

# FairPEK - Documentação

Instalação

February 6, 2024

## 1 Introdução

Esta documentação foi criada com o objetivo de guiar o Desenvolvedor de Software a entender, configurar e manter este sistema, que é dividido em 4 módulos principais:

- **Engenharia de dados:** Módulo criado com o objetivo de simular processos de transformação e limpeza de dados.
- **Módulo de Machine Learning:** Módulo que executa um Pipeline capaz de automatizar uma aplicação de Machine Learning (ML), com estágios de preparação de dados (Pré-processamento), treinamento (Processamento) e avaliação dos resultados (Pós-processamento) para a geração de um modelo final.
- **Gerenciador Autônomo:** Módulo contendo um loop, baseado na arquitetura MAPE-K, que controla o Módulo de ML como um Elemento Gerenciado para automatizar parte das atividades a serem executadas.
- **Interface:** : Módulo cujo objetivo é prover uma experiência de usuário mais simples e intuitiva para configurar e iniciar o Gerenciador Autônomo. É composto por dois componentes:
  - **Frontend:** Componente visual, exibido em um navegador.
  - **Backend:** Componente no qual o Frontend estabelece comunicação para obter os dados e montar o visual corretamente, de forma que corresponda a configurações utilizadas pelo Gerenciador Autônomo.

## 2 Programas necessários para instalação

- **Python:** É a linguagem de programação utilizada para montar e executar todos os módulos com a exceção do Frontend da interface. É disponível no site <https://www.python.org/> e é necessária a versão **3.8**, **3.9** ou **3.10**. Versões inferiores ou iguais a **3.7** a superiores ou iguais a **3.11** apresentaram problemas de compatibilidade devido a mudanças de design.

- **Node.js:** É o programa necessário para montar o Frontend da interface. É disponível no site <https://nodejs.org/> e foi testado na versão **16.14.2**, embora outras versões podem ser executadas sem problemas de compatibilidade.
- **Git:** É o programa necessário para realizar o download do código-fonte e realizar atualizações no mesmo. É disponível no site <https://git-scm.com/> e foi testado na versão **2.35.1**, embora outras versões podem ser executadas sem problemas de compatibilidade.
- **CUDA Toolkit:** É a biblioteca necessária para rodar alguns dos algoritmos presentes no Módulo de ML. É disponível no site <https://developer.nvidia.com/cuda-toolkit-archives> e foi testado na versão **11**, mas não houve testes se versões posteriores são compatíveis. É compatível apenas para GPUs Nvidia, caso não tiver há uma versão apenas para CPUs disponível.

### 3 Instalação do sistema

A partir desta parte, os exemplos serão realizados utilizando o Git Bash no Sistema Operacional Windows. Entretanto, no Linux e no Mac os passos são semelhantes por ambos também utilizarem esta linha de comando.

#### 3.1 Obtenção do código-fonte

O sistema se encontra no repositório <https://github.com/tenazatto/MsC>. Para obter seu código-fonte, basta digitar o seguinte comando:

```
git clone https://github.com/tenazatto/MsC.git
```

O Git baixará todos os arquivos e após o download é possível ver a pasta e seus arquivos na pasta **MsC**

#### 3.2 Montagem de ambiente

Para evitar problemas de versão com bibliotecas de outros projetos instalados, é possível criar um ambiente virtual para realizar a instalação das bibliotecas separadamente. Para criar, é necessário o **virtualenv** instalado no Python. Caso ele não esteja instalado, ele é obtido através do comando:

```
pip install virtualenv
```

Para criar um novo ambiente virtual, é preciso digitar o comando

```
python3 -m venv ./(nome do ambiente)
```

```
MINGW64:/c/Users/tenaz/OneDrive/Documentos/Pós

tenaz@DELL-G15-THA MINGW64 ~/OneDrive/Documentos/Pós
$ git clone https://github.com/tenazatto/MsC.git
Cloning into 'MsC'...
remote: Enumerating objects: 833, done.
remote: Counting objects: 100% (108/108), done.
remote: Compressing objects: 100% (90/90), done.
remote: Total 833 (delta 30), reused 55 (delta 18), pack-reused 725
Receiving objects: 100% (833/833), 42.62 MiB | 6.11 MiB/s, done.
Resolving deltas: 100% (367/367), done.

tenaz@DELL-G15-THA MINGW64 ~/OneDrive/Documentos/Pós
$
```

