A Inteligência Artificial na Previsão da Produtividade do Agronegócio: Um Estudo de Caso

Lucas O. Santos, Gabriel de S. Kelly, Raul S. de Paiva, Ricardo P. Mesquita

Ciência da Computação – Centro Universitário Carioca (UNICARIOCA)
29735-160 – Rua Venceslau – Méier – Rio De Janeiro – RJ – Brasil

Lucasoriental@gmail.com, raul.sena15@gmail.com,
gabrielkellyone@gmail.com, rmesquita@unicarioca.edu.br

Abstract. *In this article, the application of the Decision Tree Regressor model was explored to predict agricultural production (in tons per hectare) based on categorical and continuous variables. The study used a public dataset collected from Kaggle and popular libraries such as Pandas, NumPy, and Scikit-learn for data analysis and modeling. The model was trained and evaluated using historical data, which included information on soil type, climatic conditions, and geographic region. The developed code was hosted on GitHub. This work highlights the potential of Artificial Intelligence in agribusiness, contributing to planning and decision-making in the sector.*

Resumo. *Neste artigo, foi explorada a aplicação do modelo Decision Tree Regressor na previsão da produção agrícola (em toneladas por hectare), com base em variáveis categóricas e contínuas. O estudo utilizou um dataset público coletado no Kaggle e bibliotecas populares como Pandas, NumPy e Scikit-learn para a análise e modelagem dos dados. O modelo foi treinado e avaliado com dados históricos, que incluíram informações sobre tipo de solo, condições climáticas e região geográfica. O código desenvolvido foi hospedado no GitHub. Este trabalho destaca o potencial da Inteligência Artificial no agronegócio, contribuindo para o planejamento e a tomada de decisões no setor.*

1. Introdução

Nos últimos anos, a crescente demanda por alimentos tem desafiado a capacidade do setor agropecuário, e o Brasil, como uma das principais potências agrícolas, desempenha um papel crucial nesse cenário. Em 2022, o agronegócio representou 25,2% do PIB brasileiro (CEPEA, 2023), reforçando sua importância econômica. Para manter sua posição no mercado internacional, é essencial adotar inovações tecnológicas que otimizem a produção agrícola.

Nesse contexto, a Inteligência Artificial (IA) surge como uma ferramenta estratégica, com o aprendizado de máquina (*machine learning*) destacando-se como um recurso relevante para o agronegócio. Essa abordagem permite avanços em áreas como a previsão de rendimento e o controle de insumos, entre outras aplicações. O uso de algoritmos como o *Decision Tree Regressor* têm se mostrado promissores na tarefa de prever a produtividade agrícola, considerando variáveis como condições climáticas, características do solo e práticas de manejo.

Este trabalho busca explorar o uso do *Decision Tree Regressor* na previsão da produtividade agrícola, considerando a relação entre os principais fatores que influenciam os rendimentos e fornecendo informações relevantes para apoiar a tomada de decisões no setor. A pesquisa apresenta uma revisão sobre Inteligência Artificial (IA) e aprendizado de máquina, descreve a implementação do modelo e discute os resultados obtidos, com o propósito de destacar o potencial da tecnologia para oferecer soluções viáveis e sustentáveis aos desafios do agronegócio.

2. Aprendizado de Máquina

2.1. Inteligência Artificial

A Inteligência Artificial (IA) é um campo multidisciplinar da ciência da computação dedicado à criação de sistemas capazes de realizar tarefas que exigem inteligência humana, tais como, reconhecimento de voz, tomada de decisões, tradução de idiomas e análise de grandes volumes de dados, reconhecimento de imagens, diagnósticos médicos, automação de processos industriais, personalização de recomendações, detecção de fraudes, condução autônoma e desenvolvimento de agentes inteligentes para jogos e simulações (RUSSELL & NORVIG, 2010). Desde suas origens, a IA tem evoluído, influenciada por marcos históricos significativos, como o Teste de Turing proposto por Alan Turing e os primeiros desenvolvimentos em processamento de linguagem natural (PINAR SAYGIN, 2000).

A IA pode ser classificada em duas categorias principais: IA fraca (*Weak AI*) e IA forte (*Strong AI*) (FLOWERS, 2019). A IA fraca, ou IA estreita, é projetada para realizar tarefas específicas em um determinado domínio e não possui consciência ou entendimento. Todos os sistemas de IA existentes, desde o processamento de linguagem natural até a triagem de imagens, podem ser classificados como IA fraca, pois operam dentro de limites definidos (BORY *et al.*, 2024). Exemplos incluem assistentes virtuais, como Siri e Alexa, que executam comandos com base em algoritmos de reconhecimento de voz (HAIKONEN, 2020). Por outro lado, a IA forte se refere a sistemas que teriam a capacidade de entender, aprender e aplicar inteligência de maneira semelhante à humana, tais como, um robô capaz de realizar diagnósticos médicos complexos ou aprender novas habilidades de forma autônoma. No entanto, essa forma de IA ainda está

longe de ser alcançada, permanecendo no campo da teoria e sendo amplamente explorada no campo da ficção científica (BORY *et al.*, 2024).

Nos últimos anos, o avanço em áreas como aprendizado de máquina e processamento de linguagem natural permitiu que a IA se integrasse em diversas aplicações práticas, impactando setores como saúde, finanças, transporte e entretenimento (STAHL *et al.*, 2023). No setor de saúde, por exemplo, algoritmos de IA têm sido utilizados para diagnosticar doenças com precisão, analisando grandes quantidades de dados médicos (AERTS & HOSNY, 2019). Na indústria financeira, sistemas de IA são usados para prever tendências de mercado e detectar fraudes, demonstrando a versatilidade e a importância da IA na vida cotidiana (ASHTIANI & RAAHEMI, 2021). As aplicações práticas da IA, como essas, serão abordadas mais adiante neste trabalho.

2.2. O que é Aprendizado de Máquina?

O aprendizado de máquina (*machine learning*) é uma área importante da inteligência artificial que permite que os sistemas de computador aprendam sozinhos e melhorem suas habilidades a partir de dados, sem que precisem ser programados para cada tarefa específica (MITCHELL, 1997). Esse processo ocorre por meio da análise de grandes volumes de dados, na qual os sistemas identificam padrões e realizam previsões, aprimorando continuamente seu desempenho com base nas informações adquiridas (MITCHELL, 1997).

