Mineração de dados: uma visão introdutória

301795 – Mineração de Dados, turma A

Prof. Marcelo Ladeira – CIC/UnB

Mestrado Profissional em Computação Aplicada

Sumário

1. Introdução
   - Conceituação de KDD
   - Conceituação de Mineração de Dados
   - Conceituação de Mineração de textos
2. Tarefas de Mineração de Dados
3. Modelo de Referência CRISP-DM
4. Aplicações Desenvolvidas
5. Estudo de Caso
6. Conclusões

1. Introdução
1.1 KDD – Áreas de Interação

- Estatística
- Reconhecimento de padrões e aprendizagem de máquina
  - Extração de padrões e construção de modelos
- Inteligência artificial (conhecimento simbólico)
  - Representação e interpretação de conhecimento
- Inteligência computacional (conhecimento numérico)
  - Aprendizagem e generalização
- Banco de dados

1. Introdução
1.1 Conceituação de KDD

- Quanto ao resultado
  - "Processo, não trivial, de extração de informações, implícitas, previamente desconhecidas e úteis, a partir dos dados armazenados em um banco de dados."
    - [Frawley, Piatetsky-Shapiro & Matheus,1991]
- Quanto ao processo
  - "Tarefa de descoberta de conhecimento intensivo, consistindo de interações complexas, feitas ao longo do tempo, entre o homem e uma grande base de dados, possivelmente suportada por um conjunto heterogêneo de ferramentas."
    - [Brachman & Anand,1995]

KDD lida com grandes massas de dados.

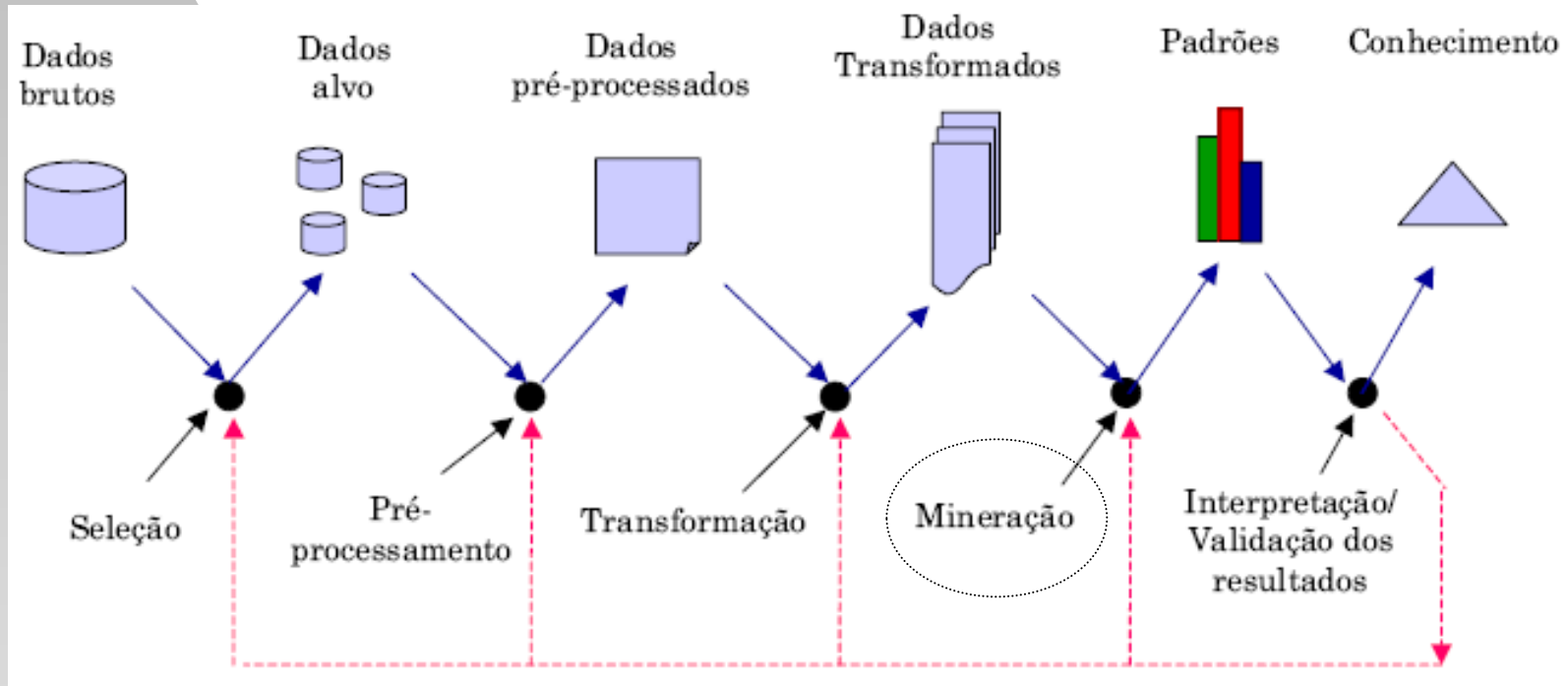
1. Introdução
1.1 KDD – Definição Formal

- Processo não trivial de identificação de padrões em um conjunto de dados que possuam as seguintes características:
  - validade: a descoberta de padrões deve ser válida em novos dados com algum grau de certeza ou probabilidade.
  - novidade: os padrões são novos (pelo menos para o sistema em estudo), ou seja, ainda não foram detectados por nenhuma outra abordagem.
  - utilidade potencial: os padrões devem poder ser utilizados para a tomada de decisões úteis, medidas por alguma função.
  - assimiláveis: um dos objetivos do KDD é tornar os padrões assimiláveis ao conhecimento humano.

FAYYAD, U.M., PIATETSKY, S.G., SMYTH, P. & UTHURUSAMY, R. From Data Mining to Knowledge Discovery: An Overview In: Advances in Knowledge Discovery and Data Mining, 1 ed. California: AAAIPress, The MIT Press, 1996, p. 1-34.

1 Introdução

1.1 Etapas do Processo de KDD



O processo é iterativo e cíclico e a saída de uma etapa pode requerer uma revisão em uma etapa anterior.

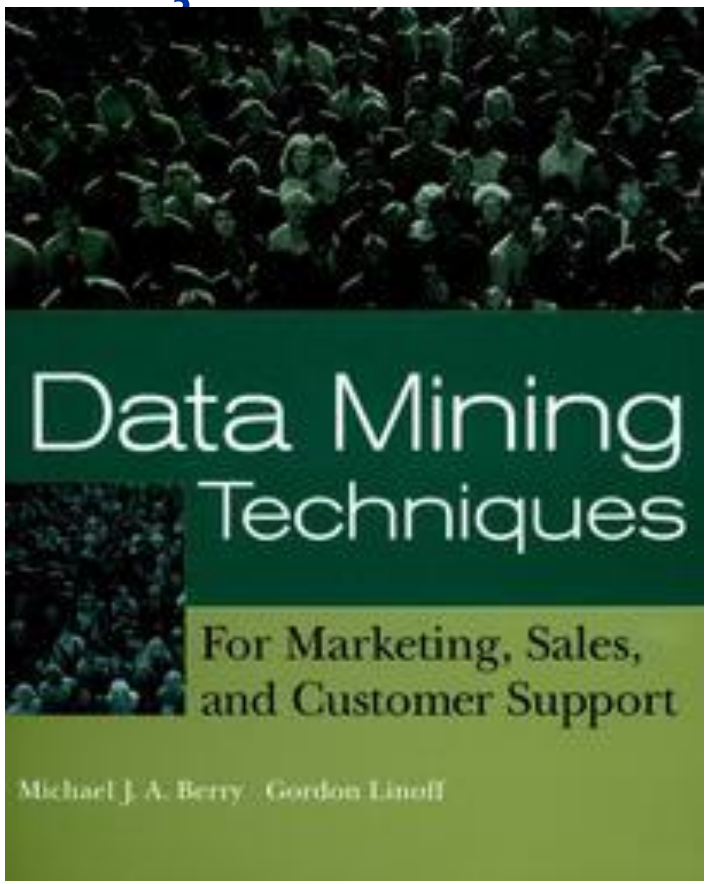
1. Introdução
1.2 Mineração de Dados - Definição Formal

- Avaliação de dados eletrônicos com a ajuda de técnicas de aprendizagem para que se possa encontrar relações ou padrões entre eles, visando:
  - descobrir novos fatos, regularidades, restrições ou relacionamentos, a partir da análise dos dados.
  - encontrar e descrever padrões estruturais (modelos) nos dados, como uma ferramenta que ajuda a explicar e fazer previsões.
    - **Entrada: conjunto de treinamento (envolve algum conceito a ser aprendido).**
    - **Saída: modelo (representa forma de predizer novos dados).**
      - Podem existir muitas descrições alternativas (modelos) que explicam os dados: em geral, opte pelo mais simples.
  - testar a validar de hipóteses (idéias pré-formuladas)
    - **Entrada: idéias e conjunto de treinamento que permita avaliá-las.**

1. Introdução

1.2 Mineração de Dados - Definição

- "Mineração de dados é a exploração e a análise, por meio automático ou semi-automático, de grandes quantidades de dados, a fim de descobrir padrões e regras significativos" (Berry e Linoff, 1997)



1. Introdução
1.3 Mineração de Textos

- Processo de descoberta de conhecimento em bases de textos.
  - recuperação de informação
  - processamento de linguagem natural
  - aprendizado de máquina
  - mineração de dados
  - Estatística
- Extrai descritores para uso em DM
  - grande dimensionalidade é problemática

1. Introdução
1.4 Mineração vs Aprendizagem de Máquina

- Aprendizagem de Máquina
  - Área de estudo das técnicas de aprendizagem.
    - O aprendizado de máquina envolve a idéia de treinamento (casos positivos e casos negativos) e uma posterior avaliação do quanto foi aprendido destes dados.
      - Fortemente baseada nos raciocínios lógico, analógico e de casos.
        - Em geral, trabalha com quantidade de dados menor.
  - Pode ser compreendida como um problema de busca para encontrar a hipótese correta.
- Mineração de Dados
  - Junção das áreas de estatística, aprendizagem de máquinas e banco de dados
    - **O raciocínio típico é o estatístico.**

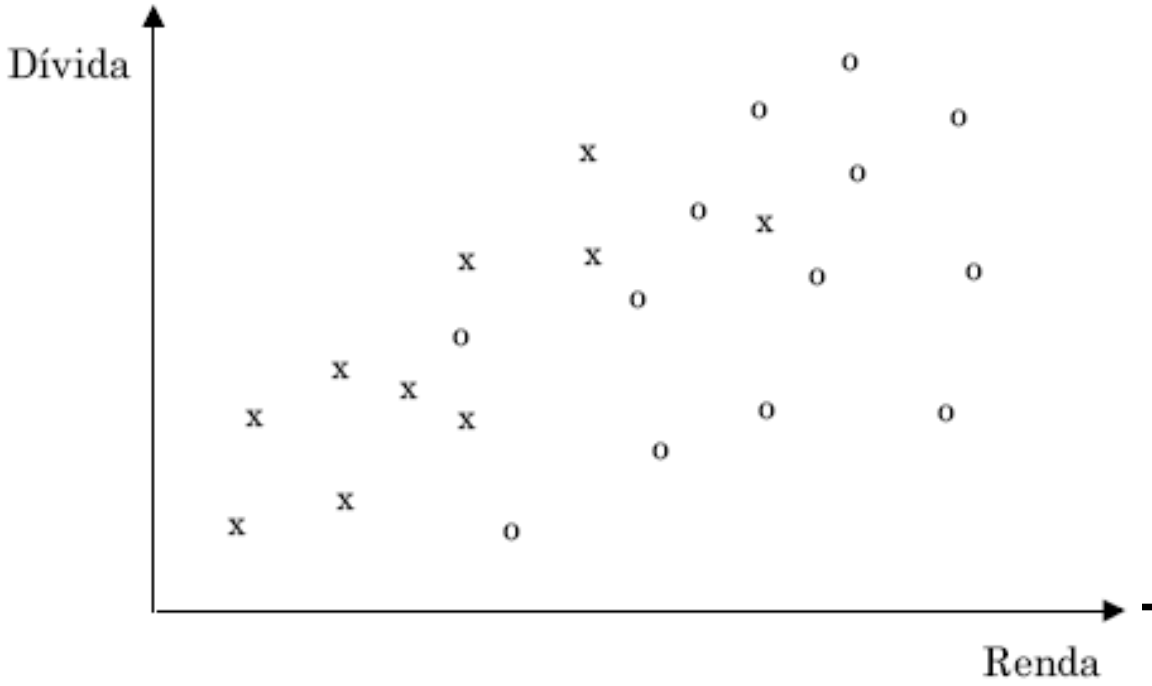
2. Tarefas de Mineração de Dados

- Os principais objetivos de alto nível são a previsão e a descrição.
  - A previsão envolve usar algumas variáveis ou campos da base de dados para prever valores desconhecidos ou futuros de variáveis de interesse.
  - A descrição se concentra em encontrar padrões que descrevem os dados, que sejam interpretáveis pelos seres humanos.
    - **No contexto de DM, a descrição tende a ser mais importante que a previsão, ao contrário das aplicações de aprendizado de máquina e reconhecimento de padrões.**
- Os objetivos de previsão e descrição são alcançados através da realização das tarefas básicas de mineração.