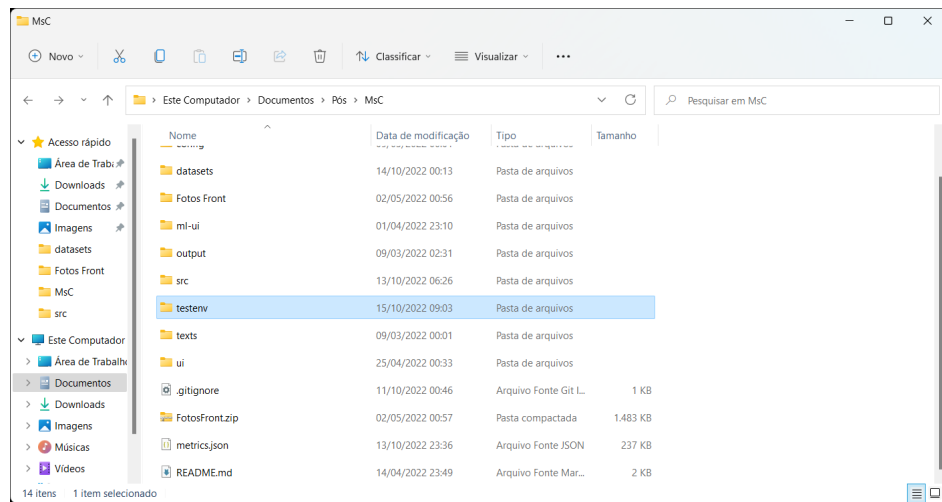
Como exemplo, nesta documentação foi criado o documento **testenv**

```
MINGW64:/c/Users/tenaz/OneDrive/Documentos/Pós/MsC

tenaz@DELL-G15-THA MINGW64 ~/OneDrive/Documentos/Pós/MsC (master)
$ python3 -m venv ./testenv

tenaz@DELL-G15-THA MINGW64 ~/OneDrive/Documentos/Pós/MsC (master)
$
```

Após o término, aparecerá uma nova pasta de mesmo nome



Após criar o ambiente virtual, é preciso ativá-lo para utilizar

`.\(nome do ambiente)\Scripts\activate.bat` (Windows - Prompt de Comando)

`source ./(nome do ambiente)/Scripts/activate` (Windows - Bash)

`source ./(nome do ambiente)/bin/activate` (Linux)

Para verificar se o ambiente foi ativado, é possível verificar, ao digitar qualquer comando no bash, que o nome do ambiente virtual aparece logo abaixo.

```

MINGW64:/c/Users/tenaz/OneDrive/Documentos/Pós/MsC
tenaz@DELL-G15-THA MINGW64 ~/OneDrive/Documentos/Pós/MsC (master)
$ python3 -m venv ./testenv

tenaz@DELL-G15-THA MINGW64 ~/OneDrive/Documentos/Pós/MsC (master)
$ source ./testenv/Scripts/activate
(testenv)
tenaz@DELL-G15-THA MINGW64 ~/OneDrive/Documentos/Pós/MsC (master)
$ python --version
Python 3.8.10
(testenv)
tenaz@DELL-G15-THA MINGW64 ~/OneDrive/Documentos/Pós/MsC (master)
$ echo $VIRTUAL_ENV
C:\Users\tenaz\OneDrive\Documentos\Pós\MsC\testenv
(testenv)
tenaz@DELL-G15-THA MINGW64 ~/OneDrive/Documentos/Pós/MsC (master)
$ ls
'Fotos Front'/'  README.md  datasets/  ml-ui/  src/  texts/
FotosFront.zip  config/  metrics.json  output/  testenv/  ui/
(testenv)
tenaz@DELL-G15-THA MINGW64 ~/OneDrive/Documentos/Pós/MsC (master)
$

```

Para desativar o ambiente virtual, é preciso digitar o comando

### **deactivate**

Para verificar se o ambiente foi desativado, é possível verificar, ao digitar qualquer comando no bash, que o nome do ambiente virtual não irá mais aparecer até ser ativado novamente.

No caso do Node.js não é necessário realizar tais etapas, pois a instalação das bibliotecas nesta documentação é realizada de maneira local

## **3.3 Instalação das bibliotecas**

### **3.3.1 Python**

Com o ambiente virtual criado e ativado, é possível utilizar os arquivos **src/requirements.txt**, dependendo das configurações da sua máquina, para instalar todas as bibliotecas necessárias através do comando

```
pip install -r ./src/requirements-nvidia.txt (GPU Nvidia)
```

```
pip install -r ./src/requirements-cpu.txt (CPU)
```

Estes arquivos foram preparados apenas para rodar de acordo com o hardware apresentado. A execução de ambos os comandos não é necessária e nem recomendável.

