Enterprise Challenge - Sprint 3 - Ingredion

Participantes:

Marco Antonio Franzoi Pereira
CIRO HENRIQUE VIOTO DE CAMARGO
Rodrigo Mazuco

Sumário

1) Objetivo da Sprint 3

Nesta terceira e última Sprint, o foco será a **validação prática do modelo de Inteligência Artificial** desenvolvido na Sprint 2, correlacionando as previsões de produtividade realizadas com dados reais históricos obtidos de bases públicas.

A tarefa principal é analisar se o modelo criado é confiável e identificar eventuais ajustes necessários para melhorar sua precisão.

Para isso, serão comparados os resultados do **NDVI** e da **produtividade predita** com a **produtividade agrícola real**, aplicando métodos estatísticos de **análise de correlação e regressão** para embasar as conclusões.

Avaliação do cumprimento do objetivo

O objetivo da Sprint 3 foi **alcançado com sucesso**. Executamos a integração dos dados de produtividade, NDVI e clima, aplicamos testes estatísticos (correlação de Pearson e Spearman), e realizamos regressões lineares para validar a força das relações.

Identificamos os limites do NDVI como preditor isolado e sugerimos melhorias com base nos dados observados.

Estrutura dos relatórios

A seguir, os resultados e análises da Sprint estão documentados em arquivos `markdown` organizados por etapa (e2*, e3*, e4_), e armazenados no diretório:

Etapa 1 – Coleta de Dados de Produtividade Histórica

Nesta primeira etapa da Sprint 3, o foco foi a **pesquisa e consolidação de bases públicas** com dados históricos de produtividade agrícola.

Fontes de dados consultadas:

- **IBGE (Instituto Brasileiro de Geografia e Estatística)** – SIDRA
- **CONAB (Companhia Nacional de Abastecimento)** – (não utilizada diretamente nesta Sprint)
- **MAPA** e **CEPEA/USP** – considerados, mas sem dados diretos no escopo desta Sprint

Cultura selecionada:

- **Cana-de-açúcar**
  Mantivemos a mesma cultura agrícola analisada nas Sprints anteriores, garantindo consistência na abordagem.

Informações coletadas:

- **Produtividade média (toneladas por hectare)** por município
- **Ano agrícola:** 2020, 2021, 2022, 2023
- **Códigos e nomes dos municípios** para integrar com outras bases (NDVI, clima)
- **Condições regionais:** contextualizadas na Etapa 4 (interpretação crítica)

⚙ Organização dos scripts utilizados nesta etapa

Para padronizar os dados de produtividade histórica da cana-de-açúcar e transformá-los em um formato compatível com análise de séries temporais e integração com NDVI, desenvolvemos os seguintes scripts:

`convert_full_csv_to_long_final.py`

- Responsável por converter a base original da produtividade (formato wide, com colunas por ano) para o formato **long**, com uma linha por município por ano.
- O script:
  - Renomeia colunas com nomes padronizados dos anos
  - Trata valores ausentes e inválidos (ex: '-', 'X', '..')
  - Converte dados para ponto flutuante (`float`)
  - Padroniza o código do município para 7 dígitos
  - Exporta o resultado com o nome `canadeacucar_long_final_<timestamp>.csv`

Validação

- Verificamos se os dados estavam coerentes em relação a anos, municípios e formato
- O dataset final serviu de base para a etapa seguinte de correlação com NDVI e variáveis climáticas

Este processo garantiu que os dados estivessem prontos para unificação e análise estatística, mantendo qualidade e estrutura adequada para regressão, correlação e visualizações futuras.

Etapa 2 – Tratamento e Preparação dos Dados

Nesta etapa, o objetivo foi organizar e padronizar os dados para viabilizar análises estatísticas comparáveis entre produtividade real e NDVI médio, com foco na cultura da cana-de-açúcar no estado de São Paulo.

