porcentagem**
Instâncias Corretamente Classificadas:	**21**	**100 %**
Instâncias Incorretamente Classificadas:	**0**	**0 %**
Quadratic loss function:	**0**	**-**
Número Total de Instâncias:	**21**	**-**

Tabela I.80 - Informações sobre o detalhamento da classe executado pelo UnBMiner aplicando o modelo de Árvore de Decisão

=== Acurácia Detalhada por Classe ===

VP Rate –Taxa Verdadeiro Positivo	FP Rate – Taxa Falso Positivo	VN Rate-Taxa Verdadeiro Negativo	FN Rate– Taxa Falso Negativo	Classe
1	0	1	0	NÃO
1	0	1	0	SIM

Tabela I.81 - Informações da matriz de confusão executado pelo UnBMiner aplicando o modelo de Árvore de Decisão.

=== Matriz de Confusão ===

a	**b**	**<-- classificado como:**	
12	0	**a =**	**NÃO**
0	9	**b =**	**SIM**

Interpretação dos dados:
1 - Através da base dos dados foram identificados que a instância ‘SINDROME DE TURNER ’ assumiu valores ‘NAO’ em 12 casos.

2 – Propagando as evidências usando a classificação por árvores de decisão (ID3) identifica-se que foram gerados 12 casos com os valores ‘NÃO’ para a instância ‘SINDROME DE TURNER ’, ou seja, valores verdadeiros positivos. Assim, obtém-se um resultado de 1 ou 100 %.

3 – Em geral as evidências probabilísticas geradas pelo modelo árvores de decisão (ID3) usando como referências dados para ocorrências em Síndromes de Turner foram identificadas com 21 casos, 100 %, dos casos como instâncias corretamente classificadas, e com nenhum caso com instâncias incorretamente classificadas.

4 – Para uma confirmação mais exata dos resultados devemos ressaltar a importância do número de amostra. Neste caso, considera-se um número pequeno de 84 casos.

Acurácia dos casos de Síndromes de Turner aplicando modelo de Classificador ID3, do inglês (decision tree classifier) – Rede bayesiana– Grupo IV – Conteúdo para treinamento.

Quadro XLIII - Informações sobre variáveis executado pelo UnBMiner aplicando o modelo de Árvore de Decisão

=== Informações Gerais ===		
Modelo Aplicado:		**Árvore de Decisão – ID3**
Relacionamentos:		**Null**
Instancias:		**63 casos**
Número de Atributos:		**9**
Tipo	**Sigla**	**Descrição**
Atributos:	**SF**	**- Sexo feminino**
Atributos:	**BE**	**- Baixa estatura**
Atributos:	**TE**	**- Tórax em escudo**
Atributos:	**DG**	**- Disgenesia gonadal**
Atributos:	**UH**	**- Unhas hipoplásicas**
Atributos:	**CV**	**- Cúbito valgo**
Atributos:	**PA**	**- Pescoço alado**
Atributos:	**HMa**	**- Hipertelorismo de mamilos**
Atributos:	**TO**	**- Tendência à obesidade**
Classe:	**TURNER**	**- Síndrome de TURNER**

Quadro XLIV - Informações modelo de teste executado pelo UnBMiner aplicando o modelo de Árvore de Decisão

Teste do Modelo: === Classificação do Modelo ===
PA = 1: TURNER = 1 (13.0) PA = 0 | DG = 1 | | BE = 1: TURNER = 1 (5.0) | | BE = 0: TURNER = 0 (2.0) | DG = 0 | | CV = 1: TURNER = 1 (1.0) | | CV = 0 | | | HMa = 1 | | | | BE = 1: TURNER = 1 (1.0) | | | | BE = 0: TURNER = 0 (1.0) | | | HMa = 0: TURNER = 0 (42.0)

Tabela I.82 Informações do resumo executado pelo UnBMiner aplicando o modelo de Árvore de Decisão