O aprendizado de máquina pode ser classificado em três categorias principais: aprendizado supervisionado, aprendizado não supervisionado e aprendizado por reforço (ESCALANTE & MORALES, 2022). No aprendizado supervisionado, modelos são treinados com um conjunto de dados rotulados, onde a saída é conhecida, permitindo que o sistema aprenda a mapear entradas para as saídas. Por exemplo, em um sistema de reconhecimento de imagem, o modelo é treinado com imagens de gatos e cães, identificando cada uma corretamente para prever a categoria de novas imagens (LUDERMIR, 2021).

Em contraste, no aprendizado não supervisionado, os dados fornecidos ao algoritmo não possuem rótulos. O algoritmo analisa os exemplos e identifica padrões ou similaridades em seus atributos, agrupando-os em *clusters* (grupos de dados que compartilham características semelhantes). Esse processo permite que o algoritmo descubra estruturas subjacentes nos dados sem orientação externa. Após a formação dos agrupamentos, geralmente é necessária uma análise para interpretar o significado de cada *cluster* no contexto do problema em questão, ajudando a entender como os dados se relacionam entre si (LUDERMIR, 2021).

O aprendizado por reforço é um processo em que um agente interage com um ambiente e aprende a tomar decisões com base em recompensas e punições (LUDERMIR, 2021). Em vez de receber respostas corretas, o algoritmo avalia a qualidade de suas ações, permitindo que ele ajuste seu comportamento ao longo do tempo. Esse método é amplamente utilizado em áreas como jogos e robótica, sendo um exemplo notável o *AlphaGo*, que derrotou campeões mundiais no jogo de *Go* (CHEN, 2021; OTTERLO & WIERING, 2012; BARTO & SUTTON, 2018).

Além disso, o aprendizado de máquina possui uma ampla gama de algoritmos e técnicas, incluindo regressão, árvores de decisão, máquinas de vetor de suporte (*SVM*) e

redes neurais, cada uma com suas próprias características e aplicações (BHATIA *et al.*, 2007).

2.3. Para que é usado o Aprendizado de Máquina e Exemplos de Aplicações

O aprendizado de máquina é utilizado em diversas aplicações práticas, desde a saúde até o entretenimento, demonstrando sua capacidade de resolver problemas complexos e otimizar processos (STAHL *et al.*, 2023).

Na área da saúde, sistemas baseados em aprendizado de máquina têm se mostrado eficazes na análise de imagens médicas, como radiografias e ressonâncias magnéticas, ajudando a identificar condições como câncer de pele e doenças oculares com precisão comparável à dos especialistas humanos (BELTRAMI, 2022; JI, 2022). Esses avanços não apenas melhoram a precisão dos diagnósticos, mas também reduzem custos e tempo na análise de exames (RAJPURKAR, 2022). A integração dessas tecnologias na prática clínica pode diminuir a carga de trabalho dos médicos, permitindo que se concentrem mais em cuidados diretos ao paciente (RAJPURKAR, 2022).

Na indústria financeira, o aprendizado de máquina é amplamente utilizado para prever flutuações de mercado, gerenciar riscos e detectar fraudes (ASHTIANI & RAAHEMI, 2021). Sistemas de monitoramento em tempo real analisam transações em busca de padrões incomuns que possam indicar atividades fraudulentas, garantindo proteção tanto para consumidores quanto para instituições financeiras (KOTAGIRI, 2023). Além disso, algoritmos de aprendizado de máquina estão sendo utilizados para desenvolver sistemas de crédito que avaliam a solvência de clientes com base em um conjunto mais amplo de dados, promovendo decisões mais informadas e justas (EL MAKNOUZI *et al.*, 2022).

No setor de transporte, empresas como Uber utilizam aprendizado de máquina para otimizar rotas e prever a demanda por serviços de carona, melhorando a eficiência operacional e a experiência do usuário (SRINIVAS *et al.*, 2021). No ambiente de varejo, algoritmos de aprendizado de máquina ajudam a personalizar recomendações de produtos com base no histórico de compras e preferências dos clientes, resultando em uma experiência de compra mais satisfatória e aumento nas taxas de conversão (ADEBAYO & KONGAR, 2021).

Além disso, o aprendizado de máquina é encontrado em veículos autônomos, onde empresas como Tesla dependem de algoritmos para o reconhecimento de objetos e a tomada de decisões em tempo real, permitindo que os veículos naveguem com segurança em ambientes complexos (AJITHA & NAGRA, 2021). Redes neurais convolucionais (*CNNs*) são utilizadas para reconhecer e classificar objetos em imagens, sendo crucial para a operação eficaz desses carros (JURASZEK, 2014; LIMA, 2019; VARGAS *et al.*, 2016).

Assistentes pessoais como Siri e Google Assistant também utilizam aprendizado de máquina para compreender comandos de voz e responder a perguntas, melhorando a interação do usuário com a tecnologia (BELLEGARDA, 2013; MICHAELY *et al.*, 2017). No marketing digital, plataformas de anúncios aplicam aprendizado de máquina para segmentar públicos e otimizar campanhas, analisando dados de comportamento do consumidor para criar mensagens personalizadas e eficazes (VAN ESCH & STEWART BLACK, 2021).

3. Árvore de Decisão

3.1. Aprendizado Supervisionado

O aprendizado supervisionado é um método de aprendizado de máquina onde o modelo é treinado com dados que já possuem respostas corretas definidas (CUNNINGHAM *et al.*, 2008). Esses dados incluem pares de entrada e saída com base nas características de entrada fornecidas (CUNNINGHAM *et al.*, 2008). A abordagem supervisionada se baseia em um conjunto de treinamento, onde cada exemplo contém um conjunto de características de entrada e um rótulo correspondente que o algoritmo deve prever (JIANG *et al.*, 2020).

Uma vez que o modelo é treinado, ele pode ser aplicado a novos dados não rotulados, onde a tarefa é generalizar bem para esses exemplos desconhecidos (JIANG *et al.*, 2020). Para melhorar a capacidade preditiva do modelo, técnicas como *cross-validation* e regularização (Um consiste em dividir os dados em partes (*folds*) para treinar e testar um modelo, permitindo avaliações e o outro adiciona penalizações à função de custo, respectivamente, na qual o objetivo é evitar problemas relacionados ao *overfitting*) são comumente usadas (YING, 2019).

