# AÇÕES BASEADAS EM DATA MINING - TRABALHO III - ISI

Mineração de Dados na Identificação de Empresas Irregulares Quanto ao Pagamento de Impostos

A Secretaria da Fazenda de Pernambuco, utilizou um sistema que classifica e identifica perfis de empresas com grandes potenciais de comportamento irregular no pagamento de impostos.

Através da utilização da base de dados deles, foi criado um algoritmo que é capaz de classificar e identificar as empresas com a utilização de tarefas de classificação e clusterização. Na mineração de dados, foi utilizada a biblioteca Scikit-learn, para utilizar os métodos de classificação rede neural MLP, Random Forest, SVM e ensemble.

Após os dados obtidos pela classificação, foram utilizados algoritmos de clusterização para identificar os níveis de nocividades das empresas analisadas.

Os resultados encontrados pela Secretaria da Fazenda de Pernambuco mostrou dados satisfatórios para o que eles buscavam, com a técnica Random Forest tendo um total de 90,7% de acerto na classificação das empresas. Com a utilização de clusters, se determinou o nível de nocividade entre as empresas analisadas, criando uma margem de 3 a 7 níveis de nocividade.

Referência: NASCIMENTO, Rafaella Leandra Souza et al. Mineração de dados na identificação de empresas irregulares quanto ao pagamento de impostos. Revista de Engenharia e Pesquisa Aplicada, v. 3, n. 3, 2018.

Método de mineração de dados para identificação de câncer de mama baseado na seleção de variáveis

Foi criado um método que através da análise de amostras de célula da mama de pacientes, classifica as variáveis obtidas em duas classes, benigno e maligno. As variáveis eram formalizadas através de 4 passos operacionais:

1- Dividir o banco de dados original em uma parte de treino e outra de teste, aplicando a análise de componentes principais na parte de teste;

2- Gerar índices de importância utilizando os parâmetros da análise de componentes principais;

3- Classificar a porção de treino utilizando os métodos de classificação k-vizinhos mais próximos e análise discriminante, eliminando as variáveis menos importantes;

4- Selecionar o melhor subgrupo de variáveis, classificando a porção de teste com as variáveis selecionadas;

Através desses 4 passos, com base no Wisconsin Breast Cancer Database, foi possível obter uma classificação de 97,77% de acerto das amostras analisadas.

Referência: HOLSBACH, Nicole; FOGLIATTO, Flávio Sanson; ANZANELLO, Michel Jose. Método de mineração de dados para identificação de câncer de mama baseado na seleção de variáveis. Ciência & Saúde Coletiva, v. 19, p. 1295-1304, 2014.