Ministério da Educação
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
Campus Campo Mourão
Curso de Bacharelado em Ciência da Computação

TRABALHO FINAL - OPT004

CAROLINA YUMI FUJII

PROF. EDUARDO PENA

UTFPR-2025
CAMPO MOURÃO

SUMÁRIO

RESUMO

Este relatório detalha o desenvolvimento de um projeto de ciência de dados focado na análise e previsão do crescimento da produção agrícola no Brasil. Utilizando dados da Pesquisa Agrícola Municipal (PAM) do IBGE de 2007 a 2023, o estudo abordou o desafio da volatilidade e da necessidade de previsibilidade no agronegócio. Foram desenvolvidos dois modelos de machine learning em paralelo: um de Classificação para identificar Unidades da Federação (UFs) com probabilidade de crescimento acima da média nacional, e um de Regressão para prever a taxa de crescimento anual. A metodologia incluiu um robusto pipeline de pré-processamento, uma profunda análise exploratória e uma sofisticada engenharia de features, utilizando ferramentas como Pandas, DuckDB e Scikit-learn. O modelo Random Forest demonstrou performance superior em ambas as tarefas, alcançando uma Área Sob a Curva (AUC) de 0.899 e acurácia de 81.36% na classificação, e um Coeficiente de Determinação (R^2) de 0.96 na regressão.

1. Definição do Problema e Perguntas de Pesquisa

1.1. Contextualização

O agronegócio é um pilar fundamental da economia brasileira, respondendo por aproximadamente 25% do Produto Interno Bruto (PIB) e sendo o principal motor do superávit da balança comercial do país. O Brasil figura entre os maiores produtores e exportadores de alimentos do mundo, desempenhando um papel vital na segurança alimentar global. Contudo, o setor opera em um ambiente de alta complexidade e incerteza. A sua performance é intrinsecamente ligada a fatores de difícil controle, como:

- Volatilidade Climática: Eventos extremos como secas, enchentes e geadas podem impactar drasticamente a produtividade.
- Flutuações de Mercado: Os preços das commodities são definidos em escala global e podem sofrer variações abruptas devido a fatores geopolíticos e econômicos.
- Desafios Logísticos e de Infraestrutura: Custos de transporte e capacidade de armazenamento são gargalos crônicos que afetam a rentabilidade.

Essa conjuntura de alta volatilidade cria uma necessidade crítica de previsibilidade. Para gestores, investidores e formuladores de políticas públicas, a capacidade de antecipar tendências, identificar oportunidades regionais e mitigar riscos de forma proativa não é apenas uma vantagem competitiva, mas uma condição para a sustentabilidade das operações. O desafio central que este projeto se propõe a resolver é, portanto, a transformação de dados históricos em inteligência acionável para o futuro.

1.2. Perguntas de Pesquisa

Para abordar o problema de forma estruturada e fornecer uma solução completa, a investigação foi dividida em duas perspectivas complementares, cada uma guiada por uma pergunta de

pesquisa específica:

1.Modelo de Classificação (Análise Qualitativa): "Qual a probabilidade de uma Unidade da Federação (UF) apresentar um crescimento no valor total da produção acima da média nacional em um determinado ano?"

- Objetivo: Identificar UFs com alto potencial de crescimento, funcionando como um sistema de alerta ou "selo de qualidade" para direcionar a atenção e os investimentos. A resposta a esta pergunta é essencialmente qualitativa, focando em "onde" as oportunidades são mais prováveis de surgir.

2.Modelo de Regressão (Análise Quantitativa): "Quais fatores (como tipo de produto, região, PIB e valores históricos) são mais influentes na previsão do valor da produção para o ano seguinte, e qual a magnitude de seu impacto?"

- Objetivo: Quantificar a taxa de crescimento esperada e entender os seus principais impulsionadores. O foco aqui é quantitativo, buscando explicar "quanto" uma UF deve crescer e "por quê", permitindo um planejamento financeiro e estratégico mais preciso.

A combinação dessas duas abordagens fornece uma visão 360 graus do problema, permitindo não apenas identificar oportunidades, mas também compreender os mecanismos que as geram.

2. Metodologia Completa e Limitações

O desenvolvimento do projeto seguiu um pipeline metodológico rigoroso, desde a coleta e tratamento dos dados até a avaliação dos modelos. Cada etapa foi executada com o objetivo de maximizar a qualidade e a capacidade preditiva da solução.

