UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"
FACULDADE DE CIÊNCIAS - CAMPUS BAURU
DEPARTAMENTO DE COMPUTAÇÃO
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

MODELO DE REGRESSÃO LINEAR MÚLTIPLA PARA PREVISÃO DE SAFRA AGRÍCOLA

Nome: Guilherme Lima Zanin, e João Lucas Cardoso Criveli	RA: 221026479 e 221024735
Professor: Dr. Clayton Reginaldo Pereira	Disciplina: Data Science

BAURU
Junho/2025

GUILHERME LIMA ZANIN, E JOÃO LUCAS CARDOSO CRIVELI

MODELO DE REGRESSÃO LINEAR MÚLTIPLA PARA PREVISÃO DE SAFRA AGRÍCOLA

Monografia para a disciplina de Data Science, o qual é utilizado como método de avaliação do curso de Bacharelado em Ciência da Computação da Universidade Estadual Paulista "Júlio de Mesquita Filho", Faculdade de Ciências, Campus Bauru.

BAURU
Junho/2025

Sumário

1 Introdução

A agricultura desempenha um papel fundamental na economia brasileira, sendo responsável por uma parcela significativa do Produto Interno Bruto (PIB), da geração de empregos e das exportações do país. Em particular, culturas como a soja e o milho têm se destacado como pilares da produção agrícola nacional, tanto em volume quanto em valor de mercado. Com a crescente demanda por alimentos, biocombustíveis e matérias-primas, prever com precisão a produção agrícola torna-se cada vez mais estratégico para o planejamento de políticas públicas, investimentos do setor privado e o abastecimento do mercado interno e externo.

A previsão de safras agrícolas é uma tarefa desafiadora, pois envolve uma série de variáveis que afetam diretamente o rendimento das culturas, como clima, tipo de solo, manejo agrícola, área plantada, localização geográfica, entre outros. Com o avanço da tecnologia e da disponibilidade de dados históricos, tornou-se possível utilizar técnicas de ciência de dados e modelagem estatística para estimar de forma mais acurada o desempenho das lavouras.

Neste trabalho, propõe-se o uso da Regressão Linear Múltipla — uma das técnicas estatísticas mais tradicionais e interpretáveis — para prever duas variáveis de interesse na produção agrícola: a **área colhida** e o **valor da produção**. O conjunto de dados utilizado foi retirado da Tabela 1612 do Sistema IBGE de Recuperação Automática (SIDRA), disponível em: https://sidra.ibge.gov.br/tabela/1612 (acesso em 22/06/2025). Essa tabela contém registros de lavouras temporárias no Brasil entre os anos de 2010 e 2023, com dados desagregados por estado e cultura.

O dataset original contempla variáveis como área plantada (hectares), área plantada - percentual do total, área colhida, área colhida - percentual do total, quantidade produzida (toneladas), valor da produção (reais) e valor da produção - percentual do total geral. Para este projeto, foi selecionada uma cultura agrícola de grande relevância nacional: **soja**, cujos dados foram tratados no arquivo `tabela1612_soja.csv`.

A motivação para este estudo reside na possibilidade de, com ferramentas simples e dados acessíveis, oferecer um modelo preditivo que auxilie produtores, pesquisadores e tomadores de decisão. Embora o modelo adotado não leve em consideração fatores externos como clima ou fertilidade do solo, ele ainda pode fornecer estimativas úteis, especialmente em situações de planejamento em larga escala ou quando informações adicionais não estão disponíveis.

Dessa forma, este trabalho contribui para o uso da ciência de dados na agricultura, evidenciando como métodos estatísticos podem ser aplicados na previsão de safras e como isso pode beneficiar a gestão do setor agropecuário no Brasil.

2 Fundamentação Teórica

2.1 Ciência de Dados

A ciência de dados é um campo multidisciplinar que combina estatística, ciência da computação e conhecimento de domínio para extrair informações úteis a partir de grandes volumes de dados. Ela engloba processos de coleta, limpeza, análise, modelagem e visualização de dados, com o objetivo de gerar conhecimento ou tomar decisões com base em evidências quantitativas.