2. Tarefas de Mineração de Dados

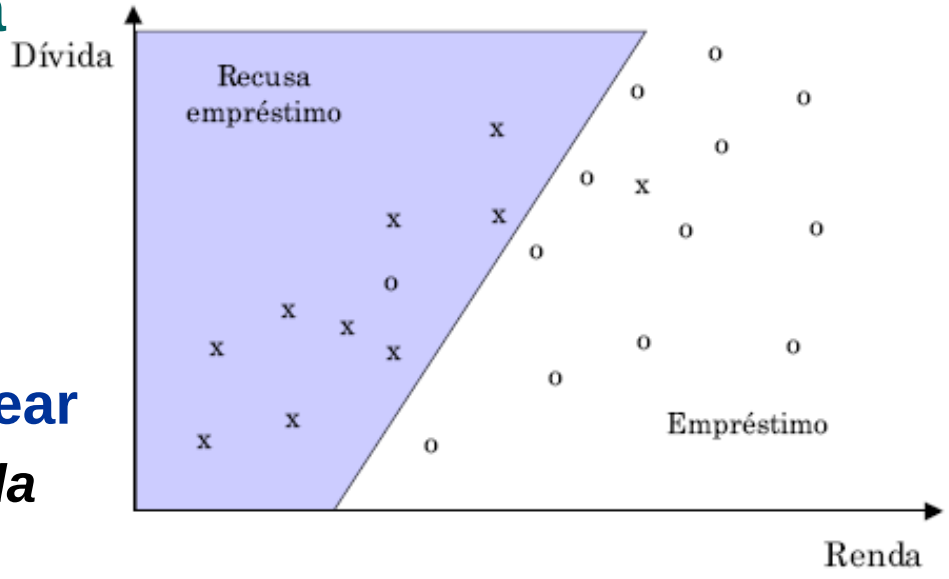
Exemplo Simples

- Considere uma distribuição correspondendo aos atributos renda e dívida de um correntista.
  - **Cada pessoa foi classificada como bom pagador (o) ou mau pagador (x)**



2. Tarefas de Mineração de Dados
2.1 Classificação

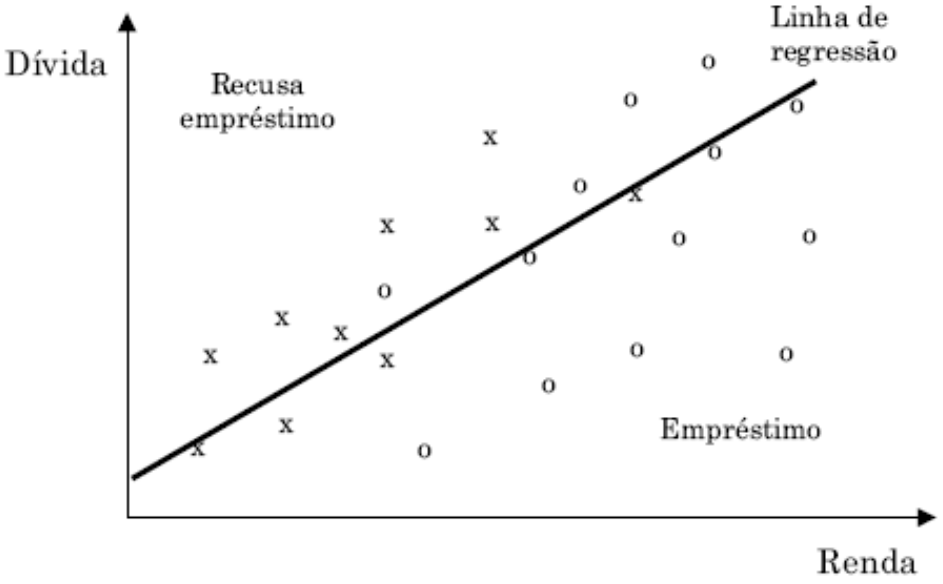
- **Consiste em aprender uma função que mapeia (classifica) um item de dado para uma entre várias classes pré-definidas.**
  - **Superfície de decisão linear**
    - **Se w_1.*renda* + w_2.*dívida* < *t*, então cliente não paga o empréstimo (x)**
    - **Possui erro associado.**



Exemplo extraído de slides preparados pelo Prof. Paulo Martins Engel - UFRGS

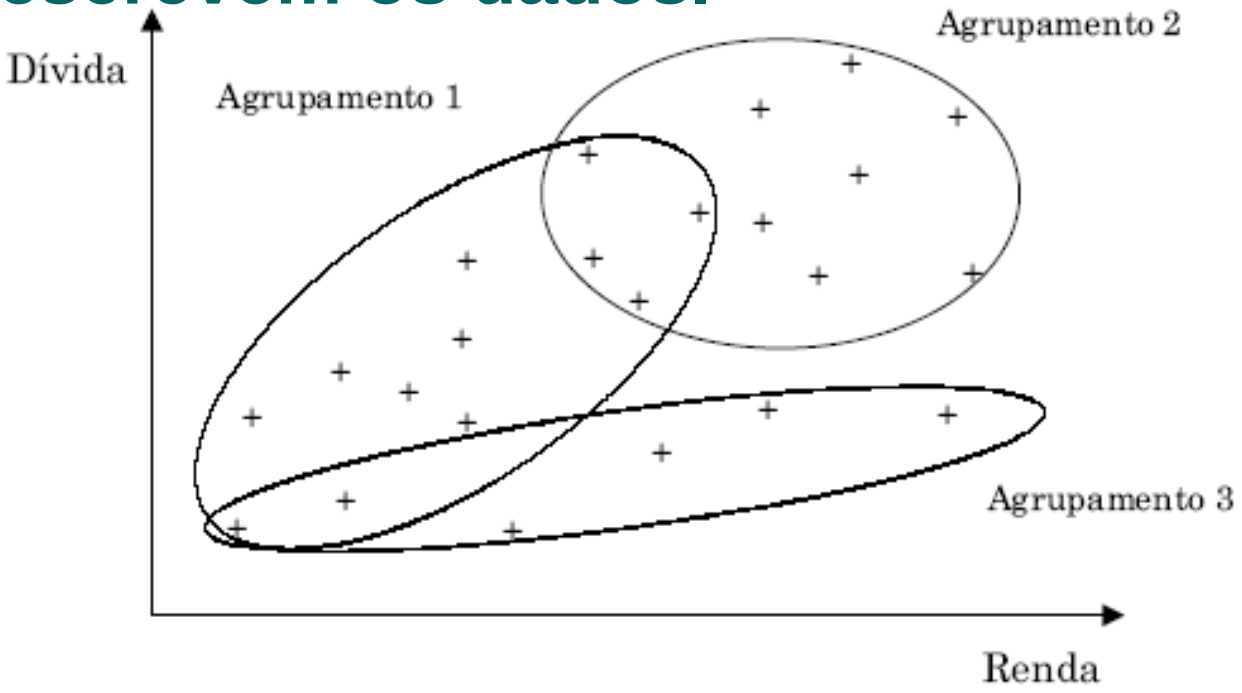
2. Tarefas de Mineração de Dados
2.2 Regressão

- Consiste em aprender uma função que mapeia um item de dado para uma variável de previsão de valor real.



2. Tarefas de Mineração de Dados
2.3 Agrupamento (*clustering*)

- **Tarefa descritiva onde se procura identificar um conjunto finito de categorias ou agrupamentos que descrevem os dados.**



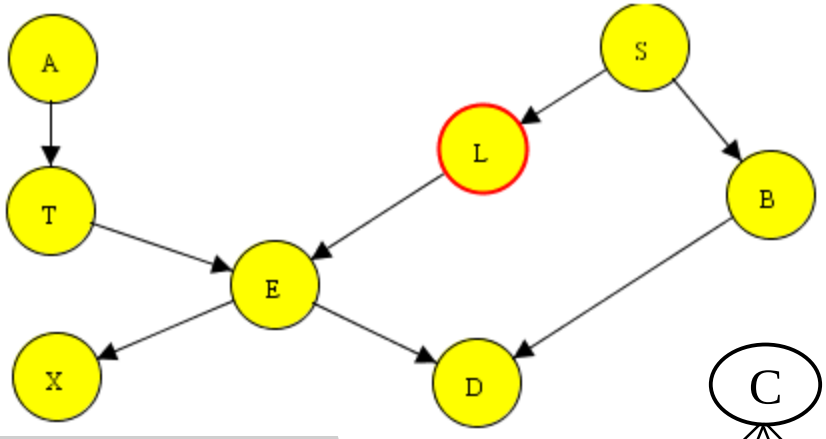
2. Tarefas de Mineração de Dados
2.4 Sumarização

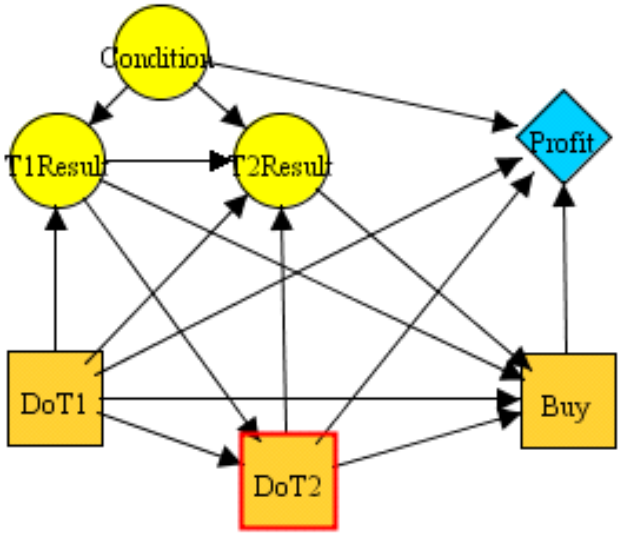
- **Envolve métodos para encontrar uma descrição compacta para um subconjunto de dados.**
  - **Um exemplo simples seria a tabulação da média e dos desvios padrões de todos os campos.**
  - **Métodos mais sofisticados envolvem derivar regras gerais, técnicas de visualização para múltiplas variáveis e a descoberta de relações funcionais entre variáveis.**
    - Estas técnicas são usadas na análise exploratória interativa e na geração automática de relatórios.

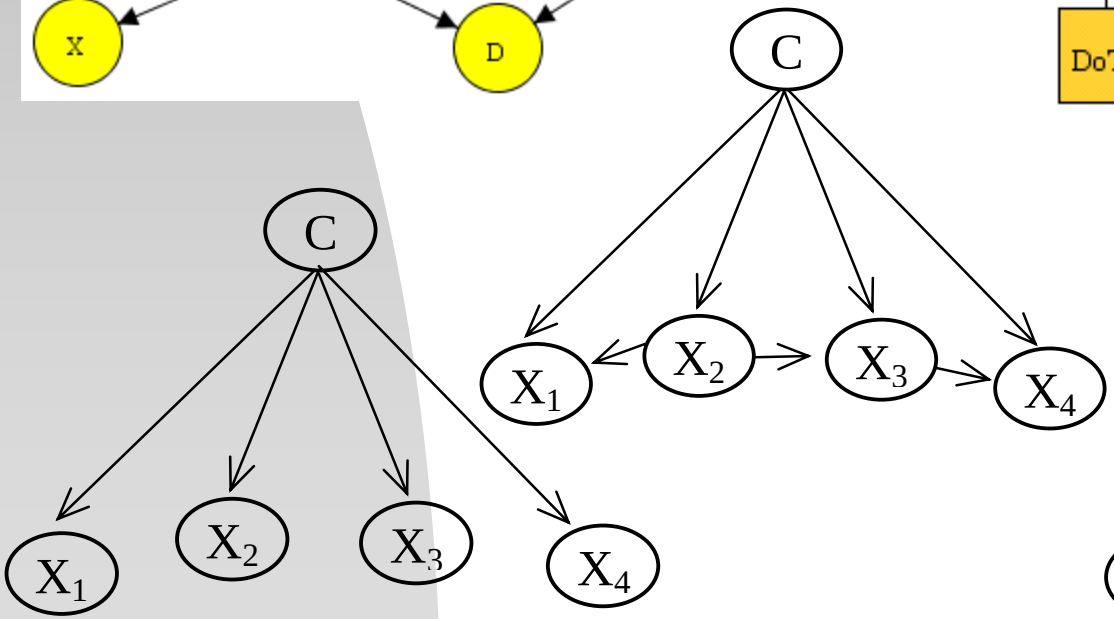
2. Tarefas de Mineração de Dados
2.5 Modelagem de Dependências

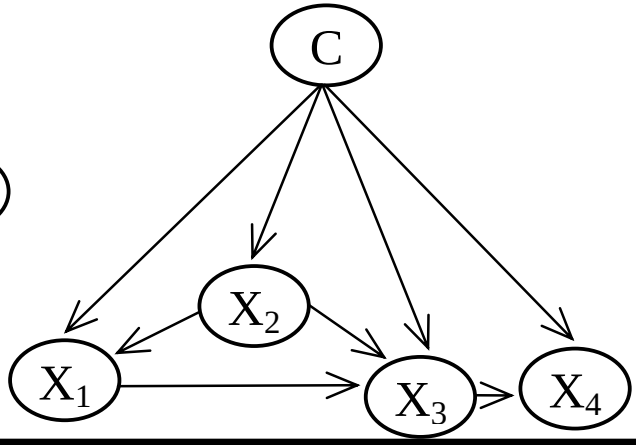
- Consiste em encontrar um modelo que descreva dependências significativas entre variáveis.
  - Modelos de dependências existem em dois níveis: o *nível estrutural* do modelo especifica quais as variáveis são localmente dependentes entre si.
  - O *nível quantitativo* especifica as intensidades das dependências usando alguma escala numérica.
- As redes probabilísticas são exemplo desta modelagem
  - Redes bayesianas, diagramas de influências, naive Bayes, TAN (Tree Augmented Naive Bayes), BAN (Bayesian Augmented Naive Bayes)

2. Tarefas de Mineração de Dados
2.5 Modelagem de Dependências









Rede bayesiana, diagrama de influências, Naive Bayes, TAN e BAN

2. Tarefas de Mineração de Dados
2.6 Detecção de Desvios

- Enfoca a descoberta das modificações mais significativas nos dados em relação aos valores médios históricos. É utilizada, por exemplo, na identificação de fraudes.

