

II SIMPÓSIO DE PESQUISA E EXTENSÃO DE CERES E VALE DE SÃO PATRÍCIO
04 a 07 de Novembro de 2014 - UEG Campus Ceres - GO

(SAÚDE, TECNOLOGIA DA INFORMAÇÃO)

**TÉCNICAS DE EXTRAÇÃO DE CONHECIMENTO POR MEIO DE DADOS
FALTANTES E MINERAÇÃO DE DADOS SOBRE AS VÍTIMAS DO CÉSIO-137: USO DE
IMPUTAÇÃO MÚLTIPLA PARA OBTENÇÃO DE MODELOS ACURADOS.**

¹Danyllo Sudário Cardoso; ²Hugo Pereira Leite Filho; ³Rafael Souto.

¹Bacharelado em Sistemas de Informação; Universidade Estadual de Goiás Câmpus Ceres; Ceres-GO; Graduando; danyllo.sudario@hotmail.com; ²Bacharelado em Sistemas de Informação; Universidade Estadual de Goiás Câmpus Henrique Santillo; Anápolis-GO; Professor Mestre; filho.hugo@gmail.com; ³Mestrado em Biologia Ac. Genética; Centro de Excelência em Ensino, Pesquisa e Projetos - Leide das Neves Ferreira; Goiânia-GO; Gerente; rsouto.775@gmail.com.

RESUMO

Introdução: Com dados massivos espalhados pelo mundo e nas nossas vidas (WITTEN, FRANK e HALL, 2011), é necessário lidar com esses dados adequadamente. Pesquisas clínicas atuais encaram um problema referenciado na literatura como dados faltantes (MD) que influencia as informações geradas (HAUKOOS e NEWGARD, 2007) pela mineração de dados (GALVÃO, 2007) e extração de conhecimento. **Objetivos:** Evidenciar o viés (HAUKOOS e NEWGARD, 2007) nos modelos estatísticos que executam deleção por lista. Comparar parâmetros estimados (PE) e R^2 nos modelos de regressão linear sobre conjuntos de dados com MD com similares submetidos a métodos de imputação múltipla (MI). **Métodos:** Mineração de dados, regressão linear e MI. **Resultados:** Utilizou-se dois conjuntos de dados de relatórios médicos dos Grupos I (GI) e II (GII) dos acidentados com o Césio-137, com 99 amostras. GI apresentou 27,78% de MD e R^2 de 0.2455, GII apresentou 62,23% de MD e R^2 de 0.4706 na análise de caso completo, ambos apresentaram subestimação dos PE. Modelos com MI apresentaram PE estimados a maior, com R^2 do GI próximo de 0.1317 denunciando que a baixa correlação no primeiro modelo, na realidade tinha um encaixe pior para a variável regressora, o R^2 do GII foi de 0.4702. **Conclusão:** Mesmo sob um baixo índice de MD há substancial viés da informação para um dado conjunto de dados. A variável regressora escolhida não explica o comportamento das demais no GI, e explica medianamente no GII.

Palavras Chave: Imputação; Dados faltantes; Mineração de dados.