Nos últimos anos, a ciência de dados tem sido amplamente aplicada em setores como saúde, finanças, marketing e, mais recentemente, na agricultura. Nesse contexto, ela permite o uso de algoritmos preditivos para estimar safras, diagnosticar doenças em plantas, otimizar o uso de insumos e prever condições climáticas, contribuindo para uma agricultura mais eficiente e sustentável.

2.2 Regressão Linear Múltipla

A regressão linear múltipla é uma técnica estatística utilizada para modelar a relação entre uma variável dependente contínua e duas ou mais variáveis independentes. Sua forma geral pode ser representada por:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

Onde:

- Y é a variável dependente (ex: área colhida ou valor da produção);
- $X_1, X_2, \ldots, X_p$ são as variáveis independentes (ex: área plantada, estado, ano);
- β_0 é o intercepto;
- $\beta_1, \ldots, \beta_p$ são os coeficientes associados a cada variável;
- ϵ é o termo de erro.

As métricas mais comuns para avaliação de modelos de regressão incluem o coeficiente de determinação (R^2), que mede o quanto da variabilidade da variável dependente é explicada pelo modelo, e o erro quadrático médio (RMSE), que indica o desvio médio das previsões em relação aos valores reais.

2.3 Indicadores Agrícolas

A Tabela 1612 do SIDRA/IBGE, utilizada neste trabalho, apresenta os seguintes indicadores:

- **Área plantada (hectares)**
- **Área plantada - percentual do total**
- **Área colhida (hectares)**
- **Área colhida - percentual do total**
- **Quantidade produzida (toneladas)**
- **Valor da produção (reais)**
- **Valor da produção - percentual do total geral**

Esses dados estão disponíveis anualmente de 2010 a 2023, para todos os estados brasileiros, e fornecem uma visão ampla e confiável sobre a evolução da produção agrícola no país.

2.4 Aplicações da Ciência de Dados na Agricultura

Diversos estudos têm demonstrado a eficácia da ciência de dados no campo agrícola. Modelos preditivos baseados em regressão, redes neurais artificiais e algoritmos de aprendizado de máquina têm sido aplicados para prever produtividade, identificar padrões de cultivo, otimizar o uso de água e fertilizantes e antecipar impactos climáticos.

Exemplos incluem o uso de modelos para estimar safras com base em séries temporais, sensoriamento remoto e dados climáticos históricos. Apesar disso, a simplicidade e interpretabilidade da regressão linear ainda a tornam uma escolha viável, especialmente quando os dados disponíveis são estruturados e a interpretação do modelo é tão importante quanto a precisão da previsão.

3 Desenvolvimento

Este capítulo descreve todas as etapas executadas no desenvolvimento do modelo preditivo em conjunto com a análise de dados para previsão de safras agrícolas. O processo foi dividido em: **aquisição, pré-processamento, transformação, análise exploratória e modelagem**.

3.1 Limpeza dos Dados

Durante a leitura dos dados, foi identificado o uso do símbolo "-" para representar valores nulos ou inexistentes. Esses valores foram substituídos por `NaN`, e as colunas numéricas foram convertidas para o tipo `float`, utilizando o parâmetro `errors='coerce'` da função `pd.to_numeric()`.

```
df['Valor'] = df['Valor'].replace('-', np.nan)
df['Valor'] = pd.to_numeric(df['Valor'], errors='coerce')
```

Esse processo garantiu a integridade dos dados numéricos para as etapas seguintes.

3.2 Transformação dos Dados

Os dados originais estavam em formato longo, ou seja, cada linha representava uma combinação de variável, estado e ano. Para facilitar a análise, foi realizada uma operação de *pivotamento*, transformando as variáveis em colunas. Assim, cada linha passou a representar uma combinação única de estado e ano, com colunas para área plantada, área colhida, produção, entre outras.

```
df_pivot = df.pivot_table(index=['Unidade da Federação', 'Ano'],
                          columns='Variável', values='Valor').reset_index()
```

Após o pivotamento, as colunas foram renomeadas para facilitar a leitura e o entendimento do dataset, adotando nomes como `Area_Plantada`, `Area_Colhida`, `Quantidade_Produzida` e `Valor_Producao`.

