# **Projeto: Análise de Fluxo Turístico Sustentável em Angola**

## **1. Preparação de Dados / Engenharia de Recursos**

### 1.1 Visão Geral

A fase de preparação de dados e engenharia de recursos é crucial em projetos de aprendizado de máquina. Ela garante que os dados estejam limpos, consistentes e no formato adequado, permitindo que os modelos aprendam padrões relevantes. Sem essa etapa, mesmo os melhores modelos podem apresentar desempenho ruim devido a problemas nos dados.

### 1.2 Coleta de Dados

* Fonte: Dados públicos de turismo em Angola, contendo informações de chegadas e saídas de turistas por continente (2019–2022) e por província (2019–2022).
* Pré-processamento durante a coleta:
  + Leitura de arquivos CSV usando pandas.
  + Padronização de nomes de colunas (str.strip(), str.lower(), remoção de acentos).
  + Conversão de colunas numéricas para o tipo correto (pd.to\_numeric).

### 1.3 Limpeza de Dados

* Remoção de linhas com valores ausentes.
* Conversão de colunas numéricas que poderiam conter strings para numeric.
* Verificação de inconsistências nos nomes de colunas e categorias.

### 1.4 Análise Exploratória de Dados (EDA)

* Objetivo: Entender padrões de chegada e saída de turistas ao longo do tempo.
* Visualizações realizadas:
  + Chegadas por continente (2020 e 2021).
  + Chegadas e saídas agregadas por ano (2019–2022).
  + Chegadas por província (Luanda, Benguela, Huíla) e tendências futuras.

### 1.5 Engenharia de Recursos

* Criação de agregados por ano para análises nacionais.
* Seleção de colunas relevantes para análise por província.
* Nenhum recurso derivado adicional foi necessário, pois a análise focou em séries temporais simples.

### 1.6 Transformação de Dados

* Não foi necessário escalonamento ou codificação, pois os modelos utilizados (Linear Regression e RandomForest) trabalham diretamente com valores numéricos contínuos.

## **2. Exploração de Modelos**

### 2.1 Seleção de Modelo

* Linear Regression: Modelo simples, permite visualizar tendência de crescimento ou queda ao longo dos anos.
* RandomForest Regressor: Captura não linearidades e padrões complexos de dados agregados.
* Justificativa: Combinação de um modelo interpretável (Linear Regression) e um modelo mais robusto (RandomForest) para comparar desempenho.

### 2.2 Treinamento de Modelo

* Dados usados: agregados por ano (2019–2022).
* Variáveis:
  + X = ano
  + y = chegadas\_turistas
* Hiperparâmetros RandomForest:
  + n\_estimators = 200
  + random\_state = 42
* Validação: in-sample (devido a poucos anos de dados).

### 2.3 Avaliação do Modelo

* Métricas utilizadas:
  + R² para ajuste in-sample.
  + RMSE para medir o erro médio absoluto da previsão.
* Resultados (agregado nacional):
  + Linear Regression: R² ≈ 0,41, RMSE ≈ 4847
  + RandomForest: R² ≈ 0,79, RMSE ≈ 2867

### 2.4 Previsões

* Período: 2022–2025
* Resultados agregados nacionais:
  + Linear Regression: 62907, 59298, 55688, 52079
  + RandomForest: 65754, 65754, 65754, 65754
* Exemplo por província:
  + Luanda: 27719, 27445, 27171, 26897
  + Benguela: 8422, 8551, 8679, 8808
  + Huíla: 5359, 5399, 5439, 5478