2.1. Fonte e Descrição dos Dados

A base de dados primária utilizada foi a Pesquisa Agrícola Municipal (PAM), disponibilizada pelo Sistema IBGE de Recuperação Automática (SIDRA). Esta fonte foi escolhida por sua abrangência, granularidade e confiabilidade, sendo o levantamento oficial da produção agrícola no Brasil. A análise compreendeu uma janela temporal de 17 anos (2007 a 2023), permitindo a captura de ciclos econômicos e agrícolas de longo prazo. O dataset original, com granularidade municipal, foi agregado por Unidade da Federação para focar em tendências macro-regionais e mitigar a alta volatilidade dos dados locais. O escopo da análise abrangeu mais de 35 produtos agrícolas em todas as 27 Unidades da Federação.

2.2. Pipeline de Pré-processamento e Limpeza

Um pipeline robusto foi construído para transformar os dados brutos em um formato adequado para modelagem:

1.Integração de Fontes: Os dados da PAM foram enriquecidos com informações de PIB per capita por estado, obtidas de fontes do IBGE. Esta etapa adicionou um contexto socioeconômico crucial à análise. A integração foi realizada com a biblioteca Pandas, utilizando a função merge

para unir os dataframes com base nas chaves "UF" e "Ano".

2.Tratamento de Valores Ausentes: A identificação de dados faltantes (realizada com a função .isna().sum()) revelou a necessidade de tratamento na variável de PIB. Em vez de descartar as linhas, o que acarretaria perda de dados de produção, foi aplicada uma estratégia de imputação. Os valores nulos foram preenchidos com a média do PIB da respectiva UF ao longo dos anos, utilizando a função transform do Pandas, que preserva a estrutura dos dados e a consistência regional.

3.Padronização e Limpeza: Foi realizada uma limpeza textual para garantir a consistência categórica. Nomes de produtos e UFs foram padronizados aplicando-se funções de string como .str.strip() e .str.title() para remover espaços e uniformizar a capitalização.

4.Tratamento de Outliers: A análise exploratória visual (descrita a seguir) revelou a presença de outliers extremos no valor da produção. Para evitar que esses pontos, representativos de eventos muito raros, distorcessem a capacidade de generalização dos modelos, foi adotada uma estratégia de corte: todos os registros com valor de produção acima do quantil de 99% foram removidos. Esta decisão foi justificada para focar o modelo nos padrões mais recorrentes do agronegócio.

2.3. Análise Exploratória de Dados (AED)

A AED foi fundamental para guiar as decisões de modelagem. Utilizando bibliotecas como Seaborn e Matplotlib, foram extraídos os seguintes insights:

•Distribuição Assimétrica: A variável "Valor da Produção" apresentou uma forte assimetria positiva (cauda longa à direita), indicando que a maior parte do valor está concentrada em poucos estados e produtos (ex: Soja e Milho no Centro-Oeste). Isso sugeriu que modelos lineares, que performam melhor com distribuições normais, poderiam ter sua performance limitada.

•Correlação com Fatores Econômicos: Foi confirmada uma correlação positiva entre o valor da produção e o PIB per capita regional, validando a decisão de incluir esta variável.

•Necessidade de Normalização: As features numéricas apresentavam escalas muito distintas (ex: valor da produção em bilhões de reais vs. PIB per capita em milhares). Para que todas as variáveis tivessem a mesma importância inicial para os modelos, foi aplicado o StandardScaler da biblioteca Scikit-learn para padronizá-las (média 0 e desvio padrão 1).

2.4. Engenharia de Features

Esta foi a etapa mais crítica para agregar poder preditivo ao modelo. Foram criadas três categorias de features:

1.Features de Lag (Inércia): Para capturar a dependência temporal, foi criada a variável Valor_Total_UF_Lag1, representando o valor total da produção da UF no ano anterior. A hipótese é que o desempenho passado é um forte indicador do futuro. A implementação utilizou a função

shift() do Pandas, com um groupby() por UF para evitar vazamento de dados entre estados.

2.Features de Crescimento (Tendência): Para que o modelo aprendesse com tendências em vez de valores absolutos, foram calculadas taxas de crescimento, como Crescimento_Anual_UF e Crescimento_Anual_Produto.

3.Features de Contexto (Estrutura): Para contextualizar a produção, foram criadas variáveis como Participacao_Produto_UF (medindo a importância relativa de um produto dentro de seu estado). A implementação desta etapa foi otimizada com o uso da ferramenta DuckDB, que permitiu a execução de consultas SQL complexas diretamente sobre os DataFrames do Pandas, de forma mais performática e declarativa.

Finalmente, para incorporar as variáveis categóricas (UF e Produto) nos modelos, foi aplicada a técnica de One-Hot Encoding (pd.get_dummies), que transforma cada categoria em uma nova coluna binária.