2. Tarefas de Mineração de Dados
2.7 Descoberta de Associações

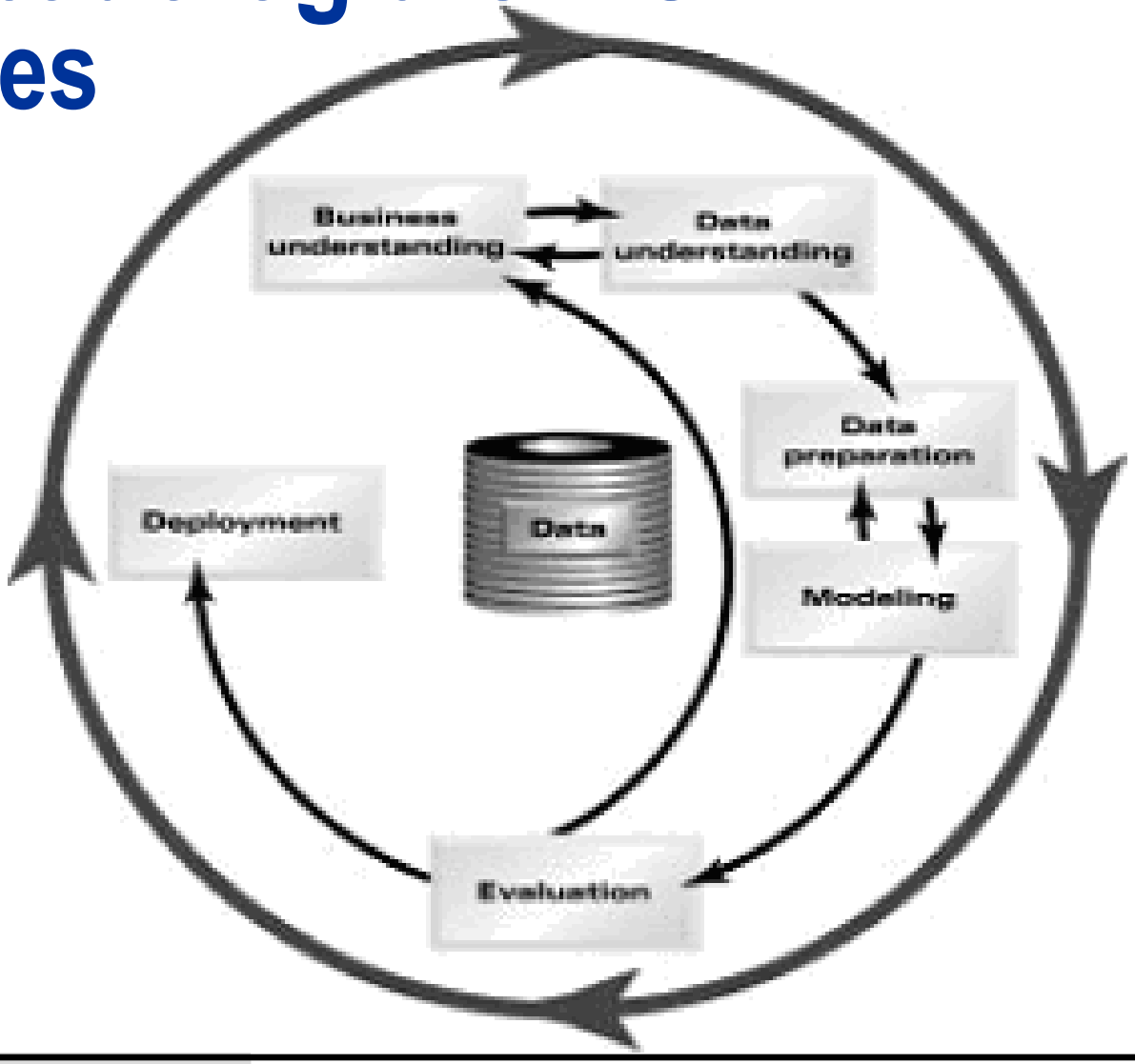
- O problema da *cesta de compras* assume que tenhamos um grande número de itens, p.ex., "*pão*", "*leite*", etc. Os clientes enchem as suas cestas de compras com um subconjunto desses itens e nós dispomos da informação sobre quais itens foram comprados *juntos* para cada cliente.
  - *Regras associativas:* $\{X_1, X_2, ..., X_n\} \Rightarrow Y$
    - se encontrarmos todos os itens $X_1, X_2, ..., X_n$ na cesta de compras, então nós temos uma boa chance de também encontrar Y.

3. Modelo de Referência CRISP-DM

Cross Industry Process Model for Data Mining

- **Modelo de processo hierárquico que parte de um conjunto de tarefas mais gerais para um conjunto de tarefas mais específicas, discriminadas em quatro níveis de abstração:**
  - a) **no topo da hierarquia, o processo de MD é organizado em *fases*;**
  - b) **as fases, por sua vez, são constituídas por diversas *tarefas genéricas*, que formam o segundo nível da hierarquia;**
  - c) **o terceiro nível, de *tarefas especializadas*, envolve a descrição de como as ações das tarefas genéricas são aplicadas em situações específicas.**
    - ☞ Por exemplo, uma tarefa genérica do segundo nível é a limpeza de dados. No terceiro nível, essa tarefa seria descrita em diferentes situações, tais como limpeza de valores numéricos ou de valores categóricos.
  - d) **o quarto nível, de *instâncias do processo*, é um registro das ações, decisões e resultados da mineração de dados de uma aplicação em particular.**

3. Metodologia CRISP-DM
Fases



www.crisp-dm.org/CRISPWP-0800.pdf

3. Metodologia CRISP-DM
3.1 Entendimento do Negócio

- Foca o entendimento dos objetivos e requerimentos do projeto, da perspectiva do domínio, a relevância do conhecimento prévio e os objetivos do usuário final.
- Nessa etapa são elaborados o plano do projeto, especificando os passos a serem executados no resto do projeto e a definição do problema.

Também denominada entendimento do domínio

3. CRISP-DM
3.2 Entendimento dos Dados

- Seleção do conjunto de dados
- Análise dos dados
  - identificar problemas de qualidade
  - descobrir os primeiros conhecimentos
  - descrição dos dados
    - **formato, quantidade de registros e campos**
  - distribuição dos atributos,
  - relacionamentos entre pares de atributos,
  - identificação de agrupamentos ou subconjuntos existentes nos dados

3. CRISP-DM

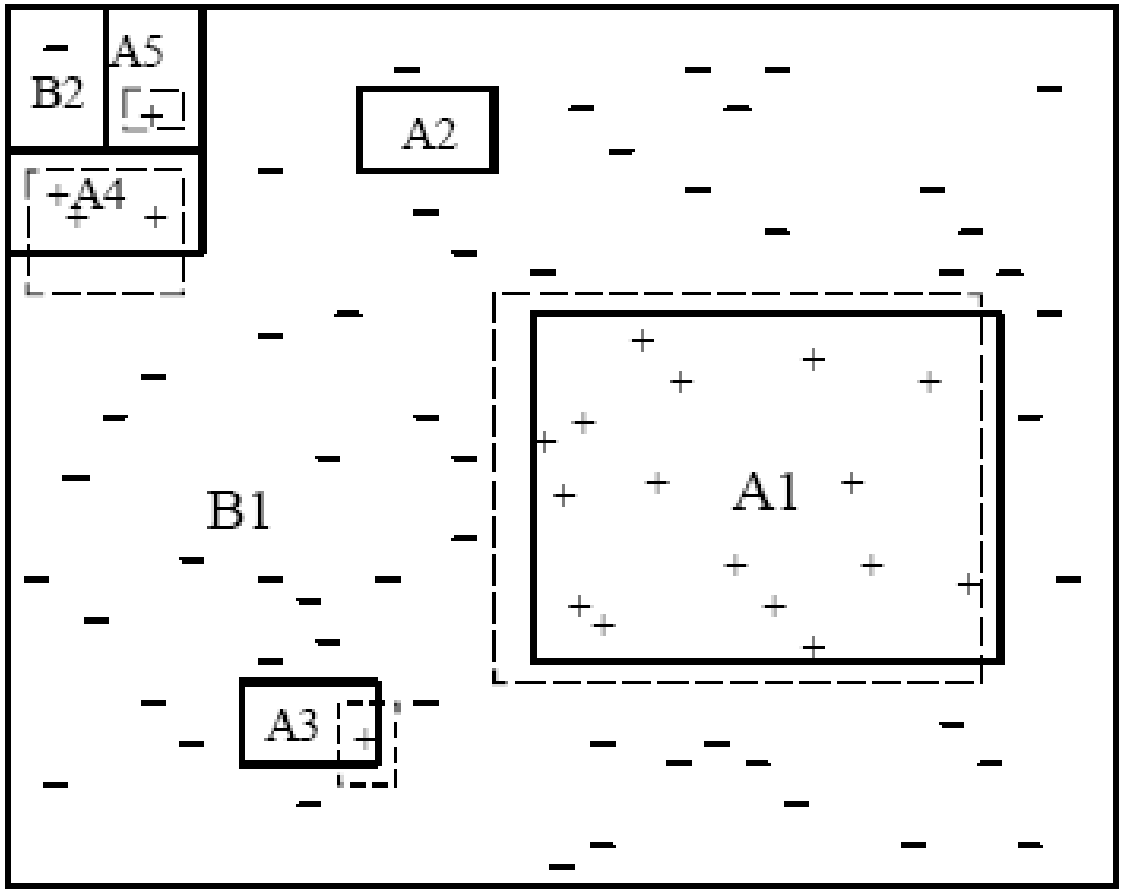
3.3 Pré-processamento dos Dados

- Seleção de atributos, limpeza, construção, integração e formatação dos dados de entrada
  - **remoção de ruído ou de dados espúrios,**
  - **estratégias para lidar com valores faltantes,**
  - **formatação dos dados para a ferramenta a usar,**
  - **criação de atributos derivados e de novos registros,**
  - **integração de tabelas,**
  - **discretização dos dados numéricos, se necessário.**

3. CRISP-DM: Pré-processamento Base de Dados Desbalanceada

- Desbalanceamento de classe (classes raras)
  - Casos de uma classe ocorrem com maior freqüência que casos de outra(s) classe(s)
    - **fraudes são menos freqüentes que transações legítimas**
- Desbalanceamento de casos dentro de uma classe (casos raros)
  - Subconjunto de estados de atributos com menor representação em vista de outros
    - **ocorrência de tipos pouco freqüentes de fraudes, por exemplo fraudes milionárias.**

3. CRISP-DM: Pré-processamento Base de Dados Desbalanceada



s da classe minoritária são representados por '+' e os da classe

3. CRISP-DM: Pré-processamento Base de Dados Desbalanceada

- Desbalanceamento de classes
  - Classificadores tendem a ignorar a classe minoritária
    - **Aproximadamente 1 fraude a cada 850 transações**
- Casos raros
  - Classificadores tendem a ignorar regiões com poucos casos.
    - **Fraudes muito específicas.**
- As duas características são problemáticas quando a classe de interesse é uma classe rara
    - **Modelo tendencioso para a classe majoritária e/ou regiões com mais casos**

3. Metodologia CRISP-DM

3.4 Modelagem

- Quais modelos e parâmetros usaremos?
  - função do tipo de dados (numéricos ou nominais).
  - problema de mineração de dados.
- Elaboração do plano de testes
  - permitir avaliar os modelos gerados.
- Divisão da massa de dados:
  - conjunto de treinamento,
  - conjunto de testes
  - conjunto de validação.

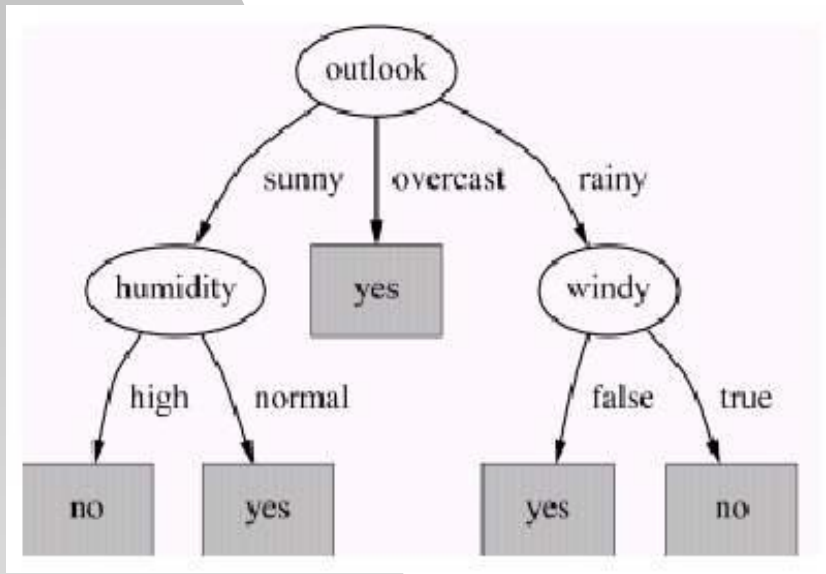
3. Modelagem
Seleção de Modelos

- Tarefa de classificação
  - Árvore de decisão
  - Classificadores neurais
  - Classificadores probabilísticos
- Tarefa de regressão
  - Regressão não linear com redes neurais
- Tarefa de descoberta de associações
  - Modelo neural combinatório

Podem ser criados com o uso do software UnBMiner

3. Seleção de Modelos – Tarefa de Classificação

Arvore de Decisão



- Vantagens:
  - Representação compacta e de fácil visualização
  - Decisão complexa decomposta em decisões elementares
  - Fácil detecção de atributos redundantes ou irrelevantes
  - Permite inferir regras de associação.
- Desvantagens:
  - Árvore pode vir a ser complexa

Espera-se que a árvore faça a classificação com o menor número de perguntas possível.