3.3 Análise Exploratória de Dados

Para o conjunto de dados de safras de soja dos estados brasileiros de 2010 a 2023, realizamos uma análise exploratória de dados (EDA) para compreender as relações entre as

variáveis. A EDA é uma etapa fundamental no processo de análise de dados, que busca resumir as principais características do conjunto de dados, frequentemente por meio de métodos visuais e estatísticos. Seu objetivo é identificar padrões, detectar anomalias, verificar suposições e formular hipóteses para análises subsequentes, fornecendo uma base sólida para a modelagem estatística e a tomada de decisões.

No contexto deste estudo, a EDA foi conduzida com foco nas variáveis relacionadas à produção de soja, incluindo "Area_Plantada" (área plantada em hectares), "Area_Colhida" (área colhida em hectares), "Quantidade_Produzida" (quantidade produzida em toneladas), "Estado" (unidade da federação) e "Ano" (ano da safra). Para explorar as relações entre essas variáveis, realizamos dois testes de hipótese utilizando modelos de regressão linear múltipla, com "Quantidade_Produzida" e "Area_Colhida" como variáveis dependentes, e "Area_Plantada", "Estado" e "Ano" como variáveis independentes. Esses testes visam avaliar a significância estatística de variáveis específicas e a adequação geral dos modelos, contribuindo para a compreensão dos fatores que influenciam a produção agrícola.

3.3.1 Testes de Hipótese Realizados

Dois testes de hipótese foram conduzidos para investigar as relações entre as variáveis:

1. **Teste 1: Efeito da Area_Plantada na Quantidade_Produzida**
   Este teste avalia se a variável "Area_Plantada" tem um efeito estatisticamente significativo na "Quantidade_Produzida". A hipótese nula (H_0) postula que o coeficiente de "Area_Plantada" no modelo de regressão linear é igual a zero, ou seja, não há efeito. A hipótese alternativa (H_1) afirma que o coeficiente é diferente de zero, indicando um efeito significativo. O teste foi realizado utilizando um teste t, que compara a estimativa do coeficiente com seu erro padrão, controlando para as variáveis categóricas "Estado" (codificada como variáveis dummy) e "Ano" (tratada como numérica). Este teste é relevante para determinar se o aumento da área plantada está associado a um aumento na produção de soja, uma questão central para o planejamento agrícola.

2. **Teste 2: Significância Geral do Modelo para Area_Colhida**
   Este teste verifica se o modelo de regressão linear para "Area_Colhida", que inclui "Area_Plantada", "Estado" e "Ano" como preditoras, é estatisticamente significativo como um todo. A hipótese nula (H_0) assume que todos os coeficientes das variáveis independentes são iguais a zero, ou seja, o modelo não tem poder explicativo. A hipótese alternativa (H_1) sugere que pelo menos um coeficiente é diferente de zero, indicando que o modelo explica uma porção significativa da variância em "Area_Colhida". O teste foi conduzido utilizando o teste F, que compara a variância explicada pelo modelo com a variância residual. Este teste é essencial para validar a utilidade do modelo na previsão

da área colhida, informando se as variáveis selecionadas são relevantes para explicar as variações observadas.

3.3.2 Resultados dos Testes

Os testes de hipótese foram realizados utilizando a biblioteca `statsmodels` em Python, com os modelos ajustados por meio de regressão linear ordinária (OLS). Os resultados são apresentados a seguir, com nível de significância adotado de 5% ($\alpha = 0,05$).

- **Teste 1: Efeito da Area_Plantada na Quantidade_Produzida**
  O teste t para o coeficiente de "Area_Plantada" no modelo de regressão para "Quantidade_Produzida" resultou em um valor-p de aproximadamente $7,76 \times 10^{-207}$. Este valor é extremamente pequeno, muito inferior ao limiar de 0,05, levando à rejeição da hipótese nula. Portanto, conclui-se que a área plantada tem um efeito estatisticamente significativo na quantidade produzida de soja, controlando para os efeitos de estado e ano. Este resultado sugere que, em média, um aumento na área plantada está associado a um aumento na produção, reforçando a importância de estratégias de expansão agrícola para maximizar a produtividade.