Estruturação dos dados

Foram organizadas colunas compatíveis para posterior unificação: - **Produtividade agrícola real** (toneladas por hectare) - **NDVI médio mensal por município** - **Ano e mês como base temporal comum**

Ajustes e padronizações realizadas

- Conversão de todos os códigos de município para **7 dígitos (formato IBGE)**
- Conversão dos dados de **NDVI e produtividade para tipo numérico (float)**
- Alinhamento da escala temporal entre os datasets:
    - Todos os dados utilizados estão no intervalo de **2020 a 2023**
- Remoção de registros com valores ausentes ou inválidos
- Garantia de consistência por meio de validações cruzadas entre os datasets

---

Integração com Google Earth Engine (NDVI via MODIS)

Um dos principais diferenciais técnicos desta etapa foi a **integração com o Earth Engine (GEE)** para coleta do NDVI mensal médio por município. Para isso:

Script utilizado:

- **`extrair_nvdi_mensal.py`**
    - Inicializa o GEE com autenticação por projeto
    - Utiliza a coleção **MODIS/061/MOD13Q1** com resolução de 250m
    - Reduz as imagens por **média mensal de NDVI por município**
    - Exporta o resultado como arquivo `.csv` para integração local
    - Dados organizados com `ano`, `mês`, `CD_MUN` e `NDVI_MEDIO`

Essa abordagem garantiu maior qualidade, periodicidade padronizada e alinhamento espacial com os dados públicos de produtividade.

⚙ Unificação e preparação final dos dados

Para consolidar os dados em um único arquivo integrado, foram utilizados os seguintes scripts:

`mapear_estacoes_inmet.py`

- Extraiu metadados (nome, código, latitude, longitude) das estações meteorológicas do INMET

`map_estacao_metereologica_ibge.py`

- Realizou o cruzamento espacial das estações com os municípios do shapefile do IBGE, associando cada estação ao código `CD_MUN`

`integrar_clima_ndvi_sp.py`

- Unificou os dados do INMET com o NDVI mensal por município
- Validou o cruzamento por ano e mês
- Exportou o resultado como `clima_ndvi_integrado.csv`

`integrar_produtividade_com_clima_ndvi.py`

- Realizou a junção final com a produtividade agrícola, resultando em:
  - **NDVI médio**
  - **Clima médio (chuva, temperatura, umidade)**
  - **Produtividade agrícola real**

Resultado

Ao final da Etapa 2, os dados foram tratados, padronizados e integrados com sucesso em um único dataset, prontos para análises estatísticas e visualizações avançadas.

Etapa 3 – Sugestões de Análise Estatística e Correlação

Nesta etapa, realizamos as análises estatísticas propostas para investigar a relação entre o **NDVI médio** e a **produtividade agrícola real** da cana-de-açúcar entre os anos de 2020 e 2023, no estado de São Paulo.

Testes de correlação aplicados

Foram aplicadas duas técnicas estatísticas para medir a correlação entre as variáveis:

- **Correlação de Pearson (linear):** r = **-0.302**, p = 0.00043
  → Interpretação: **fraca correlação negativa**
- **Correlação de Spearman (monótona):** r = **-0.456**, p < 0.00001
  → Interpretação: **moderada correlação negativa**

Regressão linear simples

Também foi ajustado um modelo de regressão linear simples para avaliar a capacidade do NDVI em prever a produtividade agrícola:

- **Equação da regressão:**
  `Produtividade = -8.46 * NDVI + 68393.12`
- **Coeficiente de determinação (R^2):**
  `0.091` → O modelo explica apenas **9,1% da variância da produtividade**

Análise por município

Executamos uma correlação de Spearman individual por município para identificar padrões regionais:

- **Total de municípios analisados:** 26

Top 5 correlação positiva:

- Avaré (SP): **0.95**
- Bebedouro (SP): **0.87**
- Rancharia (SP): 0.60
- José Bonifácio (SP): 0.45
- Bragança Paulista (SP): 0.45

Top 5 correlação negativa:

- Itapira (SP): **-0.80**
- Bauru (SP): -0.80
- São Simão (SP): -0.77
- Cachoeira Paulista (SP): -0.77
- São Luiz do Paraitinga (SP): -0.77

Gráficos produzidos

1. Correlação de Spearman por município:

Gráfico de correlação por município

2. Distribuição da produtividade por ano:

Boxplot produtividade por ano

3. Dispersão NDVI × produtividade com linha de tendência:

Dispersão NDVI vs Produtividade

Conclusão

Embora o NDVI mostre alguma correlação com a produtividade em certos municípios, ele não é um preditor confiável de forma isolada para todo o estado. O modelo linear simples apresenta limitações claras, mas os resultados fornecem uma base sólida para futuras melhorias com variáveis adicionais ou modelos mais robustos.

