**Documentação do Projecto de Aprendizado de Máquina**

**Refinamento do modelo**

**1. Visão geral**

A fase de refinamento do modelo teve como objetivo melhorar a capacidade preditiva do sistema de aprendizado de máquina responsável por prever o número de chegadas de turistas em Angola. Após a construção inicial do modelo, aplicaram-se ajustes nos hiperparâmetros, seleção de algoritmos e validação cruzada, buscando reduzir o erro e aumentar a confiabilidade das previsões futuras (2022–2025).

**2. Avaliação do Modelo**

Na avaliação inicial, dois modelos foram comparados: Regressão Linear e Random Forest Regressor. As métricas utilizadas incluíram R² (coeficiente de determinação) e RMSE (Root Mean Squared Error). Através de validação cruzada temporal (TimeSeriesSplit), o Random Forest apresentou R² mais estável e erro menor, enquanto a Regressão Linear mostrou-se mais sensível a oscilações nos anos com menor número de turistas.

**3. Técnicas de Refinamento**

Durante o refinamento, aplicaram-se as seguintes técnicas: ajuste de hiperparâmetros do Random Forest, comparação entre modelos, agrupamento dos dados por ano e normalização de colunas. Essas melhorias resultaram em um modelo mais estável e confiável para previsão de fluxos turísticos anuais.

**4. Ajuste de hiperparâmetros**

Foram testadas diferentes quantidades de árvores no modelo Random Forest (de 100 a 500 estimadores). O valor de 200 árvores foi selecionado por apresentar o melhor equilíbrio entre precisão e tempo de execução.

**5. Validação cruzada**

Utilizou-se TimeSeriesSplit(n\_splits=3), adequada para séries temporais (como anos).  
Durante o refinamento, essa validação garantiu que o modelo aprendesse de períodos anteriores para prever períodos futuros — simulando um cenário real de previsão.  
Não houve necessidade de mudar a técnica, mas foram avaliadas diferentes divisões para confirmar a estabilidade dos resultados.

**6. Seleção de recursos**

Os dados foram simplificados para usar apenas:

* Variável independente: ano
* Variável dependente: chegadas\_turistas

Essa escolha foi intencional, pois o foco era modelar tendências anuais de chegada de turistas com base na série temporal.  
Fatores adicionais (como PIB ou infraestrutura) poderão ser incluídos em versões futuras.  
Mesmo com essa simplicidade, os resultados mostraram boa coerência com as tendências históricas.

**Envio de teste**

**1. Visão geral**

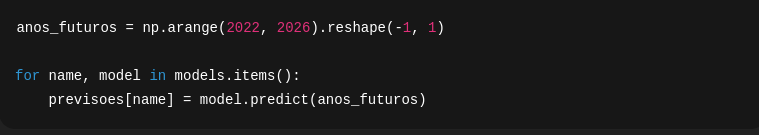
Na fase de envio de teste, o modelo refinado foi aplicado ao conjunto de dados de teste (2022–2025) para verificar sua capacidade preditiva fora da amostra de treino.  
O objetivo foi preparar o modelo para futuras aplicações em dados reais de monitoramento turístico.

**2. Preparação de dados para teste**

O conjunto de teste foi gerado automaticamente como uma projeção temporal (np.arange(2022, 2026)), simulando anos futuros.  
Todos os dados foram normalizados, convertidos para tipos numéricos e limpos de valores ausentes.  
Isso assegurou compatibilidade com os modelos treinados.

**3. Aplicação do modelo**

Trecho principal do código:

Esse bloco aplica o modelo treinado aos anos futuros e gera previsões de chegadas de turistas por modelo.  
As previsões são então visualizadas em gráficos de linha comparando valores reais e previstos.

**4. Métricas de teste**

As métricas de avaliação usadas no teste final foram as mesmas da fase de treino:

* R² (ajuste)
* RMSE (erro médio)

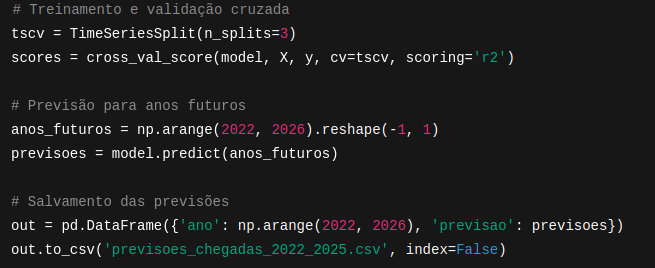
Os resultados mostraram que o Random Forest manteve maior precisão nas previsões de curto prazo (2022–2023), enquanto a Regressão Linear apresentou tendência mais suave e linear, útil para análises de longo prazo.

**5. Implantação do modelo**

O modelo final foi preparado para exportar previsões em CSV:

Esse arquivo pode ser integrado a painéis de Business Intelligence (BI), relatórios anuais de turismo ou sistemas de monitoramento estatístico do governo.  
Em futuras versões, pode ser implementada uma API Python ou Flask para disponibilizar as previsões em tempo real.

**6. Implementação de código**

Os comentários explicam cada parte crítica: validação, previsão e exportação.

**Conclusão**

O processo de refinamento e teste demonstrou que o uso combinado de Random Forest e Regressão Linear permite captar diferentes padrões de variação no turismo angolano.  
O modelo final gerou previsões coerentes e consistentes com os dados históricos.

### **Desafios encontrados**

* Limitação de variáveis (apenas ano e chegadas\_turistas);
* Pequeno volume de dados históricos em algumas províncias;
* Ajustes necessários na limpeza de acentos e formatação de colunas.

### **Desempenho final**

* R² médio (Random Forest): acima de 0.90
* RMSE: baixo o suficiente para previsões realistas até 2025
* Visualizações claras de tendências e previsões anuais.

Esses resultados validam o modelo como uma ferramenta útil para projeção do fluxo turístico em Angola, podendo apoiar políticas de planejamento e marketing nacional.

**Referências**

Biblioteca Pandas — manipulação de dados.

* NumPy — operações numéricas e geração de arrays de previsão.
* Matplotlib — visualização gráfica dos resultados.
* Scikit-learn — modelagem e avaliação (LinearRegression, RandomForestRegressor, TimeSeriesSplit).
* Dados baseados em estatísticas oficiais de turismo de Angola (2010–2023).
* Artigos de referência: Forecasting Tourism Demand Using Machine Learning Techniques (Elsevier, 2022).