Etapas para análise de dados

# 1 - Problema a ser resolvido

Tudo se inicia entendo o problema do negócio que visa ser resolvido, isso é importante para traçar a melhor rota para resolver o problema. As seguintes perguntas de negócios devem ser respondidas:

- Definir o problema a ser resolvido

- Quais tipos de dados serão utilizados

- É possível realizar análises exploratórias (não recomendado)

- Cientistas buscam padrões não descobertos

# 2 – Preparação e Exploração de Dados

Consome 50% a 70% do tempo do cientista de dados, é sub-dividida em:

- identificação das variáveis: variáveis são as colunas de um DataFrame.

Definir as variáveis de input, output e categorizá-los

- Tratamento de Valores Missing.

valores podem estar faltando em determinadas variáveis (NaN). Os valores missing podem reduzir o poder de tratamento de dados. Pode ser um problema da fonte de dados e durante o tratamento de dados.

- Tratamento de Outliers.

Outliers são valores extremos, que fogem do padrão observado em um conjunto de dados. Podem ser uni-variado ou multi-variado.

- Transformação de Variáveis.

Variáveis podem ter valores missing, outliers e dados incorretos. Podem ocorrer remoção de valores, concatenação de valores, distribuição de valores de 1 coluna para 3 colunas, etc.

- Criação de Variáveis.

Transformação de variáveis existentes ou coletando mais dados.

# 3 – Criação do Modelo Preditivo

É definido o modelo de ML para realizar as previsões. Há mais de 60 algoritmos para Machine Learning. Dados devem ser subdivididos em dados de treino e dados de teste.

Criar previsões para resolver o problema de negócio,

# 4 – Automatização do Processo

Tornar o modelo autônomo, que ao receber os dados os processos são realizados sem a interferência do Cientista de Dados.

# 5 – Apresentação dos dados

Nesta etapa é escolhido a forma como será apresentado os dados.