Existem dois tipos principais de aprendizado supervisionado, a classificação e a regressão (MAHESH, 2020).

Relativamente à classificação, esta envolve a previsão de rótulos de classes discretas ou rótulo para um dado de entrada (KOTSIANTIS *et al.*, 2007). O objetivo é dividir os dados de entrada em classes pré-definidas, atribuindo a cada amostra uma etiqueta categórica, como "sim" ou "não", ou "positivo" e "negativo" (ALBALDAWI & ALNUAIMI, 2024). Os algoritmos de classificação aprendem com um conjunto de dados rotulados, onde a variável de saída é categórica, e ajustam o modelo para que ele possa predizer corretamente o rótulo das novas entradas (ALBALDAWI & ALNUAIMI, 2024). Esse tipo de tarefa é avaliado geralmente por métricas como acurácia, precisão, *recall* e *F1-score* (AXMAN & YACOUBY, 2020). Esses algoritmos são geralmente usados em problemas como detecção de fraudes e classificação de textos (ASHTIANI & RAAHEMI, 2021).

A regressão, por outro lado, é usada para prever valores contínuos, como preços ou temperaturas (DEQIAN *et al.*, 2023; SINGH *et al.*, 2020). O modelo de regressão tenta prever valores numéricos baseados em dados de entrada, ajustando uma função que minimize a diferença entre os valores reais e previstos (BHATNAGAR & KUMAR, 2021). Diferente da classificação, em que o resultado pertence a uma classe, na regressão a previsão é um valor contínuo que pode variar dentro de um intervalo (BHATNAGAR & KUMAR, 2021). Modelos de regressão são avaliados frequentemente por métricas como erro quadrático médio (*Mean Squared Error - MSE*) e erro absoluto médio (*Mean Absolute Error - MAE*) (MURPHY, 2012).

Em termos de desempenho, a qualidade do modelo supervisionado depende fortemente da quantidade e qualidade dos dados rotulados disponíveis, bem como o modelo deve ser capaz de evitar o *overfitting* e o *underfitting*, que ocorre quando ele se ajusta excessivamente aos dados de treinamento, perdendo sua capacidade de generalização e quando o modelo é superficial demais, sendo incapaz de capturar a complexidade dos padrões nos dados (YING, 2019).

3.2. Regressão Baseado em Árvore de Decisão

O algoritmo *Decision Tree Regressor* segue um processo de particionamento recursivo, onde os dados são divididos em subconjuntos com base nas variáveis preditoras, até que uma condição de parada seja atingida (LOH, 2011). Cada divisão busca minimizar um critério de erro, frequentemente o Erro Quadrático Médio (*MSE - Mean Squared Error*) ou a soma dos erros quadráticos entre o valor previsto e o valor real (MURPHY, 2012).

Cada nó na árvore representa uma condição de divisão com base em uma variável, e cada ramo corresponde a uma das possíveis saídas dessa condição (LOH, 2011). No final, as folhas da árvore contêm previsões numéricas (geralmente a média dos valores da variável-alvo naquela região). Em termos simples, a árvore busca criar regiões nos dados onde as previsões são o mais homogêneas possível (LOH, 2011).

Os algoritmos para regressão baseado em árvore de decisão oferecem vantagens como a facilidade de interpretação e a capacidade de modelar relações não lineares, tornando-os acessíveis a usuários não técnicos (LOH, 2011). Eles também podem lidar de forma eficiente com dados mistos, incluindo variáveis categóricas e numéricas, e, dependendo da implementação, podem apresentar robustez frente a dados ausentes (LOH, 2011).

No entanto, essas árvores apresentam desvantagens, como o risco de *overfitting* (citado anteriormente) se não forem devidamente podadas, e geralmente têm desempenho inferior a modelos mais avançados, como *Random Forest* e *Gradient Boosting* (MURPHY, 2012). Outro ponto é que os algoritmos de construção de árvores tendem a produzir uma estrutura 'quase ótima', e não totalmente otimizada, o que pode limitar sua precisão e eficácia (MITROFANOV & SEMENKIN, 2021). Portanto, embora sejam uma ferramenta valiosa, é crucial avaliá-las em comparação com outras abordagens de modelagem.

3.3. Formas de Implementação do Algoritmo

A implementação de algoritmos de árvore de decisão para regressão envolve etapas essenciais que podem ser descritas da seguinte forma:

Construção da Árvore: O processo começa com a divisão do conjunto de dados em subconjuntos, com base em critérios de impureza, como a Variância ou o Erro Quadrático Médio (*MSE*) (KITTS, 2000; MURPHY, 2012). A árvore é construída recursivamente, onde cada nó representa uma condição sobre uma característica, e as folhas representam as previsões finais (KITTS, 2000).

Critério de Parada: Durante a construção da árvore, é importante definir critérios de parada para evitar *overfitting,* podendo incluir a definição de um número mínimo de amostras em um nó ou uma profundidade máxima da árvore (KITTS, 2000; YING, 2019). Essas condições garantem que a árvore não se torne excessivamente complexa, mantendo sua capacidade de generalização em dados não vistos (KITTS, 2000).

Poda da Árvore: Após a construção inicial, a poda da árvore pode ser realizada para remover nós que oferecem pouca ou nenhuma melhoria nas previsões, sendo feita no início ou no fim da construção (KITTS, 2000). A poda prévia interrompe a construção da árvore quando o critério de parada é atendido, enquanto a poda posterior envolve a remoção de partes da árvore que não contribuem significativamente para o desempenho (KITTS, 2000).

Avaliação do Modelo: A avaliação do modelo de árvore de decisão é uma etapa fundamental que determina sua precisão e capacidade de generalização (KITTS, 2000). Métodos como a validação cruzada são utilizados para medir a eficácia do modelo, permitindo uma análise mais detalhada ao dividir o conjunto de dados em múltiplos subconjuntos (KITTS, 2000). Isso possibilita a realização de testes em diferentes combinações de dados, fornecendo uma estimativa robusta do desempenho do modelo em situações do mundo real (KITTS, 2000).

