MODELAGEM E PREPARAÇÃO DE DADOS PARA APRENDIZADO DE MÁQUINA: Entendimento do domínio de problema

Professor:

Luis E. Zárate

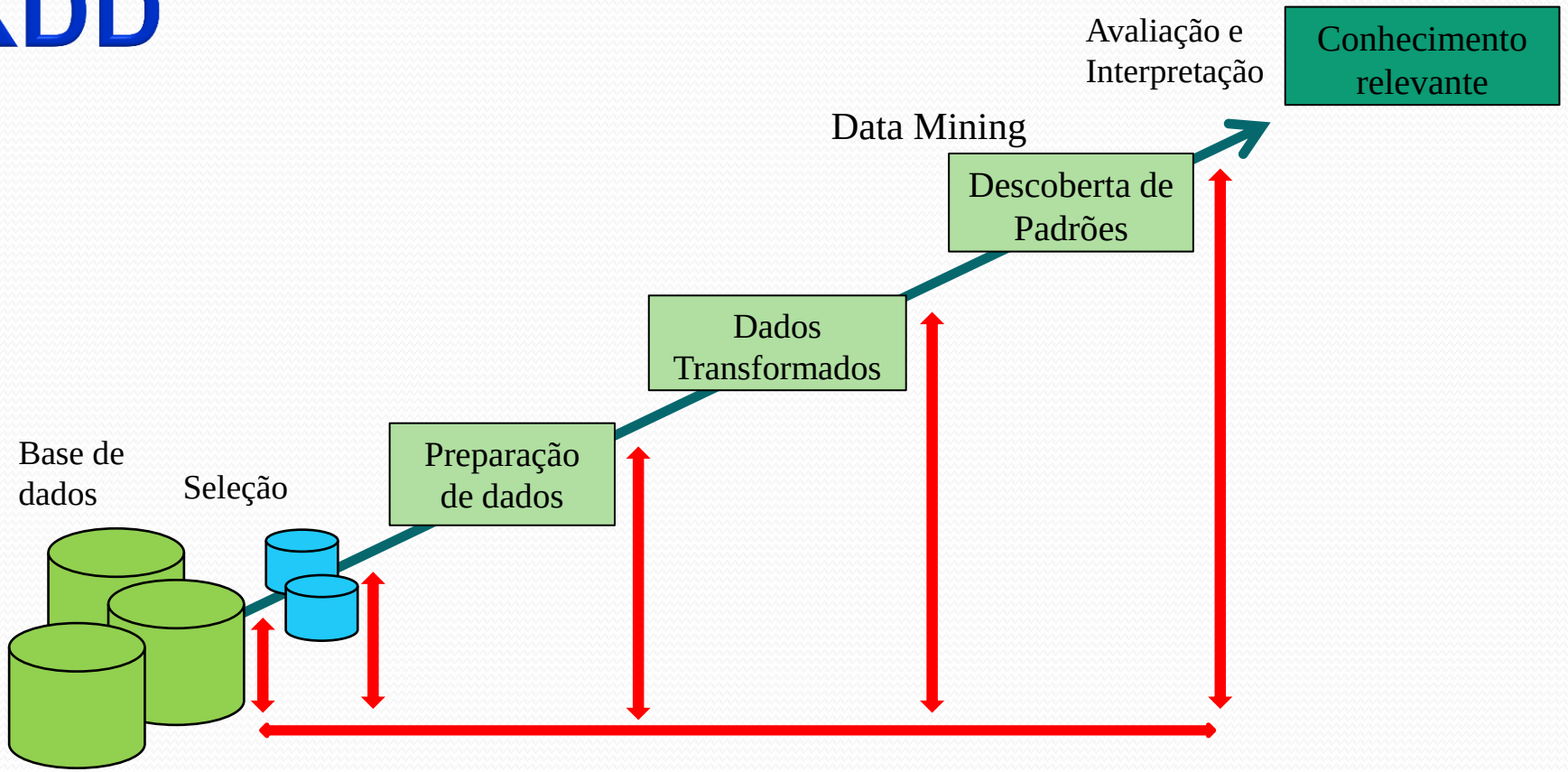
O Processo KDD

O que é Descoberta de Conhecimento em Base de Dados (KDD)?

