

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Trabalho Prático 1

Você foi contratado pelo Tribunal de Contas da União (TCU) com o objetivo de auxiliá-lo na identificação de suspeitos de fraudes em compras governamentais. Ou seja, o TCU suspeita que estejam ocorrendo fraudes diversas tais como cartel, favorecimento, superfaturamento, dentre outras, em compras realizadas a partir do processo de pregão inverso do governo federal. Como o volume de compras é muito grande, é impraticável que os auditores verifiquem cada comprador individualmente. Como o intuito de identificar um conjunto restrito de possíveis fraudadores, os auditores do TCU disponibilizaram uma base de dados para que você, um especialista em Mineração de Dados, possa auxiliá-los neste processo. Seu primeiro passo, naturalmente, consiste em tratar a base disponibilizada de forma a viabilizar uma posterior análise envolvendo técnicas tradicionais de Mineração de Dados (MD). Dessa forma, pede-se:

- Proponha e descreva quais etapas de tratamento de dados serão necessárias para a base disponibilizada (e.g., remoção de ruídos, binificação, discretização, dentre outras). Uma informação importante neste processo é que foi definido pelos auditores que, baseado em tarefas comumente realizadas por eles, um conjunto de técnicas de MD de particular interesse são as baseadas em padrões frequentes;
- Aplique as etapas propostas gerando uma base que possa servir de entrada para algoritmos tradicionais de MD. Note que na sua proposta podem haver novos atributos derivados a partir dos atributos originais associadas a cada compra;
- Identifique quão redundantes (ou correlacionadas) são as informações contidas em cada atributo associado às compras;
- Suponha que o primeiro objetivo requisitado pelos auditores seja que você defina uma única métrica, derivada a partir dos atributos das compras realizadas, capaz de ordenar os vendedores de acordo com alguma medida de ‘fraudulência’. Qual seria sua proposta? Quais atributos da base comporiam sua métrica?

Deve ser entregue uma documentação com todas as análises e decisões tomadas. Os critérios de correlação do presente trabalho serão todos embasadas na qualidade das discussões realizadas, bem como nas justificativas para as decisões tomadas. Obviamente não há uma única resposta aplicável a cenários como este. De fato, o objetivo do presente trabalho reside justamente em verificar a capacidade de alunos de MD tomarem decisões acertadas em cenários reais como o apresentado. Dessa forma, não é objetivo do trabalho o desenvolvimento de técnicas de análise de dados já consolidadas na literatura. Aconselhamos, inclusive, o uso de ferramentas bem estabelecidas para este fim, tais como *Weka*, *R*, *RapidMiner*, *Pentaho*, dentre outras. A documentação deve conter no máximo 10 páginas, com espaçamento simples entre linhas, tamanho de fonte 12 e fonte *Times New Roman*.

Data de entrega: 25 de Agosto