# Mineração de Dados e Descoberta de Conhecimento

*Profs. Heitor S. Lopes e Thiago H. Silva (UTFPR, 2023) - Exercício #5*

## A) Análise Associativa (1)

**Objetivo**: O dataset **bebê-baixo-peso** foi construído com base no *Very Low Weigth Infants Dataset* e registra dados de 671 bebês com peso ao nascer abaixo de 1600 gramas. Há dados sobre a mãe e o bebê incluindo alguns medicamentos utilizados, questões relacionadas ao parto, problemas diagnosticados no nascituro, e se o mesmo sobreviveu ou não ao parto. Na literatura médica, já são conhecidos alguns fatores que diminuem as chances de sobrevida de bebês prematuros, por exemplo: bebês com extremo baixo peso (<1000 gramas), aqueles nascidos muito prematuramente (<27 semanas de idade gestacional), e presença de comorbidades tais como pneumotórax ou complicações cardio-pulmonares resultantes da própria imaturidade pulmonar. Deseja-se, portanto, utilizar os métodos de regras de associação para descobrir associações relevantes que **não sejam óbvias**.

a. Caso seja utilizado o Orange, é necessário reformatar o dataset para o formato "Basket" (mais informacões aqui e aqui). Caso escolha o software Weka, utilize o algoritmo de associação *hotSpot* (instalável através do *Package Manager*), com *support*=1 (ou muito próximo deste valor), *outputRules=True*, *maxBranchingFactor*<5 e ajuste o consequente da regra para a variável *Dead* através da opção *target* e *targetIndex* (=1 ou 2 para *Died=no* ou *Died=yes*). Opcionalmente, pode ser utilizado Python.

b. Em vez de regras genéricas, há interesse especial em descobrir regras que mostrem alguma associação entre o uso de *beta-methasone* pela gestante e a sobrevivência, ou não, do bebê-baixo-peso. Observe as regras encontradas pelo algoritmo e investigue se o resultado tem algum fundamento (Pesquise no Google!!).

c. Fato importante: este dataset foi coletado na *Duke University Medical Center* no Estado da Georgia (USA), sendo este um dos estados americanos onde há uma grande concentração de afro-descendentes. Com isto em mente, obtenha regras de associação (com pelo menos 20% de suporte) que envolvam uma eventual relação entre raça e a prevalência, ou não, de morte de bebês-baixo-peso). Discuta as implicações sociais dos achados.

# B) Análise Associativa (II) DESAFIO!

**Objetivo**: O objetivo é utilizar dados reais sobre acidentes nas rodovias federais (fornecido pela Polícia Rodoviária Federal – PRF) que cortam o Estado do Paraná, para encontrar relacionamentos entre os atributos de modo a explorar o *dataset* e obter conhecimentos **não-óbvios** no que diz respeito a acidentes **com vítimas fatais**.

a. Importe os dados no Orange, selecione **apenas** acidentes com vítimas fatais. Mostre a visualização dos pontos de acidentes no mapa do Estado do Paraná, utilizando as coordenadas geográficas dos acidentes. Isto é feito com o *gadget Geo Map*, sendo a cor dos pontos dada pela variável "Causa Principal". Com a análise gráfica do mapa, informe quais as **causas** mais frequentes de acidentes com vítimas fatais nas três maiores regiões metropolitanas do Estado do Paraná (Curitiba, Londrina, Maringá). Compare (qualitativamente) com as causas de acidentes nas rodovias, fora das regiões metropolitanas. Quais as principais diferenças?

b. Selecione apenas as variáveis **qualitativas** e busque gerar *itemsets* frequentes com o mais alto suporte e que contenha: Classificação do acidente='Com vítimas fatais'

c. A partir dos *itemsets*, procure regras também com suporte mínimo de **40%** e confiança **>= 90%.** Destas regras, reitera-se que só interessam aquelas que tenham o consequente Classificação do acidente='Com vítimas fatais'

d. Analise os resultados obtidos e informe os conhecimentos **não-óbvios** obtidos com a análise, para as regiões metropolitanas e para as regiões fora delas.