Documentação Técnica: Mapeamento de Locais Turísticos e Análise Preditiva de Desenvolvimento Turístico em Angola com IA

1. Resumo Executivo

Este documento detalha a criação de um modelo base (baseline) de Machine Learning para prever o potencial de desenvolvimento turístico em Angola. Utilizando uma abordagem de **Regressão**, o modelo foi treinado para prever o Índice de Desenvolvimento Humano (IDH) de uma localidade com base em características geográficas, climáticas e socioeconómicas. A metodologia emprega um **do Scikit-learn** para garantir robustez e reprodutibilidade, encapsulando o pré-processamento de dados e um modelo **`RandomForestRegressor`**. Os resultados iniciais, baseados num dataset limitado e parcialmente sintético de 15 amostras, indicam um desempenho modesto (R^2=0.211). Este resultado serve como um diagnóstico honesto da necessidade de mais dados e valida a arquitetura do modelo como uma base sólida para futuros desenvolvimentos. O artefato final é um ficheiro `tourism_model.pkl`, um sistema preditivo completo e autossuficiente, cuja aplicação foi demonstrada através de um mapa de visualização interativo.

2. Introdução

2.1. Problema de Negócio

A identificação de novas áreas com alto potencial para desenvolvimento turístico em Angola é um desafio estratégico. A avaliação é muitas vezes subjetiva e carece de uma base quantitativa. Este projeto visa resolver este problema através da criação de uma ferramenta de ciência de dados para apoiar a tomada de decisão de forma objetiva.

2.2. Objetivo do Projeto

O objetivo principal é desenvolver um primeiro modelo funcional (baseline) que preveja uma métrica de "potencial turístico". Escolhemos o **IDH** como variável alvo, pois ele serve como um excelente *proxy* para as condições de infraestrutura, segurança e capital humano necessárias para o florescimento do turismo.

3. Metodologia

3.1. Fonte e Natureza dos Dados

O modelo foi treinado com o dataset `model_input.csv`, contendo 15 observações e 18 colunas. As features incluem:

- **Dados Geográficos:** Latitude, Longitude, Altitude.
- **Dados Climáticos:** Temperatura média, Precipitação anual.
- **Dados Ambientais (Satélite):** NDVI, EVI, NDWI.
- **Dados Socioeconómicos:** População, Densidade, PIB per capita, Taxa de urbanização.
- **Dados Categóricos:** Nome do Ponto Turístico, Província.

É crucial notar que, para garantir a completude do dataset nesta fase de prototipagem, foi utilizada uma combinação de **dados reais e sintéticos (fictícios)**. Esta decisão, embora necessária, tem implicações diretas na interpretação dos resultados do modelo, como será discutido na seção de análise.

3.2. Abordagem de Modelagem

Foi escolhida uma abordagem de **Regressão** para prever o valor contínuo da nossa métrica alvo. Os dados são **tabulares estruturados**, onde cada linha representa uma observação independente, tornando os modelos de ML Clássico a escolha mais adequada.

3.3. Justificativa da Variável Alvo (Target): O IDH

Para traduzir o conceito abstrato de "potencial turístico" numa métrica quantificável, escolhemos o **Índice de Desenvolvimento Humano (IDH)** como a nossa variável alvo. Esta foi uma decisão estratégica pelas seguintes razões:

- **É a "Espinha Dorsal" do Desenvolvimento:** O IDH é um indicador compósito que mede a saúde, a educação e o padrão de vida de uma região. Um IDH elevado implica a existência da infraestrutura fundamental que qualquer tipo de turismo de qualidade exige.
- **Captura Fatores Indiretos Essenciais:**
    - **Saúde e Estabilidade:** Um componente de "vida saudável" forte sugere um ambiente seguro e estável.
    - **Capital Humano:** Um componente de "conhecimento" elevado indica uma força de trabalho qualificada.
    - **Infraestrutura e Renda:** Um "padrão de vida digno" está diretamente correlacionado com a qualidade de estradas, energia, telecomunicações e serviços.
- **É um Alvo Robusto e Universal:** Ao invés de prever um indicador volátil, o IDH oferece uma visão holística e estável do nível de desenvolvimento de uma área. Ao treinar o modelo para prever o IDH, estamos a ensiná-lo a identificar locais que possuem o **ecossistema subjacente necessário para que o turismo floresça**.

