

Relatório Técnico - Challenge Ingredion – SPRINT 3

Correlação do Modelo de IA com Dados Reais de Produtividade Agrícola

Projeto Ingredion – FIAP ON 2024 IA

Grupo: Amandha Nery, Bruno Conterato, Gustavo Castro, Kild Fernandes, Luis Emidio

Data de Entrega: 26/05/2025

1. Introdução

Este relatório documenta a SPRINT 3 do Challenge Ingredion, focada na validação do modelo de Inteligência Artificial (IA) desenvolvido na SPRINT 2. O objetivo principal é correlacionar as previsões de produtividade agrícola do modelo com dados reais históricos, avaliando sua confiabilidade e precisão. A cultura analisada foi a produção de café em Manhuaçu (MG), buscando dados do IBGE.

2. Metodologia

A metodologia empregou as seguintes etapas:

2.1. Coleta de Dados de Produtividade Histórica

A coleta de dados históricos de produtividade agrícola foi realizada por meio de pesquisa em bases de dados públicas, priorizando o **SIDRA/IBGE (Sistema IBGE de Recuperação Automática/Instituto Brasileiro de Geografia e Estatística)**. Os dados extraídos representam a produção de café em Manhuaçu, Minas Gerais, para o período de 1974 a 2023.

Os dados coletados incluíram:

- Produção (kg);
- Área colhida (hectares);
- Variável: Quantidade produzida (Toneladas);
- Ano/safra dos dados (1974-2023);

2.2. Coleta de Dados NDVI

Os dados NDVI foram extraídos utilizando o Google Earth Engine (GEE) para a mesma região de Manhuaçu. O período analisado também abrange de 2000-01-01 até 2025-12-31, utilizando imagens da coleção "MODIS/061/MOD13Q1". As imagens foram filtradas pela área de interesse e pela data.

Os dados coletados incluíram:

- NDVI médio;
- Data.

2.3. Tratamento e Preparação dos Dados

Os dados coletados foram organizados em dataframes utilizando a biblioteca Pandas do Python. Os scripts de tratamento de dados incluíram:

- **Tratamento de dados de Produção (prod_manhuacu):**
 - Leitura do arquivo do IBGE "tabela1613.xlsx".
 - Filtro para dados de Manhauçu.
 - Renomeação das colunas para representar os anos.
 - Conversão da produção de toneladas para kg.
- **Tratamento de dados de Área (area_manhuacu):**
 - Leitura do arquivo do IBGE "tabela1613.xlsx".
 - Filtro para dados de Manhauçu.
 - Renomeação das colunas para representar os anos.
- **Merge dos dados de produção e área (manhuacu_df):**
 - Combinação dos dataframes de produção e área por meio da coluna "Year".
 - Cálculo da produtividade (kg/ha) dividindo a produção pela área.
- **Tratamento de dados NDVI (NDVI_df):**
 - Leitura dos dados NDVI extraídos do GEE.
 - Renomeação da coluna "date" para "Data" e "ndvi" para "NDVI".
 - Extração do ano a partir da coluna "Data".
- **Conversão para CSV:**
 - Os dataframes tratados (manhuacu_df e NDVI_df) foram salvos em arquivos CSV para uso posterior.

2.4. Análise Estatística e Correlação

Para verificar a relação entre o NDVI médio e a produtividade real, foram aplicados os seguintes testes de correlação, utilizando as bibliotecas NumPy e SciPy do Python:

- **Correlação de Spearman (não linear):** Devido à natureza não necessariamente linear da relação entre NDVI e produtividade, a correlação de Spearman foi utilizada.
- **Regressão Linear:** Foi também realizada uma regressão linear simples, para melhor visualizar os dados e obter a equação de tendência e o coeficiente de determinação (R^2).

Os resultados foram visualizados por meio de gráficos de dispersão (scatter plots) com linha de tendência e gráficos comparativos por safra.

2.5. Extração de Dados NDVI com Google Earth Engine (GEE)

A extração dos dados NDVI da área de Manhauçu foi feita utilizando o Google Earth Engine (GEE), utilizando as imagens MODIS e realizando o processamento das imagens diretamente na plataforma.