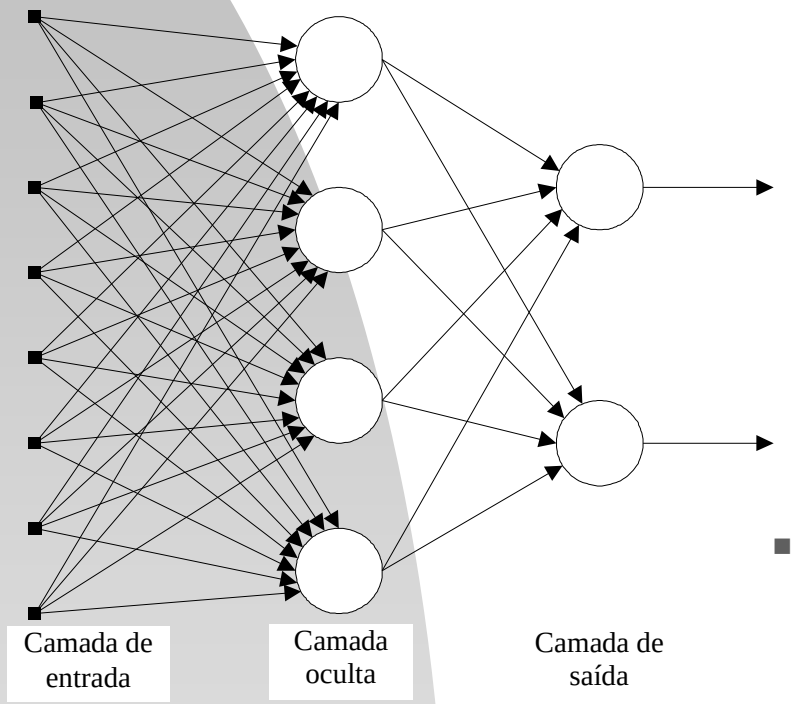
3. Seleção de Modelos – Tarefa de Classificação

Arvore de Decisão

- Algoritmos
  - ID3
    - **Atributos são considerados categóricos**
    - **Conjunto de treinamento deve ser completo**
      - Sem valores faltantes ou valores contínuos
  - C4.5
    - **Atributos podem ser contínuos**
    - **O conjunto de treinamento pode ter valores faltantes**

3. Seleção de Modelos – Tarefa de Classificação

Classificadores Neurais



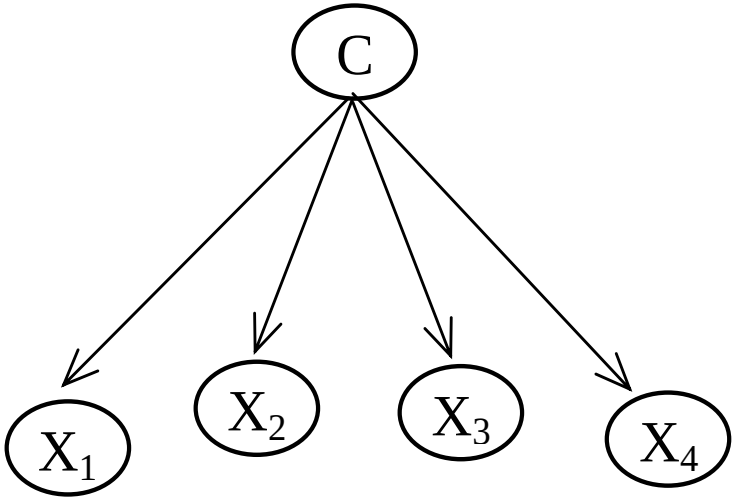
- Características
  - Usa variáveis categóricas
    - **Valores numéricos devem ser discretizados**
  - Treinamento supervisionado
    - **Diversas técnicas**
  - Não requer conhecimento prévio sobre o domínio
  - Facilidade para generalizar e tratar ruídos
- Desvantagens
  - Caixa preta
  - Ajuste de parâmetros é artesanal
    - **Pode tornar-se complexo**

Um neurônio na camada de saída para cada classe possível.

3. Seleção Modelos – Tarefa de Classificação

Classificadores Probabilísticos

- Naive Bayes
  - A classe é pai de todos os atributos
  - Associa probabilidades à classificação da classe
  - Atributos
    - **igualmente importantes**
    - **Independência condicional dado o valor da classe**
    - **assume distribuição normal para atributos numéricos**
    - **admite valores faltantes**

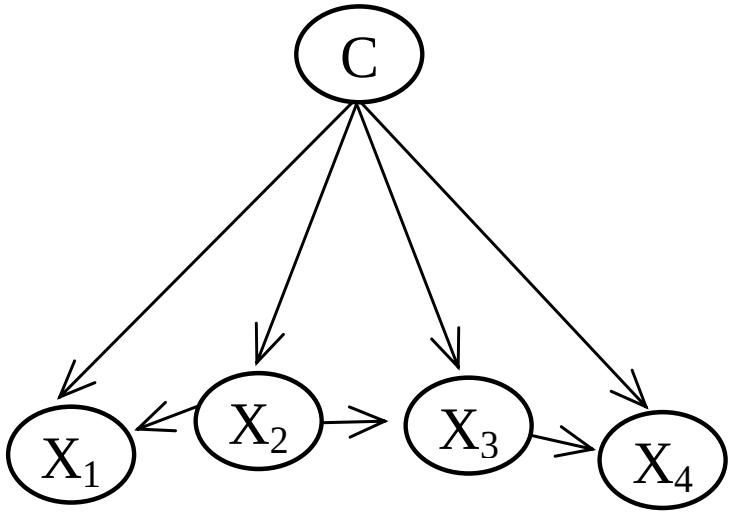


Apresentam bons resultados na prática embora a suposição de independência condicional seja muito forte.

3. Seleção de Modelos – Tarefa de Classificação

Classificadores Probabilísticos

- TAN
  - A classe é pai de todos os atributos
  - Associa probabilidades à classificação da classe
  - Atributos
    - **com dependências probabilísticas na forma de uma árvore**
    - **conjunto de treinamento deve ser completo**
    - **valores categóricos**
      - valores numéricos devem ser discretizados

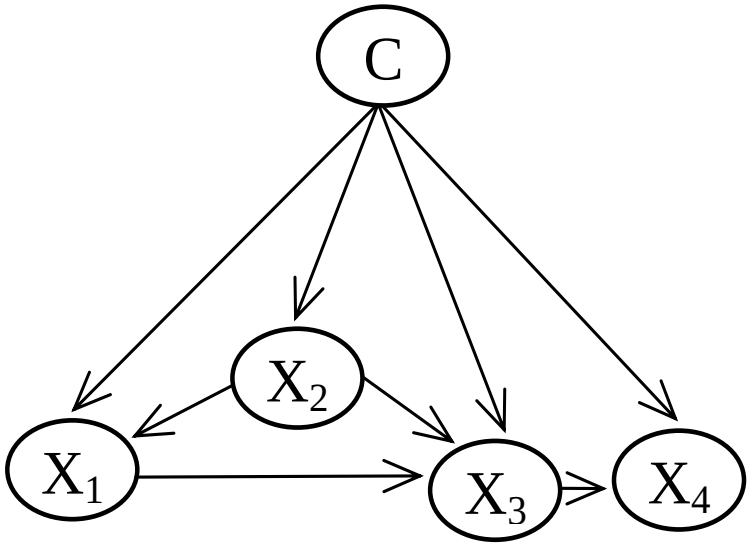


TAN - Tree Augmented Naive Bayes

3. Seleção de Modelos – Tarefa de Classificação

Classificadores Probabilísticos

- BAN
  - A classe é pai de todos os atributos
  - Associa probabilidades à classificação da classe
  - Atributos
    - **com dependências probabilísticas na forma de um grafo acíclico orientado**
    - **conjunto de treinamento deve ser completo**
    - **valores categóricos**
      - valores numéricos devem ser discretizados

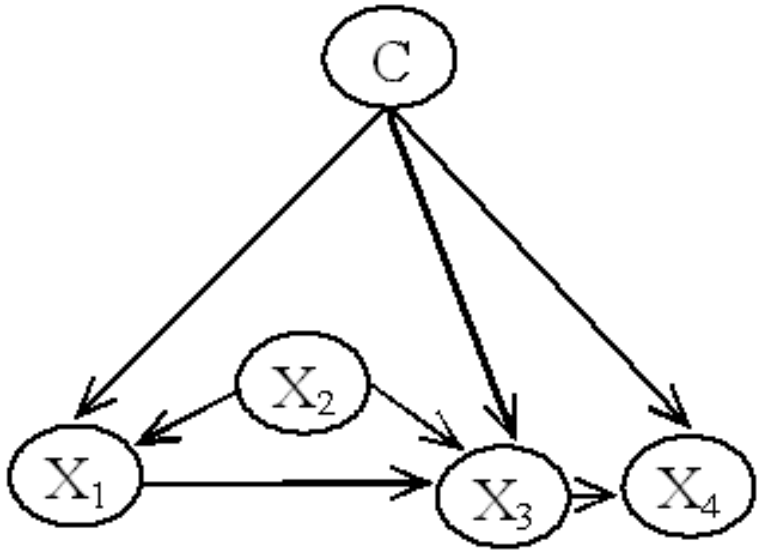


BAN - Bayesian Network Augmented Naive Bayes

3. Seleção de Modelos – Tarefa de Classificação

Classificadores Probabilísticos

- Rede bayesiana
  - Associa probabilidade à classificação da classe
  - Atributos e classe
    - **com dependências probabilísticas na forma de um grafo acíclico orientado**
    - **conjunto de treinamento deve ser completo**
    - **valores categóricos**
      - valores numéricos devem ser discretizados

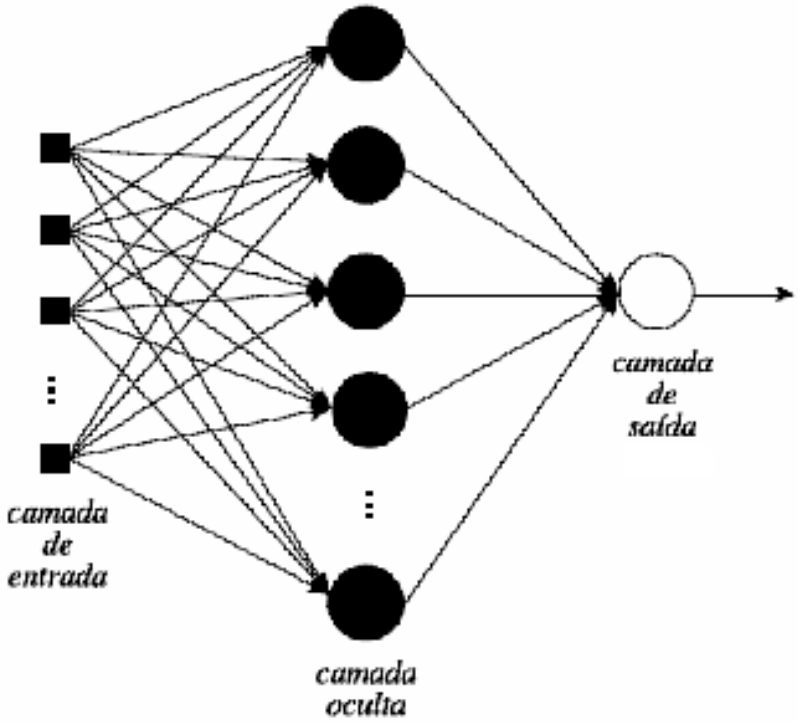


O relacionamento de dependências é o mais geral possível mas nem sempre apresenta bons resultados.

3. Seleção de Modelos – Tarefa de Regressão

Regressão com Redes Neurais

- Vantagens
  - Regressão não linear
    - **sem conhecimento prévio da forma da função**
  - Atributos e classe
    - **conjunto de treinamento completo**
    - **valores numéricos**
- Desvantagens
  - Caixa preta
  - Não permite regressão *stepwise*



3. Metodologia CRISP-DM
3.5 Avaliação

- Desempenho no conjunto de treinamento NÃO é um bom indicador de desempenho em conjuntos de testes independentes
  - Divisão dos dados: treinamento, teste e validação
  - Classificadores predizem a classe de cada instância:
    - **Taxa de sucesso**
      - proporção dos sucessos em relação a todas as instâncias
    - **Qual a relação entre a taxa de sucesso no conjunto de teste e a verdadeira taxa de sucesso?**
      - intervalo de confiança para a taxa de sucesso

3. Metodologia CRISP-DM
3.5 Avaliação

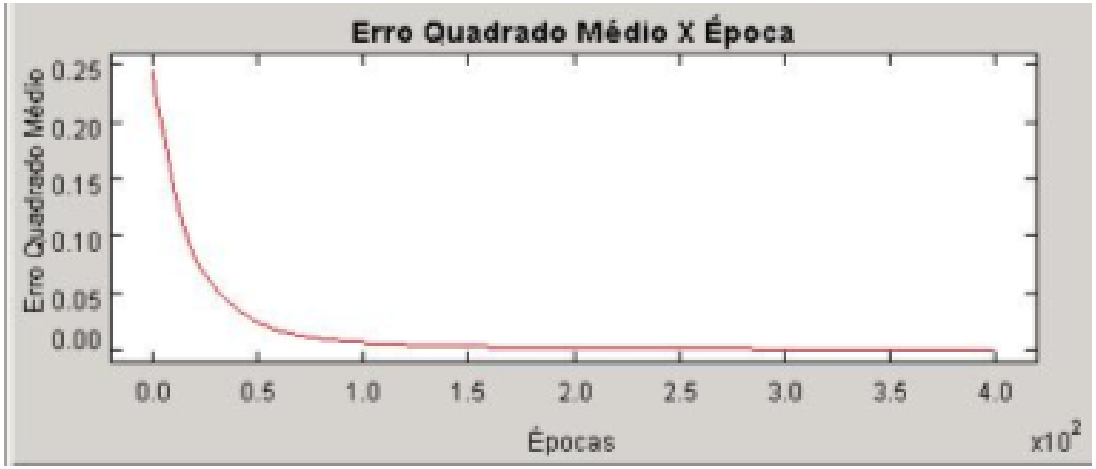
- Avaliar a qualidade dos modelos obtidos no treinamento
  - do ponto de vista de análise dos dados.
  - critério para seleção entre modelos.