- **Teste 2: Significância Geral do Modelo para Area_Colhida**
  O teste F para o modelo de regressão linear de "Area_Colhida" resultou em um valor-p de 0,0, indicando que o p-valor é menor que a precisão numérica da máquina (tipicamente $< 10^{-16}$). Este valor é significativamente inferior a 0,05, levando à rejeição da hipótese nula. Assim, conclui-se que o modelo, que inclui "Area_Plantada", "Estado" e "Ano" como preditoras, é estatisticamente significativo, ou seja, pelo menos uma das variáveis independentes contribui para explicar as variações na área colhida. Este resultado valida a relevância do modelo para previsão e análise da área colhida, sugerindo que as variáveis selecionadas capturam fatores importantes que influenciam a extensão da área efetivamente colhida.

3.3.3 Conclusão

Os resultados dos testes de hipótese fornecem insights valiosos sobre as relações entre as variáveis no conjunto de dados de safras de soja dos estados brasileiros de 2010 a 2023:

- A área plantada exerce um impacto significativo na quantidade produzida de soja, conforme evidenciado pelo valor-p extremamente baixo do teste t ($7,76 \times 10^{-207}$). Este achado implica que políticas e estratégias que promovam a expansão da área plantada podem ser eficazes para aumentar a produção de soja, desde que outros fatores, como condições climáticas e práticas agrícolas, sejam considerados.

- O modelo de regressão para a área colhida é estatisticamente significativo, com um valor-p do teste F de 0,0, indicando que as variáveis "Area_Plantada", "Estado" e "Ano" são relevantes para explicar as variações na área colhida. Este resultado sugere que o modelo é uma ferramenta útil para prever a área colhida e entender os fatores que influenciam a proporção da área plantada que é efetivamente colhida.

Esses achados são fundamentais para o planejamento agrícola, pois fornecem evidências estatísticas que podem orientar decisões sobre alocação de recursos, estratégias de plantio e políticas agrícolas. Para análises futuras, recomenda-se explorar os coeficientes individuais de cada estado e ano, bem como investigar possíveis relações não lineares ou interações entre as variáveis, utilizando modelos mais complexos, como regressão não linear ou algoritmos de aprendizado de máquina.

3.4 Tratamento de Valores Ausentes

Foram identificados valores ausentes em algumas combinações de estado e ano. Como se tratavam de registros esparsos, optou-se pela exclusão dessas linhas para evitar o comprometimento do modelo, já que representavam uma fração mínima dos dados totais.

3.5 Seleção de Variáveis

Dentre todas as variáveis disponíveis, foram utilizadas as seguintes para modelagem:

- **Variáveis independentes**: Área plantada, estado (codificado posteriormente), ano;
- **Variáveis dependentes**: Área colhida e valor da produção.

A codificação da variável `Estado` foi feita por meio de *one-hot encoding*. Essa técnica transforma uma variável categórica em múltiplas variáveis binárias (0 ou 1). Para cada categoria distinta (no caso, cada estado brasileiro), é criada uma nova coluna no conjunto de dados. Um valor 1 é atribuído à coluna correspondente ao estado de origem da observação, e 0 às demais.

3.6 Divisão entre Treinamento e Teste

Os dados foram divididos entre conjunto de treinamento e de teste, comumente utilizando uma proporção de 80% para treino e 20% para teste. Essa separação é essencial para validar a capacidade preditiva do modelo em dados não vistos.

3.7 Treinamento dos Modelos

A definição das variáveis independentes (features) e dependentes (targets) foi feita da seguinte forma:

- **Variáveis independentes (X)**: Área plantada, ano e estado (já codificado);
- **Variáveis dependentes (y)**: Quantidade produzida e área colhida.

