Estudo de Técnicas Multivariadas Para Seleção de Variáveis em Grandes Bancos de Dados: Uma Aplicação Envolvendo Dados de Inibição (IC_{50})

Jaciele de Jesus Oliveira[1], Antônio Luiz Silveira Vilanova Costa[2],João Batista filgueira costa[3], Guilherme Rocha Moreira[4] , Nivan Bezerra da Costa Júnior[5], Carlos Raphael Araújo Daniel[6]

Resumo: A análise multivariada é um meio eficiente na análise de grandes bancos de dados contendo inúmeras variáveis, pois tais técnicas podem ser utilizadas para obter um número reduzido de variáveis sem perda de informação útil. Este trabalho tem por objetivo estudar as técnicas de regressão múltipla PLS e PCR em problemas de seleção de variáveis e avaliar o desempenho destas estratégias em um banco de dados real. O banco de dados utilizado apresenta 602 estruturas e 93 variáveis buscando descrever o comportamento das variáveis resposta IC_{50} e suas transformações $\ln(IC_{50})$ e $1/ IC_{50}$. O IC_{50} é uma medida da potência de uma substância no processo de inibição de uma função química ou biológica, indicando quanto da substância é necessária para inibir um dado processo pela metade, portanto, quanto menor o IC_{50}, mais ativo é o composto. Foram ajustados modelos particionando os dados em conjuntos de treinamento e teste. Os dados também foram submetidos a uma análise de agrupamento numa tentativa de separar grupos de compostos semelhantes entre si. A presença de *outliers* e sua influência nos ajustes foram avaliadas. No geral as técnicas utilizadas tiveram um desempenho satisfatório comparando valores de erro quadrático médio, permitindo identificar um modelo que se ajustou bem ao conjunto teste e conseguiu descrever bem os dados. Na maioria dos casos, a técnica PLS apresentou melhores resultados nesse estudo. Por fim, foi possível destacar as 20 variáveis mais relevantes para o modelo.

Palavras-chave: PCR; PLS; seleção de variáveis, IC_{50}.

Departamento de Estatística e Ciências Atuariais – UFS, email: *jacioliveira416@gmail.com*
Departamento de Química – UFS, email: *antoniovilanova10@gmail.com*
Departamento de Química – UFS, email: *nbcj@ufs.br*
Departamento de Estatística – UFRPE, email: *jfilgueiracosta@gmail.com*
Departamento de Estatística – UFRPE, email: *guirocham@gmail.com*
Departamento de Estatística e Ciências Atuariais – UFS, email: *raphael_crad@yahoo.com*