R \ P	sim	não
sim	Verdadeiro Positivo (VP)	Falso Negativo (FN)
não	Falso Positivo (FP)	Verdadeiro Negativo (VN)

- Verificar se os objetivos do negócio foram atingidos
  - de acordo com os critérios de sucesso adotados .

3. Metodologia CRISP-DM
3.5 Avaliação

- Discriminação entre Regressores
  - Maior coeficiente de correlação de Pearson (r)
  - Maior coeficiente de determinação (r^2)
  - Menor erro quadrático médio (MSE)
  - Menor erro absoluto médio



3. Metodologia CRISP-DM
3.6 Colocação em Uso

- Modelo selecionado
  - incorporado ao processo de tomada de decisão da organização
- Plano de monitoração e manutenção
  - previne uso incorreto dos resultados do mineração, durante um longo período de tempo.

3. Metodologia CRISP-DM

Compreensão do domínio	Compreensão dos dados	Preparação de dados	Modelagem	Avaliação	Aplicação
Determinação dos objetivos da aplicação *Cenário* *Objetivos da aplicação* *Critérios de sucesso da aplicação* **Situação a ser avaliada** *Inventário de recursos* *Requisitos, suposições e limitações* *Terminologia* *Custos e benefícios* **Determinação das metas de MD** *Metas* *Critérios de sucesso de MD* **Produção do projeto** *Projeto* *Avaliação inicial de ferramentas e técnicas*	**Coleta de dados inicial** *Relatório da coleta de dados inicial* **Descrição dos dados** *Relatório da descrição dos dados* **Exploração dos dados** *Relatório da exploração dos dados* **Verificação da qualidade dos dados** *Relatório de qualidade dos dados*	*Conjunto de dados* *Descrição do conjunto de dados* **Seleção de dados** *Racionalizar para inclusão/ exclusão* **Limpeza de dados** *Relatório da limpeza de dados* **Construção de dados** *Atributos derivados* *Registros gerados* **Integração de dados** *Dados mesclados* **Formatação de dados** *Dados reformatados*	**Seleção da técnica de modelagem** *Técnica de modelagem* *Suposições de modelagem* **Geração do projeto de teste** *Projeto de teste* **Construção do modelo** *Configurações de parâmetros* *Modelos* *Descrição dos modelos* **Modelo a ser avaliado** *Avaliação do modelo* *Configurações de parâmetros revisadas*	**Avaliação dos resultados** *Avaliação dos resultados de MD em função dos critérios de sucesso da aplicação* *Modelos aprovados* **Revisão do processo** *Revisão do processo* **Determinação dos próximos passos** *Lista de possíveis ações* *Decisões*	**Aplicação do projeto** *Plano de aplicação* **Plano de monitoramento e manutenção** *Monitoramento e manutenção do plano* **Produção do relatório final** *Relatório final* *Apresentação final* **Revisão do projeto** *Documentação da experiência*

4. Aplicações Desenvolvidas

- Detecção de fraudes em cartões de crédito
  - Banco comercial (redes bayesianas)
- Área militar (Comando da Aeronáutica)
  - IEAv (redes bayesianas)
- Busca decadactilar de impressões digitais
  - INI/DPF
    - **Lupa Digital (mineração de dados e teorema de Bayes)**
- Diagnóstico médico
  - Cardiopatias congênitas (redes bayesianas)
    - **Instituto de Cardiologia/RS**
  - Síndrome da apnéia obstrutiva do sono (redes neurais)
    - **Laboratório do Sono – HUB**
- Pedotransferência de água
  - CPAC/Embrapa (redes neurais - regressão não linear)

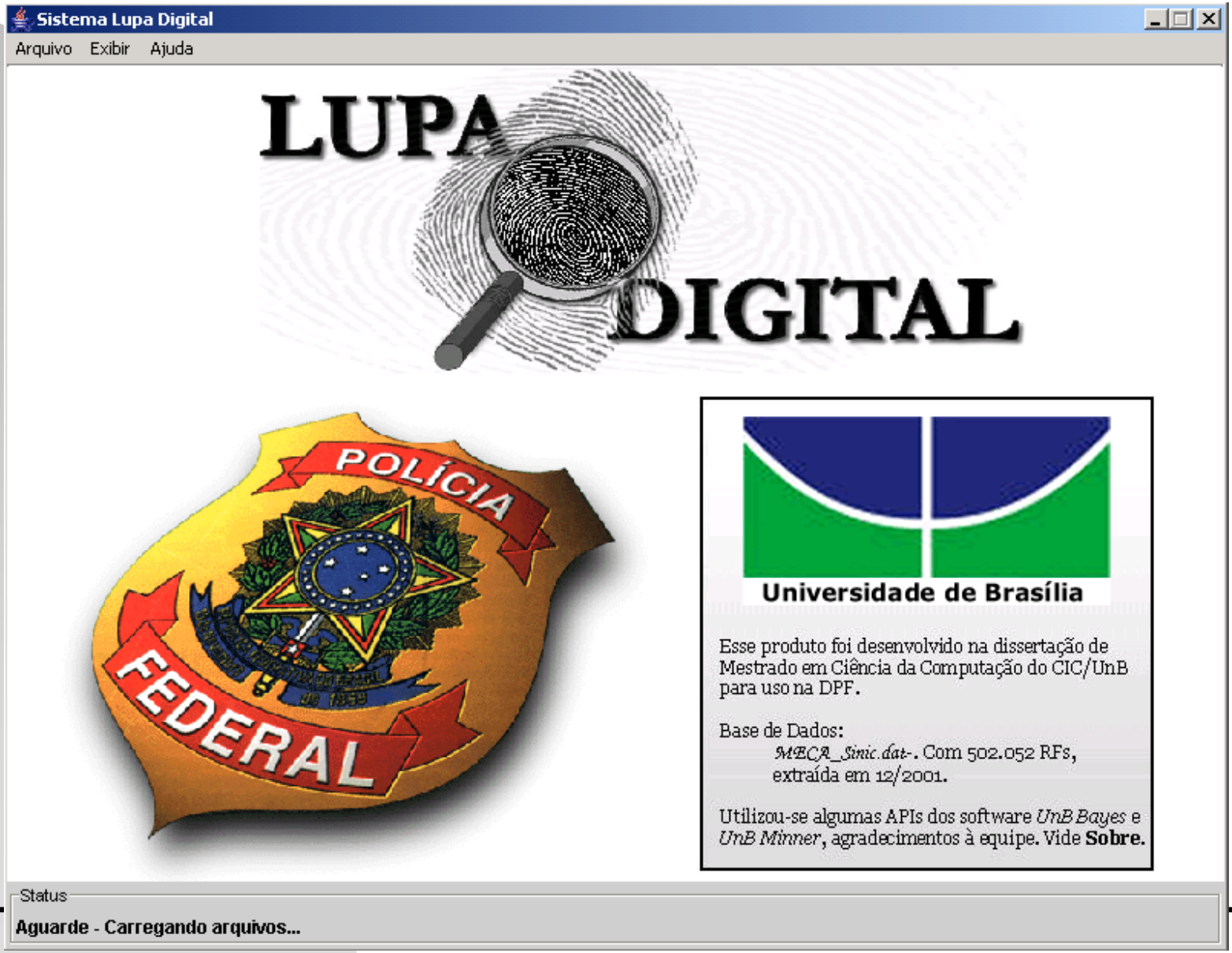
4. Aplicações Desenvolvidas

- Base de Germoplasma do SIBRARGEN
  - Embrapa CENARGEN (metodologia de DM)
- Classificação por Imagens do Acabamento de Gordura de Carcaças Bovinas
  - FAP-DF (Metodologia de DM)
- Aprendizagem estatística para recuperação da informação
  - Jurisprudências de tribunal superior (TCU)
- Recomendação de Consultores Ad Hoc Baseada na Extração de Perfis do Currículo Lattes
  - CNPq (Edital Universal)
- Classificação Automática de Páginas Web Multi-label via Support Vector Machines e MDL
  - Árvore de Huffman
- Ferramenta para Filtragem Adaptativa de Spam
  - Árvore de Huffman

4 Busca Decadatilar de Impressões Digitais
Sistema Lupa Digital

- Hipótese de pesquisa
  - Obter um, ou mais, modelos de classificador que gerasse (complementasse) os códigos *Vucetich* para as impressões digitais dos dedos faltantes e, desta forma, reduzisse o espaço de busca em pesquisa manual ou automatizada (AFIS*) de identificação de impressões digitais.

*** Automatic Fingerprint Information System**

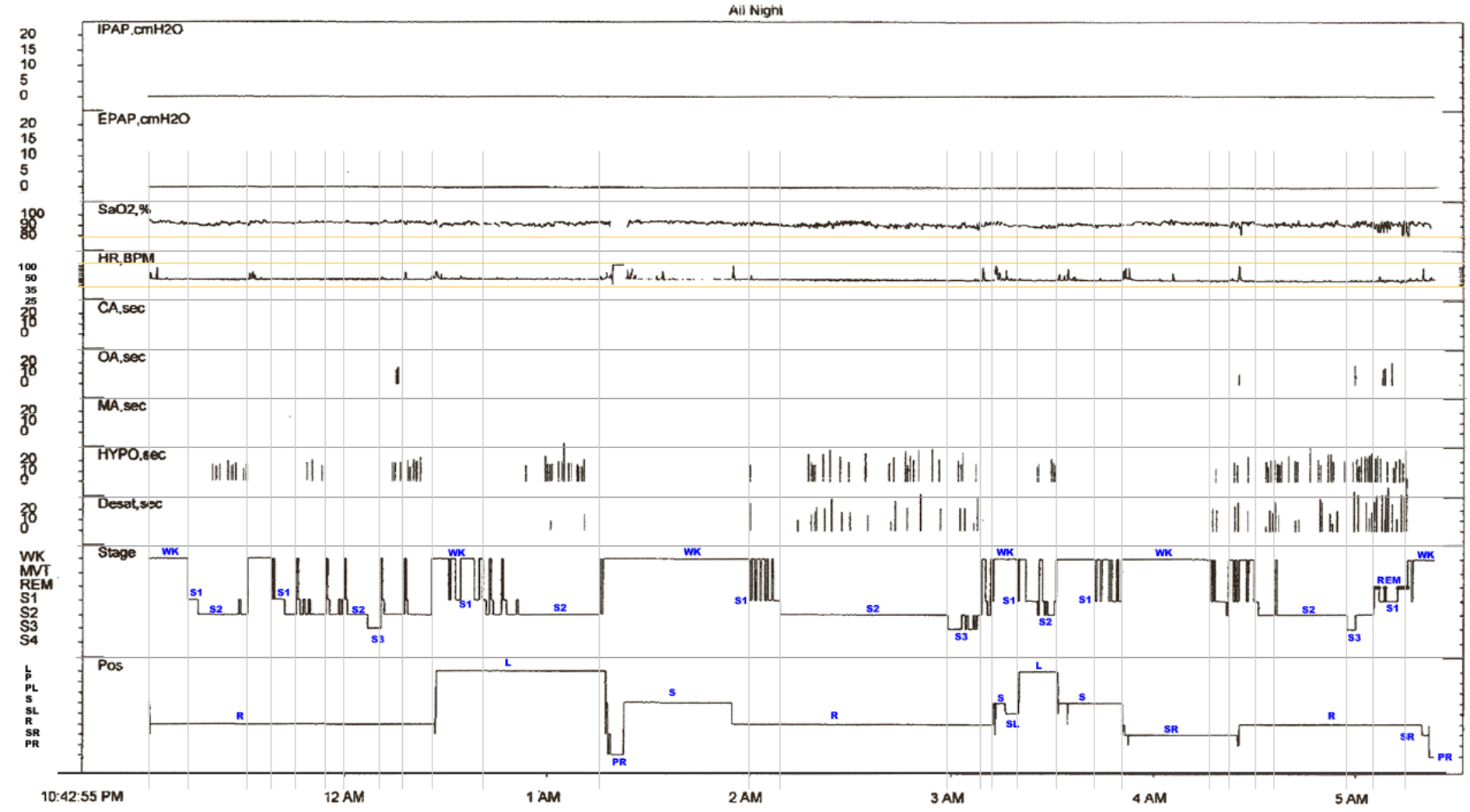

Sistema Lupa Digital
Arquivo Exibir Ajuda
LUPA
DIGITAL
POLÍCIA
FEDERAL
Universidade de Brasília
Esse produto foi desenvolvido na dissertação de Mestrado em Ciência da Computação do CIC/UnB para uso na DPF.
Base de Dados:
MECA_Sinic.dat-. Com 502.052 RFs, extraída em 12/2001.
Utilizou-se algumas APIs dos software UnB Bayes e UnB Minner, agradecimentos à equipe. Vide Sobre.
Status
Aguarde - Carregando arquivos...

