# Mineração de Dados - Previsão de Evasão Escolar

## 1. Técnicas de Pré-processamento

Objetivo: Garantir que os dados estejam limpos, completos e prontos para a aplicação de algoritmos de mineração.

a. Tratamento de valores ausentes:  
- Verificar colunas com valores nulos ou faltantes.  
- Técnicas: Imputação por média/mediana/moda; exclusão de linhas; uso de modelos preditivos.

b. Codificação de variáveis categóricas:  
- Técnicas: Label Encoding, One-Hot Encoding, Ordinal Encoding.

c. Normalização/Escalonamento:  
- Aplicável a variáveis numéricas: MinMaxScaler, StandardScaler.

d. Remoção de colunas irrelevantes ou redundantes:  
- Identificadores únicos ou colunas altamente correlacionadas.

e. Transformações:  
- Conversão de datas, agrupamento de categorias raras.

## 2. Ferramentas

- Linguagens: Python  
- Bibliotecas Python: pandas, scikit-learn, matplotlib, seaborn, plotly, xgboost, lightgbm, SHAP, lime.

## 3. Técnicas e Algoritmos de Mineração de Dados

Prever abandono ou sucesso acadêmico

#### 3.1Tecnicas de Mineração Vistas em Aula que serão usadas:

* Classificação
  + Random Forest
  + Logistic Regression
* Regressão
  + Linear Regression
  + Gradient Boosting Regressor
* Detecção de Desvios
  + Detecção de Anomalias
  + Local Outlier Factor (LOF)

## 4. Visualização de Dados

**Durante a Análise Exploratória (Pré-modelagem):**

* **Histogramas**: Para visualizar a distribuição de atributos numéricos.
* **Boxplots**: Para detectar outliers e comparar valores numéricos entre grupos (ex: idade por status do aluno).
* **Gráficos de barras**: Para variáveis categóricas (ex: tipo de curso, estado civil, situação).
* **Heatmaps** Para analisar relações entre variáveis numéricas.

**Durante/Pós-modelagem (Avaliação de Modelos):**

* **Importância das Features**: Para entender quais variáveis mais influenciam o modelo (especialmente em Random Forest, Gradient Boosting).
* **Matriz de Confusão**: Para avaliar classificadores — mostra acertos e erros por classe.
* **Curva ROC / AUC**: Para verificar a capacidade discriminatória do modelo de classificação binária/multiclasse.