



**Protótipo de um ambiente de aprendizado  
de máquina, utilizando a técnica de  
*Knowledge Discovery in Databases (KDD)*  
para reconhecimento de padrões em  
ocorrências atendidas pela Polícia Militar**

EMERSON TENFEN

Prof. Jomi Fred Hubner

Orientador



# Roteiro da Apresentação

- Objetivos do Trabalho
- KDD
- Mineração de Dados (Data Mining)
- Árvores de Decisão
- Algoritmo C4.5
- Atributos numéricos
- Protótipo
- Conclusões
- Extensões



# Objetivos do Trabalho

- Criação do Protótipo
- Aplicação de técnicas
  - KDD
  - Classificação
  - Árvores de decisão
  - C4.5



## **KDD – *Knowledge Discovery in Databases***

- **“Processo não trivial de identificação de padrões em dados que sejam válidos, novos, potencialmente úteis e compreensíveis.” (FAYYAD, 1996)**



# KDD – *Knowledge Discovery in Databases*

- **Processos de KDD:**

- Desenvolvimento e entendimento do domínio da aplicação
- Criação de um grupo de dados alvo
- Limpeza dos dados
- Redução e projeção dos dados
- Mineração
- Interpretação
- Consolidação do conhecimento descoberto



# MD - Mineração de Dados

- Técnica que permite extrair conhecimento de uma massa de dados que, de outra maneira, permaneceria escondido nas grandes bases de dados (QUONIAM, 2002).



# Mineração de Dados

- **Tarefas de MD:**

- Estimativa;
- Segmentação;
- Associação;
- Classificação.



# Árvores de Decisão

- Ferramentas analíticas usadas para descobrir automaticamente regras e relacionamento entre dados, subdividindo a informação em subconjuntos (BERRY, 1998).





# Árvores de Decisão

- **Etapas da construção de uma Árvore de Decisão:**
  - 1 - Escolher um atributo;
  - 2 - Estender a árvore adicionando um ramo para cada valor do atributo;
  - 3 - Passar exemplos para as folhas;
  - 4 - Para cada folha:
    - 4.1 - Se todos os exemplos são da mesma classe, associar essa classe à folha;
    - 4.2 - Se não, repetir os passos 1 a 4.



# Árvores de Decisão

- Entropia

- Medida da impureza do conjunto de dados.
- Usada para estimar a aleatoriedade da variável alvo.



# Árvores de Decisão

- Ganho de informação;
  - Usado para escolher o melhor atributo, de forma a diminuir o número de nodos da árvore