"Processo não-trivial de identificação de padrões válidos, novos, potencialmente úteis e finalmente compreensíveis a partir de dados"

(Usama Fayyad)

Processo de Descoberta de conhecimento em Base de Dados - KDD



Etapas de um processo KDD - Adaptado de Fayyad et al (1996)

Detalhamento do Processo KDD

- Entendimento e modelagem do domínio de problema
- Montagem da base de dados
- Enriquecimento e Melhoramento da base de dados
- Limpeza de dados
- Análise de Outliers e de dados aussentes
- Integração e combinação de dados
- Discretização, Codificação ou Transformação
- Data Mining
- Validação de Padrões
- Visualização e Apresentação do Conhecimento

Entendimento do Domínio de Problema

- O cientista de dados, com auxílio do especialista de domínio, deverá entender o domínio do problema e caracterizá-lo utilizando modelos de ontologia ou mapas conceituais.

- O objetivo desta etapa é identificar as características (atributos) que possam enriquecer a base de dados levando a conhecimento não óbvio e útil.

- A experiência mostra que o conhecimento não óbvio é resultado muitas vezes de características consideradas ‘julgamentos’.

- A descoberta de conhecimento em banco de dados baseado unicamente em informações de ‘fatos’ pode não levar a conhecimento relevante.

  - Por exemplo, consideremos a busca de padrões para os acidentes de trânsitos. Dar atenção somente aos **fatos** <imprudência, efeito do álcool, velocidade excessiva e falha mecânica> não traz conhecimento útil. Variáveis de **Julgamento:** <perfil dos condutores, perfil dos acompanhantes, etc> pode enriquecer a base de dados trazendo conhecimento não óbvio.

Técnicas para a Entendimento de Problemas

- Re-definição precisa do problema
- Observação de Fatos e Julgamentos
- Análise Divergente - Convergente
- Pró-Contra-e-Fixação
- Mapas Cognitivos
- Resolução de Ambigüidades
- Diagramas de Causa-Efeito
- Mapas conceituais

1) Re-definição precisa do Problema

- Para a definição precisa do problema é necessário procurar todas as informações que podem ser úteis.

- Exemplo: Falha na linha de produção
- Informações adicionais

  - Elementos constituem a falha
  - Situações onde a falha é detectada
  - Componentes da falha devem ser observados: equipamentos, pessoal, meio ambiente, etc.

2) Identificando Fatos e Julgamentos

- **Fatos:**
  - Circunstância que causam diretamente o problema
- **Julgamentos:**
  - Observações a serem disputados ou decididos

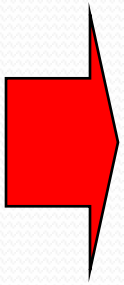
Exemplo 1- Acidente de tránsito

Imprudência

Efeito de álcool

Excessiva velocidade

Falha mecânica



Qual é o perfil do motorista de cada veículo?

Quais são os perfis dos acompanhantes dos veículos?

Quais eram as condições meteorológicas durante o acidente?

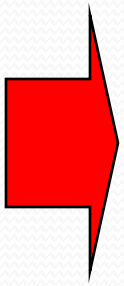
Exemplo 2- Devolução de Produtos

Defeito na envoltura

Fora da validade

Produto estragado

Fora de especificação



Quem fabrica a envoltura?

Quem comercializa o produto?

Como o produto foi estocado?

Como o produto é embalado?

Como o produto é transportado?

- **Confiança Analítica:**

  - Os problemas podem ser categorizados pelo papel que os **FATOS** e os **JULGAMENTOS** têm haver na análise do problema.

