DATA SCIENCE

AULA 12 – PROJETO – ETAPA III

PROFª. ANA CAROLINA B. ALBERTON

MÉTODO CRISP - DM

Etapa1 – **Entendimento do Negócio**

(Ouvindo a História)

Etapa 2 – **Entendimento dos Dados**

(Buscando o Mapa)

Etapa 3 – **Preparação dos Dados**

(Entrando na Mina)

Etapa 4 – **Modelagem**

(Vamos ver se é possível extrair algo)

Etapa 5 – **Avaliação dos Modelos**

(A Hora da Verdade)

Etapa 6 – **Publicação**

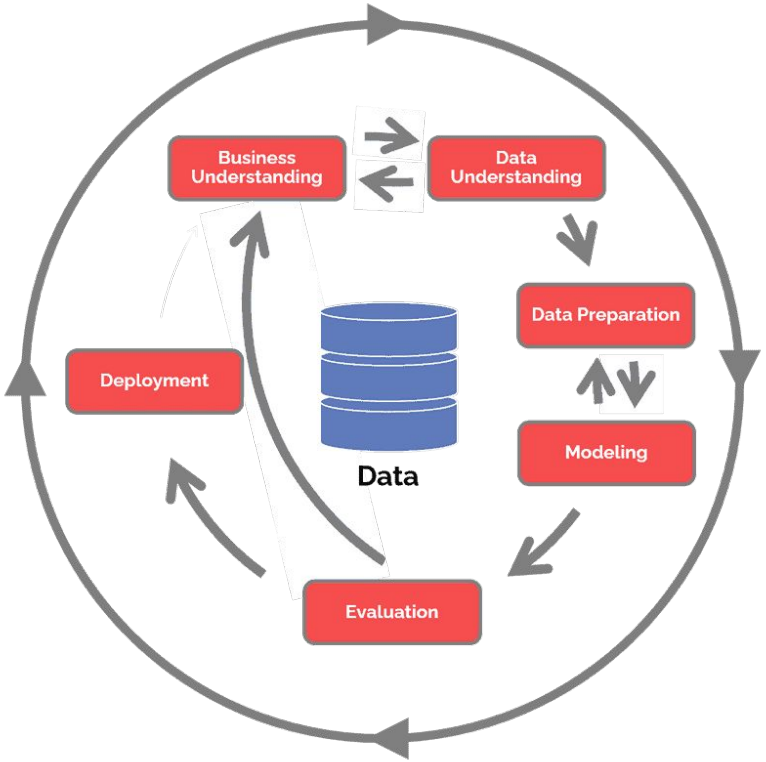
(Festa ou Aprendemos com a Experiência)

CRISP –DM ETAPA 3

Após entender as coisas, tanto os dados quanto o negócio, partimos para o que seja a ideal preparação dos dados, definir o que será utilizado e não, como será tratado os seus dados já visando os modelos que serão produzidos logo mais nas próximas etapas.

Aqui já iremos “minerar nossos dados” e todos os objetivos dessa etapa estão ligados à eles:

- **Seleção;**
- **Limpeza;**
- **Construção;**
- **Integrar;**
- **Formatar.**



O QUE DEVO FAZER?

Após entender as coisas, tanto os dados quanto o negócio, partimos para o que seja a ideal preparação dos dados.
Essa preparação exige uma análise de cada dado, ou seja, cada tabela, incluindo:

- Seleção;
- Limpeza;
- Construção;
- Integração;

Usando os códigos vistos na ultima aula, aplicar para cada dataset.

FAZER PARA CADA TABELA

- Visualizar os dados;
- Dar nomes as colunas, caso necessário;
- Explorar dados categóricos e dados numéricos, identificando erros e corrigindo-os, caso necessário:
  - lembrar que, para remover nan ou trocar dados categóricos, não é regra, mas normalmente usado:

    Dados categóricos: Usar moda

    Dados numéricos: Usar mediana
- Visualizar dados que podem estar fora do domínio considerado correto;
- Caso trabalhe com dados temporais, você deve definir uma escala de tempo que melhor atenda a todos os dados e aplicá-la a todos as tabelas. Essa escala pode ser dias, semanas, quinzenas, meses... Garantir que vamos iniciar e finalizar no mesmo ponto, e que todas as tabelas tenham a mesma quantidade de linhas.

Caso tenha mais de uma tabela: Insira todos os dados no mesmo dataframe, usando a função pd.merge().
Assim teremos um dataframe final e único com todos os nossos dados

O QUE DEVE SER ENTREGUE?

- Como entrega nessa etapa, as equipes deverão entregar um dataframe final (.xlsx ou .csv) e um relatório contendo as alterações que foram feitas em cada dado, justificando as suas escolhas e suas ações. Caso não seja feito nenhuma alteração, justificar também.
- A data de entrega é para o próximo domingo, 11/05 ate as 23:59h no ava, na tarefa destinada a isso.
- Caso dê algum contratempo, enviar o trabalho para o email **anacarolina.alberton@satc.edu.br**
- Dúvidas podem ser tiradas por email ou pelo nosso grupo de WhatsApp