2.5. Modelagem e Avaliação

- Definição dos Targets: Para a Classificação, o target foi a variável binária Crescimento_Acima_da_Media (1 se o crescimento da UF superou a média nacional, 0 caso contrário). Para a Regressão, o target foi a variável contínua Crescimento_Anual_UF.
- Seleção de Modelos: Para cada tarefa, foi estabelecido um modelo de baseline (simples e linear) e um modelo avançado (complexo e não-linear):
- Classificação: Regressão Logística (baseline) vs. Random Forest Classifier (avançado).
- Regressão: Regressão Linear (baseline) vs. Random Forest Regressor (avançado). A escolha do Random Forest foi informada pela AED, dada sua capacidade de capturar relações não-lineares, sua robustez a outliers e sua habilidade de calcular a importância das features.
- Métricas de Avaliação:
- Classificação: Foram utilizadas a Acurácia, AUC (Área sob a Curva ROC), que mede a capacidade de discriminação do modelo, e o F1-Score, importante para lidar com classes desbalanceadas.
- Regressão: Foram utilizados o R^2 (Coeficiente de Determinação), que indica a proporção da variância explicada pelo modelo, e os erros RMSE (Raiz do Erro Quadrático Médio) e MAE (Erro Médio Absoluto).

2.6. Limitações da Metodologia

É importante reconhecer as limitações deste estudo:

- Dados Climáticos: O modelo não incluiu variáveis climáticas (ex: precipitação, temperatura), que são conhecidamente um dos principais fatores de influência na produção agrícola.
- Nível de Agregação: Ao agregar os dados por UF, perde-se a granularidade municipal,

que poderia revelar heterogeneidades importantes dentro de um mesmo estado.
- Fatores Externos: O modelo não incorpora explicitamente choques externos, como crises econômicas globais, pandemias ou mudanças abruptas em políticas comerciais.
- Tratamento de Outliers: A remoção de dados acima do quantil de 99% pode, em alguns casos, eliminar registros de eventos extremos válidos que seriam importantes para uma análise de risco mais completa.

3. Resultados das Análises

Os resultados quantitativos demonstraram a eficácia da abordagem, com destaque para a performance superior dos modelos de Random Forest em ambas as tarefas.

3.1. Resultados do Modelo de Classificação

O objetivo era prever se uma UF teria um crescimento acima da média nacional. A tabela abaixo compara a performance do baseline (Regressão Logística) com o modelo avançado (Random Forest).

Métrica	Regressão Logística (Baseline)	Random Forest (Avançado)
AUC	0.767	0.899
Acurácia	72.4%	81.36%
F1-Score	0.71	0.81

O Random Forest Classifier alcançou uma performance excelente, com uma AUC próxima de 0.9, indicando um alto poder de discriminação entre as classes. A acurácia de 81.36% significa que o modelo acerta sua previsão em mais de 8 a cada 10 casos.

3.2. Resultados do Modelo de Regressão

Na tarefa de prever o valor exato da taxa de crescimento, a superioridade do Random Forest foi ainda mais pronunciada.

Métrica	**Regressão Linear (Baseline)**	**Random Forest (Avançado)**
R^2	**0.87**	**0.96**
MAE	**0.08**	**0.02**
RMSE	**0.15**	**0.07**

Um R^2 de 0.96 é um resultado excepcional, indicando que o modelo consegue explicar 96% da

variância na taxa de crescimento das UFs. O Erro Médio Absoluto (MAE) de apenas 0.02 demonstra que, em média, as previsões do modelo erram por apenas 2 pontos percentuais, conferindo alta precisão para fins de planejamento.

3.3. Análise de Importância de Features

A análise de importância de features do Random Forest revelou quais fatores mais influenciaram as previsões. Os 5 principais preditores foram:

1. Valor_Total_UF_Lag1 (15.8%): O valor da produção da UF no ano anterior. Este foi, de longe, o fator mais importante, confirmando a forte inércia e dependência temporal do setor.
2. Ano (9.5%): A variável ano, indicando a presença de tendências de longo prazo (tecnológicas, econômicas) que afetam o crescimento.
3. Valor_Total_Producao (8.1%): O valor total da produção no ano corrente, que serve como uma base para o crescimento.
4. Participacao_Produto_UF (6.7%): A importância relativa de um produto dentro de seu estado, uma feature criada na engenharia de features.
5. PIB_Per_Capita (5.2%): O contexto econômico da região.

A presença de features criadas no topo da lista valida a importância da etapa de engenharia de features para o sucesso do projeto.

4. Discussão dos Resultados

A superioridade consistente dos modelos de Random Forest sobre os baselines lineares é a primeira grande conclusão. Isso confirma a hipótese inicial, levantada na AED, de que as relações que governam o crescimento agrícola são fundamentalmente não-lineares. Modelos lineares, embora mais simples e interpretáveis, não conseguem capturar as interações complexas entre as variáveis, como o efeito combinado de um determinado produto em uma região específica com um certo nível de PIB. O Random Forest, por ser um ensemble de árvores de decisão, é inerentemente capaz de modelar essas interações.

