ANÁLISE DE DADOS MULTIDIMENSIONAIS PARA PREDIÇÃO DE SAFRAS

Gustavo Guse Martins (PIBIC/CNPq), Emili Everz Golombiéski, Maria Salete Gomes Vaz (Orientador), e-mail do orientador: salete@uepg.br.

Universidade Estadual de Ponta Grossa/Departamento de Informática.

Ciência da computação / Banco de Dados

Palavras-Chave: *Previsão de safras, Agricultura, Dados Multidimensionais, Análise de Dados.*

Resumo

Este trabalho apresenta o desenvolvimento de um modelo de *machine learning*, utilizando o algoritmo *Random Forest*, para a predição de safras no Estado do Paraná. Utilizando dados públicos do IBGE e INMET, o modelo alcançou um R^2 ajustado de média 96%. Apesar do sucesso do protótipo, a principal barreira para a criação de modelos robustos é a baixa qualidade dos dados públicos, especialmente os climáticos e de solo, destacando a necessidade de uma melhor infraestrutura de dados para futuras pesquisas.

Introdução

A agricultura enfrenta desafios crescentes devido às mudanças climáticas, tornando a previsibilidade das safras um fator importante para a segurança alimentar e estabilidade econômica. Este trabalho foca no desenvolvimento de um modelo de predição de safras para o Estado do Paraná, utilizando técnicas de *machine learning* para analisar dados multidimensionais e fornecer uma ferramenta que auxilie na tomada de decisões no setor agrícola.

Materiais e Métodos

O modelo foi desenvolvido utilizando dados públicos do Instituto Brasileiro de Geografia e Estatística (IBGE) referentes à produção agrícola e do Instituto Nacional de Meteorologia (INMET) para variáveis climáticas. O conjunto de dados abrangeu os anos de 2011 a 2024, com uma lacuna entre 2017 e 2019 devido à dados corrompidos.

O algoritmo escolhido foi o *Random Forest*, implementado na Linguagem R. O conjunto de dados foi dividido em 80% para treinamento e 20% para teste. A validação do modelo foi realizada por meio de uma análise de regressão linear entre os valores previstos e os reais, utilizando o coeficiente de determinação R^2 como métrica de desempenho.

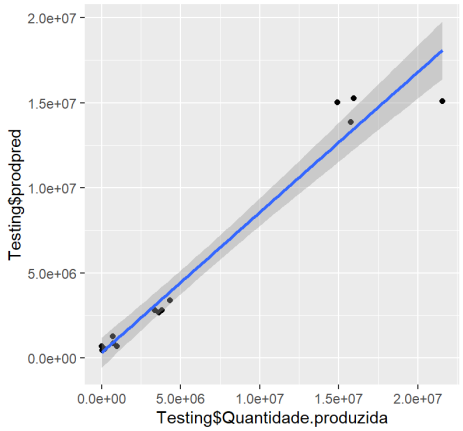
Resultados e Discussão

Após os testes, a análise estatística resultou em um coeficiente de determinação R^2 ajustado médio de 0,96, isso indica que o modelo consegue explicar cerca de 96% da variabilidade da produção agrícola.

O P-valor encontrado, de $4{,}124 \times 10^{-14}$ confirmou que a significância estatística do modelo não ocorreu ao acaso.

Visualmente, o ajuste pode ser observado na Figura 1, que demonstra a forte correlação entre os valores reais de produção e os valores previstos.

Figura 1. Análise em regressão linear sobre o ajuste dos dados previstos em relação aos reais.



Apesar do ajuste, foi observado que o modelo tende a cometer desvios em culturas com volumes de produção nos extremos (muito altos ou muito baixos). Esta inconsistência é atribuída à ausência de dados mais refinados, como do solo, que não estavam disponíveis de forma consistente nas fontes públicas.

Conclusão

O estudo demonstrou a viabilidade da utilização do algoritmo *Random Forest* para a predição de safras, com o protótipo alcançando precisão estatística de 96%. No entanto, a principal conclusão é que o avanço e a confiabilidade de tais modelos são severamente limitados pela disponibilidade de dados públicos. A carência de informações climáticas completas e, principalmente, de dados de solo uniformes, emergiu como o principal obstáculo. O aprimoramento futuro de modelos preditivos na agricultura depende, portanto, mais da criação de uma infraestrutura de dados públicos mais íntegra e detalhada do que do avanço dos algoritmos em si.

Referências

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. IBGE, 2025. Página inicial. Disponível em: <https://www.ibge.gov.br/>. Acesso em: 14 de mar. de 2025.

MINISTÉRIO DA AGRICULTURA E PECUÁRIA. Instituto Nacional de Meteorologia, 2025. Página inicial. Disponível em: <https://portal.inmet.gov.br>. Acesso em: 14 de mar. de 2025.