# Resumo do Projeto de Mineração de Dados

**Autores:**
João Augusto S. Pacolla
Maikon F. Gino

Este documento resume as atividades desenvolvidas no Google Colab para atender aos requisitos da disciplina de Mineração de Dados. O projeto foi estruturado em etapas conforme solicitado, incluindo justificativa teórica, pré-processamento, resumo estatístico e análise inicial de dados.

## Tema do Projeto

O tema definido foi: **Desafios no uso de nuvem pública, privada e de governo em ciências de dados**. A proposta é conectar a prática de ciência de dados com as dificuldades reais enfrentadas no uso de diferentes modelos de nuvem, considerando aspectos de custo, segurança, privacidade e conformidade regulatória.

## Fonte de Dados

O dataset escolhido foi o **Bank Marketing Dataset**, disponível no UCI Machine Learning Repository. Este conjunto possui 41.188 instâncias e 20 atributos, incluindo variáveis socioeconômicas e demográficas. O atributo alvo é **y**, que indica se um cliente aderiu ou não a um depósito a prazo. A escolha se justifica por atender os requisitos mínimos, ser amplamente utilizado em estudos e possuir relevância prática.

## Etapa 1 — Pré-Processamento e Resumo Estatístico

As seguintes atividades foram realizadas no Colab:

1. Carregamento do dataset a partir do repositório UCI.
2. Verificação de tipos de dados e valores ausentes.
3. Tratamento de valores ausentes utilizando imputação.
4. Codificação de variáveis categóricas (One-Hot Encoding).
5. Normalização de variáveis numéricas.
6. Divisão dos dados em treino e teste (80%/20%).
7. Resumo estatístico (describe) para variáveis numéricas e categóricas.
8. Geração de visualizações: histogramas, gráficos de barras e matriz de correlação.

## Etapa 2 — Modelagem (Planejada)

Na segunda etapa, será realizada a aplicação de algoritmos de classificação. Os modelos definidos foram:

1. Decision Tree Classifier — escolhido pela facilidade de interpretação e visualização.
2. Random Forest Classifier — escolhido por ser mais robusto e reduzir overfitting.

## Discussão sobre os Desafios de Nuvem

O projeto também discute os desafios de execução de ciência de dados em diferentes modelos de nuvem:

1. Nuvem Pública: escalabilidade e baixo custo inicial, mas risco de custos variáveis, dependência de fornecedores e exigência de compliance com LGPD/GDPR.
2. Nuvem Privada: maior controle e segurança dos dados, mas exige alto investimento em infraestrutura e equipe especializada.
3. Nuvem de Governo: garante conformidade regulatória e governança, mas pode trazer burocracia e menor agilidade de inovação.

## Conclusão

Com a Etapa 1 concluída, o dataset encontra-se limpo, pré-processado e explorado, além de uma justificativa teórica sólida que conecta mineração de dados com os desafios da nuvem. A próxima etapa será a aplicação prática dos algoritmos de classificação e avaliação dos resultados.