FATOS		JULGAMENTOS	
Confiança cai			**Erro**
Simplista	**Determinística**	**Aleatória**	**Indeterminado**
Existe somente uma resposta	Somente uma resposta e deve ser usada a formula correta	Várias respostas e todas podem ser identificadas	Várias respostas suposições e não todas podem ser identificadas

3) Análise Divergente / Convergente

- **Divergência:**
  - Consiste em direcionar nossa mente em diferentes direções de um simples ponto procurando novas evidências.
- **Convergência:**
  - Consiste em direcionar a nossa atenção, focalizando nossa mente sobre um simples aspecto do problema.

- Ambas são necessárias para a efetiva solução de problemas. A divergência abre a mente para criar alternativas e a convergência peneira alternativas fracas e fortalece as fortes.

- **Passos:**
  - **1) Divergente:** Brainstorm
  - **2) Convergente:** Examinar e agrupar cada idéia
  - **3) Divergente:** Voltar ao ***passo 1*** para cada grupo de idéias.

- **Exemplo:** Problemas na produção de soja
  - **Divergente:** Problemas na qualidade da semente, no plantio, no cultivo, no clima, na maquinaria utilizada, etc.
  - **Convergente:** Problema na qualidade da semente
  - **Divergente:** melhoramento genético da semente, tecnica de preparação da semente, estocagem da semente, etc.

4) Pró-contra-e-fixação

- Esta técnica é fundamentada na compulsão humana de ser críticos, especialmente quando algo é novo ou fora do convencional.

- **Passos:**
  - **1) Listar todos os prós:** aspectos positivos, benefícios, méritos e vantagens.
  - **2) Listar todos os contrários:** razões que não permitem que determinada situação aconteça (pensamento divergente).
  - **3) Revisar e consolidar os contrários:** pensamento convergente.
  - **4) Neutralizar aspectos contrários quanto possíveis**

- **Exemplo:** O que leva à compra de um veículo

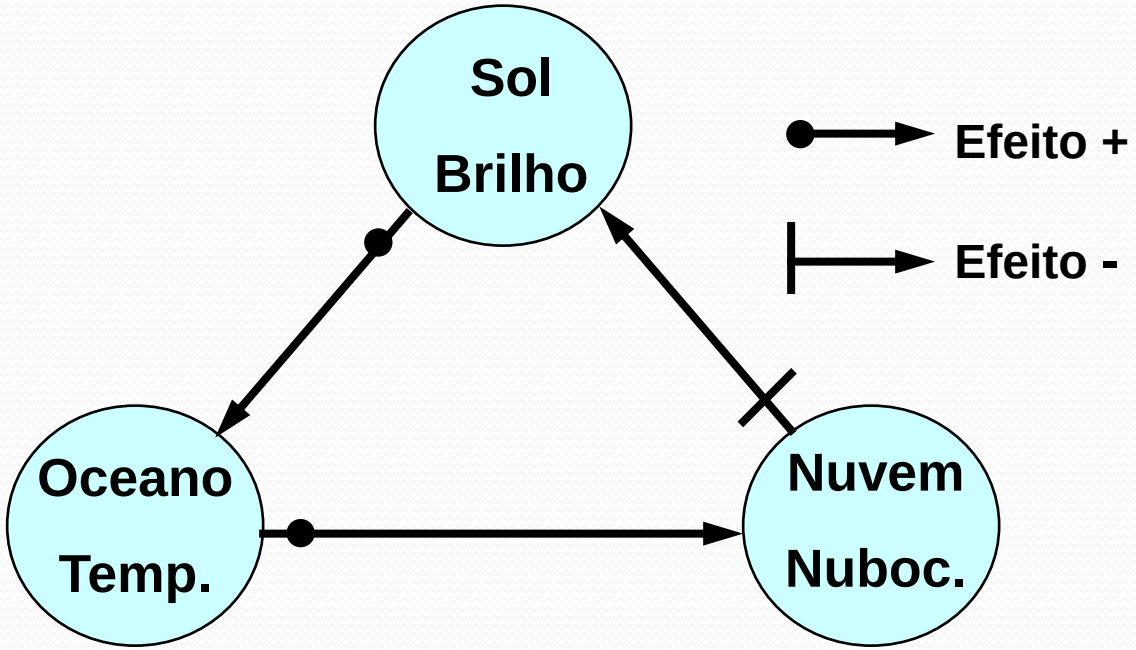
  - **1) Listar todos os prós:** preço, promoção, consumo de combustível, revenda, conforto, segurança, etc..