Exemplo do código utilizado para o modelo de soja:

```
X = df_pivot_soja[['Estado', 'Area_Plantada', 'Ano']]
y_quantidade = df_pivot_soja['Quantidade_Produzida']
y_area_colhida = df_pivot_soja['Area_Colhida']

modelo_quantidade = LinearRegression()
modelo_quantidade.fit(X_train, y_quantidade_train)

modelo_area_colhida = LinearRegression()
modelo_area_colhida.fit(X_train, y_area_colhida_train)
```

Dois modelos de regressão linear múltipla foram treinados para soja:

- Previsão da quantidade produzida;
- Previsão da área colhida.

3.8 Avaliação dos Modelos

A performance dos modelos foi avaliada por meio das métricas estatísticas:

- **Erro Quadrático Médio (RMSE)**: Mede o desvio médio quadrático entre as previsões e os valores reais;
- **Coeficiente de Determinação (R^2)**: Indica a proporção da variabilidade dos dados explicada pelo modelo.

Para o modelo de soja, foram obtidos os seguintes resultados:

- RMSE (Quantidade produzida): 2.461.662.352.880,37
- R^2 (Quantidade produzida): 0,9424

- RMSE (Área colhida): 444.540.571,60
- R^2 (Área colhida): 0,9999

Além disso, foram produzidos gráficos de dispersão comparando valores previstos e observados, bem como gráficos de resíduos, que permitem analisar visualmente o padrão dos erros do modelo e verificar suposições importantes como a homocedasticidade e a normalidade dos resíduos. Esses gráficos auxiliaram na avaliação qualitativa da adequação dos modelos às bases de dados utilizadas.