4. Previsão de Produtividade Agrícola através do uso de Algoritmo para Regressão Baseado em Árvore de Decisão

O objetivo deste artigo é realizar a previsão da produção agrícola (em toneladas por hectare) através do modelo *Decision Tree Regressor*, com base em um conjunto de variáveis categóricas e contínuas. O algoritmo foi treinado e seu desempenho avaliado por meio de métricas de validação em dados históricos que incluem características como: tipo de solo, condições climáticas e região geográfica.

4.1. Introdução

Em 2022, o PIB do Brasil foi de R$9,9 trilhões, com o agronegócio representando 25,2% deste valor (CEPEA 2023), sendo um setor crucial da economia brasileira. De acordo com a Organização Mundial do Comércio (OMC) em seu relatório "*World Trade Statistical Review 2023",* o Brasil representou 6.4% das exportações de produtos agrícolas no ano de 2022, sendo o terceiro maior exportador atrás dos Estados Unidos e da União Europeia.

Para atender às crescentes demandas globais e melhorar a competitividade do setor, é necessário que o agronegócio se mantenha atento às novas tecnologias que possibilitem o aprimoramento da produção. A Inteligência Artificial (IA) surge como uma ferramenta poderosa, oferecendo novas abordagens para otimizar processos agrícolas, desde a gestão de recursos naturais até a previsão de resultados produtivos. Nesse contexto, o presente estudo propõe a construção de um modelo preditivo utilizando o algoritmo *Decision Tree Regressor*, com o intuito de prever o volume final da colheita, considerando a relação entre fatores de cultivo e sua importância no rendimento da safra.

4.2. Etapas de Implementação da Solução

4.2.1. *Conjunto de dados*

Neste estudo foi utilizado *dataset* sintético de dados agrícolas, obtidos através da plataforma *Kaggle*, contendo 1 milhão de amostras. As variáveis incluem: Região (*Region - North, South, East and West*); Tipo de Solo (*Soil_type - Chalky, Clay, Loam, Peaty, Sandy and Silt*); Cultivo (*Crop - Barley, Cotton, Maize, Rice, Soybean and Wheat*); Volume de Chuva em mm (*Rainfall_mm*); Temperatura Média (*Temperature_Celsius*) durante o cultivo; Uso de Fertilizante (*Fertilizer_Used - True or False*); Uso de Irrigação (*Irrigation_Used - True or False*); Condição Climática (*Weather_Condition - Sunny, Cloudy and Rainy)*; Dias até a Colheita (*Days_to_Harvest*); e Colheita Final em toneladas/hectare (*Yield_tons_per_hectare*).

4.2.2. *Pré-processamento de dados*

Para preparar o *dataset*, foi realizado um pré-processamento que incluiu etapas essenciais para garantir a qualidade dos dados antes de aplicar o modelo de aprendizado de máquina. Primeiramente, as variáveis categóricas foram transformadas em variáveis binárias utilizando a técnica de *One-Hot Encoding,* que transforma categorias em colunas binárias, permitindo que o modelo interprete adequadamente esses dados. Essa técnica foi aplicada às variáveis Região (*Region*), Tipo de Solo (*Soil_type*), Cultura (*Crop*) e Condição Climática (*Weather_Condition*).

Em seguida, para melhorar a qualidade dos resultados, foram removidos *outliers* da variável Rendimento Final (*Yield_tons_per_hectare*). A remoção foi realizada com base no intervalo interquartil (IQR), excluindo os valores que estavam acima do limite superior ou abaixo do limite inferior, definidos como três vezes o IQR. Essa abordagem é amplamente utilizada na detecção de *outliers*, pois ajuda a identificar valores extremos que podem prejudicar o desempenho do modelo, considerando que, em distribuições normais, a maioria dos dados está dentro de três desvios-padrão da média. Além disso, foram eliminadas as linhas em que o rendimento final fosse menor que 0 e 0.1, uma vez que esses valores não são consistentes no contexto do estudo.

Além disso, devido ao grande volume de dados originais (cerca de um milhão de amostras), foi aplicado a técnica de Amostragem Estratificada, reduzindo a base para aproximadamente cem mil amostras e preservando a representatividade dos dados. Esse procedimento contribuiu para diminuir o tempo de processamento e garantir que o modelo pudesse ser treinado de forma mais eficiente, sem comprometer a qualidade das previsões.

Esta etapa foi fundamental para garantir que os dados estivessem prontos para serem utilizados nas previsões do modelo, permitindo uma análise mais adequada entre as variáveis e contribuindo para uma melhor performance do modelo de regressão.

4.3. Implementação, Análise e Avaliação do Modelo Preditivo

A partir dos dados disponíveis, o modelo de predição foi implementado utilizando a classe *DecisionTreeRegressor* da biblioteca *sklearn*. O processo de análise e otimização seguiu três etapas principais: avaliação inicial com o *dataset* original, avaliação após a amostragem estratificada e remoção de outliers, e avaliação após otimização do modelo. A seguir, discutiremos os resultados obtidos em cada uma dessas etapas.

4.3.1. *Avaliação Inicial com o Dataset Original*

Inicialmente, foi realizada uma avaliação das métricas de desempenho do modelo com o *dataset* original, sem remoção dos outliers e nenhuma otimização, apenas com a transformação das variáveis categóricas em contínuas, pois se trata de um algoritmo para regressão. Os resultados podem ser observados na Figura 1.

```
MSE dataset Original: 0.53
R²  dataset Original: 0.81
RMSE dataset Original: 0.73
MAE  dataset Original: 0.58
```

Figura 1. Resultado das métricas de desempenho do modelo com o dataset original.

Além disso, foi gerado um gráfico de dispersão linear (Figura 2) que ilustra a relação entre as previsões e os valores reais.

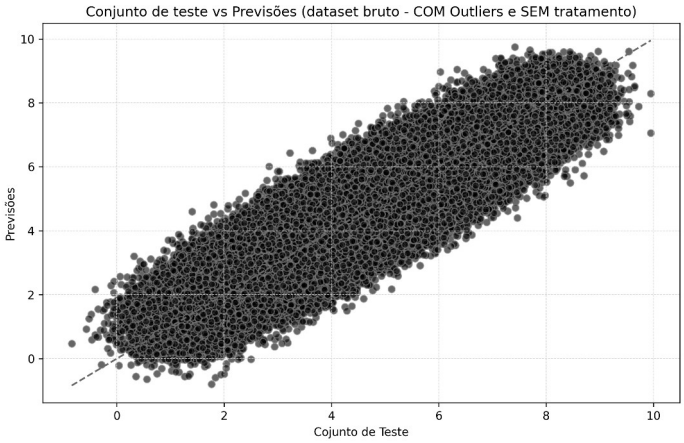

Figura 2. Gráfico de dispersão demonstrando a relação linear entre os valores reais e as previsões do Modelo com o dataset original.