  - **2) Listar todos os contrários:** importado, visado, raro serviço de oficina, manutenção cara, mecânica especializada, etc..

  - **3) Revisar e consolidar os contrários:** importado, **visado**, raro serviço de oficina, **manutenção cara**, mecânica especializada, etc..

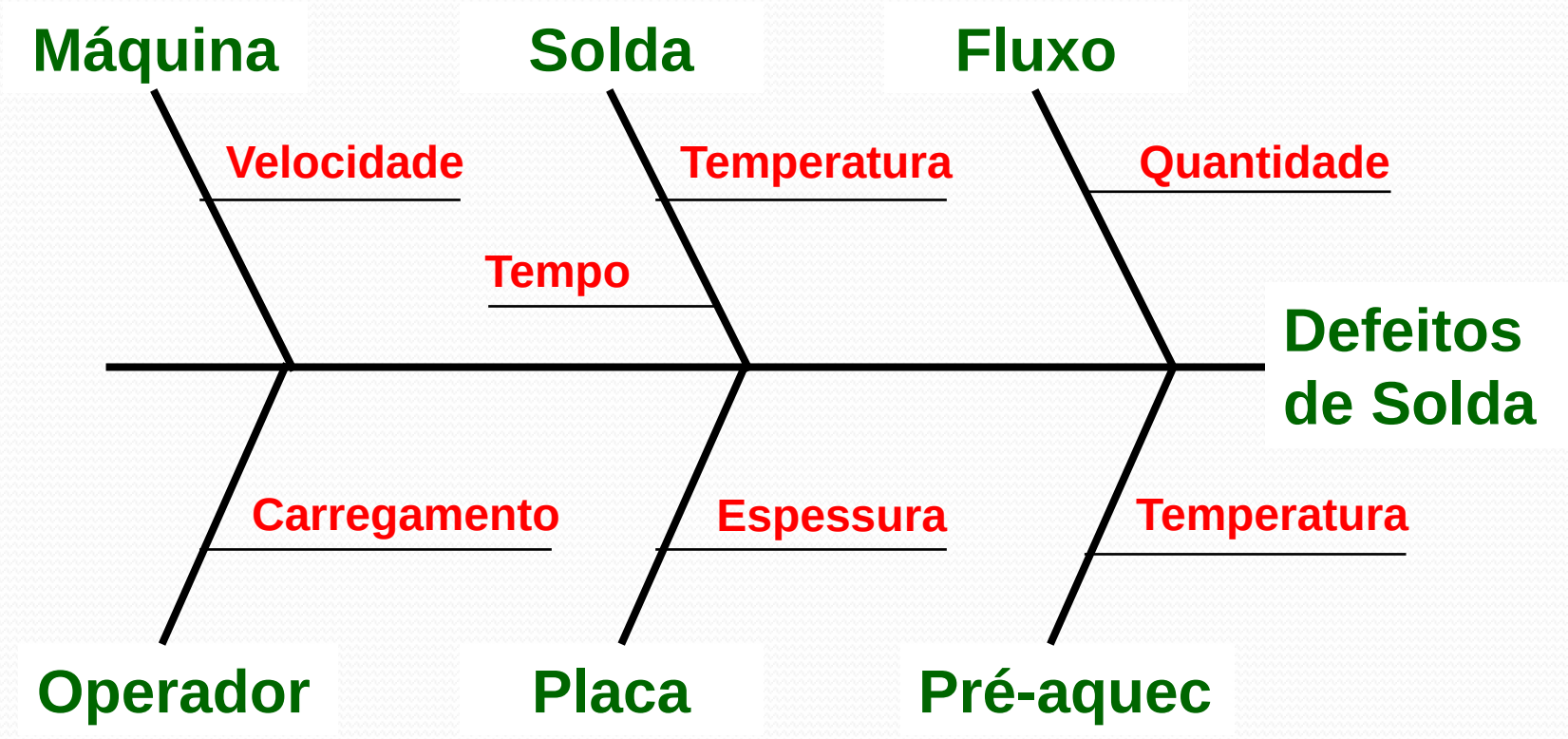
5) Mapas Cognitivos

- São usados quando os problemas são difíceis de serem entendidos ou alguns fatos são difíceis de serem inseridos na estrutura do problema.

