awari.

# Projeto Final - Orientações

*Diretrizes para a criação do documento de códigos do seu projeto final do Intensivo de Data Science Awari.*

Projetos de ciência de dados geralmente seguem uma sequência de passos e este template tem como objetivo guiar os estudantes da Awari na criação de seus projetos finais.
Vale ressaltar que, embora as etapas que assim forem assinaladas sejam obrigatórias, a maneira com que você vai utilizar o presente template é como um guia, não como um conjunto de regras.
Sendo assim, as bibliotecas e procedimentos internos às etapas demonstradas devem ser encarados como **sugestões**, o que faz com que os alunos, em conjunto com seus mentores, tenham liberdade de optar por outras ferramentas e divergir de alguns dos procedimentos sugeridos abaixo.

## Onde devo realizar meu projeto?

Recomendamos que o código do projeto seja implementado majoritariamente em um ambiente do Jupyter Notebook. No entanto, você poderá fazer uso de outros editores que permitam a programação em Python 3. Além disso, as etapas devem ser documentadas em um arquivo PDF ou em um post de Blog (sugerimos Medium ou FastPages), maiores detalhes na etapa VI. O código de algumas etapas, poderão ser implementados externamente ao Jupyter como, por exemplo, a etapa II, que **poderá** ser realizada fora desse ambiente, e a etapa VII (opcional), que, caso venha a ser realizada, **deverá** ser fora de tal ambiente (informações mais detalhadas nas seções referentes a cada uma dessas etapas). Todos os documentos deverão estar presentes em um repositório no seu GitHub.

## Meu projeto pode envolver dados confidenciais?

Dados privados ou proprietários somente devem ser utilizados se houver permissão expressa. Idealmente, a base de dados deve ser proveniente de algum acesso público. Atualmente existem diversas bases de dados públicas na internet que podem ser utilizadas em um projeto de ciência de dados. Alguns exemplos são Kaggle Datasets, UCI Machine Learning Repositories e Google Dataset Search (mais exemplos na etapa II).

awari.

# Etapas

*Etapas que devem constar no seu projeto final.*

## I - Exposição do problema (obrigatória)

Esta etapa consiste em apresentar um panorama da situação em que seu projeto se encaixa, demonstrar o problema que você buscará resolver através da utilização de dados e justificar sua relevância.
É recomendada cautela com a formulação de problemas muito amplos e generalistas (ex.: o impacto da internet na vida humana)! Esse tipo de problema, caso não seja desmembrado em múltiplos problemas menores, acaba tornando-se muito vago e de resolução impossível com apenas um projeto.

## II - Coleta ou Importação dos dados (obrigatória)

Como veremos ao longo do curso, para que possamos realizar um projeto de ciência de dados, é necessário coletar dados. E para isso, possuímos basicamente duas opções:

**Coleta de dados primária:** neste modelo, você será responsável pela criação de uma base de dados coletando-os ativamente. Por exemplo, você possui uma empresa e gostaria de compreender em que turno ela é mais eficiente. Você não encontrará dados publicados a respeito disso e precisará gerá-los a partir de alguma forma de mensuração interna.
Em uma outra forma de coleta primária, você pode realizar um procedimento de *Web Scraping* para coletar dados diretamente de sites da internet, dados esses que não estarão necessariamente organizados para você.*

**Coleta de dados secundária:** neste modelo, você coletará os dados de alguma base já criada por outra pessoa, empresa ou organização. Para isso, você pode baixar arquivos de excel, JSON, csv, entre outros de algum repositório online. Você pode também conectar-se a uma um mais API's que façam a comunicação com alguma base de dados que seja útil para a resolução do problema que você visa resolver.

Algumas possíveis fontes de coleta de dados secundária são:

- Kaggle

awari.

- Google Dataset Search
- UCI - Machine Learning Repository
- Plataforma de dados abertos do governo brasileiro
- Plataforma de dados abertos do Banco Central
- Portal europeu de dados abertos
- Plataforma de dados abertos do governo americano
- Data.world

**Bibliotecas sugeridas para esta etapa:**
Pandas, Requests, JSON, Selenium, Beautiful Soup entre outras.

* Caso o projeto envolva Web Scraping com Selenium, você poderá optar por sair do Jupyter Notebook criar um arquivo .py à parte para a realização do procedimento de coleta de dados. Certifique-se de fazer uso de websites em que coleta de dados não violem seus termos de uso.

