## Projeto Integrador 4: Análise de fechamentos de empresas a partir das informações do CNPJ usando IA

Tema do PI

**Desenvolver análise de dados em escala utilizando algum conjunto de dados existentes ou capturados por IoT e aprendizagem de máquina. Preparar uma interface para visualização dos resultados.**

## Etapas da análise:

# 1. ETL – Extração, tratamento e limpeza dos dados

## 1.1 Coleta dos Dados

Foram coletados os arquivos disponíveis publicamente no site do Portal Dados Abertos do Governo Federal. Os arquivos estavam em formato compactado (“zip”). Eles foram extraídos para a pasta dados\_csv do projeto e renomeados.

Vários algoritmos foram desenvolvidos em python para tratar os dados e enviá-los para um banco de dados Postgres.

Geramos um novo arquivo com os dados segmentados da cidade de Macatuba

# 2. Exploração dos dados

## 2.1 Filtrar empresas inativas

## 2.2 Analisar Evolução Temporal dos Fechamentos

- Identificar colunas com a data de abertura e fechamento das empresas.

- Criar coluna com o ano de fechamento e o tempo de funcionamento (data de fechamento – data de abertura)

- Criar coluna com o ano de abertura

- Criar coluna com o ano de fechamento

- Criar gráficos:

- Aberturas x Fechamentos por Ano

- % do total de empresas total fechadas por ano

- evolução no número de CNPJs por tipo de empresa (MEI, Simples, Grande Porte)

## 2.3 Identificar setores mais afetados

- agrupar empresas por CNAE principal e número de fechamentos por ano (10 primeiros)

- gráfico de setores com maior número de fechamentos

## 2.4 – Distribuição geográfica dos fechamentos

Gráfico de calor por UF com mais fechamentos

# 3 – Aplicação de Machine Learning

## 3.1 – Predição de fechamento de empresas

* **Modelo:** Construir um modelo de classificação para prever se uma empresa irá fechar ou não com base em suas características.
* **Variáveis Preditivas:** Utilizar informações como CNAE, porte da empresa, capital social, tempo de atividade, número de sócios, etc.
* **Avaliação:** Empregar métricas como acurácia, precisão, recall e F1-score para avaliar a performance do modelo.
* **Insights:** Identificar os principais fatores que contribuem para o fechamento de empresas e, potencialmente, criar um sistema de alerta para empresas em risco.

## 3.2 – Agrupamento de Empresas (clustering):

* **Modelo:** Aplicar algoritmos de agrupamento para identificar grupos de empresas com características semelhantes em relação ao fechamento.
* **Variáveis de Agrupamento:** Utilizar as mesmas variáveis preditivas do modelo de classificação, além de outras como motivo de fechamento, se disponível.
* **Análise dos Clusters:** Analisar as características de cada cluster para entender os diferentes perfis de empresas que tendem a fechar.
* **Insights:** Identificar grupos de empresas com maior risco de fechamento e suas características comuns, possibilitando ações preventivas direcionadas.

## 3.3 – Análise de Sobrevivência

* **Modelo:** Utilizar modelos de análise de sobrevivência para estimar o tempo até o fechamento de uma empresa.
* **Variáveis Preditivas:** As mesmas utilizadas nos modelos anteriores, além da data de início da atividade da empresa.
* **Visualização:** Construir curvas de sobrevivência para visualizar a probabilidade de uma empresa sobreviver ao longo do tempo, comparando diferentes grupos de empresas (por CNAE, porte, etc.).
* **Insights:** Analisar como diferentes fatores afetam a longevidade das empresas e identificar perfis de empresas com maior ou menor tempo de sobrevivência.