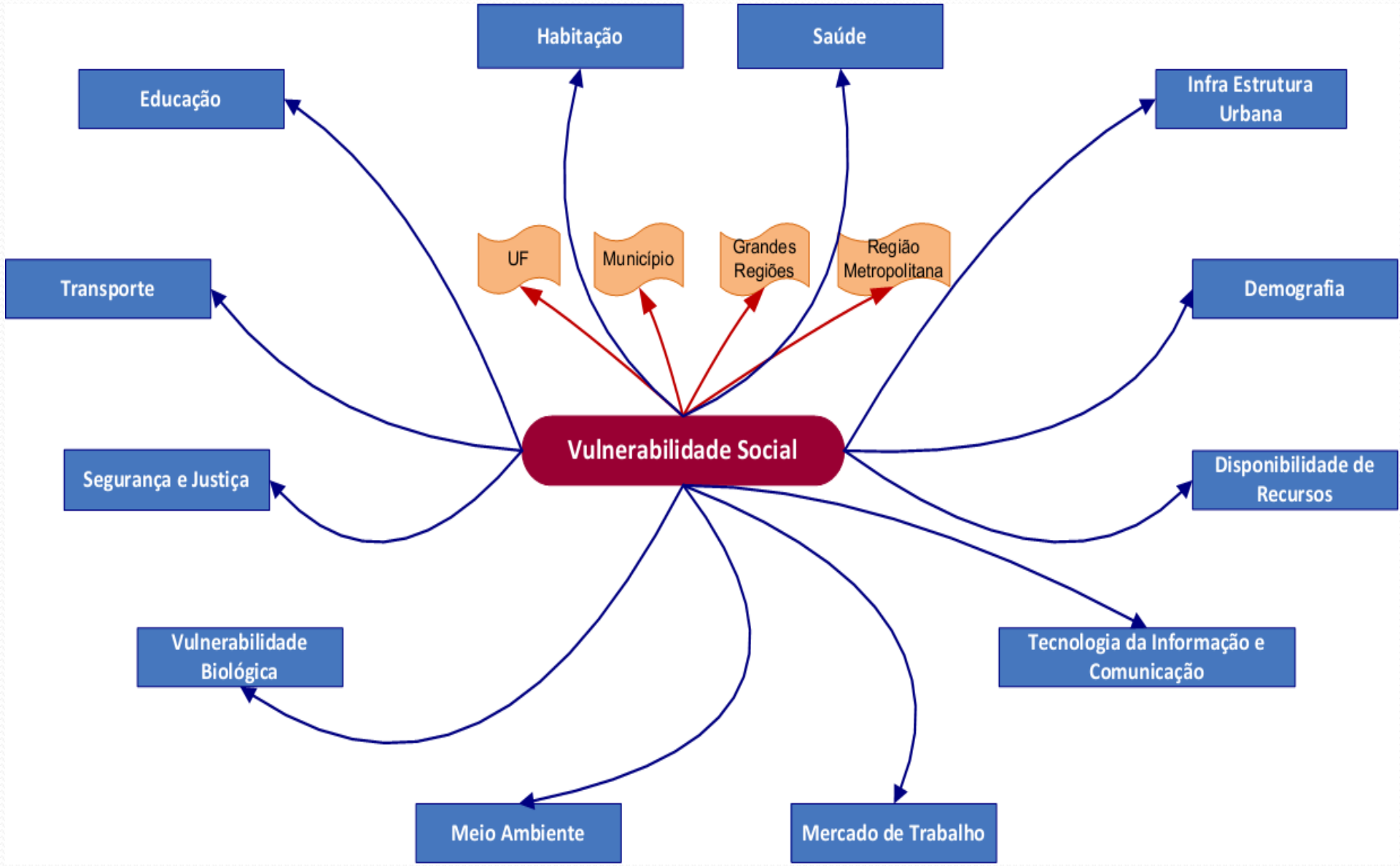


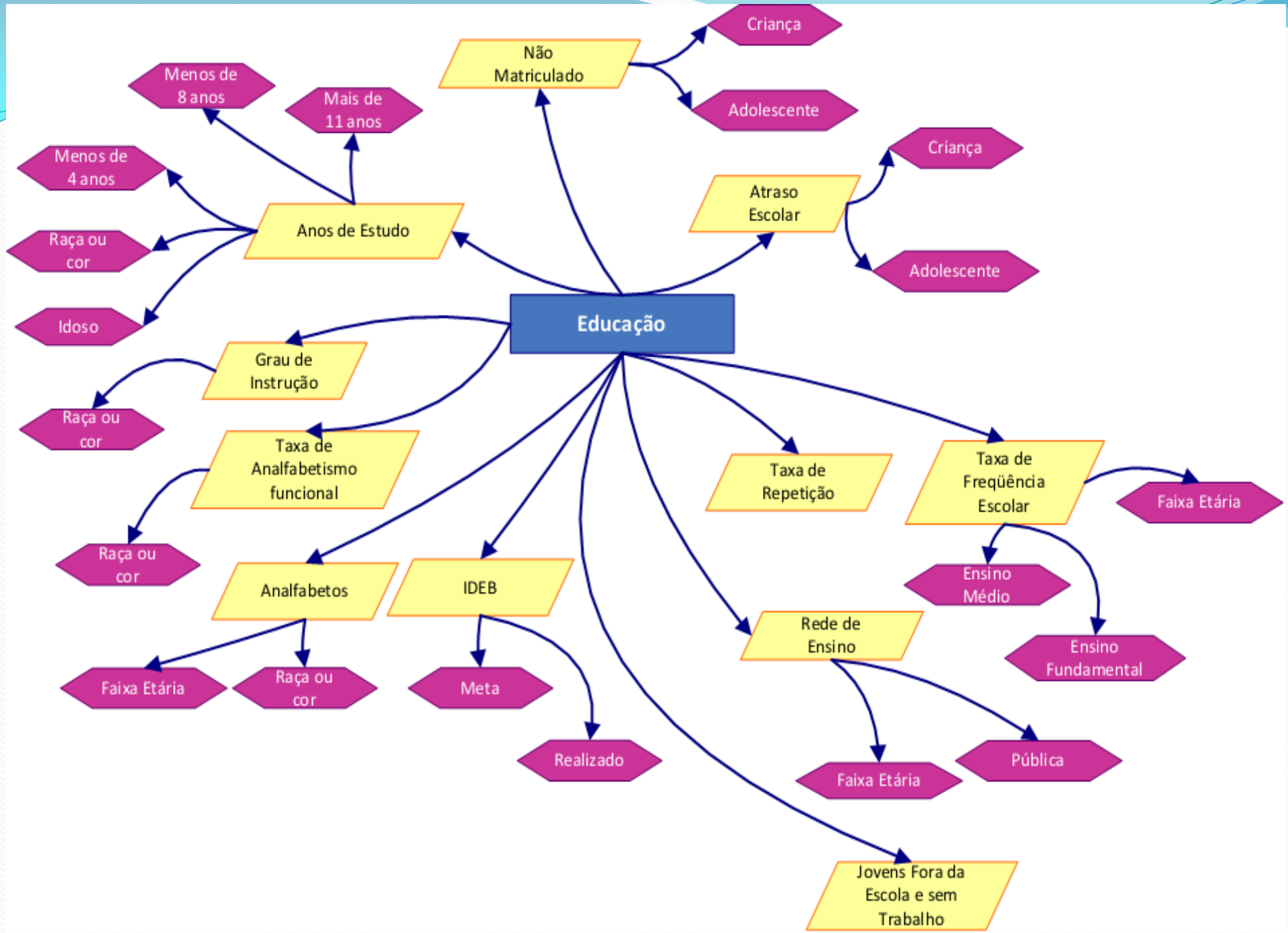
6) Diagramas de Causa-Efeito

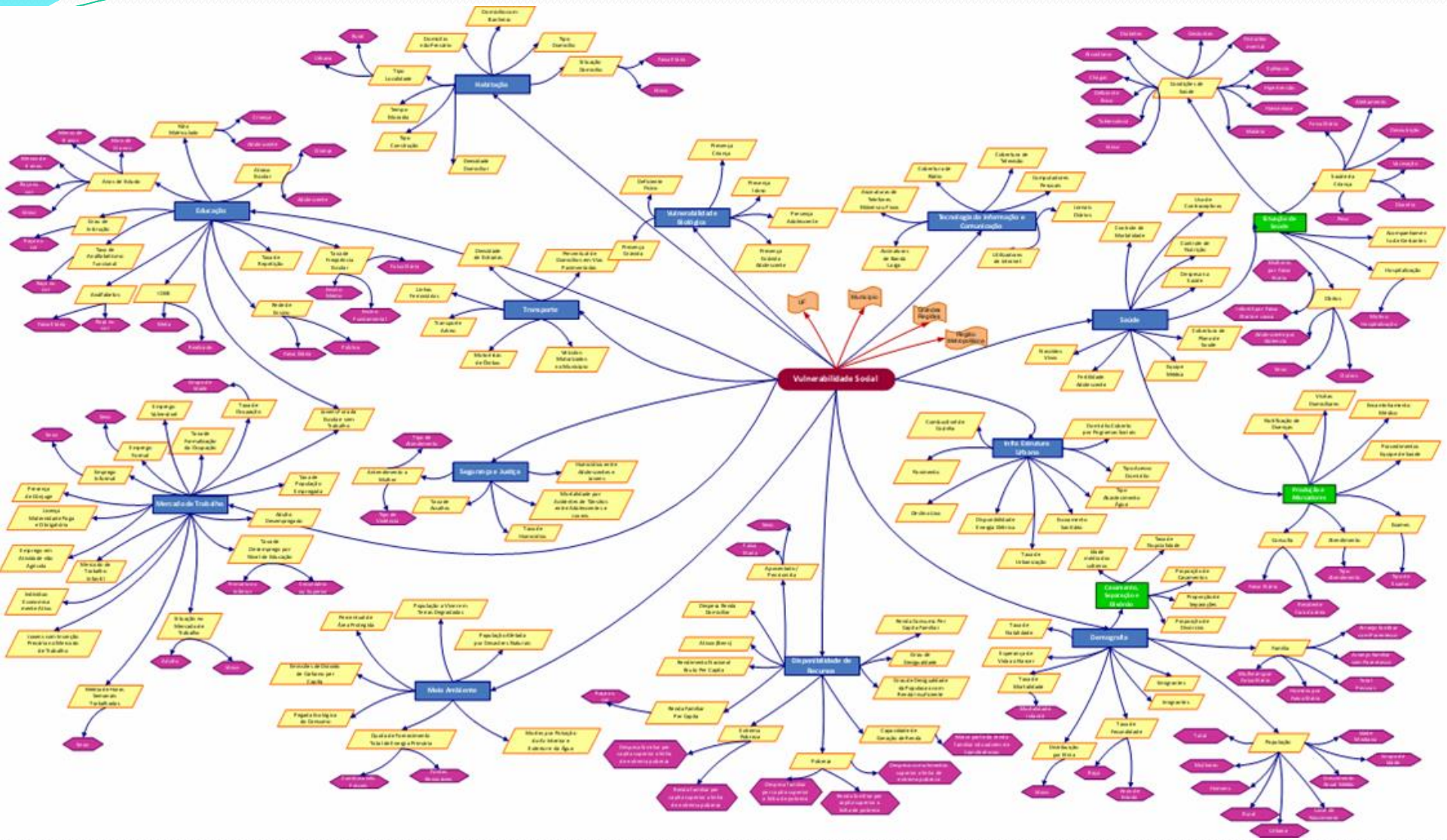
Diagrama de Ishikawa o de causa-efeito



7) Mapa conceitual




Educação
Não Matriculado
Criança
Adolescente
Anos de Estudo
Menos de 8 anos
Mais de 11 anos
Menos de 4 anos
Raça ou cor
Idoso
Atraso Escolar
Criança
Adolescente
Grau de Instrução
Raça ou cor
Taxa de Analfabetismo funcional
Raça ou cor
Analfabetos
Faixa Etária
Raça ou cor
IDEB
Meta
Realizado
Taxa de Repetição
Taxa de Freqüência Escolar
Faixa Etária
Ensino Médio
Ensino Fundamental
Rede de Ensino
Faixa Etária
Pública
Jovens Fora da Escola e sem Trabalho


Vulnerabilidade Social
UF
Município
Grandes Regiões
Regiões Metropolitanas
Habitação
Educação
Vulnerabilidade Biológica
Tecnologia da Informação e Comunicação
Saúde
Transporte
Segurança e Justiça