A análise de importância de features fornece insights de grande valor estratégico. A dominância da feature Valor_Total_UF_Lag1 revela que o setor agrícola possui uma forte inércia. Regiões que foram produtivas no passado tendem a continuar sendo, devido a investimentos consolidados, conhecimento local e infraestrutura existente. Isso sugere que mudanças estruturais no ranking de produção são lentas e graduais. A importância da variável Ano aponta para a existência de tendências macro que transcendem fatores locais, como avanços tecnológicos em sementes e maquinário, ou mudanças em políticas agrícolas nacionais.

O sucesso das features de engenharia, como Participacao_Produto_UF, demonstra que não basta olhar para os valores brutos. O contexto estrutural é fundamental. Um aumento na produção de soja tem um significado diferente em Mato Grosso, onde é a cultura dominante, versus no Paraná, onde a produção é mais diversificada. O modelo foi capaz de aprender essas

nuances.

Finalmente, a alta precisão do modelo de regressão (R^2 de 0.96) valida a abordagem como uma ferramenta viável para planejamento e orçamentação (budgeting). Um erro médio de apenas 2 pontos percentuais permite que as previsões sejam usadas com alto grau de confiança para estimar receitas, planejar investimentos e gerenciar expectativas de stakeholders.

5. Recomendações Práticas

Os resultados deste projeto se traduzem em recomendações estratégicas acionáveis para diferentes agentes do agronegócio:

1.Para Empresas de Insumos e Investidores:

- Alocação Inteligente de Recursos: Utilizar o modelo de classificação para criar um "ranking de potencial" das UFs. Direcionar equipes de vendas, esforços de marketing e investimentos em distribuição para as regiões com maior probabilidade de crescimento acima da média, maximizando o ROI.
- Análise de Risco e Diversificação: Usar as previsões do modelo de regressão para simular cenários e avaliar o risco de concentração em determinadas culturas ou regiões. A ferramenta pode auxiliar na decisão de diversificar o portfólio de produtos ou a presença geográfica.

2.Para Gestores e Produtores Rurais:

- Benchmarking de Performance: Comparar o crescimento previsto para sua cultura e região com o desempenho real, identificando gaps de eficiência e oportunidades de melhoria.
- Planejamento Estratégico: Utilizar as previsões de crescimento como um input para o planejamento de safras, negociação de contratos futuros e decisões de investimento em tecnologia e infraestrutura.

3.Para Formuladores de Políticas Públicas:

- Monitoramento de Indicadores-Chave: Focar o monitoramento nos fatores que o modelo apontou como mais importantes (ex: PIB regional, tendências de produtos específicos) para antecipar a necessidade de intervenções, como linhas de crédito ou programas de incentivo.
- Fomento a Regiões Promissoras: Identificar, através do modelo, regiões com alto potencial latente e direcionar políticas de desenvolvimento de infraestrutura e capacitação para destravar esse crescimento.

6. Trabalhos Futuros

Este projeto estabelece uma base sólida, mas abre diversas avenidas para expansão e aprimoramento:

- Inclusão de Dados Climáticos: A adição de variáveis como precipitação, temperatura média, umidade e ocorrência de eventos climáticos extremos é o próximo passo mais lógico e com maior potencial de aumentar a acurácia do modelo.
- Modelos de Séries Temporais: Testar abordagens específicas para séries temporais, como ARIMA, SARIMA ou modelos baseados em redes neurais (LSTM), que podem capturar dependências temporais e sazonalidades de forma ainda mais sofisticada.
- Aumento da Granularidade: Descer a análise para o nível municipal. Embora computacionalmente mais desafiador, isso permitiria recomendações muito mais precisas e localizadas, identificando "bolsões" de oportunidade dentro de uma mesma UF.

REFERÊNCIAS

1. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32. Disponível em: https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf Referência fundamental que introduziu o algoritmo Random Forest, detalhando sua construção e os princípios teóricos por trás de sua eficácia.

2. Zheng, A., & Casari, A. (2018). Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists. O'Reilly Media. Uma guia prático e abrangente sobre a arte e a ciência da engenharia de features, cobrindo técnicas para transformar dados brutos em preditores poderosos para modelos de machine learning.

3. Scikit-learn Developers. (n.d.). 1.11. Ensemble methods. Scikit-learn Documentation. Disponível em: https://scikit-learn.org/stable/modules/ensemble.html Documentação oficial da biblioteca Scikit-learn para métodos de ensemble, incluindo a implementação e os parâmetros do Random Forest Classifier e Regressor utilizados no projeto.