### **3.3.2 Node.js**

Para o Node.js, como o arquivo **package.json** está dentro da pasta **ml-ui**, é possível acessar essa pasta e digitar o comando

```
npm install
```

## **4 Execução do sistema**

### **4.1 Engenharia de Dados**

Dentro da pasta **MsC** e com o ambiente virtual criado e ativado, para rodar o módulo de Engenharia de Dados basta digitar o seguinte comando

```
python -m src.data_engineering.data_engineering_start -data (Opção)
```

No momento, há 3 opções disponíveis:

- **GERMAN\_CREDIT:** Manipula o German Credit Dataset, cujo arquivo está na localização **datasets/german.data**, para utilização no Módulo de ML.
- **LENDINGCLUB:** Baixa e manipula o Lendingclub Dataset para utilização no Módulo de ML.
- **METRICS:** Obtém o maior valor, menor valor e a média de cada métrica para cada execução já realizada no Módulo de ML.

## 4.2 Módulo de ML

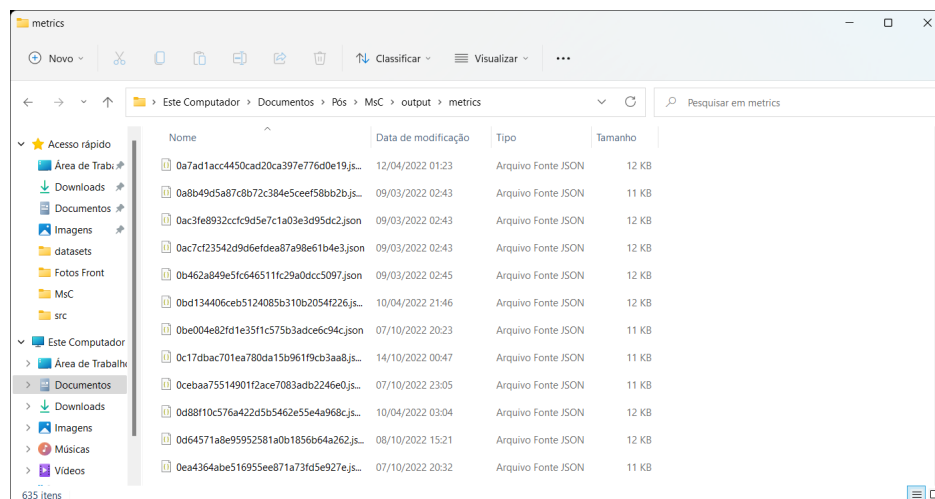
Dentro da pasta **MsC**, com o ambiente virtual criado e ativado e com dados já tratados pelo módulo de Engenharia de Dados, para rodar todos os Pipelines presentes no Módulo de ML basta digitar o seguinte comando

```
python -m src.pipeline.pipeline_start --dataset (Opção)
```

Há 4 opções disponíveis para execução:

- **ADULT\_INCOME\_SEX:** Executa os Pipelines do Módulo de ML para o Adult Income Dataset, cujo arquivo está na localização **datasets/adult.csv**, utilizando Sexo (Masculino/Feminino) como atributo protegido.
- **GERMAN\_CREDIT\_FOREIGN:** Executa os Pipelines do Módulo de ML para o German Credit Dataset, cujo arquivo é manipulado na etapa anterior, utilizando Nacionalidade (Alemão/Estrangeiro) como atributo protegido.
- **GERMAN\_CREDIT\_AGE:** Executa os Pipelines do Módulo de ML para o German Credit Dataset, cujo arquivo é manipulado na etapa anterior, utilizando Idade (-25 anos/25 ou + anos) como atributo protegido.
- **LENDINGCLUB\_INCOME:** Executa os Pipelines do Módulo de ML para o Lendingclub Dataset, cujo arquivo é manipulado na etapa anterior, utilizando Renda (-1 salário mínimo/1 ou + salários mínimos) como atributo protegido.

Após a execução, é possível ver a geração das métricas dentro da pasta **output/metrics**, necessárias para a execução do Gerenciador Autônomo.



### 4.3 Gerenciador Autônomo

Dentro da pasta **MsC**, com o ambiente virtual criado e ativado e com pelo menos uma execução do Módulo de ML realizada, é possível verificar como o Gerenciador Autônomo funciona com o seguinte comando

```
python -m src.mapek.mapek_start
```

Nesta etapa, ele escolhe o Pipeline que apresentou as melhores métricas, porém a filtragem por conjunto de dados foi desenvolvida apenas na Interface. Como ele roda ininterruptamente, é preciso interromper sua execução.