Infra Estrutura Urbana
Mercado de Trabalho
Meio Ambiente
Disponibilidade de Recursos
Demografia
Condições Sanitárias e Ambiental
Produção e Rendimentos

Prática 1 : Considere os seguintes domínios de problemas:

a) Condições ambientais que podem levar a uma pessoa a ter um AVC.

b) Perfil de clientes com tendência a abandonar um sistema de Marketplace.

– Para os domínios utilize a técnica dos mapas conceituais para representar as principais variáveis que podem ser consideradas.

Caracterização do Problema por meio de atributos

- Após o entendimento do domínio do problema, das expectativas e resultados esperados, é necessário identificar as variáveis (futuros atributos) relevantes para compor a base de dados.

- Cada atributo deve ser avaliado pelo analista de domínio e cientista de dados e selecionado de acordo com a sua relevância em relação ao problema. Este procedimento é chamado Seleção Conceitual

- Será documentado cada atributo, seu tipo, faixa de valor e a sua relevância (Fato ou Julgamento) esperada em relação ao domínio do problema.

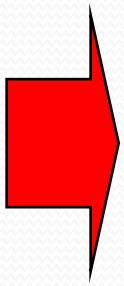
Exemplo 1- Acidente de tránsito

Imprudência

Efeito de álcool

Excessiva velocidade

Falha mecânica



Qual é o perfil do motorista de cada veículo?

Quais são os perfis dos acompanhantes dos veículos?

Quais eram as condições climáticas durante o acidente?

Atributos: **{**

FATOS

- Dados pessoais: Idade, sexo, ocupação dos envolvidos;
- Tipo de manobra prévia ao acidente dos envolvidos;
- Manobra declarada por testemunhas para os envolvidos;
- Nível alcoólico dos envolvidos;
- Velocidade no momento do acidente dos veículos;
- Tipo de falha mecânica dos veículos;
- Tempo de carteira dos envolvidos;

JULGAMENTOS

- Dados pessoais dos acompanhantes: idade, sexo, ocupação;
- Condições climáticas durante o acidente;
- Pontos na carteira dos motoristas; .....

...}

Documentando os atributos:

Atributos: **{**

Dados pessoais: Idade, sexo, ocupação dos envolvidos;
Tipo de manobra prévia ao acidente dos envolvidos;

FATOS Manobra declarada por testemunhas para os envolvidos;
Nível alcoólico dos envolvidos;
Velocidade no momento do acidente dos veículos;
Tipo de falha mecânica dos veículos;
Tempo de carteira dos envolvidos;

JULGAMENTOS Dados pessoais dos acompanhantes: idade, sexo, ocupação;
Condições climáticas durante o acidente;
Pontos na carteira dos motoristas; .....

...}

Atributo	Tipo	Faixa	Relevância
Idade	Numérico	21 a 60	Fato
Sexo	Categórico	M,F	Fato
Ocupação	Categórico	Profissional/ Desocupado/ Estudante/	Julgamento

Obrigado

Professor:

Luis E. Zárate