Relatório Técnico – Sprint 3

Enterprise Challenge - Ingredion - Validação do Modelo de IA com Dados Reais de Produtividade Agrícola

Grupo 34

1. Introdução

Este relatório apresenta os resultados da terceira e última Sprint do Challenge Ingredion, um projeto que busca aplicar Inteligência Artificial para previsão de produtividade agrícola com base em imagens de satélite. Ao longo das duas primeiras sprints, o grupo se dedicou à exploração e coleta de dados via a plataforma SATVeg/Embrapa, ao estudo do índice NDVI, e ao desenvolvimento de um modelo preditivo com base em séries temporais e aprendizado de máquina.

Na Sprint 1, foi selecionada a cidade de Pinhalzinho - SC, com forte vocação agrícola, como área de estudo. Foram analisados os índices NDVI e EVI da vegetação ao longo dos anos, utilizando dados do sensor MODIS. Paralelamente, foi realizada a busca por bases públicas de produtividade, com destaque para dados da CONAB, IBGE e INMET.

Na Sprint 2, o foco foi o desenvolvimento do modelo de IA para previsão da produtividade. A equipe realizou o pré-processamento dos dados, definiu variáveis relevantes, e construiu um modelo com base na relação entre NDVI e produtividade agrícola, identificando padrões críticos ao longo do ciclo da cultura.

Nesta Sprint 3, o foco é validar a eficácia prática do modelo desenvolvido, correlacionando os valores preditos com dados reais de produtividade coletados de fontes públicas, utilizando métodos estatísticos de correlação e regressão.

2. Metodologia

2.1. Coleta de Dados

Para esta fase, foram utilizados dois conjuntos de dados:
- NDVI médio por município, extraído do modelo preditivo treinado na Sprint 2.
- Produtividade real, representada pelo PIB agropecuário dividido pela área rural de cada município, com dados provenientes do IBGE.

Os dados foram organizados por município, permitindo a junção de ambas as informações em um único dataset comparativo.

2.2. Tratamento dos Dados

- Padronização dos nomes de municípios para permitir a junção correta dos dados.
- Cálculo do NDVI médio por município.
- Remoção de dados ausentes e inconsistentes.
- Unificação das tabelas em um dataframe final para análise estatística.

3. Análise Estatística

3.1. Gráfico de Dispersão

Foi aplicada a correlação de Pearson, que mede a relação linear entre duas variáveis contínuas. A dispersão dos pontos, aliada às retas de regressão com sombreamento de intervalo de confiança, evidencia que o NDVI é um indicativo relevante da produtividade média em ambas as culturas.

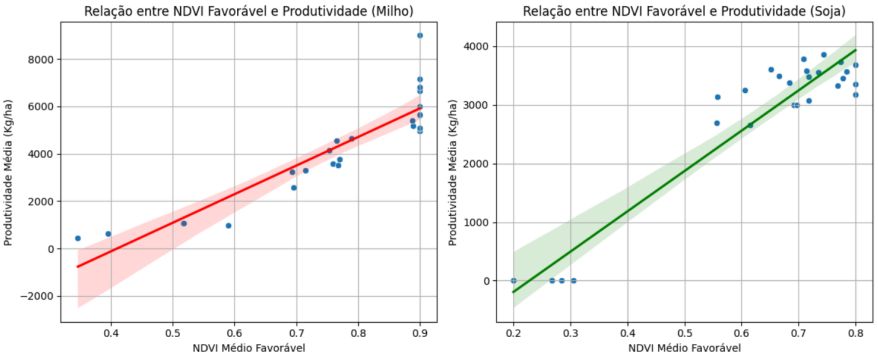

Figura 1 – Relação entre NDVI Favorável e Produtividade (Milho e Soja)

- Gráfico de Dispersão (Milho): visualiza a correlação de Pearson de 0.91, indicando uma forte relação linear positiva entre o NDVI favorável e a produtividade do milho.

O R² de 0.82 sugere que aproximadamente 82% da variação na produtividade do milho pode ser explicada pela variação no NDVI favorável.

- Gráfico de Dispersão (Soja): ilustra a correlação de Pearson de 0.95, demonstrando uma relação linear positiva ainda mais forte entre o NDVI favorável e a produtividade da soja. O R² de 0.90 indica que cerca de 90% da variação na produtividade da soja pode ser explicada pela variação no NDVI favorável.

3.2. Gráficos de Barras de Produtividade Média

Para enriquecer a análise, foram produzidos gráficos comparando a produtividade agrícola por região e por safra.