=== Resumo ===		
Descrição	**Qtde**	**Porcentagem**
Instâncias Corretamente Classificadas:	**63**	**100 %**
Instâncias Incorretamente Classificadas:	**0**	**0 %**
Quadratic loss function:	**0**	**-**
Número Total de Instâncias:	**63**	**-**

Tabela I.83 - Informações sobre o detalhamento da classe executado pelo UnBMiner aplicando o modelo de Árvore de Decisão

=== Acurácia Detalhada por Classe ===

VP Rate –Taxa Verdadeiro Positivo	FP Rate – Taxa Falso Positivo	VN Rate-Taxa Verdadeiro Negativo	FN Rate– Taxa Falso Negativo	Classe
1	0	1	0	NÃO
1	0	1	0	SIM

Tabela I.84 - Informações da matriz de confusão executado pelo UnBMiner aplicando o modelo de Árvore de Decisão.

=== Matriz de Confusão ===

a	**b**	**<-- classificado como:**	
44	0	**a =**	**NÃO**
0	19	**b =**	**SIM**

Interpretação dos dados:
1 - Através da base dos dados foram identificados que a instância ‘SINDROME DE TURNER ’ assumiu valores ‘NAO’ em 44 casos.

2 – Propagando as evidências usando a classificação por árvores de decisão (ID3) identifica-se que foram gerados 44 casos com os valores ‘NÃO’ para a instância ‘SINDROME DE TURNER ’, ou seja, valores verdadeiros positivos. Assim, obtém-se um resultado de 1 ou 100 %.

3 – Em geral as evidências probabilísticas geradas pelo modelo árvores de decisão (ID3) usando como referências dados para ocorrências em Síndromes de Turner foram identificadas com 63 casos, 100 %, dos casos como instâncias corretamente classificadas, e com nenhum caso com instâncias incorretamente classificadas.

4 – Para uma confirmação mais exata dos resultados devemos ressaltar a importância do número de amostra. Neste caso, considera-se um número pequeno de 84 casos.

ANEXO II

Este anexo apresenta um quadro que representa o arquivo utilizado como base de dados pelo algoritmo de aprendizagem de Redes bayesianas. É representada também uma outra tabela que fornece o significado de cada valor assumido pelas variáveis.

Quadro XLV - Todos os casos com os diagnósticos e sintomas primários

SF Sexo feminino	**BE** Baixa estatura	**TE** Tórax escudo	**DG** Disgenesia gonadal	**UH** Unhas hipoplásicas	**CV** Cúbito valgo	**PA** Pescoço alado	**Hma** Hipertelorismo de mamilos	**TO** Tendência obesidade	**TURNER** Turner
1	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	1	0	0
1	0	0	0	0	0	1	0	0	0
1	0	0	1	0	0	0	0	0	0
1	0	0	1	0	0	0	0	0	0
1	0	0	1	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	1	0	1	1
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	1	0	1
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	1	0	0	0	1
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0

Continuação do Quadro do Anexo II.									
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	1	0	0	0	0	0	1
1	1	0	1	0	0	0	0	0	1
1	1	0	1	0	0	0	0	0	1
1	1	0	1	0	0	0	0	0	1
1	1	0	1	0	0	0	0	0	1
1	1	0	1	0	0	0	1	0	1
1	1	0	1	0	0	0	0	0	1
1	1	1	0	0	0	1	0	0	1
1	1	1	0	0	0	1	0	0	1
1	1	1	0	0	0	1	0	0	1
1	1	1	0	0	0	1	0	0	1
1	1	1	0	0	0	1	0	0	1
1	1	1	0	0	0	1	0	0	1
1	1	1	0	0	0	1	0	0	1
1	1	1	0	0	0	1	0	0	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1

ANEXO III

Este anexo apresenta um quadro que representa os valores assumidos pelas variáveis para a identificação da classe – TURNER.

Quadro XLVI - Valores de variáveis associados a sua descrição.

Valor inicial da variável	Valor convertido em numérico (categóricos)
SIM	1
NAO	0