Os principais passos foram:

- **Configuração da área de interesse:** Definição das coordenadas geográficas da área de Manhauçu.
- **Seleção de imagens:** Utilização de imagens da coleção "MODIS/061/MOD13Q1".
- **Filtragem por data:** Seleção de imagens para o período de 2000-01-01 até 2025-12-31.
- **Cálculo do NDVI médio:** Cálculo do NDVI médio para a região a cada data.
- **Exportação dos dados:** Exportação dos dados para um arquivo CSV.

3. Resultados e Discussão

3.1. Análise Exploratória dos Dados

A análise exploratória dos dados (EDA) identificou as seguintes características:

- **Dados de produção:** A produção de café em Manhauçu apresenta variabilidade ao longo dos anos, com picos e quedas que podem estar relacionados a fatores climáticos ou de manejo.
- **Dados de área:** A área colhida também apresenta variabilidade, indicando expansão ou contração da área plantada ao longo do tempo.
- **Dados NDVI:** Os dados NDVI mostram padrões sazonais, com valores mais altos durante os períodos de maior atividade vegetativa e valores mais baixos durante os períodos de menor atividade.

3.2. Correlação entre NDVI e Produtividade

Os resultados da correlação de Spearman indicaram uma **correlação moderada** entre o NDVI médio e a produtividade real. Isso sugere que o NDVI pode ser um indicador útil da produtividade, mas não é o único fator determinante.

Na análise da regressão linear, o coeficiente de determinação (R^2) indicou que o modelo linear explica aproximadamente a variância da produtividade em função do NDVI.

3.3. Fatores Externos Influenciadores

A análise revelou que fatores externos podem ter influenciado a produtividade, como eventos climáticos atípicos (secas, geadas, enchentes), a presença de pragas e doenças agrícolas, e a qualidade das imagens NDVI utilizadas.

4. Melhorias para o Modelo de IA

Com base nos resultados e na discussão, sugiro as seguintes melhorias para o modelo de IA:

- **Inclusão de novos tipos de dados:** Incorporar dados climáticos (precipitação, temperatura), dados de manejo (fertilização, irrigação) e dados de solo para aumentar a precisão do modelo.
- **Ajuste do período de coleta de NDVI:** Realizar testes com diferentes períodos de coleta de NDVI para identificar o período mais relevante para a previsão da produtividade.

- **Melhoria do tratamento de imagens:** Aplicar técnicas de pré-processamento mais robustas para remover ruídos e artefatos nas imagens NDVI, aumentando a qualidade dos dados de entrada.
- **Uso de séries temporais:** Usar modelos de séries temporais para analisar a evolução da produção agrícola ao longo do tempo e identificar tendências.
- **Machine Learning para prever a produção:** Utilizar modelos de Machine Learning que considerem o NDVI, dados climáticos e de manejo, para prever a produção com maior precisão.

5. Limitações da Análise

A análise apresenta as seguintes limitações:

- **Tamanho da amostra:** O tamanho da amostra de dados históricos pode ser considerado limitado, o que pode afetar a generalização dos resultados.
- **Qualidade das bases de dados públicas:** A qualidade e a precisão das bases de dados públicas podem variar, o que pode introduzir erros na análise.
- **Resolução espacial:** A resolução espacial dos dados NDVI pode ser um limitante, pois o NDVI médio para uma região pode não representar a variabilidade local.

6. Conclusão

Este relatório apresentou a metodologia e os resultados da SPRINT 3 do Challenge Ingredion, focada na validação do modelo de IA para previsão de produtividade agrícola. A análise de correlação e regressão revelou uma correlação moderada entre o NDVI médio e a produtividade real, além de identificar fatores externos que podem influenciar a produtividade. As sugestões de melhorias para o modelo de IA visam aumentar sua precisão e robustez, permitindo a integração em sistemas de apoio à decisão agrícola.

7. Referências

- IBGE (Instituto Brasileiro de Geografia e Estatística): <https://www.ibge.gov.br/>
- Sistema IBGE de Recuperação Automática - SIDRA: <https://sidra.ibge.gov.br/>
- Google Earth Engine: <https://earthengine.google.com/>

8. Anexos

8.1. Código-fonte no GitHub

O código-fonte utilizado para o tratamento dos dados, análise estatística e geração de gráficos está disponível no seguinte repositório GitHub:

github.com/luisfuturist/ec-ingredion-3