# Algoritmo C4.5

**FormaArvore(T)**

**(1) CalculaEntropia(T);**

**(2) Se todos os exemplos são de uma mesma classe**

**Retorna uma folha com o nome da classe;**

**Cria um nodo Raiz;**

**(3) Para cada atributo A**

**Calcula Ganho de informação(A);**

**(4) Raiz = Atributo com maior ganho de informação;**

**(5) Se Atributo = Continuo**

**EncontraPontoReferencia;**

**(6) Para Cada T' possível em T**

**(7) Se T' é vazio**

**Retorna folha com valor mais comum**

**Else**

**(8) AdicionaArvore = FormaArvore(T');**



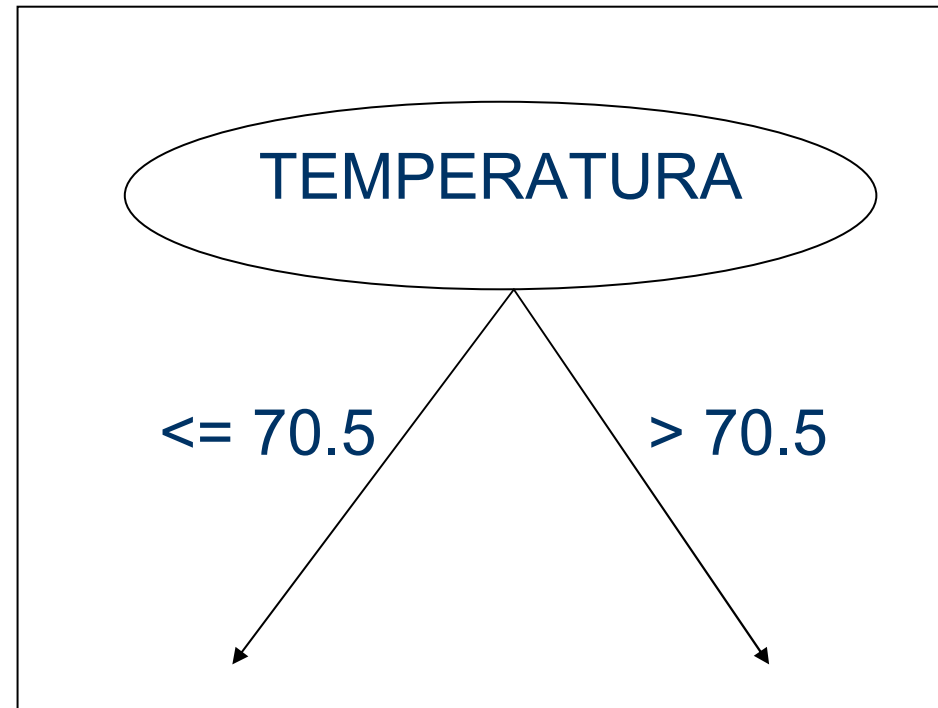
## Atributos Numéricos

| TEMPERATURA | JOGA |
|-------------|------|
| 64          | Sim  |
| 65          | Não  |
| 68          | Sim  |
| 69          | Não  |
| 70          | Sim  |
| 71          | Não  |

- Para cada valor único
  - Calcular  $V_{(i)} + V_{(i+1)} / 2$
- Para cada Ponto
  - Calcular ganho