4. Síndrome SAOS

Exame por Polissonografia

4. A Síndrome SAOS

Resultado da Polissonografia



4. Germoplasma

Definição do Problema

- SIBRARGEN – base de germoplasma atualmente com mais de 100.000 tipos característicos de vegetais (acessos)
  - visa auxiliar no processo decisório, atuando na pesquisa em recursos genéticos
    - **o suporte atual à decisão é incipiente**
      - consultas SQL e emissão de relatórios pré-determinados
        - filtros (expressões booleanas)
  - Usuários altamente especializados
    - **sem formação específica em informática**
  - Hipótese
    - **uso de técnicas de mineração pode prover aumento potencial da exploração dos dados**

4. Germoplasma

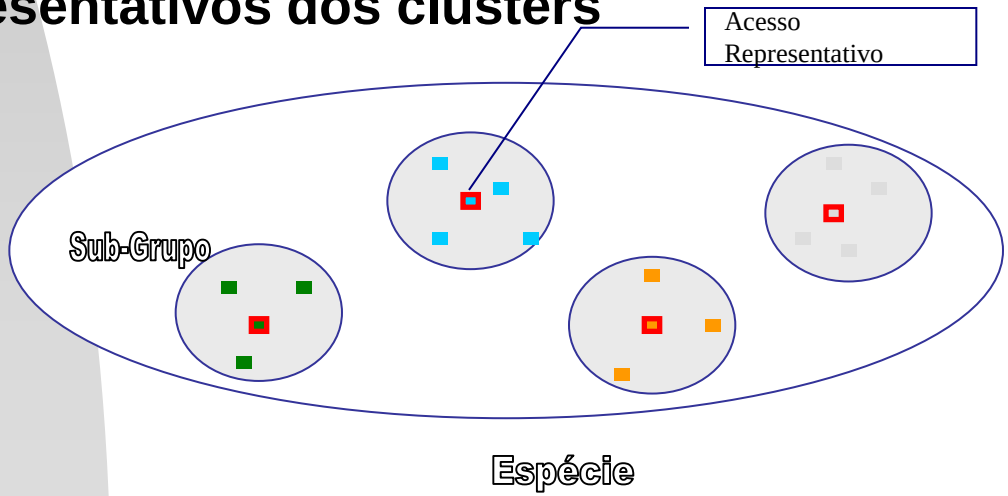
Objetivo

- **Propor metodologia de mineração de dados de germoplasma, aplicada às bases de *passaporte*, *caracterização* e *avaliação*, para facilitar aos pesquisadores, em geral leigos em informática, aplicarem técnicas de mineração aos dados armazenados no SIBRARGEN.**
  - baseada em metodologia de mineração de dados aceita mundialmente
  - implemente todo o ciclo de mineração de dados
- **Prover um ferramental para materializar a metodologia proposta, sem requerer programação**
  - integrado ao SIBRARGEN
  - acesso via Web
  - interface intuitiva e de fácil uso
- **Validar experimentalmente com estudo de caso**

4. Germoplasma

Acessos representativos

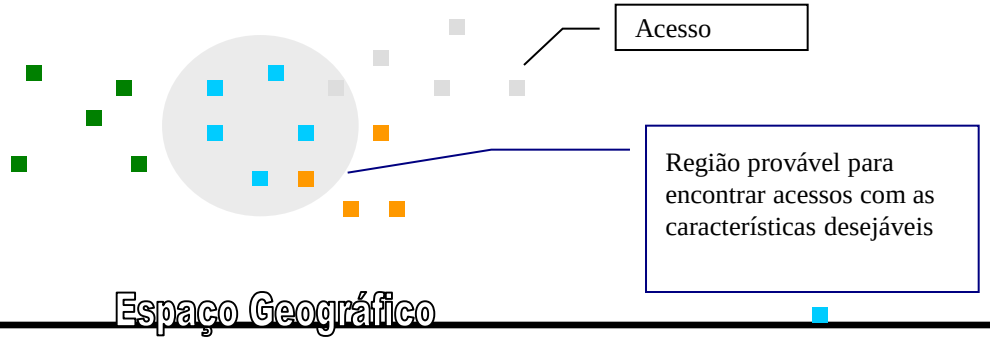
- Problema real na Embrapa
  - Selecionar dentro de um grupo os acessos que são mais representativos por conterem características marcantes e diversas de outros sub-grupos formados no grupo
    - **análise de agrupamentos e seleção dos indivíduos mais representativos dos clusters**



4. Germoplasma

Coleta direcionada

- Problema real na Embrapa no planejamento das expedições de coleta de acessos
  - A expedição tem o objetivo de encontrar acessos com certas características. A base tem o registro dos acessos já coletados e a localização da coleta.
    - **a partir de dados geográficos de latitude e longitude oferecer sugestões de regiões onde é mais provável encontrar os acessos desejados.**



4. Classificação do acabamento de gordura em carcaças bovinas

- A classificação realizada conforme Portaria Ministerial nº 612/89 do MAPA envolve subjetividade por parte do técnico habilitado.