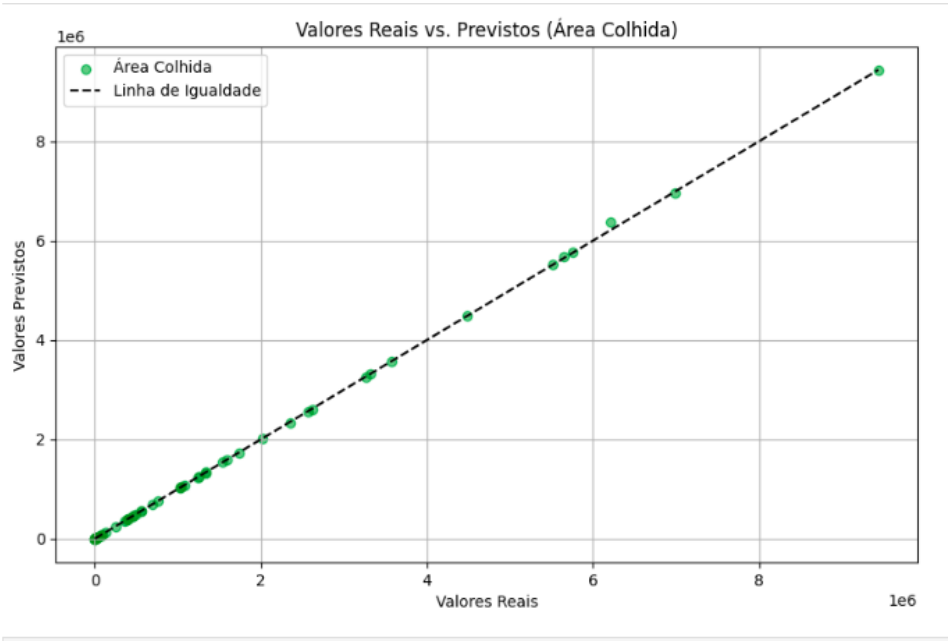

Figura 1 – Valores Reais x Previstos - Área Colhida

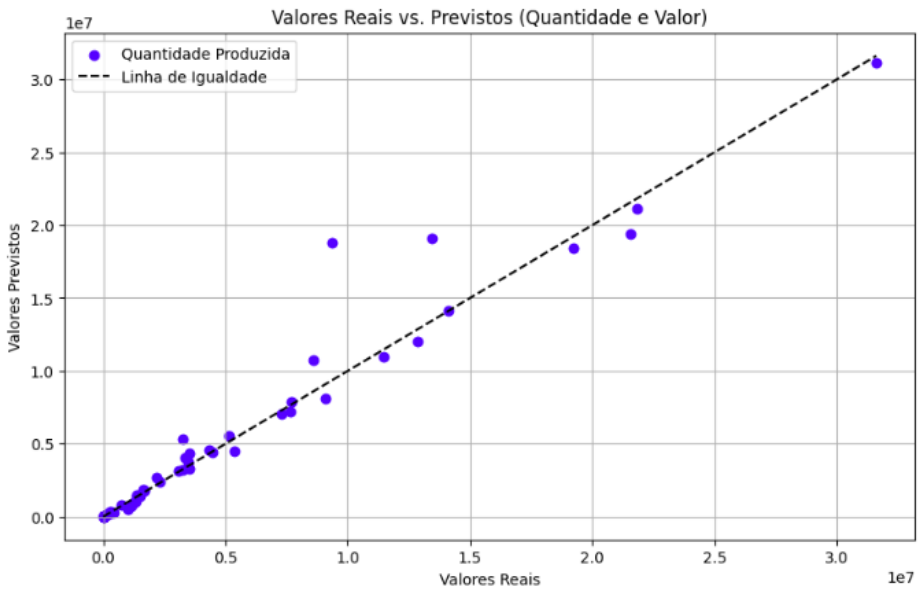

Figura 2 – Valores Reais x Previstos - Quantidade e Valor

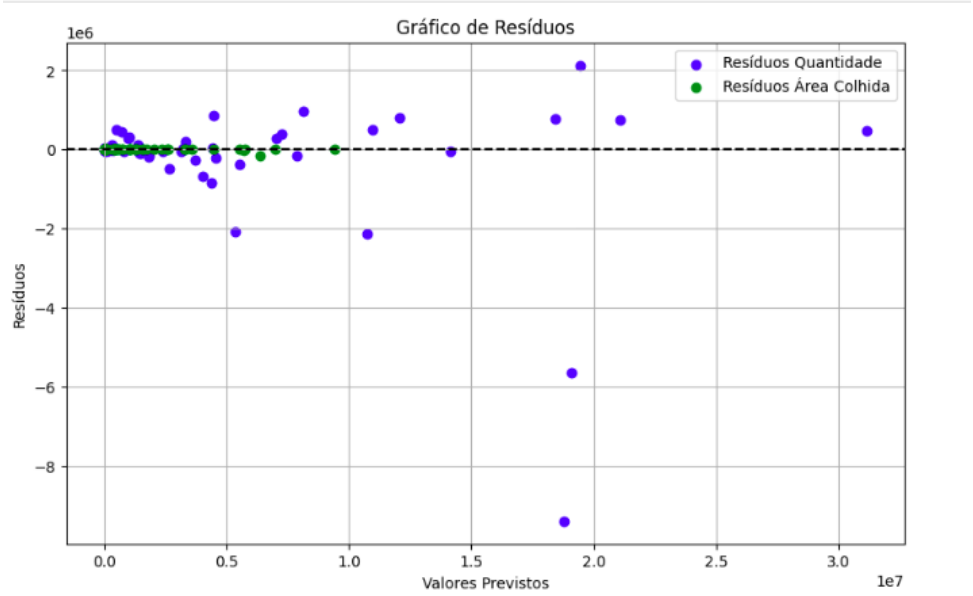

Figura 3 – Resíduos

3.9 Exemplo de Aplicação dos Modelos

Para ilustrar a aplicação prática dos modelos desenvolvidos, foram realizadas previsões de quantidade produzida e área colhida para diferentes estados brasileiros, considerando o mesmo ano e área plantada.