Etapa 4 – Interpretação dos Resultados e Discussão Crítica

---

1. O NDVI foi um bom preditor da produtividade?

De forma geral, **o NDVI não se mostrou um preditor confiável** da produtividade agrícola da cana-de-açúcar quando utilizado isoladamente.
Os testes estatísticos revelaram:

- **Correlação de Pearson:** -0.302 → fraca
- **Correlação de Spearman:** -0.456 → moderada (negativa)
- **Regressão linear simples (R^2):** 0.091 → o modelo explica apenas 9,1% da variação da produtividade

---

2. Em que situações o modelo teve melhor ou pior desempenho?

Melhor desempenho:

- Municípios como **Avaré** e **Bebedouro** apresentaram **forte correlação positiva**, indicando que em certas regiões o NDVI pode estar mais alinhado com a produtividade real.

Pior desempenho:

- Municípios como **Itapira**, **Bauru** e **São Simão** tiveram correlação negativa forte, mostrando que NDVI alto não necessariamente significou produtividade alta — o que pode indicar presença de outras variáveis críticas não modeladas.

---

3. Fatores externos que podem ter influenciado os resultados

- **Eventos climáticos:**
  Geadas, estiagens e chuvas irregulares registradas em SP durante os anos de 2021 e 2022 não foram modeladas diretamente, mas certamente impactaram a produtividade agrícola.
- **Pragas e doenças agrícolas:**
  A ausência de dados fitossanitários impede avaliar o impacto de infestações ou doenças sobre os rendimentos.
- **Pandemia de COVID-19:**
  Durante 2020 e 2021, o Brasil — e especialmente o estado de São Paulo — sofreu com os efeitos da pandemia, que podem ter influenciado negativamente a produtividade agrícola devido a:
  - Redução da força de trabalho no campo
  - Dificuldade logística de insumos e transporte
  - Possíveis atrasos ou falhas no manejo da cultura
- **Qualidade das imagens NDVI:**
  O modelo utilizou a coleção **MODIS/061/MOD13Q1** com 250m de resolução, o que pode ter causado **mistura de pixels agrícolas com vegetação nativa ou áreas urbanas**, distorcendo o NDVI médio por município.

---

4. Sugestões de melhorias para o modelo de IA

- **Incluir novos dados climáticos e ambientais** como:
  - Precipitação acumulada
  - Temperatura média
  - Umidade do solo
  - Índices como EVI ou evapotranspiração
- **Melhorar o tratamento das imagens**:
  - Utilizar imagens Sentinel-2 com 10m de resolução
  - Aplicar filtros para remoção de nuvens e correção atmosférica
- **Ajustar o período de coleta do NDVI**:
  - Ao invés da média mensal, usar o **NDVI máximo da safra**
  - Ou aplicar uma **média acumulada entre meses críticos da cultura**