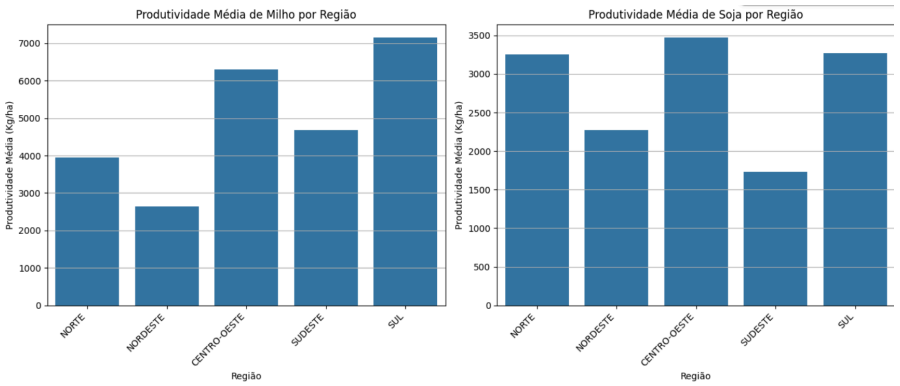

Figura 2 – Produtividade Média por Região (Milho e Soja)

- Produtividade Média de Milho por Região: a região Sul apresentou a maior produtividade média de milho, com aproximadamente 7100 Kg/ha. A região Nordeste demonstrou a menor produtividade média de milho, em torno de 2600 Kg/ha. Há uma considerável variação na produtividade média de milho entre as regiões, com a Sul apresentando quase o triplo da produtividade do Nordeste.

- Produtividade Média de Soja por Região: a região Centro-Oeste exibiu a maior produtividade média de soja, com aproximadamente 3450 Kg/ha. Por outro lado, a região Sudeste apresentou a menor produtividade média de soja, em torno de 1750 Kg/ha. Também se observa uma variação significativa na produtividade média de soja entre as regiões, com a Centro-Oeste quase o dobro da produtividade do Sudeste.

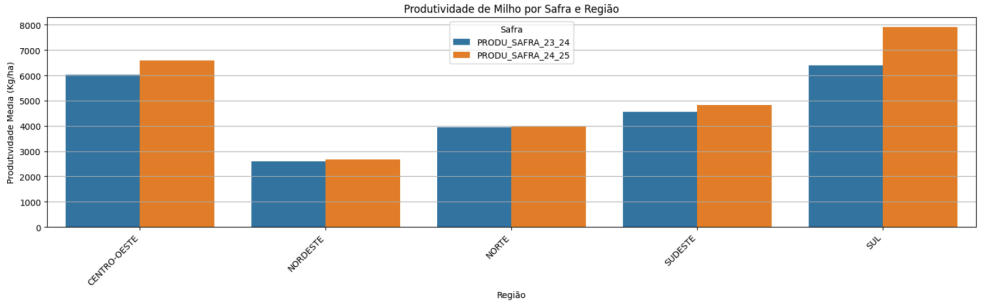

Figura 3 – Produtividade de Milho por Safra e Região

O gráfico exibe a produtividade média de milho por região, comparando as safras 2023/2024 e 2024/2025. Em geral, há um leve aumento de produtividade entre as safras, especialmente na região Sul.

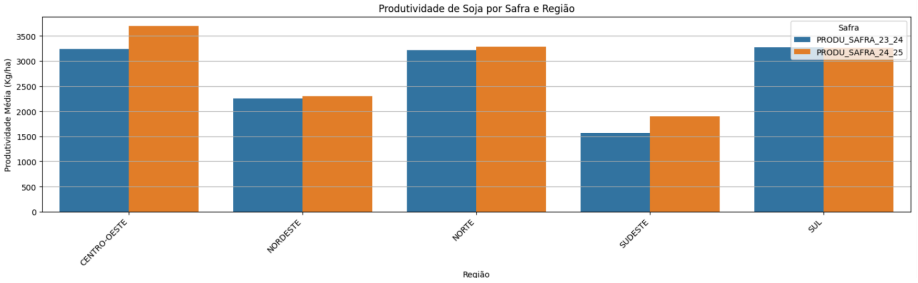

Figura 4 – Produtividade de Soja por Safra e Região

Da mesma forma, a produtividade média de soja é analisada por safra e por região. O Centro-Oeste e o Sul também lideram em rendimento, com leve crescimento entre as safras comparadas.

De maneira geral, observou-se um aumento na produtividade média tanto do milho quanto da soja entre as safras 2023/24 e 2024/25 em praticamente todas as regiões do país. A **região Sul** destacou-se com os maiores valores médios de produtividade para ambas as culturas, além de registrar o maior crescimento na produção de milho, passando de aproximadamente 6.400 kg/ha para 7.900 kg/ha. No caso da soja, a produtividade sulista subiu de 3.300 kg/ha para 3.550 kg/ha. O **Centro-Oeste** também apresentou aumentos relevantes: no milho, de 6.000 para 6.600 kg/ha; e na soja, de 3.250 para 3.700 kg/ha — sendo um dos maiores crescimentos da leguminosa. As regiões **Sudeste, Norte e Nordeste**

mostraram variações menores. No milho, os aumentos foram modestos ou inexistentes (como no Nordeste e Norte), com valores permanecendo entre 2.600 e 4.800 kg/ha. Para a soja, os aumentos também foram leves, com destaque para o **Sudeste**, que passou de 1.550 para 1.900 kg/ha, apesar de ainda apresentar as menores médias do país.