O exemplo a seguir demonstra a geração de previsões para os estados de São Paulo, Mato Grosso e Paraná, para o ano de 2025 e uma área plantada de 3.694.468 hectares:

```
[21]: # Testa previsões para diferentes estados com mesmos valores
estados_teste = ['São Paulo', 'Mato Grosso', 'Paraná']
novos_dados = pd.DataFrame({
    'Estado': estados_teste,
    'Ano': [2025] * len(estados_teste),
    'Area_Plantada': [3694468.0] * len(estados_teste),
})

print(novos_dados.head(3))
novos_dados_encoded = preprocessor.transform(novos_dados)
quantidade_prevista = modelo_quantidade.predict(novos_dados_encoded)
area_colhida_prevista = modelo_area_colhida.predict(novos_dados_encoded)

# Imprime previsões para diferentes estados
print("\nPrevisões para diferentes estados (Ano=2023, Área Plantada=100000, Área Colhida=100000):")
for estado, qtd, area in zip(estados_teste, quantidade_prevista, area_colhida_prevista):
    print(f"{estado}: Quantidade = {qtd:.2f} toneladas, Área Colhida = {area:.2f} hectares")
```

```
        Estado   Ano  Area_Plantada
0    São Paulo  2025      3694468.0
1  Mato Grosso  2025      3694468.0
2       Paraná  2025      3694468.0

Previsões para diferentes estados (Ano=2023, Área Plantada=100000, Área Colhida=100000):
São Paulo: Quantidade = 14392360.36 toneladas, Área Colhida = 3696166.86 hectares
Mato Grosso: Quantidade = 7798158.80 toneladas, Área Colhida = 3677696.02 hectares
Paraná: Quantidade = 11019632.94 toneladas, Área Colhida = 3691825.15 hectares
```

Figura 4 – Exemplo de Previsão Agrícola

Saída obtida:

	Estado	Ano	Area_Plantada
0	São Paulo	2025	3694468.0
1	Mato Grosso	2025	3694468.0
2	Paraná	2025	3694468.0

Previsões para diferentes estados (Ano=2025, Área Plantada=3694468):

- São Paulo: Quantidade = 14392360.36 toneladas, Área Colhida = 3696166.86 hectares
- Mato Grosso: Quantidade = 7798158.80 toneladas, Área Colhida = 3677696.02 hectares
- Paraná: Quantidade = 11019632.94 toneladas, Área Colhida = 3691825.15 hectares

Esse exemplo demonstra como os modelos são capazes de diferenciar a produção e área colhida estimadas conforme o estado, mesmo para uma mesma área plantada e ano, evidenciando a importância da variável categórica `Estado` na predição.

4 Conclusão

Este trabalho apresentou o desenvolvimento de modelos preditivos para a previsão de safras agrícolas, com foco nas culturas de soja e milho, utilizando técnicas de ciência de dados e aprendizado de máquina.

O processo iniciou-se com a aquisição e limpeza dos dados, onde foram tratados valores ausentes e dados inconsistentes para garantir a qualidade da base utilizada. A transformação dos dados possibilitou a reorganização das informações, facilitando a análise e o desenvolvimento dos modelos.

A análise exploratória dos dados permitiu compreender melhor as características das variáveis envolvidas, identificando padrões e correlações importantes que fundamentaram a escolha das variáveis independentes e dependentes no processo de modelagem. Testes estatísticos auxiliaram na validação dessas escolhas.

Dois modelos de regressão linear múltipla foram treinados, um para prever a quantidade produzida e outro para a área colhida, considerando variáveis como área plantada, ano e o estado da federação, este último codificado por meio de *one-hot encoding*. A avaliação dos modelos indicou bom desempenho, com altos coeficientes de determinação (R^2) e baixos erros quadráticos médios (RMSE), especialmente para os estados com maior volume e consistência histórica de dados. A análise gráfica, incluindo gráficos de dispersão e de resíduos, reforçou a adequação dos modelos e auxiliou na verificação das suposições estatísticas.

Por fim, foi apresentado um exemplo prático de aplicação dos modelos, demonstrando a capacidade de gerar previsões específicas para diferentes estados, considerando a mesma área plantada e ano, evidenciando a importância das variáveis categóricas na predição.

Como trabalhos futuros, sugere-se a exploração de modelos não lineares e técnicas avançadas de aprendizado de máquina, bem como a incorporação de variáveis climáticas e socioeconômicas que possam influenciar a produção agrícola, buscando aprimorar a capacidade preditiva e a robustez dos modelos.

Este projeto contribuiu para a compreensão da dinâmica das safras agrícolas e apresentou um caminho viável para a utilização de dados históricos na previsão de produção, podendo servir de base para planejamento agrícola e tomada de decisão em políticas públicas e privadas.