



## Enunciado do T2 – Projeto de Mineração de Dados

---

O T2 da disciplina de IN consiste no desenvolvimento de um projeto de mineração de dados. Para tanto, os alunos deverão se reunir em grupos de **3 a 4 alunos (mesmos grupos do T1)** e definir qual tarefa de mineração de dados estão dispostos a realizar: (i) classificação; (ii) agrupamento; ou (iii) associação.

Uma vez definida a tarefa de mineração, os alunos deverão encontrar um problema interessante capaz de ser trabalhado sob a ótica da tarefa de mineração escolhida. Esta etapa deve ser comunicada ao professor para ver se o mesmo está de acordo com o problema sugerido. É imperativo que os alunos busquem **dados reais** públicos para realização do trabalho. Em casos de exceção, serão permitidos dados provenientes de empresas onde o aluno trabalhe, desde que o mesmo tenha autorização por escrito para utilização de tais dados. A base de dados deve apresentar necessariamente **pelo menos 5 (cinco) atributos e pelo menos 100 (cem) instâncias**.

### Entrega:

O grupo deverá entregar um relatório em PDF com capa e as seguintes seções:

1. Justificativa do Problema e Tarefa Escolhida;
2. Pré-processamento dos dados;
3. Algoritmo(s) Escolhido(s);
4. Resultados Encontrados;
5. Conclusões sobre as Descobertas.

Na Seção 1, os alunos deverão descrever detalhadamente a abordagem: o problema do mundo real que desejam resolver e os detalhes da base de dados (descrição dos objetos e atributos). Na Seção 2, deverão detalhar o que foi necessário realizar de pré-processamento para que os dados pudessem ser utilizados como entrada para o algoritmo de mineração. Na Seção 3, deverão relatar qual algoritmo(s) utilizado(s) para resolução do problema, com um breve relato do seu funcionamento e a justificativa para sua escolha. Na Seção 4, espera-se que o grupo descreva o modelo encontrado (exs: árvores, grupos, regras), bem como os resultados das medidas quantitativas que avaliam qualidade do modelo (exs: acurácia, SSE, confiança). No caso de problemas de classificação, indicar o protocolo de separação dos dados (*holdout*, *k-fold cross-validation*, etc.). Sugere-se que o grupo inclua *print-screens* da ferramenta de mineração utilizada para reportar cada uma das exigências previamente descritas.

**Ferramentas sugeridas:** Weka (Java), scikit-learn (Python), Orange (Python).