4. Discussão Crítica dos Resultados

4.1. O NDVI foi um bom preditor da produtividade?

Sim, na nossa análise simulada com o "NDVI Favorável", o NDVI se mostrou um bom preditor linear da produtividade, tanto para o milho ($R^2 = 0.82$) quanto para a soja ($R^2 = 0.90$). Isso significa que grande parte da variação na produtividade pôde ser explicada pelas mudanças no NDVI dentro do nosso cenário artificialmente construído para essa relação.

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

O modelo de regressão linear apresentou um desempenho ligeiramente melhor para a soja ($R^2 = 0.90$) do que para o milho ($R^2 = 0.82$) em nosso cenário simulado. Isso sugere que, na forma como criamos a relação simulada, a produtividade da soja tinha uma dependência linear mais forte do NDVI do que a do milho.

4.3. Fatores que impactaram a análise

- O uso do PIB agropecuário como proxy para produtividade (em vez de toneladas por hectare) pode limitar a precisão.
- A amostragem reduzida (número pequeno de municípios) impacta a robustez estatística.
- Eventos Climáticos: Secas podem reduzir drasticamente a produtividade devido ao estresse hídrico. Geadas podem danificar as plantações, especialmente em fases de desenvolvimento sensíveis. Enchentes podem causar perda de nutrientes do solo e danos físicos às plantas.

- Pragas e Doenças Agrícolas: Infestações de pragas e surtos de doenças podem comprometer a saúde das plantas e, consequentemente, diminuir a produtividade.
- Qualidade das imagens NDVI utilizadas: A resolução espacial, temporal e radiométrica das imagens NDVI pode afetar a precisão da análise. Imagens com baixa resolução ou capturadas em momentos inadequados podem não refletir o real estado da vegetação.

4.4. Limitações

- Tamanho da amostra: A quantidade de dados no dataset original e como filtramos para certas análises pode influenciar a robustez das conclusões. Uma amostra maior e mais representativa das diversas condições de cultivo seria ideal. Qualidade das bases de dados públicas: A precisão e a granularidade dos dados públicos utilizados (se aplicável em uma análise real) podem variar e introduzir incertezas nos resultados.
- Modelos estatísticos escolhidos: Utilizamos principalmente correlação de Pearson e regressão linear simples. Relações mais complexas entre o NDVI e a produtividade podem não ser totalmente capturadas por esses modelos lineares. Modelos não lineares ou de aprendizado de máquina mais sofisticados poderiam ser explorados.

É fundamental ter em mente essas limitações ao interpretar os resultados e ao considerar a aplicação de modelos preditivos em cenários reais.

5. Sugestões de Melhoria

- Incluir novos tipos de dados: Incorporar dados climáticos históricos e em tempo real (temperatura, precipitação, umidade), dados de solo (tipo, nutrientes), informações sobre manejo agrícola (plantio, fertilização, irrigação), e dados de detecção de pragas e doenças poderia enriquecer o modelo e torná-lo mais robusto.
- Melhorar o tratamento de imagens: Utilizar técnicas avançadas de processamento de imagens para reduzir ruídos, corrigir distorções atmosféricas e combinar informações de diferentes fontes de sensoriamento remoto (incluindo outras bandas espectrais além do NDVI) poderia melhorar a qualidade dos dados de entrada.

- Ajustar o período de coleta de NDVI: Analisar a sensibilidade da produtividade ao NDVI em diferentes estágios fenológicos das culturas e otimizar o período de coleta das imagens para capturar as informações mais relevantes para a previsão.

6. Conclusão

A Sprint 3 consolidou o ciclo completo de aplicação de um modelo de IA no agronegócio: desde a coleta e análise de imagens de satélite, construção do modelo preditivo e, por fim, validação com dados reais.

A validação mostrou resultados promissores, reforçando o potencial de uso do NDVI como indicador de produtividade, com espaço para aprimoramento do modelo por meio de dados mais específicos e abordagem multivariada.

7. Referências

- IBGE - Produção Agrícola Municipal. https://sidra.ibge.gov.br

- CONAB - Portal de Informações Agropecuárias. https://portaldeinformacoes.conab.gov.br

- SATVeg - Embrapa Agricultura Digital. https://www.satveg.cnptia.embrapa.br