---

5. Limitações da análise

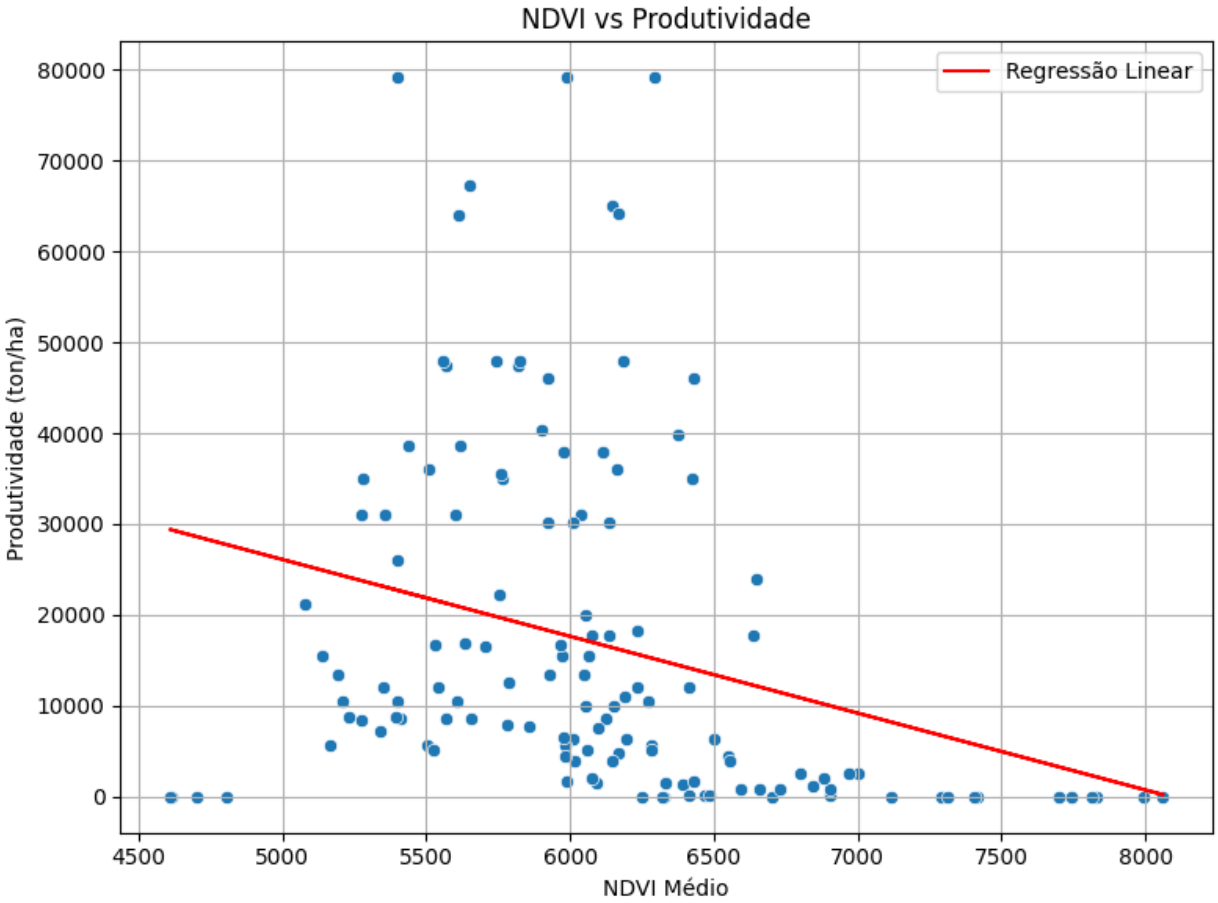
- **Tamanho da amostra:**
  A análise foi baseada em apenas **132 registros válidos**, o que limita a robustez estatística.
- **Qualidade das bases públicas:**
  Estações do INMET com falhas de cobertura e inconsistências em alguns arquivos CSV demandaram pré-processamento manual.
- **Modelo estatístico simples:**
  Foi utilizada **regressão linear simples**, o que não captura relações complexas entre múltiplas variáveis.

Conclusão

Mesmo com as limitações, esta Sprint cumpriu seu papel de **validar criticamente a proposta de modelo de IA aplicada à agricultura de precisão**, destacando caminhos promissores para futuras melhorias e demonstrando na prática a importância do rigor estatístico na modelagem preditiva.