Cada ponto no gráfico representa uma previsão, com sua posição baseada na correspondência entre o valor predito e o real. A linha de perfeição (onde as previsões são iguais aos valores reais) serve como referência: quanto mais próximos os pontos dessa linha, melhor o desempenho do modelo. A dispersão dos pontos reflete a precisão das previsões: menor dispersão indica previsões mais precisas. Embora haja uma correspondência aceitável entre as previsões e os valores reais, nota-se uma dispersão maior, o que indica que o modelo ainda pode ser melhorado.

4.3.2. *Avaliação Após Amostragem Estratificada e Remoção de Outliers*

Após a aplicação da Amostragem Estratificada, foi realizado uma avaliação das métricas de desempenho com e sem a presença de *outliers* (Tabela 1), usando o *dataset* reduzido, resultando em um tempo de carregamento significativamente mais rápido e ilustrando o impacto dos valores atípicos na capacidade preditiva do modelo.

Tabela 1. Comparação de antes e depois da remoção dos Outliers.

Métrica	Com Outliers	Sem Outliers
MSE	0.5238	0.5163
R^2	0.8201	0.8210
RMSE	0.7238	0.7186
MAE	0.5785	0.5737

Os resultados demonstraram que a remoção dos *outliers* teve um impacto mínimo nas métricas de desempenho, com pequenas variações nos valores de *MSE*, *RMSE*, R^2 e *MAE*. No entanto, os gráficos de dispersão linear gerados (Figura 3 e 4) indicaram uma leve melhoria na qualidade da previsão após a remoção dos *outliers*, sugerindo que, embora sua presença não tenha comprometido significativamente a performance do modelo, a remoção deles pode ajudar a evitar possíveis distorções nas previsões e a reduzir o risco de *overfitting* e *underfitting* em modelos de aprendizado de máquina.

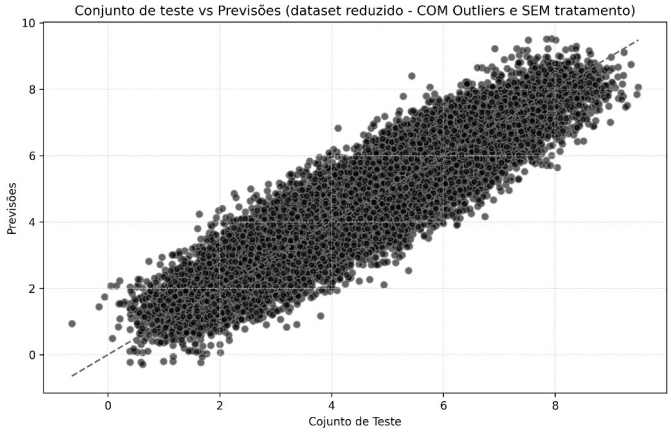

Figura 3. Gráfico de dispersão demonstrando a relação linear entre os valores reais e as previsões do modelo com o dataset reduzido e com presença de outliers.

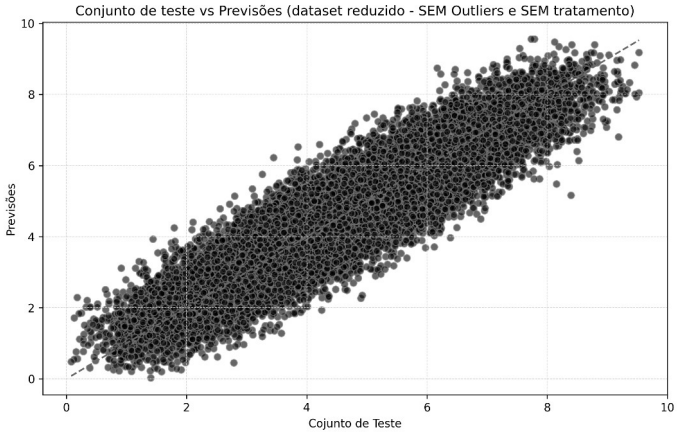

Figura 4. Gráfico de dispersão demonstrando a relação linear entre os valores reais e as previsões do modelo com o dataset reduzido e sem presença de outliers.

4.3.4. *Avaliação e Análise do Modelo Após Otimização*

A etapa de otimização do modelo *DecisionTreeRegressor* foi conduzida visando a melhoria do desempenho preditivo, ajustando os hiperparâmetros principais. Para isso, foi utilizado o método *Grid Search*, pertencente à função *GridSearchCV* da biblioteca *scikit-learn*, em conjunto com validação cruzada, com o objetivo de explorar combinações de valores para os principais hiperparâmetros do modelo. Entre eles, os parâmetros: *max_depth*, que controla a profundidade máxima da árvore, desempenha um papel crucial ao evitar *overfitting* (quando a árvore é excessivamente profunda e captura ruídos dos dados) e *underfitting* (quando a árvore é superficial demais para capturar padrões relevantes); *min_samples_split*, que define o número mínimo de amostras necessárias para que um nó possa ser dividido, influenciando diretamente o crescimento da árvore e sua complexidade; e *min_samples_leaf*, que determina o número mínimo de amostras que um nó terminal deve conter, sendo essencial para evitar a criação de folhas com poucas amostras, podendo ser altamente influenciadas por *outliers* e comprometer a capacidade de generalização do modelo.

Os melhores valores obtidos para os hiperparâmetros foram: *max_depth: 10; min_samples_split: 10; min_samples_leaf: 4.* Com o modelo otimizado, as métricas de avaliação foram recalculadas com base no conjunto de teste, e os resultados são apresentados na Tabela 2.

Tabela 2. Métricas de desempenho do modelo após otimização.