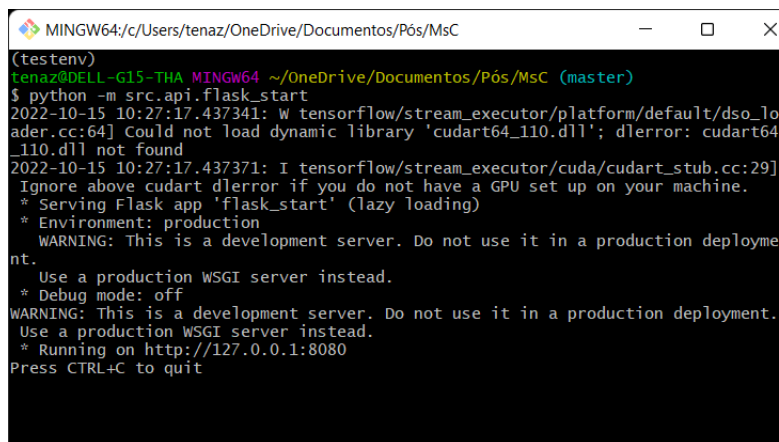
### 4.4 Interface

#### 4.4.1 Backend

Dentro da pasta **MsC**, com o ambiente virtual criado e ativado e com pelo menos uma execução do Módulo de ML realizada, é possível rodar o Backend da interface com o seguinte comando

```
python -m src.api.flask_start
```

Ele vai iniciar um servidor na porta 8080, necessário para rodar as requisições que o Frontend vai solicitar

A screenshot of a Windows terminal window titled 'MINGW64:/c/Users/tenaz/OneDrive/Documentos/Pós/MsC'. The terminal shows the execution of the command 'python -m src.api.flask\_start'. The output includes a warning about the development server, the Flask app name 'flask\_start', the environment 'production', and the server running on 'http://127.0.0.1:8080'. There is also a warning about the development server and a message to use a production WSGI server instead. The terminal text is as follows:

```
(testenv)
tenaz@DELL-G15-THA MINGW64 ~/OneDrive/Documentos/Pós/MsC (master)
$ python -m src.api.flask_start
2022-10-15 10:27:17.437341: W tensorflow/stream_executor/platform/default/dso_loader.cc:64] Could not load dynamic library 'cudart64_110.dll'; dlderror: cudart64_110.dll not found
2022-10-15 10:27:17.437371: I tensorflow/stream_executor/cuda/cudart_stub.cc:29] Ignore above cudart dlerror if you do not have a GPU set up on your machine.
* Serving Flask app 'flask_start' (lazy loading)
* Environment: production
  WARNING: This is a development server. Do not use it in a production deployment.
  Use a production WSGI server instead.
* Debug mode: off
WARNING: This is a development server. Do not use it in a production deployment.
Use a production WSGI server instead.
* Running on http://127.0.0.1:8080
Press CTRL+C to quit
```

#### 4.4.2 Frontend

Dentro da pasta **MsC/ml-ui**, é possível rodar o Frontend da interface com o seguinte comando

```
npm start
```

Ele vai iniciar o navegador acessando um servidor na porta 3000, e deverá iniciar a tela no menu de Análise

The screenshot displays the FairPEK web application interface. On the left is a sidebar menu with the following items: 'Configuração' (Configuration), 'Análise' (Analysis), 'Planejamento' (Planning), 'Execução' (Execution), 'Pipeline Manual', and 'Pipeline Autônomo'. The main content area is titled 'FairPEK' and 'Análise'. It features two tabs: 'MÉTRICAS DE AVALIAÇÃO' (Evaluation Metrics) and 'MÉTRICAS DE FAIRNESS' (Fairness Metrics). Under 'Métricas de Avaliação', there are sliders for 'Peso p/ Avaliação' (Weight for Evaluation) and a dropdown menu for 'Métricas p/ uso' (Metrics for use). The dropdown menu is open, showing 'Accuracy, Precision, Recall, F1-Score, AUC Score'. Below this, there are sliders for 'Accuracy', 'Precision', 'Recall', and 'AUC Score', each with a 'Peso p/ Avaliação' (Weight for Evaluation) slider set to 20%.

Métrica	Peso p/ Avaliação
Accuracy	20%
Precision	20%
Recall	20%
AUC Score	20%