Gordura Magra
Tipo 1

Gordura Escassa
Tipo 2

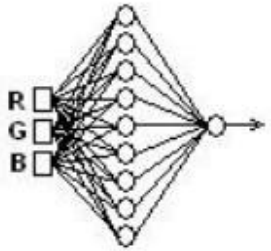
Gordura Mediana
Tipo 3

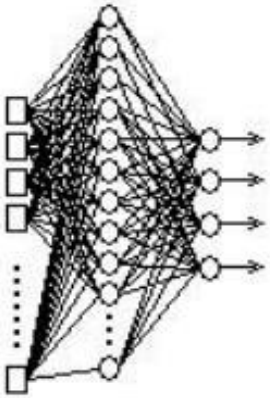
Gordura Uniforme
Tipo 4

Gordura Excessiva
Tipo 5

4. Classificação do Acabamento de Gordura

Solução Proposta





1 – Magra

2 – Gordura escassa

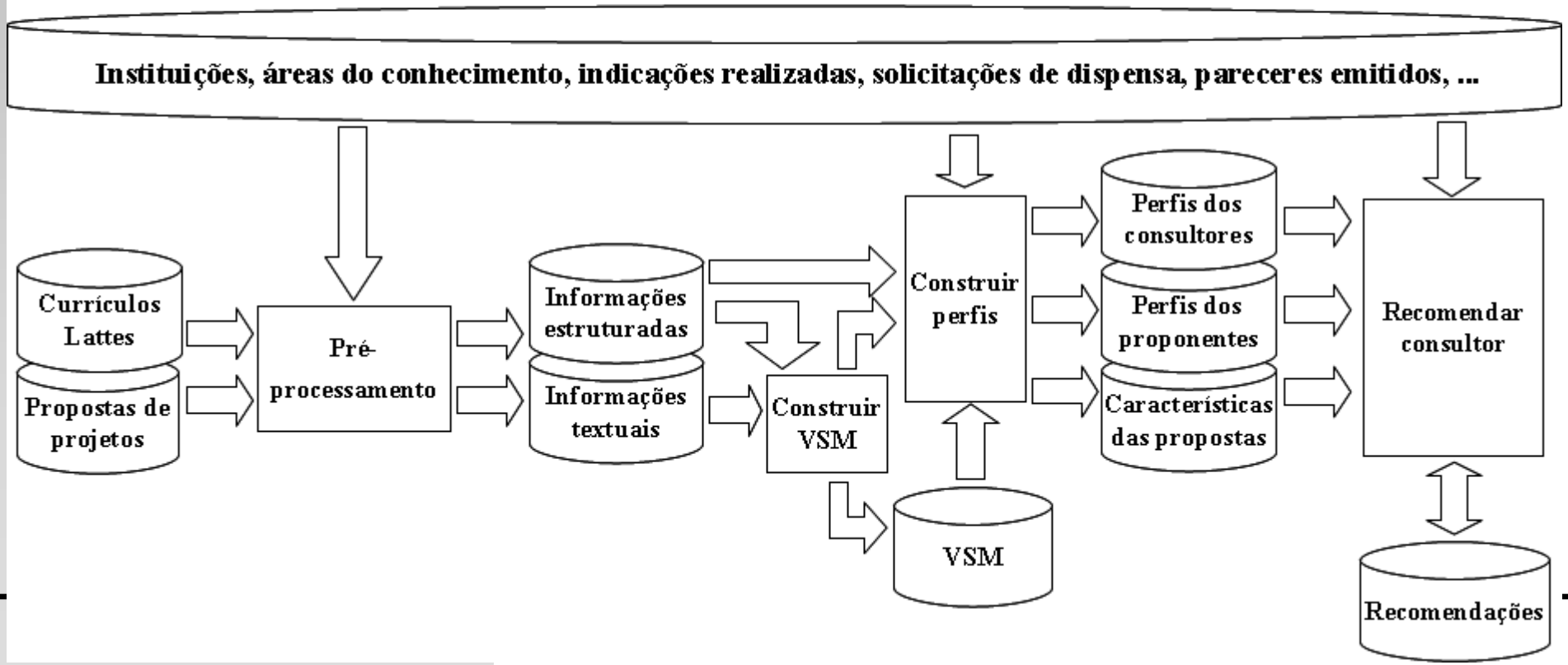
3 – Gordura mediana

4 – Gordura uniforme

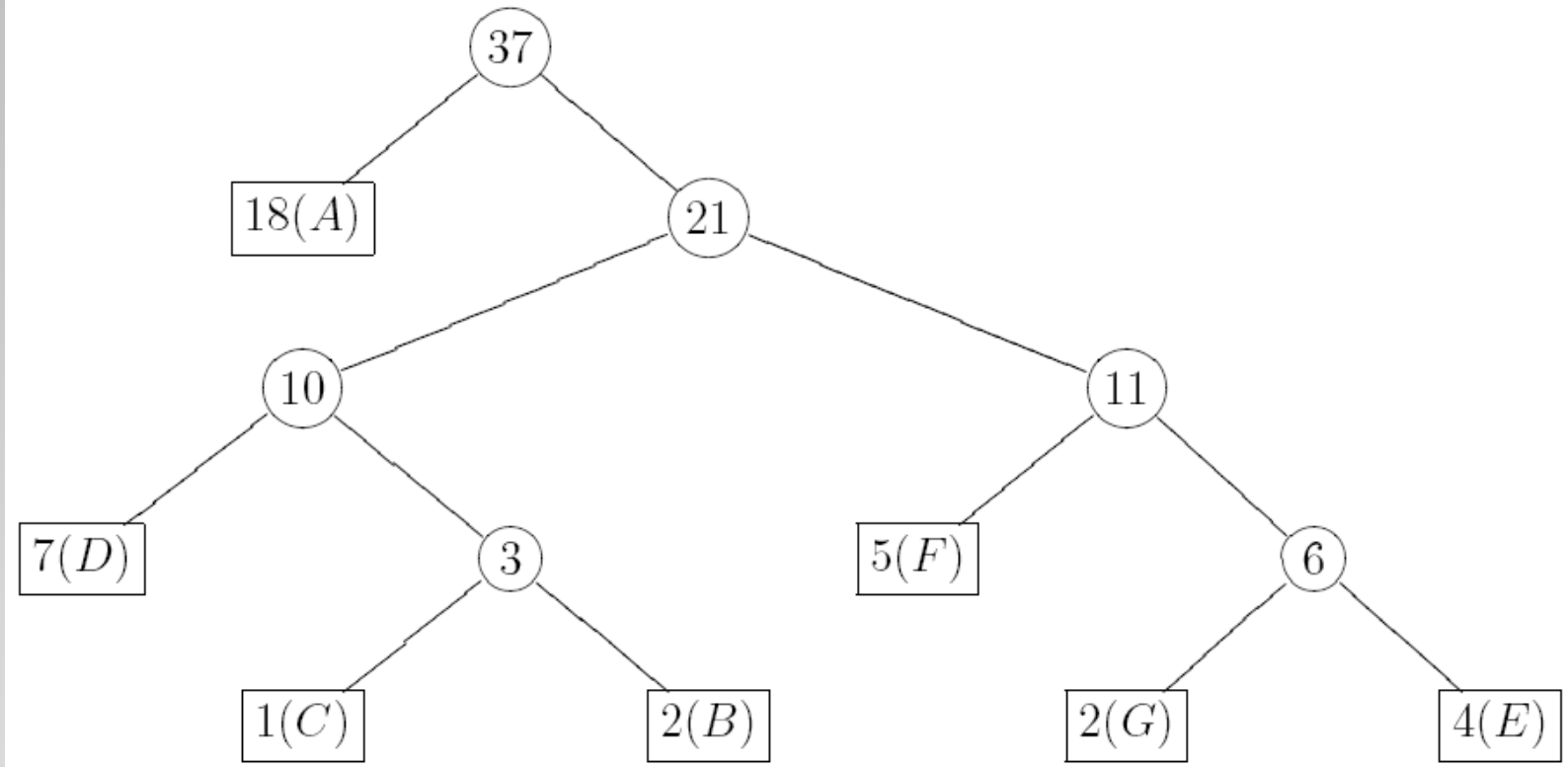
Máscara de 50 x 50 pixels

4. Recomendação de Consultores Ad Hoc Baseada na Extração de Perfis do Currículo Lattes

Metodologia Proposta

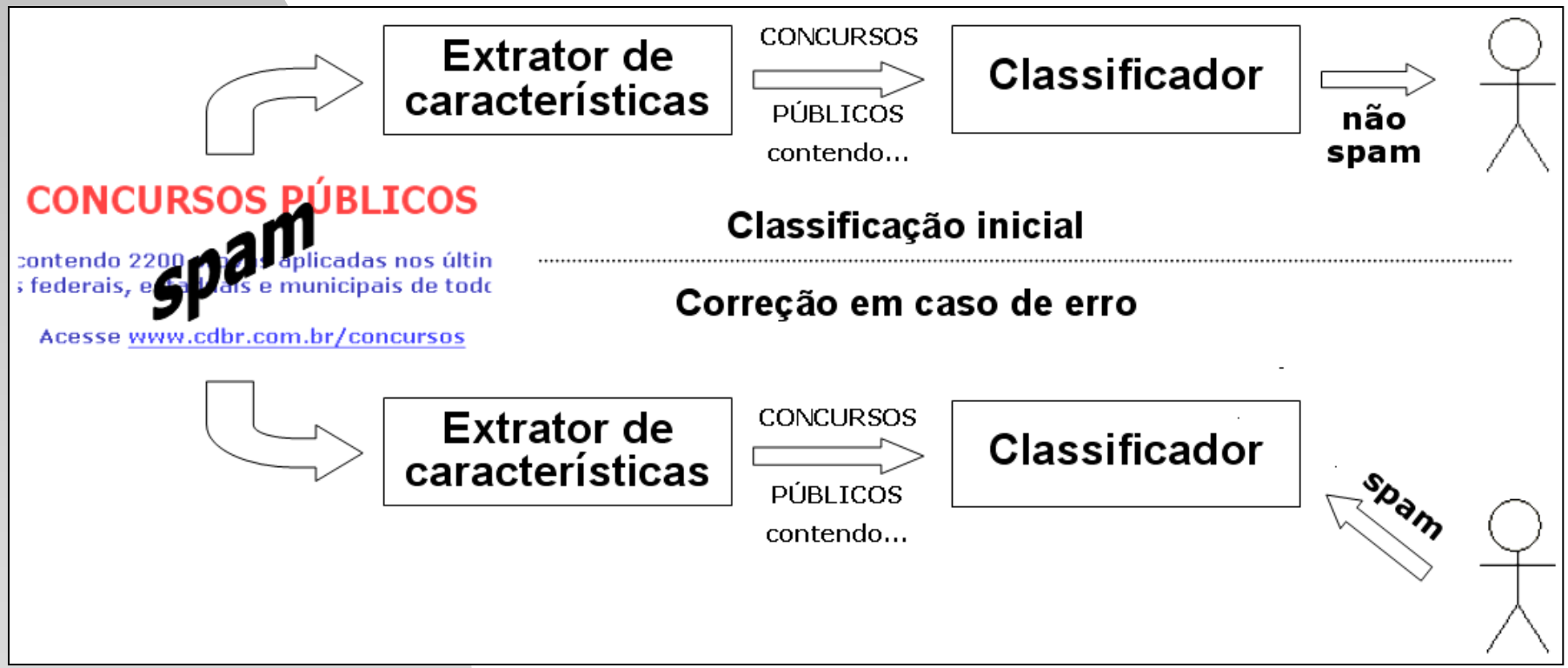


4. Classificação Automática de Páginas Web Multi-label



Árvore de Huffman

4. Ferramenta para Filtragem Adaptativa de Spam



5. Estudo de Casos

- Base de dados balanceada
  - Diagnóstico médico
    - **Síndrome da apnéia obstrutiva do sono**
- Base de dados não balanceada
  - Detecção de fraudes
    - **Transações com cartões de crédito**

5. Estudo de Casos – Base de Dados Balanceada

A Síndrome SAOS

- A Síndrome da Apnéia Obstrutiva do Sono
  - É caracterizada por 5 ou mais paradas respiratórias do fluxo aéreo durante o sono e hipersonolência diurna.
  - Está fortemente associada à ocorrência de:
    - **HAS, obesidade, hipersonolência diurna, coronariopatias, DPOC, AVC, infartos, acidentes de trânsito.**
  - É pouco reconhecida mas sua importância cresceu rapidamente nos últimos anos

5. Estudo de Casos – Base de Dados Balanceada

A Síndrome SAOS

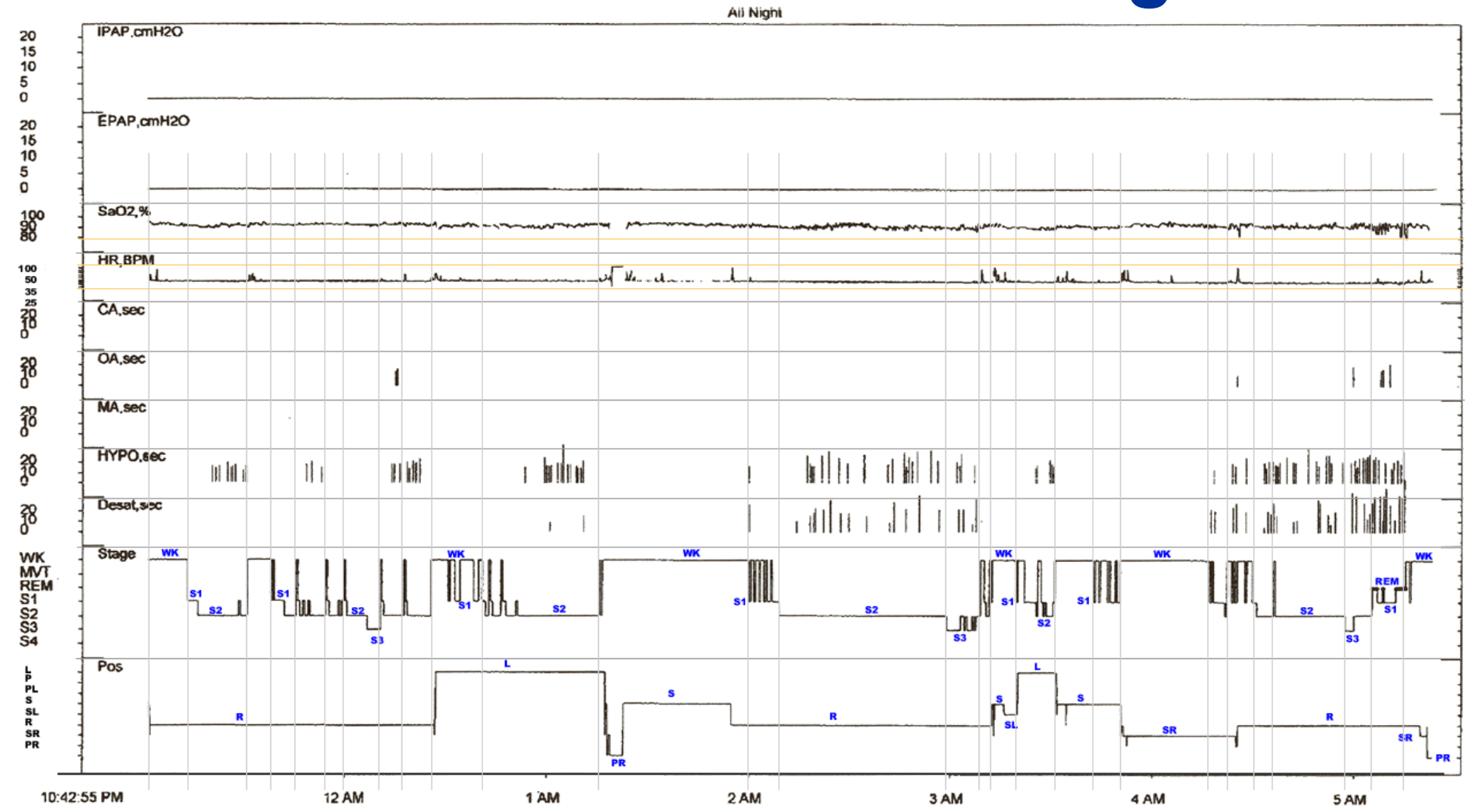
- Prevalência de 3% na população
- Custo estimado nos EUA: US$ 3,1 bilhões ao ano.
- Diagnóstico feito por polissonografia
  - aparelhos caros e apoio de técnicos especializado,
  - exige internação por uma noite e um leito,
  - exame disponível em poucos centros públicos
    - **requer a existência de um laboratório do sono**
  - filas de espera são crescentes e cronológicas.

5. Estudo de Casos – A Síndrome SAOS

Exame por Polissonografia

5. Estudo de Casos – A Síndrome SAOS

Resultado da Polissonografia



5. Estudo de Casos – A Síndrome SAOS

Diagnóstico por Polissonografia

- Principal parâmetro para o diagnóstico é o índice de apnéias e hipopnéias (IAH)
  - razão entre o total de apnéias e hipopnéias registradas e o tempo total de registro em horas.
  - categorização da SAOS pela intensidade do IAH, segundo a Sociedade Americana do Sono
    - **sem SAOS (<5), leve (5-15), moderada (15-30), severa (>30).**

5. Estudo de Casos – A Síndrome SAOS
Necessidade de Diagnóstico Efetivo

- Diagnóstico clínico de SAOS
  - acurácia máxima (probabilidade de acerto máximo) até 60%
- Os custos de não diagnosticar e tratar a doença precocemente justificam as polissonografias
- Já existem grandes filas de espera e sobrecarga dos laboratórios de sono do SUS

5. Estudo de Casos – A Síndrome SAOS

Objetivos

- Propor classificador para IAH (portador ou não portador da SAOS) a partir de variáveis clínicas e antropométricas com acurácia, sensibilidade e especificidade melhores que os critérios atuais.
- Implementar o melhor classificador na clínica do sono do HUB para otimizar a fila para polissonografias, diminuindo a possibilidade de ocorrência de complicações e os custos.

5. Estudo de Casos – A Síndrome SAOS
Método

- Aplicação do modelo de referência CRISP-DM
- Dados de treinamento: 1000 pacientes
  - Dados reais do Laboratório do Sono do HUB
- Dados de avaliação: 157 (total 1157)
- Ferramentas: UnBBayes, UnBMiner e SAS
- Formalismos: redes bayesianas, TAN, BAN, *naïve Bayes e backpropagation*.

5. Estudo de Casos – A Síndrome SAOS
Entendimento dos Dados e Pré-Processamento

- Seleção de atributos (variáveis do domínio)
  - revisão bibliográfica médica e opinião de especialista (inicialmente 11 atributos).

Variável	Tipo	Média	Desvio padrão	Mínimo	Máximo
Altura (em cm)	**Numérico**	**166.573**	**9.299**	**140**	**188**
Choque noturno (Apnéia)	**Nominal**	**0.771**	**-**	**0 (não)**	**1 (sim)**
Circunferência do pescoço (CP)	**Numérico**	**38.745**	**3.997**	**29**	**49**
Droga miorrelaxante (DrMio)	**Nominal**	**0.280**	**-**	**0 (não)**	**1 (sim)**
Gênero masculino (Masc)	**Nominal**	**0.586**	**-**	**0 (não)**	**1 (sim)**
Hipertensão arterial sistêmica	**Nominal**	**0.299**	**-**	**0 (não)**	**1 (sim)**
IAHSN	**Nominal**	**0.600**	**-**	**0 (não)**	**1 (sim)**
Idade (em anos)	**Numérico**	**47.866**	**14.631**	**13**	**88**
Peso (em quilos)	**Numérico**	**81.159**	**20.927**	**40**	**167**
Roncos noturnos freqüentes	**Nominal**	**0.892**	**-**	**0 (não)**	**1 (sim)**
Tabagismo	**Nominal**	**0.102**	**-**	**0 (não)**	**1 (sim)**

Base de dados balanceada.

5. Estudo de Casos – A Síndrome SAOS
Entendimento dos Dados e Pré-Processamento

- Seleção de atributos (variáveis do domínio)
  - filtro com regressão linear com seleção *stepwise* para a variável IAH numérica e regressão logística para IAH dicotômica:
    - **eliminação dos atributos ronco e tabagismo.**
    - **enriquecimento**
      - altura e peso normalizados como índice de massa corporal IMC = peso / altura²

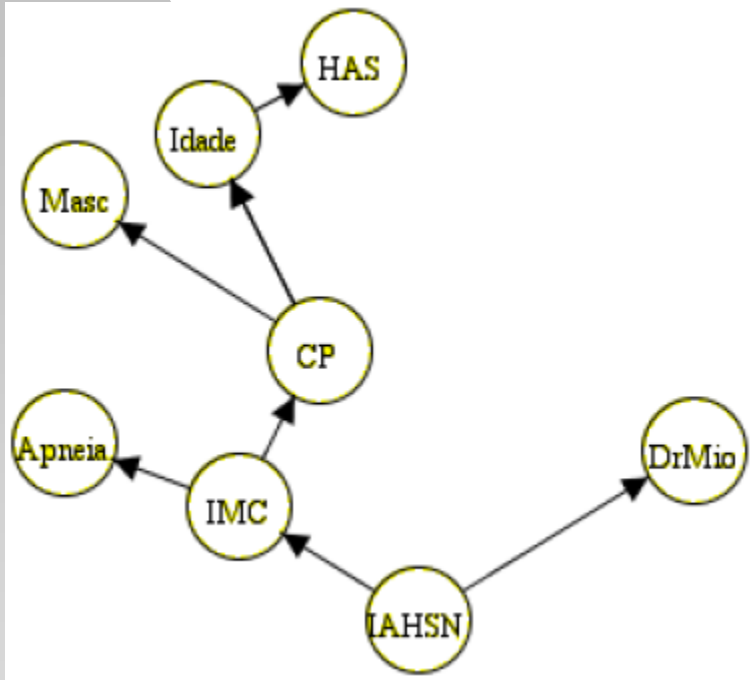
Após as fases de entendimento dos dados e pré-processamento restaram 8 atributos.