3.4. Justificativa da Escolha do Modelo

Optou-se pelo `RandomForestRegressor`, um modelo da família Bagging, em detrimento de abordagens de Deep Learning (Redes Neurais), pois modelos baseados em árvores são o estado-da-arte para dados tabulares de pequena escala e são mais robustos contra o overfitting nestas condições.

3.5. Arquitetura do Pipeline

A implementação foi feita através de um `Pipeline` do Scikit-learn para garantir as melhores práticas.

1. Aplica transformações distintas a diferentes tipos de colunas: `StandardScaler` para as numéricas e `OneHotEncoder` para as categóricas.
2. O modelo de regressão que recebe os dados já processados.

4. Resultados e Análise

4.1. Desempenho do Modelo

O pipeline foi treinado em 12 amostras e avaliado em 3 amostras de teste.

- **R^2 (Coeficiente de Determinação):** 0.211
- **RMSE (Raiz do Erro Quadrático Médio):** 0.058

A baixa pontuação do R^2 é um resultado esperado, indicando que o modelo, com a quantidade e natureza atuais dos dados, tem um poder preditivo limitado.

4.2. Análise de Importância das Variáveis

As features mais influentes, segundo o modelo, foram `taxa_urbanizacao`, `lat_clima`, e `distancia_cidade_km`. Esta análise oferece uma hipótese inicial, mas a sua fiabilidade aumentará com a melhoria do dataset.

4.3. Limitações e Análise Crítica dos Resultados

Uma parte fundamental da validação do modelo é a análise crítica das suas previsões. Um exemplo notável emergiu ao comparar a previsão para o **Cristo Rei (Huambo)** com as **Quedas de Calandula (Malanje)**. O modelo atribuiu um potencial predito mais baixo ao primeiro (0.586) do que ao segundo (0.609), um resultado contraintuitivo, dado o maior nível de infraestrutura da província do Huambo.

Esta discrepância não é uma falha do algoritmo, mas um sintoma direto das limitações do nosso dataset de treino:

1. **Impacto de Dados Sintéticos:** Os dados fictícios, criados para preencher lacunas, podem introduzir vieses e não capturar perfeitamente as complexas realidades locais.
2. **Sensibilidade a um Pequeno Dataset:** Com apenas 12 amostras de treino, o modelo é altamente suscetível a "ruído", e uma única observação com características ligeiramente enviesadas pode influenciar significativamente as regras aprendidas.

Este resultado prático serve como a justificação mais forte para a principal recomendação deste projeto: a necessidade de expandir o dataset com **dados reais e verificados**.

5. Artefato Produzido e Aplicação Prática

O resultado final do processo de treino é o ficheiro `tourism_model.pkl`. Este ficheiro contém o objeto `Pipeline` completo. Como demonstração, este pipeline foi carregado no notebook `mapa_visualization.ipynb` e usado para gerar previsões que foram plotadas num mapa interativo de Angola. Esta visualização serve como a **prova de conceito final** do projeto, transformando um modelo de Machine Learning num produto de análise geoespacial tangível.

6. Conclusão e Próximos Passos

O projeto atingiu com sucesso o seu objetivo de criar um **modelo base funcional, metodologicamente robusto e com uma aplicação prática demonstrada**. A principal conclusão é que a arquitetura do modelo é sólida, mas o seu desempenho e fiabilidade estão diretamente limitados pela qualidade e quantidade dos dados de entrada.

Os próximos passos recomendados são:

1. **Expansão e Verificação do Dataset (Prioridade Máxima):** Priorizar a coleta de dados de mais pontos turísticos e a substituição de todos os dados sintéticos por dados reais.
2. **Experimentação com Modelos:** Após a expansão do dataset, testar outros algoritmos de ML Clássico (como XGBoost) para otimizar a precisão.
3. **Regressão Multi-Output:** Evoluir o modelo para prever múltiplos indicadores de desenvolvimento simultaneamente (ex: IDH, emprego, PIB per capita, Mobilidade etc).
4. **Exploração de Deep Learning:** Com um dataset substancialmente maior, investigar o uso de Redes Neurais para dados tabulares ou substituir o dataset por dados não tabulares(Imagens, áudio etc.).