Métrica	Resultado
MSE	0.2633
R^2	0.9087
RMSE	0.5132
MAE	0.4093

Os resultados indicam uma melhoria perceptível no desempenho do modelo em relação às etapas anteriores. O coeficiente de determinação *(R^2)* de 0,9087 sugere que o modelo é capaz de explicar aproximadamente 90,87% da variância nos dados analisados, enquanto métricas como *RMSE* (0,5132) e *MAE* (0,4093) apontam para níveis reduzidos de erro absoluto. Esses indicadores sugerem que o modelo apresenta um bom desempenho para o contexto do estudo, dentro das limitações dos dados e da abordagem utilizada. A Figura 5 ilustra um alinhamento mais consistente entre os valores previstos e os observados, evidenciando a eficácia dos ajustes realizados durante a etapa de otimização.

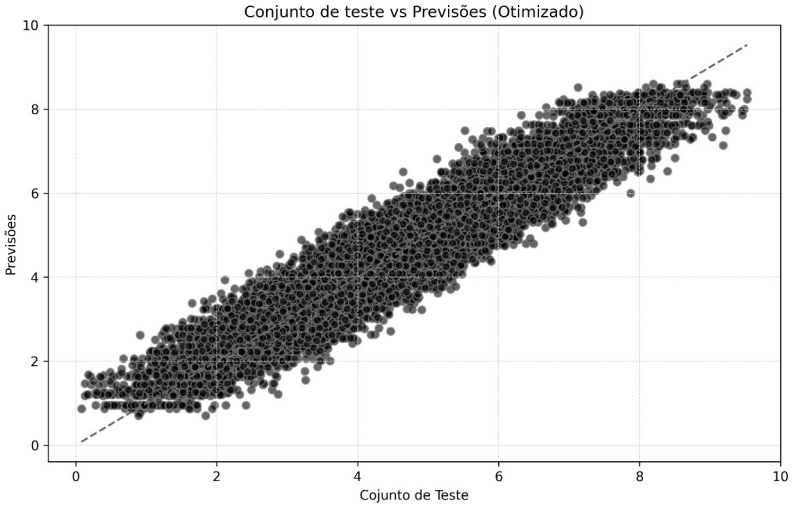

Figura 5. Gráfico de dispersão demonstrando a relação linear entre os valores reais e as previsões do modelo otimizado.

Adicionalmente, a Figura 6 apresenta o gráfico de importância das variáveis, mostrando o impacto relativo de cada variável independente na previsão do modelo.

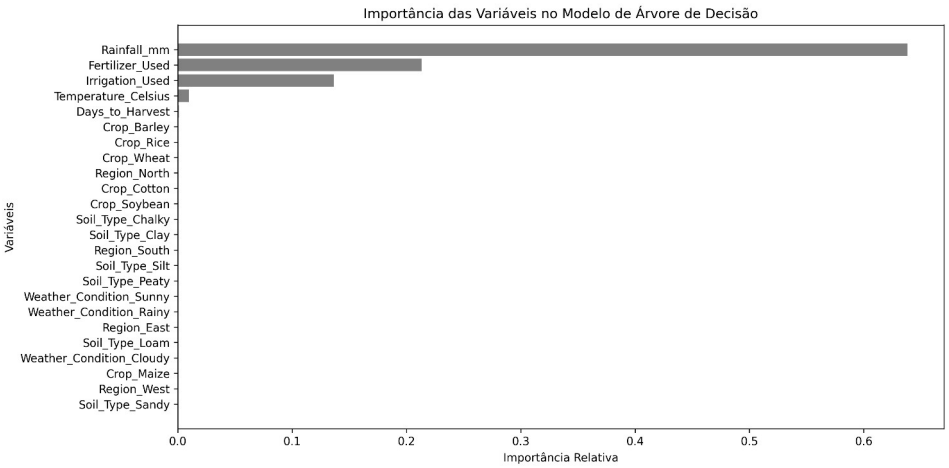

Figura 6. Gráfico ilustrando a importância das variáveis no Modelo de Árvore de Decisão.

A análise de importância das variáveis revelou que *Rainfall_mm* teve a maior influência nas previsões do modelo, com uma importância de 0,6383. As variáveis *Fertilizer_Used* (0,2131) e *Irrigation_Used* (0,1367) também se mostraram relevantes. Já variáveis como *Temperature_Celsius* (0,0097) e *Days_to_Harvest* (0,0011), assim como alguns tipos de cultivo, apresentaram pouca relevância, sugerindo que fatores climáticos e de manejo são os mais determinantes para a previsão da produtividade, levando em conta as informações do *dataset* utilizado.

4.3.5. Análise e Discussão dos Resultados

Na Tabela 3, apresentam-se os resultados de 10 amostras selecionadas aleatoriamente a partir do conjunto de teste, evidenciando os valores reais, as previsões geradas pelo modelo, a diferença absoluta em pontos e a margem de erro percentual correspondente. Esses dados ilustram a capacidade preditiva do modelo *Decision Tree Regressor* no contexto avaliado, destacando o desempenho em relação às disparidades entre os valores reais e previstos.

Tabela 3. Resultados das previsões do modelo criado, comparando os valores reais do conjunto de teste com as previsões geradas para dez amostras selecionadas aleatoriamente.

Índice	Valor Real	Valor Previsto	Diferença em Pontos	Margem de Erro (%)
38910	5.593787	5.434603	0.159184	2.845733
579	5.561166	5.721038	-0.159872	2.874784
72367	7.226636	7.522519	-0.295882	4.094327
27489	4.090851	4.391450	-0.300600	7.348100
88053	4.034822	4.221497	-0.186675	4.626603
45951	4.494672	4.927712	-0.433039	9.634500
85447	3.144771	3.707483	-0.562712	17.893573
18758	2.635296	3.134606	-0.499310	18.947018
26220	3.003560	3.426717	-0.423157	14.088504
43221	5.012430	4.884306	0.128124	2.556125

A média geral da margem de erro percentual foi de 11,2249%, enquanto a média da diferença em pontos ficou em 0,0013, evidenciando que, no geral, o modelo apresentou um desempenho aceitável para diversas previsões com margens de erro abaixo de 10%.

No entanto, observa-se que há previsões em que a margem de erro ultrapassa 15%, atingindo valores como 17,89% e 18,94%, o que indica uma dificuldade do modelo em capturar corretamente as relações complexas presentes em algumas regiões específicas do espaço de dados.

Uma das possíveis razões para o erro elevado está relacionada ao fato de o *dataset* ter sido gerado de forma sintética e posteriormente amostrado. Esse processo pode ter introduzido padrões artificiais ou relações que não refletem a realidade, dificultando a capacidade do modelo de generalizar para entradas que se afastam dos padrões mais frequentes.