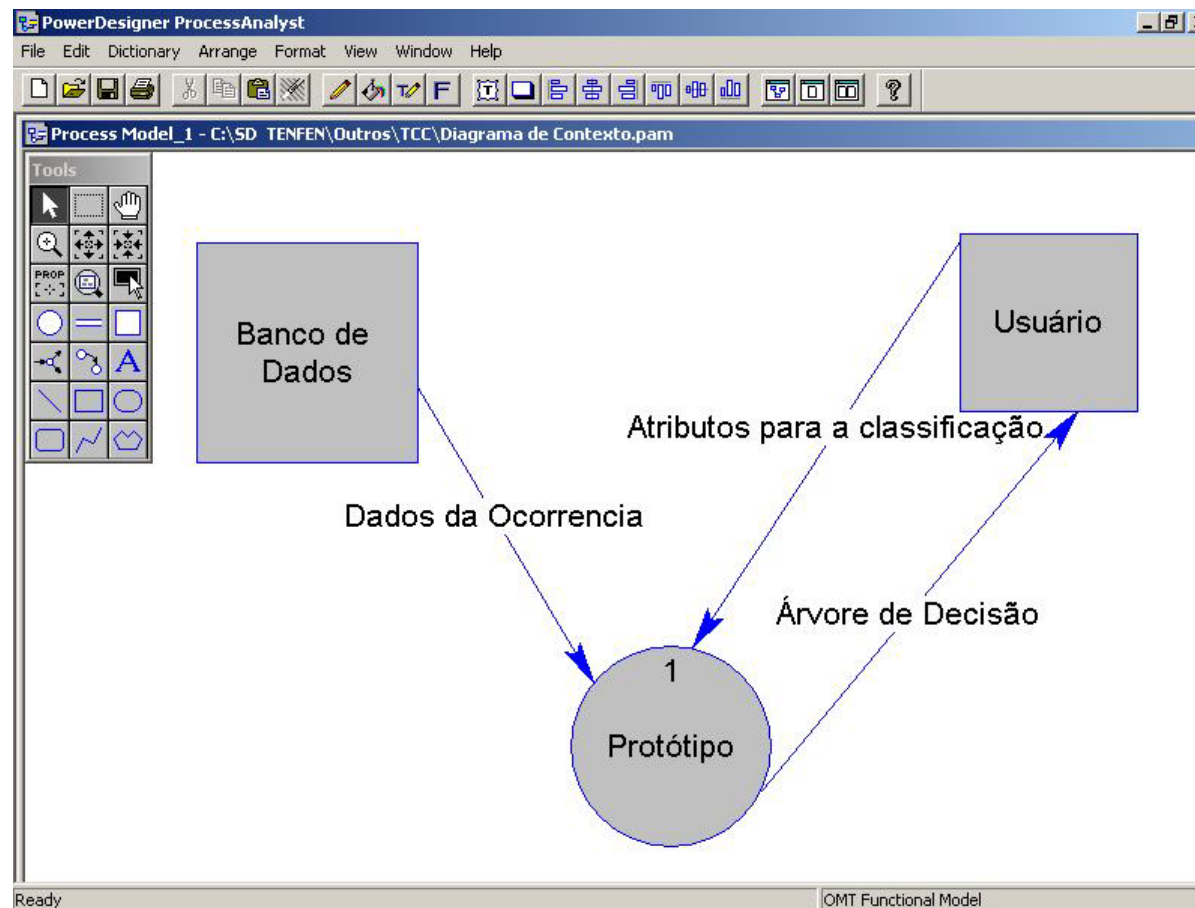


# Atributos Numéricos



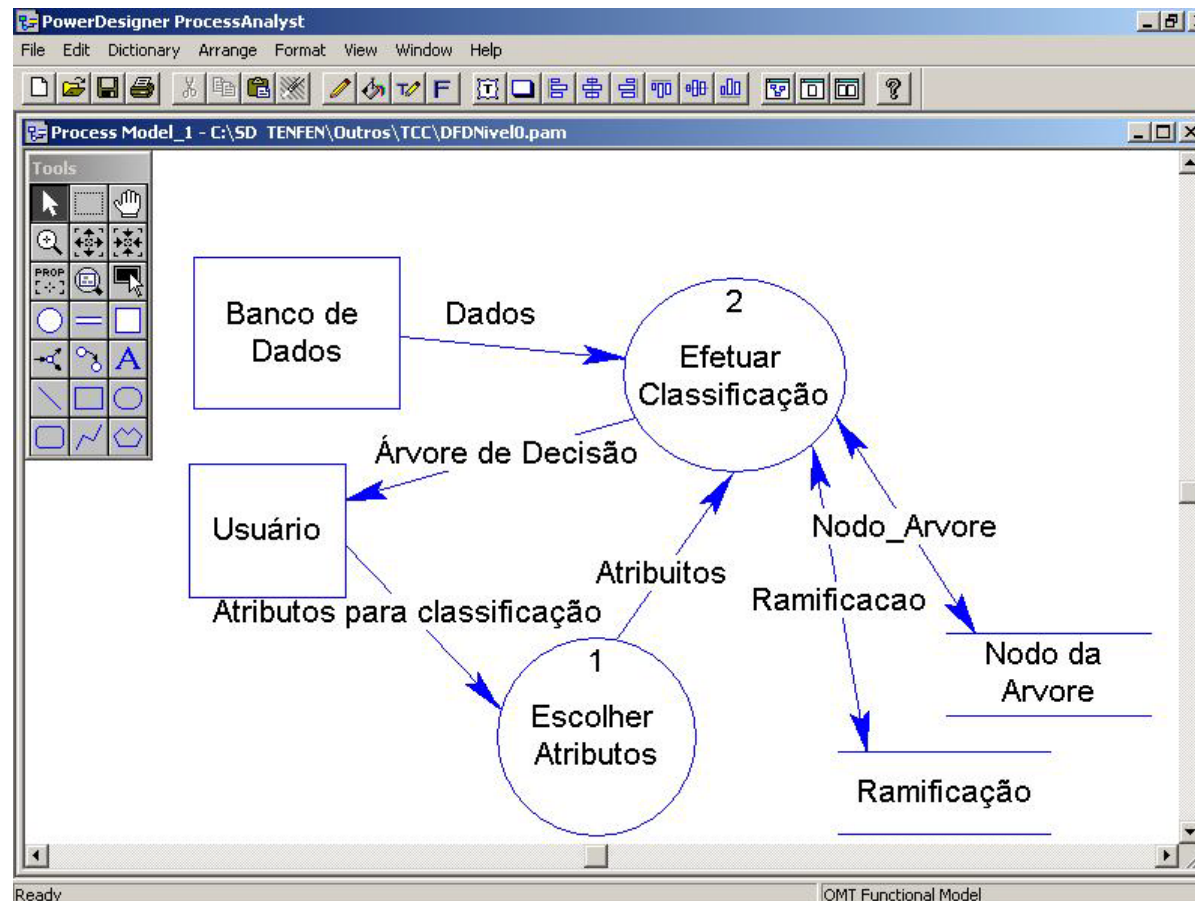


# Protótipo – Diagrama de contexto





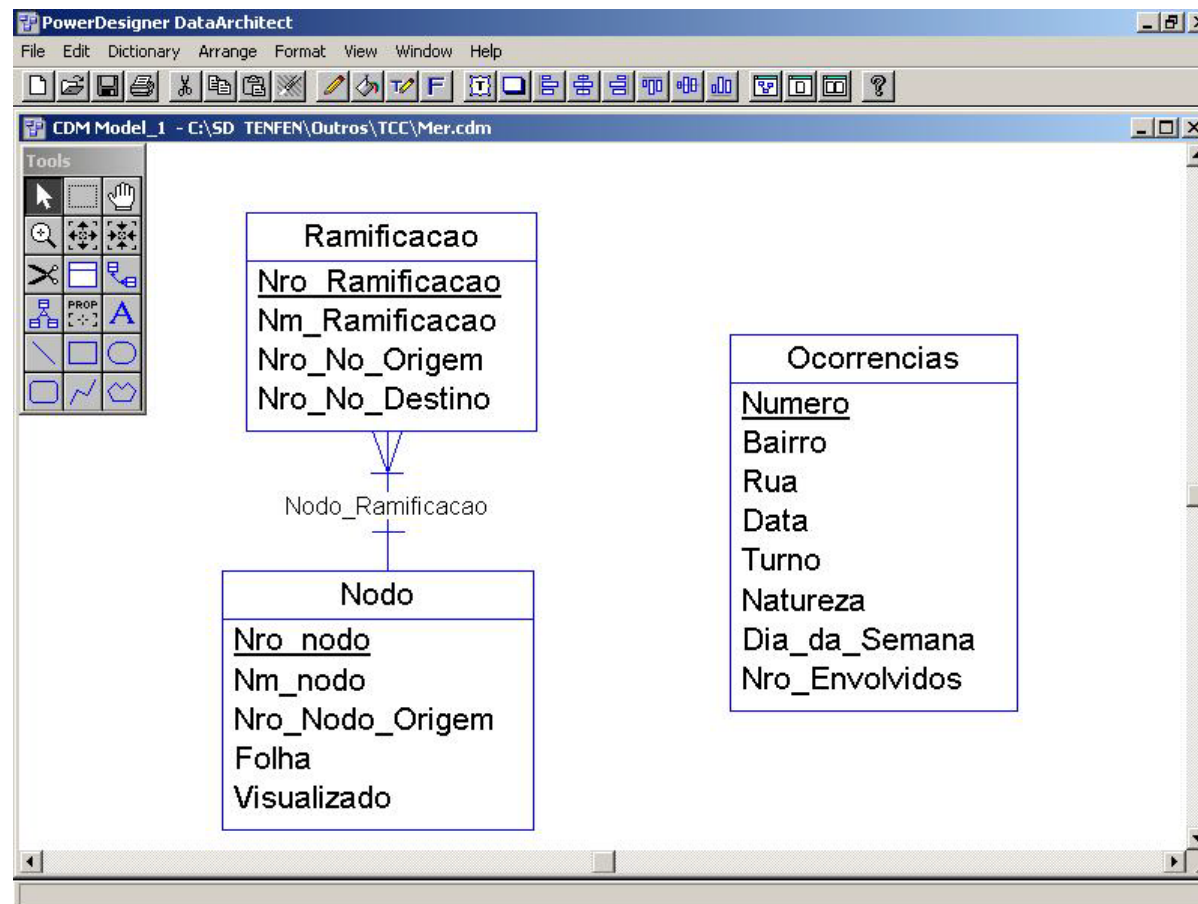
# Protótipo – DFD Nível 0







# Protótipo – MER





## Conclusões

- Eficiência da tarefa de classificação para Mineração de Dados;
- Objetivo principal do trabalho foi atingido.



# Extensões

- Estudo de MD com aplicação de outras tarefas;
- Desenvolvimento de protótipo com base de dados variável;
- Possibilidade de escolha do atributo alvo.