## III - Preparação dos dados (obrigatório)

Após a coleta de dados, é preciso organizá-los de forma a facilitar sua exploração e modelagem. Lembre-se, geralmente este é o procedimento mais trabalhoso de todo o processo. Durante essa etapa, alguns dos procedimentos que podem ser realizados são:

- Remoção de colunas e dados indesejados
- Manipulações do(s) índice(s) do dataframe
- Remoção de dados duplicados
- Tratamento de outliers
- Tratamento de dados ausentes
- Ajustes dos tipos de dados (*datetime, floats, integers, strings*, etc...)
- Tratamento de Strings
- Combinação de dados de diferentes fontes

**Bibliotecas sugeridas para esta etapa:**
Pandas, Numpy, Scikit Learn, (módulo preprocessing) entre outras.

## IV - Análise exploratória (obrigatória)

Após preparar os dados, você deverá buscar entendê-los. Para isso, você fará majoritariamente o uso de procedimentos de **estatística descritiva.**
Durante essa etapa, alguns dos procedimentos que podem ser realizados são:

- Verificar medidas de tendência central e de dispersão de variáveis que você julgar importantes para a resolução do problema proposto.
- Verificar a existência de correlações entre variáveis.
- Verificar a distribuição dos dados através da plotagem de histogramas ou mesmo da aplicação de testes estatísticos com essa finalidade.
- Gerar visualizações como gráficos de barras, gráfico de setores, histogramas, box plots, gráficos de linha (sequência), etc...

**Bibliotecas sugeridas para esta etapa**
Pandas, Matplotlib, Seaborn, entre outras.

## V - Modelagem (obrigatória)

Depois de compreender e ganhar intuições acerca de seus dados, chega o momento de você de fato modelá-los.
Durante essa etapa, alguns dos procedimentos que podem ser realizados são:

- Divisão dos dados em dados de treino e teste
- Criação de um *benchmark* (modelo inicial para comparações futuras)
- Triagem de modelo(s) para utilização
- Utilização de métricas de mensuração de performance dos algoritmos
- Calibração dos hiperparâmetros do(s) algoritmo(s)

**Bibliotecas sugeridas para esta etapa:**
Scikit Learn, XGBoost, LightGBM, Pandas, entre outras.

## VI - Comunicação e visualização (obrigatória)

Esta etapa consiste na documentação do seu projeto. Você pode criá-la através de um post no *Medium*, no *ReadMe* do seu repositório no *GitHub* ou em um documento PDF.
Alguns pontos importantes nessa etapa:

- Apresentação detalhada do problema, de sua relevância e do porquê de ser um problema resolvível por um projeto de ciência de dados
- Apresentação e justificativa da escolha dos procedimentos de coleta de dados utilizados

- Apresentação e justificativa da escolha dos procedimentos de manipulação de dados utilizados
- Apresentação de *insights* retirados da análise exploratória de dados
- Apresentação e justificativa da escolha dos procedimentos de modelagem utilizados
- Avaliação do modelo final e comparação com o *benchmark* utilizado
- Reflexões sobre o quão eficaz foi todo o processo para a resolução do problema proposto
- Apresentação de melhorias possíveis em seu projeto (da aquisição de dados à modelagem)

Além dos pontos apresentados, é aconselhável que você traga visualizações de seus dados. Essas visualizações podem ser referentes à análise exploratória de dados ou mesmo à performance do(s) algoritmo(s) utilizados por você durante a realização do projeto. Utilize visualizações que auxiliem os leitores na compreensão das decisões tomadas por você para realizar o projeto.
Caso você deseje ou julgue necessário, pode também criar *dashboards* com múltiplas visualizações se o tipo de projeto que estiver realizando for adaptável a essa ferramenta.

**Bibliotecas sugeridas para esta etapa:**
Matplotlib, Seaborn, Plotly, Bokeh, Altair entre outras.

## VII - Implementação e manutenção (opcional)

Esta etapa deve ser feita fora do ambiente do Jupyter Notebook.
Caso se aplique, este é o momento de fazer com que todo o processo que você desenvolveu se torne utilizável por outras pessoas.
Durante essa etapa, alguns dos procedimentos que podem ser realizados são:

- Exportação do modelo
- Criação e conexão das funções em uma página web
- Criação de um arquivo que indique as dependências do projeto
- Criação de um repositório (Git)
- Implementação na plataforma escolhida

**Bibliotecas sugeridas para esta etapa:**
Flask, Streamlit entre outras.

awari.

**Plataformas sugeridas para esta etapa:**
Heroku, Amazon AWS, Google Cloud, Azure Cloud