Além disso, a qualidade dos dados também pode ter influenciado os resultados. Embora tenham sido aplicadas técnicas de pré-processamento e remoção de *outliers*, os dados sintéticos originais podem não possuir a riqueza e a consistência necessárias. A ausência de correlações reais e naturais entre as variáveis provavelmente limitou o potencial de aprendizado do modelo.

Outro ponto relevante é a complexidade inerente à tarefa de prever a produtividade agrícola. Este problema é fortemente influenciado por diversos fatores ambientais, climáticos e de manejo, cujas interações são frequentemente não lineares e desafiadoras de modelar. O *Decision Tree Regressor*, sendo uma abordagem relativamente simples, pode não ser suficientemente robusto para capturar essas interações mais complexas.

Por fim, destaca-se a possibilidade de desequilíbrio nos dados. Apesar do uso de amostragem estratificada, algumas combinações de variáveis (como tipos específicos de solo associados a determinadas condições climáticas) podem estar sub-representadas no conjunto final, comprometendo o desempenho preditivo em cenários menos comuns.

5. Conclusão

Este trabalho apresentou a aplicação de um modelo de regressão baseado em árvore de decisão (*Decision Tree Regressor*) para a previsão de produtividade agrícola, utilizando um *dataset* sintético disponibilizado na plataforma Kaggle. O estudo abordou desde a análise e pré-processamento dos dados até a avaliação e otimização do modelo, seguindo um fluxo estruturado que incluiu: criação de variáveis *dummies*, remoção de *outliers*, amostragem estratificada, ajuste de hiperparâmetros via *GridSearchCV*, e validação cruzada.

Os resultados obtidos demonstraram que o modelo conseguiu produzir previsões com desempenho aceitável na maioria dos casos, evidenciado por uma média da margem de erro de 11,22% e uma diferença média de 0,0013 pontos entre os valores reais e previstos. No entanto, também foi observado que erros mais elevados ocorreram em casos isolados, sobretudo devido às limitações impostas pela natureza sintética do *dataset*, pela ausência de correlações reais e pela complexidade intrínseca do problema, como a interação de múltiplos fatores climáticos, ambientais e de manejo agrícola.

Apesar das limitações, o estudo alcançou seu objetivo principal de implementar, avaliar e discutir o desempenho de um modelo preditivo para produtividade agrícola, fornecendo uma base inicial para trabalhos futuros nessa área.

5.1. Trabalhos Futuros

A continuação deste trabalho pode explorar diversas direções para aprimorar os resultados obtidos e expandir a compreensão sobre a previsão de produtividade agrícola. Primeiramente, seria interessante realizar uma análise mais detalhada sobre a qualidade dos dados utilizados, considerando a possibilidade de gerar ou utilizar um conjunto de dados mais representativo e realista, que ofereça uma base mais sólida para o treinamento dos modelos.

Outra área que pode ser abordada é a experimentação com modelos de aprendizado de máquina mais complexos, como redes neurais ou modelos de *ensemble* (abordagens que combinam múltiplos modelos de aprendizado para melhorar as previsões), que possuem maior capacidade de capturar interações não lineares e padrões complexos nos dados. Além disso, seria relevante integrar dados ambientais e climáticos reais, como informações sobre temperatura, umidade e precipitação, para melhorar ainda mais a precisão das previsões.

Também é importante explorar o impacto de diferentes abordagens de balanceamento de dados, a fim de melhorar a performance do modelo em cenários menos representados. Por fim, uma possível linha de pesquisa seria aplicar técnicas de aprendizado profundo (*deep learning*) para explorar novas maneiras de modelar as relações entre as variáveis e aumentar a robustez da previsão.

6. Referências Bibliográficas

Ajitha, P. V., & Nagra, A. (2021) “An overview of artificial intelligence in automobile industry–a case study on Tesla cars”. Solid State Technology, 64(2), 503-512.

Alnuaimi, A. F., & Albaldawi, T. H. (2024) “An overview of machine learning classification techniques”. In BIO Web of Conferences (Vol. 97, p. 00133). EDP Sciences.

Ashtiani, M. N., & Raahemi, B. (2021) “Intelligent fraud detection in financial statements using machine learning and data mining: a systematic literature review”. Ieee Access, 10, 72504-72525.

Bellegarda, J. R. (2013) “Spoken language understanding for natural interaction: The siri experience”. Natural Interaction with Robots, Knowbots and Smartphones: Putting Spoken Dialog Systems into Practice, 3-14.

Beltrami, E. J., Brown, A. C., Salmon, P. J., Leffell, D. J., Ko, J. M., & Grant-Kels, J. M. (2022) “Artificial intelligence in the detection of skin cancer”. Journal of the American Academy of Dermatology, 87(6), 1336-1342.

Bory, Paolo & Natale, Simone & Katzenbach, Christian. (2024). “Strong and weak AI narratives: an analytical framework”. AI & SOCIETY. 1-11. 10.1007/s00146-024-02087-8.

Centro de Estudos Avançados em Economia Aplicada (CEPEA). (2023). PIB do Brasil. Available at https://cepea.esalq.usp.br/upload/kceditor/files/C%C3%B3pia%20de%20PIB%20do%20Agroneg%C3%B3cio_Sum%C3%A1rio%20Executivo%20(1).pdf. Accessed on 3 December 2024.

Chen, J. X. (2016) “The evolution of computing: AlphaGo”. Computing in Science & Engineering, 18(4), 4-7.

Cunningham, P., Cord, M., & Delany, S. J. (2008) “Supervised learning”. In Machine learning techniques for multimedia: case studies on organization and retrieval (pp. 21-49). Berlin, Heidelberg: Springer Berlin Heidelberg.

Deqian, Li & Hu, Shujuan & Jinyuan, Guo & Wang, Kai & Gao, Chenbin & Wang, Siyi & He, Wen-Ping. (2023) “A New Hybrid Machine Learning Model for Short-Term Climate Prediction by Performing Classification Prediction and Regression Prediction Simultaneously”. Journal of Meteorological Research. 36. 853-865. 10.1007/s13351-022-1214-3.

Flowers, J. C. (2019, March) “Strong and Weak AI: Deweyan Considerations”. In AAAI spring symposium: Towards conscious AI systems (Vol. 2287, No. 7).

