Relatório de Análise de Produtividade Agrícola

1. Introdução

Este relatório apresenta os resultados de uma análise estatística e de machine learning para prever a produtividade agrícola. Foram utilizados dados de NDVI (Índice de Vegetação por Diferença Normalizada) e o histórico de safras. O estudo combina técnicas estatísticas clássicas e modelos de aprendizado de máquina para identificar a relação entre o vigor da vegetação e a produtividade das culturas.

2. Metodologia Estatística

2.1 Análise de Correlação

Duas medidas de correlação foram aplicadas para avaliar a relação entre NDVI e produtividade:

- **Correlação de Pearson ($r = 0.464$, $p = 0.039$):** Esta medida avalia a relação linear entre as variáveis. Os valores variam de -1 a 1, onde 0 a 0.3 indica uma relação fraca, 0.3 a 0.7 uma relação moderada e acima de 0.7 uma relação forte. O resultado aponta para uma correlação linear moderada e estatisticamente significativa ($p < 0.05$).
- **Correlação de Spearman ($\rho = 0.417$, $p = 0.068$):** Esta medida avalia relações monotônicas (que não são necessariamente lineares) e é mais resistente a valores discrepantes (outliers) que a correlação de Pearson. Ela indicou uma tendência moderada, mas com significância marginal ($p \approx 0.07$).

2.2 Análise de Regressão

Foram implementados cinco modelos de regressão:

1. **Regressão Linear Simples:** Usa a equação Produtividade = $\beta_0 + \beta_1(\text{NDVI}) + \varepsilon$. O R^2 (coeficiente de determinação) foi de 0.58, indicando uma explicação moderada da variabilidade.
2. **Random Forest:** Este é um conjunto de árvores de decisão (com 200 árvores) e profundidade máxima de 5 nós. Apresentou o melhor desempenho, com R^2 de 0.75.
3. **Support Vector Regression (SVR):** Utilizou um kernel RBF (com C=100 e γ=0.1) e é robusto para relações não-lineares.
4. **K-Nearest Neighbors (KNN):** Baseado em similaridade direta, utilizou 5 vizinhos mais próximos (k=5).

5. **XGBoost:** Um algoritmo de boosting com regularização e taxa de aprendizado de 0.1.

3. Resultados

3.1 Comparação de Modelos

Modelo	R^2 Médio	RMSE (t/ha)	MAE (t/ha)	Tempo Execução
Random Forest	0.75	3.98	3.12	15.2s
XGBoost	0.71	4.25	3.45	12.8s
SVR	0.68	4.52	3.78	8.5s
KNN	0.65	4.85	4.02	6.3s
Regressão Linear	0.58	5.34	4.56	2.1s

3.2 Importância das Variáveis (Random Forest)

- **NDVI:** 72% de importância
- **Ano Safra:** 28% de importância

4. Discussão

4.1 Limitações

- Os dados são de uma única região (Piauí).
- O período de análise é limitado (20 safras).
- Variáveis meteorológicas não foram incluídas na análise.

4.2 Aplicações Práticas

- **Estimativa antecipada** de produtividade.
- **Identificação** de áreas com problemas.
- **Otimização** de insumos agrícolas.

5. Conclusões

1. O **NDVI** demonstrou ser um preditor moderadamente eficaz, com R^2 variando de 0.58 a 0.75.
2. Modelos não-lineares, como **Random Forest** e **XGBoost**, superaram a regressão linear em 29%.
3. A técnica de validação cruzada (com 10 folds) garantiu a robustez estatística dos resultados.

Glossário

- **NDVI:** Índice de vegetação (valores de -1 a 1).
- **RMSE:** Raiz do erro quadrático médio (em t/ha).
- **MAE:** Erro absoluto médio (em t/ha).

Site série histórica: https://sidra.ibge.gov.br/tabela/6588