IFSP - Instituto Federal de Educação, Ciência e Tecnologia

Câmpus São Paulo

RELATÓRIO CICLO DE VIDA DO DADO - ICDA6

EDILENE SHIZUE ONIZUKA REGES - SP3039005

MARCO ANTONIO DE SOUZA REIS JUNIOR - SP3041671

VICTOR HUGO SAMPA HAMAGUTI - SP3038998

São Paulo - SP - Brasil

2022

1. **Produção:** Os dados foram coletados de *datasets* já previamente estruturados.

2. **Armazenamento:** Os dados estavam armazenados em repositórios digitais, com o cunho de ser informações públicas e de fácil acesso a todos. Além do mais, os dados estavam estruturalmente em formato XLS.

3. **Transformação:** Os dados passaram por uma transformação de estrutura através do Excel, onde eles foram tabulados e estruturados em suas respectivas descrições.

4. **Análise:** A etapa de análise de dados consiste na execução de qualquer operação para extrair informação e conhecimento dos dados.

   Análise baseada no modelo CRISP-DM:

   - **Entendimento do negócio (Business Understanding):** Os *datasets* disponibilizados para a análise trazem dados bem intrigantes sobre a criminalidade de uma forma geral no Mundo e como afeta uma das cidades mais seguras como Montreal.

   - **Entendimento dos dados (Data Understanding**): Os dados começaram analisados através de simples visualizações e filtros, com o intuito de identificar quais seriam os dados mais propícios ao estudo entendimento dos casos. Além do mais, foram cerificadas as respetivas frequências de cada dado.

   - **Preparação de dados (Data Preparation - data munging, data wrangling):** Os dados começaram a ser preparados após o seu entendimento, principalmente com o uso de algoritmos para remover valores, além da troca de suas variáveis, para melhor entendimento das análises.

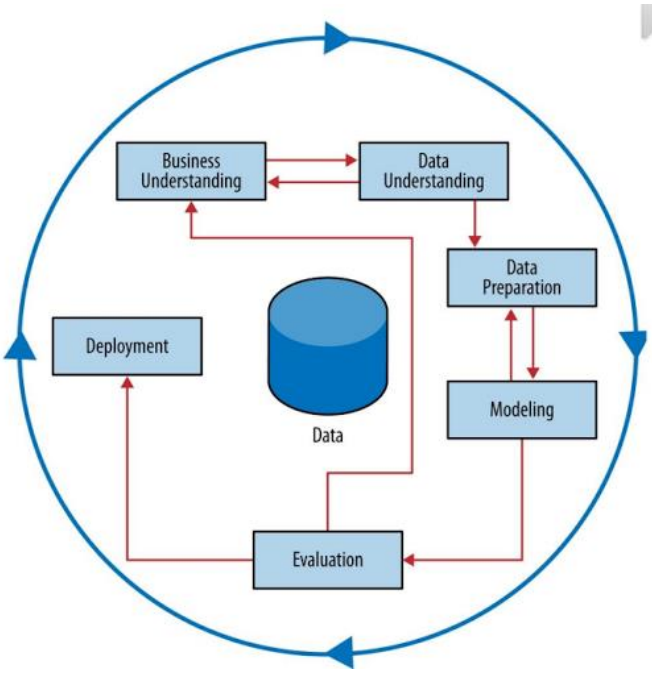
   - **Modelagem (Modeling):** A partir da mineração, foi possível modelar os dados da melhor forma para de compreensão, facilitando as plotagens.

   - **Avaliação (Evaluation):** Após a modelagem,os dados estão prontos para serem aplicados a *machine learning,* que no caso foi utilizado a técnica PCA. Uma das técnicas mais utilizadas na redução de dimensionalidade é um método estatístico designado por *Principal Component Analysis* (PCA). O PCA é caracterizado por identificar as dimensões ao longo das quais os dados se encontram mais dispersos. Desta forma, conseguimos identificar as dimensões que melhor diferenciam o conjunto de dados em análise, ou seja, os seus componentes principais.

     Usando esta técnica, é possível realçar as semelhanças e diferenças neles existentes através da identificação de padrões. A sua identificação em

dados caraterizados por grandes dimensões é difícil, uma vez que a sua representação gráfica não é viável, logo uma análise visual aos dados não é possível. Quando identificados os padrões no conjunto, o número de dimensões a analisar pode ser reduzido sem que haja uma perda significativa de informação, pois o foco recai sobre a análise das dimensões principais que caracterizam o conjunto de dados.

- **Implantação (Deployment)**: os resultados da mineração de dados são colocados em uso real para obter algum retorno sobre o investimento. A implantação do modelo preditivo em algum sistema de informação ou processo de negócio.



5. **Descarte:** Além disso, o dado precisará passar por um processo seguro e legal de descarte.