Haikonen, P. O. (2020) “On artificial intelligence and consciousness”. Journal of Artificial Intelligence and Consciousness, 7(01), 73-82.

Hosny, A., & Aerts, H. J. (2019) “Artificial intelligence for global health”. Science, 366(6468), 955-956.

Ji, Y., Liu, S., Hong, X., Lu, Y., Wu, X., Li, K., ... & Liu, Y. (2022) “Advances in artificial intelligence applications for ocular surface diseases diagnosis”. Frontiers in Cell and Developmental Biology, 10, 1107689.

Jiang, T., Gradus, J. L., & Rosellini, A. J. (2020) “Supervised machine learning: a brief primer”. Behavior therapy, 51(5), 675-687.

Juraszek, G. D. (2014) “Reconhecimento de produtos por imagem utilizando palavras visuais e redes neurais convolucionais”.

Kitts, B. (2000). Regression trees. Technical Report, http://www.appliedaisystems.com/papers/RegressionTrees.doc.

Kongar, E., & Adebayo, O. (2021) “Impact of social media marketing on business performance: A hybrid performance measurement approach using data analytics and machine learning”. IEEE Engineering Management Review, 49(1), 133-147.

Kotagiri, A. (2023) “Mastering Fraudulent Schemes: A Unified Framework for AI-Driven US Banking Fraud Detection and Prevention”. International Transactions in Artificial Intelligence, 7(7), 1-19.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007) “Supervised machine learning: A review of classification techniques”. Emerging artificial intelligence applications in computer engineering, 160(1), 3-24.

Kumar, Sunil & Bhatnagar, Vaibhav. (2021) “A Review of Regression Models in Machine Learning”. Journal of Intelligent Systems and Computing. 2. 40-47. 10.51682/JISCOM.00202005.2021.

Lima, K. K. D. S. (2019) “Desenvolvimento e comparação de redes neurais convolucionais para classificação de objetos”.

Loh, W. Y. (2011) “Classification and regression trees”. Wiley interdisciplinary reviews: data mining and knowledge discovery, 1(1), 14-23.

Ludermir, T. B. (2021) “Inteligência Artificial e Aprendizado de Máquina: estado atual e tendências”. Estudos Avançados, 35, 85-94.

Mahesh, B. (2020) “Machine learning algorithms-a review”. International Journal of Science and Research (IJSR). [Internet], 9(1), 381-386.

Michaely, A. H., Zhang, X., Simko, G., Parada, C., & Aleksic, P. (2017, December) “Keyword spotting for Google assistant using contextual speech recognition”. In 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (pp. 272-278). IEEE.

Mitchell, T. M., & Mitchell, T. M. (1997) “Machine learning”. (Vol. 1, No. 9). New York: McGraw-hill.

Mitrofanov, S. A., & Semenkin, E. S. (2021, February) “Tree retraining in the decision tree learning algorithm”. In IOP Conference Series: Materials Science and Engineering (Vol. 1047, No. 1, p. 012082). IOP Publishing.

Morales, E. F., & Escalante, H. J. (2022) “A brief introduction to supervised, unsupervised, and reinforcement learning”. In Biosignal processing and classification using computational learning and intelligence (pp. 111-129). Academic Press.

Murphy, K. P. (2012) “Machine learning: a probabilistic perspective”. MIT press.

OpenAI. (2024, December). ChatGPT [Assistive tool]. Available at https://chat.openai.com/. Accessed on 3 December 2024.

Oriental dos Santos, L., Kelly, G. S., & Paiva, R. S. (2024). TCC_DecisionTreeRegressor_Agricultura. GitHub. https://github.com/s1ngk/TCC_DecisionTreeRegressor_Agricultura. Acesso em: 3 dez. 2024.

Otiattakorah, S. (2024). Agriculture Crop Yield Dataset. Kaggle. https://www.kaggle.com/datasets/samuelotiattakorah/agriculture-crop-yield/data. Acesso em: 3 dez. 2024.

Pinar Saygin, A., Cicekli, I., & Akman, V. (2000) “Turing test: 50 years later”. Minds and machines, 10(4), 463-518.

Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022) “AI in health and medicine”. Nature medicine, 28(1), 31-38.

Russell, S. J., & Norvig, P. (2016) “Artificial intelligence: a modern approach”. Pearson.

Sadok, H., Sakka, F., & El Maknouzi, M. E. H. (2022) “Artificial intelligence and bank credit analysis: A review”. Cogent Economics & Finance, 10(1), 2023262.

Singh, Y., Bhatia, P. K., & Sangwan, O. (2007) “A review of studies on machine learning techniques”. International Journal of Computer Science and Security, 1(1), 70-84.

Srinivas, R., Ankayarkanni, B., & Krishna, R. S. B. (2021, May) “Uber related data analysis using machine learning”. In 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 1148-1153). IEEE.

Stahl, B. C., Antoniou, J., Bhalla, N., Brooks, L., Jansen, P., Lindqvist, B., ... & Wright, D. (2023) “A systematic review of artificial intelligence impact assessments”. Artificial Intelligence Review, 56(11), 12799-12831.

Sutton, R. S. (2018) “Reinforcement learning: An introduction”. A Bradford Book.

Van Esch, P., & Stewart Black, J. (2021) “Artificial intelligence (AI): revolutionizing digital marketing”. Australasian Marketing Journal, 29(3), 199-203.

Vargas, A. C. G., Paes, A., & Vasconcelos, C. N. (2016, October). “Um estudo sobre redes neurais convolucionais e sua aplicação em detecção de pedestres”. In Proceedings of the xxix conference on graphics, patterns and images (Vol. 1, No. 4). Sn.

Wiering, M. A., & Van Otterlo, M. (2012) “Reinforcement learning”. Adaptation, learning, and optimization, 12(3), 729.

World Trade Organization. (2023). World Trade Statistical Review 2023. Available at https://www.wto.org/english/res_e/booksp_e/wtsr_2023_e.pdf. Accessed on 3 December 2024.

Yacouby, R., & Axman, D. (2020, November) “Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models”. In Proceedings of the first workshop on evaluation and comparison of NLP systems (pp. 79-91).

Ying, X. (2019, February) “An overview of overfitting and its solutions”. In Journal of physics: Conference series (Vol. 1168, p. 022022). IOP Publishing.