Checklist para Projetos de Data Science

Checklist para aplicar aos seus projetos de Data Science. Baseado nos *frameworks* CRISP-DM e na metodologia proposta por Aurélien Géron, esta é a rotina que se adequa à maior parte dos meu projetos de Data Science.

Lembre-se que não é um *checklist* rígido ou imutável. Pelo contrário!

Este é um guia para você não sair do zero. Você pode (e deve) adaptar ele a sua realidade quando trabalhando em um projeto de Ciência de Dados.

1. Entender o Problema

- Olhar o todo e delimitar o escopo do projeto
- Como a solução vai ser usada?
- Quais são as soluções já existentes?
- Qual abordagem usar?
    - Aprendizado Supervisionado
    - Aprendizado Não Supervisionado
    - Aprendizado Por Reforço
- Qual é a métrica de performance?
- Qual a performance mínima esperada para atingir o objetivo?
- Liste as premissas básicas do projeto

2. Explorar os Dados

- Criar uma cópia dos dados para a exploração
- Criar um Jupyter Notebook para documentar a exploração
- Estudar cada atributo e suas características:
    - Nome
    - Tipo
        - Categórica
        - Numérica
            - int
            - float
        - Estruturada

        - Não Estruturada
        - etc
    - % de valores ausentes
    - Ruído nos dados e tipo de ruído (outliers, estocásticos, erros de arredondamento)
    - Tipo de distribuição
        - Gaussiana
        - Uniforme
        - Logarítmica
        - etc
- Identificar a variável alvo (target)
- Visualizar os dados
- Estudar a correlação entre os dados
- Identificar as transformações que podem ser aplicadas
- Identificar os dados extras que podem ser úteis

3. Preparar os Dados

- Trabalhar em cópias dos dados
- Escrever funções para todas as transformações

1. Limpeza dos Dados
    - Consertar ou remover outliers
    - Preencher os valores faltantes ou eliminar as linhas/colunas
        - Zero
        - Média
        - Mediana
        - etc
2. Seleção de atributos
    - Eliminar os atributos (*features*) que não contêm informações úteis
3. *Feature Engineering*
    - Discretizar variáveis contínuas
    - Decompor *features* (categóricas, data, tempo)
    - Aplicar transformações às variáveis
    - Agregar *features* para gerar novas
4. *Feature Scaling*
    - Normalizar ou padronizar *features*

4. Construção do Modelo

- Automatizar o maior número de passos possíveis
- Treinar mais de um modelo e comparar as performances
- Analisar as variáveis mais significativas para cada algoritmo
- *Fine-Tune* dos *hyperparameters*
- Uso de *cross-validation*
- Verificar o desempenho dos métodos *Ensemble*, combinando os modelos que tiveram os melhores desempenhos individuais
- Testar o desempenho do mesmo com o *dataset* de teste.

5. Apresentação da Solução e Deploy

- Documentar todos as etapas
- Tornar todos os passos replicáveis (download de arquivos, uso da API do Kaggle)
- Lembrar do Storytelling
    - Decisores e Diretores provavelmente desconhecem a parte técnica
- Ver qual o melhor gráfico para contar cada *insight* descoberto
- Escrever testes unitários
- Criar rotinas de monitoramento e alertas
- Determinar quando atualizar o modelo