

# Guia de Análise de Propostas

O pensamento analítico de dados eficaz permitirá **avaliar potenciais projetos de mineração de dados** de forma sistemática e potenciais falhas nas propostas.

Essa habilidade pode ser aplicada como uma **autoavaliação para as propostas** e como ajuda na avaliação de propostas de equipes internas de data science ou consultores externos.

## Negócios e Compreensão dos Dados

1. Qual é **exatamente o problema** a ser resolvido?
2. A solução de Data Science foi adequadamente formulada para resolver este problema de negócio?
3. A **qual entidade de negócios** o exemplo corresponde?
4. O problema é **supervisionado** ou **não supervisionado**?
  1. **Problema supervisionado:**
    1. Tem uma **variável alvo (target) definida**?
      1. A variável alvo é **definida com precisão**?
      2. Pensar sobre valores que podemos adotar.
5. Os **atributos são bem definidos**?
  1. Pensar sobre os valores que podemos adotar.
6. Para problemas supervisionados: **modelar a variável alvo melhorará o problema de negócios**? Um subproblema importante? No último caso, o restante do problema de negócios foi abordado?
7. Dispor o problema em termos de valor esperado ajuda a estruturar subtarefas que precisam ser resolvidas?
8. Caso não supervisionado, existe um caminho de "análise exploratória de dados" bem definido? (onde está indo a

análise?)

## Preparação da Análise

1. **Será prático obter valores para atributos?** Criar vetores de características e colocá-los em uma única tabela?
2. Em caso negativo, um formato de dados alternativo foi definido com clareza e precisão? Isso é levado em conta nas fases posteriores do projeto? (Muito dos métodos posteriores assumem que o conjunto de dados está no formato de vetor de característica).
3. Se a modelagem será supervisionada, a **variável alvo está bem definida**? Está claro como obter valores para a variável alvo (treinamento e teste) e colocá-los na tabela?
4. Como exatamente os valores para a variável alvo são adquiridos? Existem custos envolvidos? Caso positivo, os custos são levados em consideração na proposta?
5. Os dados estão sendo extraídos de uma população semelhantes em que o modelo será aplicado? Se houver discrepâncias, alguma tendência de seleção foi claramente observada? Existe um plano para saber como compensá-las?

## Modelagem

1. A **escolha do modelo é adequada** para a escolha da variável alvo?
  1. Classificação, estimativa de probabilidade de classe (regressão logística), avaliação, regressão, agrupamento, etc.
2. Vários modelos precisam ser testados e comparados (na avaliação)?
3. Para agrupamento, existe uma métrica de semelhança definida? Ela faz sentido para o problema de negócios?

# Avaliação e Implantação

1. Existe um plano para validação do domínio de conhecimento?
  1. Especialistas no domínio ou investidores querem vetar o modelo antes da implantação? Em caso positivo, o modelo estará em um formato que eles podem compreender?
2. A configuração e a métrica de avaliação são adequadas para a tarefa de negócios?
  1. Os custos de negócios e os benefícios são levados em consideração?
  2. Para classificação, **como os limiares de classificação são escolhidos?**
  3. As estimativas de probabilidade são diretamente usadas?
  4. **A classificação é a mais adequada?**(para um orçamento fixo)?
  5. Para regressão, como avaliariamos a qualidade das previsões numéricas? Por que este é o caminho certo no contexto do problema?
3. A avaliação usa dados de retenção?
  1. **Validação cruzada** é uma técnica.
4. Contra quais bases os resultados serão comparados?
  1. Por que isso faz sentido no contexto do problema real e a ser resolvido?
  2. Existe um plano para avaliar os métodos de base também de forma objetiva?
5. Para agrupamento, como ele será entendido?
6. A implantação conforme o planejado vai realmente resolver (melhor) o problema de negócios declarado?

7. Se a despesa de projeto tem que ser justificada para os investidores, **qual é o plano para medir o impacto de negócio final** (implantado)?