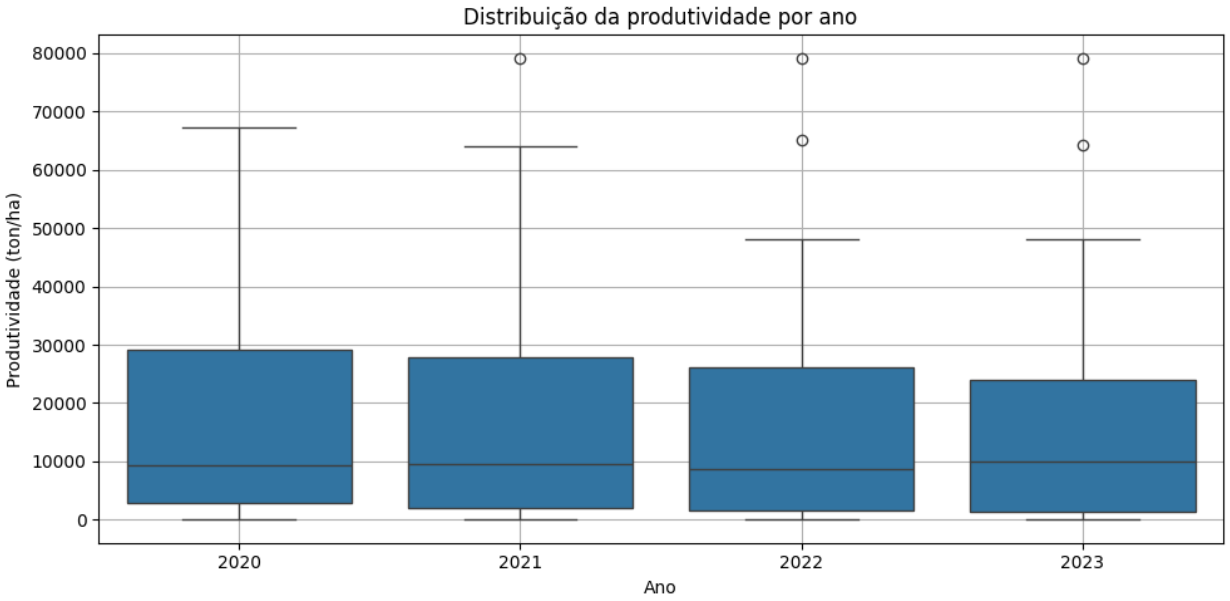
Gráficos Gerados

Dispersão NDVI vs Produtividade:



Este gráfico mostra a relação entre o NDVI médio mensal e a produtividade agrícola real por município. A linha vermelha representa a regressão linear ajustada.

Boxplot da produtividade por ano:



Distribuição da produtividade agrícola da cana-de-açúcar por ano, permitindo visualizar a variabilidade e mediana por safra entre 2020 e 2023.

Correlação de Spearman por município:

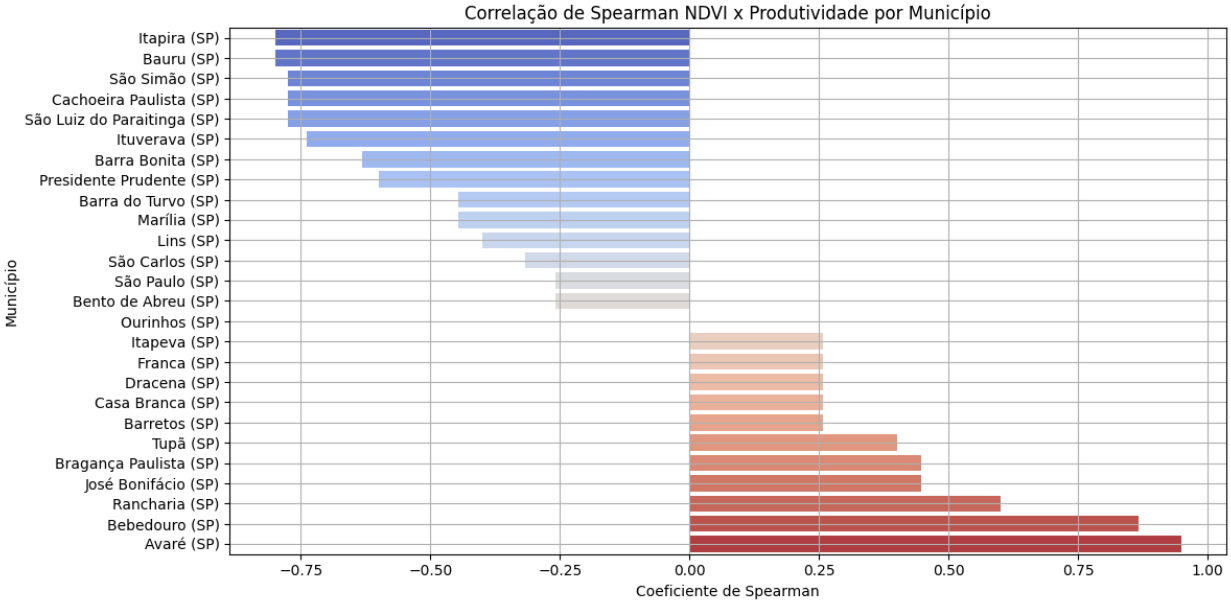


Gráfico de barras com o coeficiente de correlação de Spearman entre NDVI e produtividade em cada município. Correlações positivas e negativas são destacadas por cor.

Referências

- IBGE/SIDRA – Sistema de Recuperação Automática de Dados
- Google Earth Engine – MODIS/061/MOD13Q1
- INMET – Instituto Nacional de Meteorologia (dados climáticos por estação)
- Scripts próprios desenvolvidos em Python com apoio do ChatGPT

Conclusão Final

O projeto da Sprint 3 foi concluído com êxito, cumprindo o objetivo proposto de validar a aplicação do NDVI e de dados climáticos na previsão da produtividade agrícola da cana-de-açúcar. Com os dados integrados, análises estatísticas aplicadas e limitações bem definidas, o grupo está apto a sugerir novos caminhos de evolução para modelos de IA aplicados ao agronegócio brasileiro.