# **Relatório – Trabalho Final (Parte 1)**

**Disciplina:** Ciência de Dados  
 **Especialização:** [Nome do curso]  
 **Equipe:** [Nomes dos integrantes]  
 **Base de dados utilizada:** [Nome da base]

## **1. Business Understanding**

**1.1 Negócio por trás dos dados** Descreva qual área ou setor o dataset representa (ex.: educação, saúde, marketing, redes sociais, finanças).

**1.2 Objetivos estratégicos da organização** Liste pelo menos 2 objetivos estratégicos relacionados ao negócio da organização detentora dos dados.

**1.3 Problema de negócio** Defina claramente o problema que será tratado.

**1.4 Tradução para Ciência de Dados** Explique como o problema de negócio pode ser transformado em problema de ciência de dados (ex.: classificação, regressão, clusterização).

**1.5 Hipóteses iniciais** Formule pelo menos 2 hipóteses a serem testadas com os dados.

**1.6 Restrições** Liste limitações identificadas (tempo, orçamento, qualidade/disponibilidade de dados, questões éticas/legais).

**1.7 Critério de sucesso** Defina qual será o critério para considerar o projeto bem-sucedido (ex.: acurácia acima de 85%, modelo interpretável, redução de falsos positivos).

**1.8 Métricas de avaliação** Liste quais métricas serão usadas (ex.: precisão, recall, F1-score, RMSE, ROI).

## **2. Data Understanding**

### **2.1 Coleta**

* Quais dados estão disponíveis? (número de registros e colunas).
* Qual a origem dos dados (bases públicas, sistemas internos, redes sociais etc.)?
* O volume, variedade e granularidade são adequados?
* Há limitações éticas, legais ou de privacidade?
* Quais as principais variáveis e como estão distribuídas?

### **2.2 Exploração**

* Existem valores faltantes, inconsistentes ou duplicados?
* Há outliers que precisam ser analisados?
* Os dados estão balanceados em relação à variável-alvo (se houver)?
* Existem correlações relevantes entre variáveis? Explique.
* O que os dados sugerem preliminarmente em relação ao problema de negócio?
* É necessário enriquecer a base com fontes externas?
* Há indícios de viés nos dados (ex.: sub-representação de grupos)?
* Quais dados precisam ser limpos, transformados ou descartados?
* O conjunto atual permite definir claramente a variável-alvo (label)?

## **3. Data Preparation (Parte 1)**

**3.1 Seleção de atributos** Quais variáveis são realmente relevantes para o problema?

**3.2 Tratamento de registros** Quais linhas devem ser mantidas ou descartadas (ex.: duplicadas, incompletas)?

**3.3 Subconjuntos de dados** Haverá divisão em treino, validação e teste?

**3.4 Outliers** Como serão tratados? (remoção, ajuste, manutenção).

**3.5 Criação de novas variáveis (feature engineering)** Liste variáveis derivadas que podem enriquecer a análise.

**3.6 Codificação de variáveis categóricas** Será necessário aplicar one-hot encoding, label encoding ou outra técnica?

**3.7 Normalização/padronização de variáveis numéricas** As variáveis precisam ser escaladas para os algoritmos a serem usados?

**3.8 Formatação dos dados** Os dados estão no formato adequado para os algoritmos?

**3.9 Balanceamento da variável-alvo** Será necessário aplicar técnicas como oversampling, undersampling ou SMOTE?

**3.10 Conjunto final** O dataset preparado reflete fielmente o problema de negócio e está pronto para aplicação de Machine Learning?