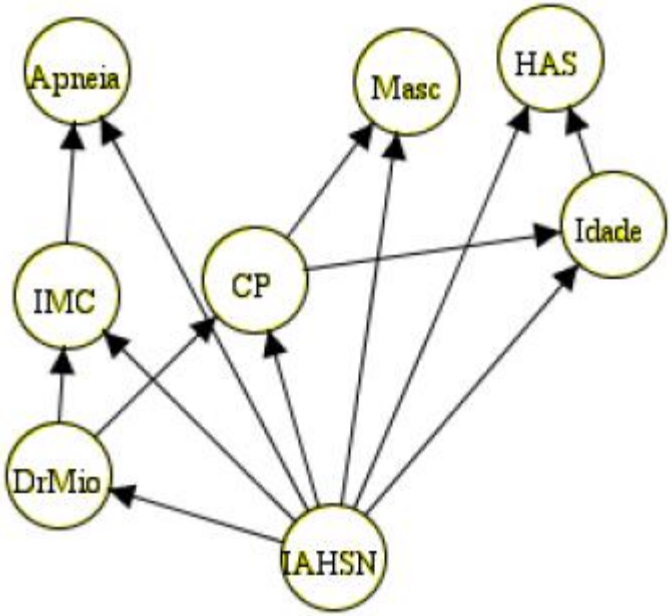
5. Estudo de Casos – A Síndrome SAOS
Atributos Selecionados

Atributo	Sigla	Tipo
Apnéia (choque noturno)	Apnéia	Dicotômica
Circunferência do pescoço	CP	Numérica
Gênero Masculino	Masc	Dicotômica
Hipertensão arterial sistêmica	HAS	Dicotômica
IAH (atributo de classe)	IAHSN	Dicotômica
Idade	Idade	Numérica
IMC	IMC	Numérica
Usa drogas miorrelaxantes	DrMio	Dicotômica

5. Estudo de Casos – A Síndrome SAOS

Modelos Gerados

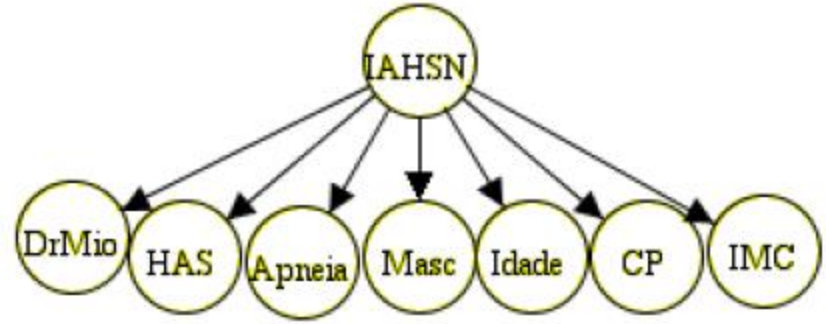


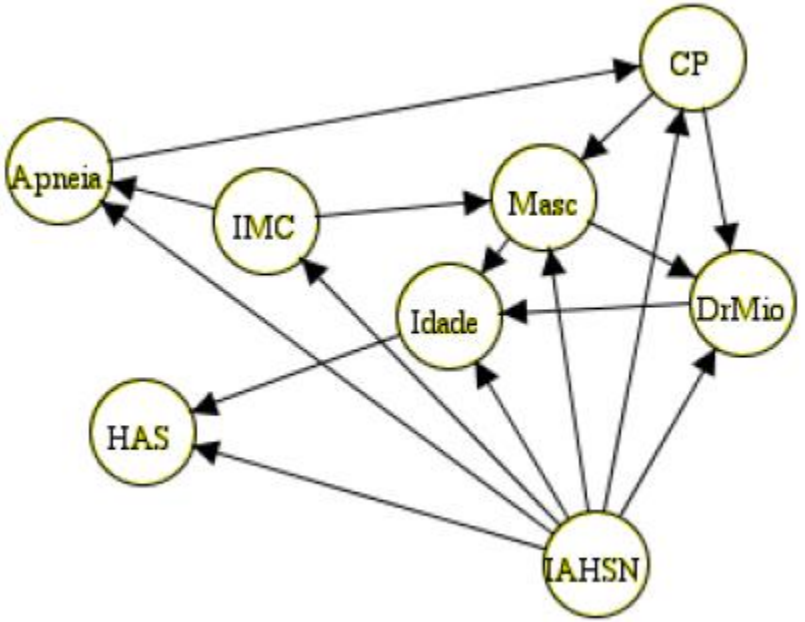


Rede bayesiana e TAN

5. Estudo de Casos – A Síndrome SAOS

Modelos Gerados





Naïve Bayes e BAN

5. Estudo de Casos – A Síndrome SAOS

Modelos Gerados

- Rede Neural MLP Backpropagation

Parâmetro	Valor	Parâmetro	Valor
Neurônios na camada de entrada	7	Taxa de aprendizagem	0,3
Neurônios na camada escondida	5	Momento	0,2
Neurônios na camada de saída	1	Função de ativação	sigmóide
Épocas	400	Normalização	linear

5. Estudo de Casos – A Síndrome SAOS
Análise dos Resultados

- Seleção de modelos

Matriz de Confusão

R \ P	sim	não
sim	Verdadeiro Positivo (VP)	Falso Negativo (FN)
não	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Índices para discriminação entre modelos

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

5. Estudo de Casos – A Síndrome SAOS
Análise dos Resultados

- Avaliação dos Modelos Gerados

Modelo	Sensibilidade	Especificidade	Acurácia	SE
Rede Bayesiana	0,769	0,697	0,7388	0,5360
Naive Bayes	0,879	0,712	0,8089	0,6258
TAN	0,714	0,470	0,6114	0,3356
BAN	0,758	0,697	0,7324	0,5283
Rede Neural	1,000	0,955	0,9800	0,9550

5. Estudo de Casos – A Síndrome SAOS

Conclusões

- Foi alcançado o objetivo de criar um método de diagnóstico barato, rápido e não invasivo com grande acurácia (98%) que poderá diminuir a falha e a demora no diagnóstico da SAOS devido às grandes filas de espera para polissonografia.
- Os resultados estão sob avaliação do Comitê de Ética Médica em Pesquisa da UnB para avaliar como será a aplicação prática do modelo.
- O fato da rede neural ser o melhor modelo é uma evidência experimental da existência de uma relação não linear entre IAH e as demais variáveis

5. Estudo de Casos – Base de Dados Desbalanceada

Fraudes em Cartões de Crédito

- Base de dados
  - real e possui classe rara e casos raros.
  - os dados correspondem às transações em cartões de crédito efetuadas nos meses de outubro a dezembro.
    - **6,4 milhões de transações**
      - apenas 4.810 dessas transações são casos de fraude.

5. Estudo de Casos – Fraudes em Cartões de Crédito
Objetivo

- Desenvolver um classificador que auxilie na identificação das transações fraudulentos
  - O *critério de sucesso*
    - **aumentar em, no mínimo, 10% o índice S.E do modelo obtido, sem preparação específica dos dados.**
      - objetivo final almejado: S.E $\geq$ 0,77.
  - Porque usar o índice S.E?
    - **Deseja-se ferramenta que identifique o maior número possível de fraudes**
      - com o menor número possível de lixo associado (ou seja, menor falso positivo).
      - poucos funcionários na “área de detecção de fraudes”
        - pode haver a necessidade de se entrar em contato com o cliente para certificar a veracidade da fraude.

5. Estudo de Casos – Fraudes em Cartões de Crédito

Método

- Aplicação do modelo de referência CRISP-DM
- 1.240.185 transações no mês de outubro
  - 1.238.708 transações legítimas
  - 1.477 transações fraudulentas,
    - **uma fraude para aproximadamente 839 não fraudes.**
  - Dados de treinamento: 826.790 (2/3 dos dados)
  - Dados de avaliação: 413.395 (1/3 dos dados)
- Ferramentas: UnBBayes, UnBMiner e SAS
- Formalismos: redes bayesianas, TAN, BAN, *naïve Bayes e backpropagation*

As distribuições das classes dos conjuntos de treinamento e avaliação são similares.

5. Estudo de Casos – Fraudes em Cartões de Crédito

Entendimento dos Dados

- Base de dados original
  - 150 atributos obtidos automaticamente e uma classe (*fraude*)
  - Todas as variáveis, exceto *hora* e *fraude*, foram renomeadas por razões de sigilo.
- Seleção de variáveis
  - facilita o desenvolvimento de modelos úteis e inteligíveis
  - selecionados os dez atributos que mais contribuem para explicar a classe (*fraude*).
    - **dados discretizados e condensados com variável de contagem (*Qtd*) que serve para agrupar casos repetidos.**
      - base de dados com 12 variáveis

5. Estudo de Casos – Fraudes em Cartões de Crédito

Entendimento dos Dados

- Descrição Estatística dos Dados

1.240.185 transações

	v1	v2	v3	v4	V5	v6	v7	v8	v9	hora	fraude	Qtd
Num. Estados	**4**	**4**	**6**	**5**	**4**	**2**	**5**	**4**	**8**	**24**	**2**	**490**
Valor Min	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**
Valor Max	**3**	**3**	**5**	**4**	**3**	**1**	**4**	**3**	**7**	**23**	**1**	**43318**

5. Estudo de Casos – Fraudes em Cartões de Crédito
Pré-processamento dos Dados

- Técnicas para redução do desbalanceamento
  - Não há garantia de que a distribuição original seja a mais adequada para a construção de classificadores
    - a distribuição que maximiza a performance do classificador deve ser determinada de forma empírica
  - **amostragem undersampling e oversampling**
    - visa mudar a distribuição dos dados de treinamento de modo a aumentar a acurácia dos modelos treinados a partir deles.
    - undersamplig
      - eliminação de casos da classe majoritária
    - oversamplig
      - replicação de casos da classe minoritária

Deseja-se reduzir as diferenças entre as distribuições das classes.

5. Estudo de Casos – Fraudes em Cartões de Crédito

Pré-processamento dos Dados

- No presente estudo
  - *undersampling* consiste em diminuir o valor de *Qtd* para as transações não fraudulentas.
  - *oversampling* consiste em incrementar o valor de *Qtd* para as transações fraudulentas.
- Variações possíveis
  - aplicar *oversampling* nos casos de *fraude* e, em seguida, aplicar *undersampling* nos casos de *não fraude*.
  - Baseline
    - **limitar os valores do atributo *Qtd* do conjunto de treinamento a um valor limite**
      - depois pode-se ou não aplicar as técnicas de amostragens citadas.

5. Estudo de Casos – Fraudes em Cartões de Crédito

Pré-processamento dos Dados

- Amostragens do conjunto de treinamento com distribuições variadas de casos de fraude em relação ao total de casos.
  - *oversampling*
    - **construídas cinco amostras com as seguintes distribuições de fraudes: 10%, 20%, 30%, 40% e 50%**
  - *oversampling* seguida de *undersampling*.
    - **construídas cinco amostras com as seguintes distibuições de fraudes: 10%, 20%, 30%, 40% e 50%.**
  - *baseline* com *Qtd* limitada a 200 (para não fraude)
    - **resultou em 175.936 casos de não fraude e 1.477 transações fraudulentas.**
      - Houve uma redução de 1.062.772 casos de não fraude.
    - **a partir desta amostra, outras dez amostras foram construídas utilizando *oversampling* e "*oversampling* seguida de *undersamplig*"**

5. Estudo de Casos – Fraudes em Cartões de Crédito

Modelagem

Performance S.E de Classificador *Naive Bayes* para a Classe Fraude

Porcentagem de fraudes	Sem *baseline*				*Baseline* (*Qtd* limitada a 200)			
	Oversampling		Oversampling undersampling		Oversampling		Oversampling Undersampling	
	N	**PR**	**N**	**PR**	**N**	**PR**	**N**	**PR**
	1	2	3	4	5	6	7	8
Original	**0,16**	**0,71**	**0,16**	**0,71**	**0,16**	**0,75**	**0,16**	**0,75**
10%	**0,48**	**0,71**	**0,47**	**0,71**	**0,25**	**0,75**	**0,24**	**0,75**
20%	**0,58**	**0,71**	**0,47**	**0,71**	**0,43**	**0,75**	**0,41**	**0,75**
30%	**0,63**	**0,71**	**0,63**	**0,71**	**0,57**	**0,75**	**0,57**	**0,75**
40%	**0,67**	**0,71**	**0,67**	**0,71**	**0,69**	**0,75**	**0,68**	**0,73**
50%	**0,71**	**0,71**	**0,71**	**0,71**	**0,75**	**0,75**	**0,74**	**0,71**

5. Estudo de Casos – Fraudes em Cartões de Crédito
Modelagem

- Colunas PR (determinação da classe com uso de relações entre probabilidades)
  - dada a evidência E, a classe dominante é determinada como:

    $\textbf{arg-max}_{\{\text{fraude=sim, fraude=não}\}}\textbf{p(}\textit{fraude}\textbf{|}E\textbf{)/p(}\textit{fraude}\textbf{),}$
  - onde p(*fraude*) é uma probabilidade a priori de fraude.
- Colunas N (determinação da classe com o uso de probabilidades)
  - a classe dominante é determinada como:

    $\text{arg-max}_{\{\text{fraude=sim, fraude=não}\}}\text{p}(\textit{fraude}|E).$
- Linha *Original*
  - modelos gerados a partir do conjunto de treinamento original e da amostra deste conjunto com *Qtd* limitada a 200.

Como determinar qual a classe se o classificador é probabilístico?

5. Estudo de Casos – Fraudes em Cartões de Crédito

Avaliação

- O ganho máximo no desempenho
  - 7% em relação ao índice S.E original
    - **amostra com *baseline* (S.E = 0,75%).**
  - Linhas *40*% e *50*% da coluna 8 destoam dos outros resultados desta coluna devido
    - ***undersampling* pode ter reduzido de forma excessiva a variável *Qtd* de alguns casos de não fraude que já haviam sido reduzidas a 200 repetições (com *baseline*).**

5. Estudo de Casos – Fraudes em Cartões de Crédito
Análise dos Resultados e Conclusões

- Há aumento de S.E apenas quando as técnicas de amostragem foram combinadas com *baseline*.
  - mesmo que se tenha um conjunto de treinamento balanceado, casos que ocorrem com grande freqüência (como o caso de não fraude que repete 43.318 vezes) degradam o desempenho de modelos gerados com tais dados.

6. Conclusões

- Navegar é preciso* minerar não é preciso ...
  - como para sambar é necessário inspiração e suor!
- Minerar requer experimentação e uso de bom senso
  - **não existe uma receita de bolo a se seguir**
    - Cada caso é um caso!
  - **requer bom conhecimento de Estatística e do domínio da aplicação!**
- Base de dados desbalanceadas requerem tratamento especulativo (muita experimentação)

Preciso no sentido de exato e não de necessário!

6. Conclusões

- A mineração somente atinge sucesso quando os conhecimentos descobertos são úteis do ponto de vista de uso no negócio
- Há uma tendência das grandes empresas de desenvolvimento de DBMS (em especial a IBM e a Oracle) de integrarem seus produtos com suas ferramentas de DM
  - BI – Business Intelligence
    - **mercado muito promissor para informatas porque muitas empresas compram produtos de DM e não sabem interpretar os resultados obtidos com o uso deles ...**