UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Influência de fatores climáticos sobre a produtividade do café utilizando técnicas de aprendizado de máquina

Fernando Leite Moreira

Trabalho de Conclusão de Curso
MBA em Inteligência Artificial e Big Data

ICMC USP
SÃO CARLOS

UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Influência de fatores climáticos sobre a produtividade do café utilizando técnicas de aprendizado de máquina

Fernando Leite Moreira

USP - São Carlos

2025

Fernando Leite Moreira

Influência de fatores climáticos sobre a produtividade do café utilizando técnicas de aprendizado de máquina

Trabalho de conclusão de curso apresentado ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial.

Orientador: Prof. Dr. Kelton Augusto Pontara da Costa

USP - São Carlos

2025

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

Leite Moreira, Fernando
L835i Influência de fatores climáticos sobre a produtividade do café utilizando técnicas de aprendizado de máquina / Fernando Leite Moreira; orientador Kelton Augusto Pontara da Costa. -- São Carlos, 2025.
65 p.

Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2025.

1. Inteligência Artificial. 2. Aprendizado de Máquina. I. Pontara da Costa, Kelton Augusto, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

DEDICATÓRIA

Para Pituca, Pepita e Jaguar.

AGRADECIMENTOS

Aos meus pais Carlos Antônio e Liliane por todo o apoio e incentivo.

Ao meu irmão Vinícius pelo companheirismo.

Ao professor Kelton por sua paciência, tranquilidade e sugestões que enriqueceram o trabalho.

RESUMO

MOREIRA, F. L. **Influência de fatores climáticos sobre a produtividade do café utilizando técnicas de aprendizado de máquina**. 2025. 72 f. Monografia (MBA em Inteligência Artificial e Big Data) – Centro de Ciências Matemáticas Aplicadas à Indústria, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2025.

O Brasil destaca-se como o maior produtor global de café, item que está entre os mais exportados pelo país. O monitoramento e antevisão da produtividade desta cultura, frente a desafios como o aquecimento global, é crucial para a manutenção da competitividade do segmento. Desta forma este trabalho teve como objetivo criar modelos de previsão da produtividade do café a partir de dados meteorológicos (temperatura, umidade e precipitação média) utilizando dados advindos do INMET e da pesquisa PAM do IBGE. Para isso, foram utilizados os algoritmos de aprendizado de máquina Random Forest, Gradient Boosting e XGBoost. Os modelos foram treinados com alocação de 70% dos dados para treino e 30% para teste. Foram estudados municípios do estado de São Paulo com produção de café e dados meteorológicos disponíveis de 2010 a 2019. Os modelos obtidos tiveram desempenho semelhante para os três algoritmos, porém o XGBoost apresentou assertividade levemente superior, com maior coeficiente de determinação (R^2) e menor raiz do erro quadrático médio (RMSE). Foram obtidos valores de R^2 de 0,12 (XGBoost), 0,11 (Random Forest) e 0,10 (Gradient Boosting); já os de RMSE foram de 412,6 (XGBoost), 415,1 (Random Forest) e 416,6 kg/ha (Gradient Boosting). Embora as três variáveis possuam semelhante significância, a umidade aparece como a mais relevante quando usados os métodos Random Forest e Gradient Boosting; já a temperatura possui maior relevância para o XGBoost. Conclui-se que, embora os modelos apresentados não tenham apresentado elevado poder preditivo, foi possível obter entendimentos relevantes acerca da influência das variáveis meteorológicas sobre a produtividade do café.

Palavras-chave: Café. Produtividade Agrícola. Fenologia. Meteorologia.

ABSTRACT

MOREIRA, F. L. **Influence of climatic factors on coffee productivity using machine learning techniques**. 2025. 72 f. Monografia (MBA em Inteligência Artificial e Big Data) – Centro de Ciências Matemáticas Aplicadas à Indústria, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2025.

Brazil stands out as the world's largest coffee producer, with the crop ranking among the country's top export commodities. Monitoring and forecasting coffee productivity, in the face of challenges such as global warming, is crucial to maintaining the competitiveness of the sector. Thus, this study aimed to create predictive models of coffee yield based on meteorological data (average temperature, humidity, and precipitation), using data from INMET and the IBGE's PAM survey. For this purpose, the machine learning algorithms Random Forest, Gradient Boosting, and XGBoost were applied. The models were trained with 70% of the data allocated for training and 30% for testing. Municipalities in the state of São Paulo with coffee production and available meteorological data from 2010 to 2019 were considered. The models obtained showed similar performance across the three algorithms; however, XGBoost achieved slightly higher accuracy, with a greater coefficient of determination (R^2) and a lower root mean squared error (RMSE). The R^2 values obtained were 0.12 (XGBoost), 0.11 (Random Forest), and 0.10 (Gradient Boosting), while RMSE values were 412.6 (XGBoost), 415.1 (Random Forest), and 416.6 kg/ha (Gradient Boosting). Although the three variables had comparable significance, humidity appeared as the most relevant factor for Random Forest and Gradient Boosting, while temperature was more relevant for XGBoost. It is concluded that, although the models did not exhibit high predictive power, they provided valuable insights into the influence of meteorological variables on coffee yield.

Keywords: Coffee. Agricultural productivity. Phenology. Meteorology.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AM – Aprendizado de Máquina

ANN – Redes Neurais Artificiais

GBR – Gradient Boosting

IBGE – Instituto Brasileiro de Geografia e Estatística

INMET – Instituto Nacional de Meteorologia

LAI – Índice de Área Foliar

LOO – Leave-one-out

MAE – Mean Absolute Error

MAPE – Erro Percentual Absoluto Médio

ME – Mean Error

MLR – Regressão Linear Múltipla

MPE – Mean Percentage Error

MSE – Mean Square Error

NEAT – Neuroevolução de Topologias Aumentadas

PAM – Produção Agrícola Municipal

PCA – Análise de Componentes Principais

PLSR – Regressão por Mínimos Quadrados Parciais

RFR – Floresta Aleatória

RMSE – Raiz do Erro Quadrático Médio

SQres – Soma dos Quadrados dos Resíduos

SQtot – Soma Total dos Quadrados

SVM – Máquinas de Vetores de Suporte Linear

XGBoost – Extreme Gradient Boosting

LISTA DE SÍMBOLOS

ºC	Graus Celsius
k	Número de subconjuntos
y_i	Valores reais
$\hat{y}_i$	Valores previstos
n	Número de observações
R^2	Coeficiente de Determinação
kg	Quilograma
ha	Hectare

SUMÁRIO

1 INTRODUÇÃO

A agricultura se destaca no Brasil como importante geradora de riquezas. Segundo dados do Ministério do Desenvolvimento, Indústria, Comércio e Serviços, em 2024 o valor exportado pelo setor agrícola correspondeu a 72,5 bilhões de dólares, o que representou 22% do total do país. Dentre os itens, os de maiores valores exportados foram soja (US$ 42,9 bi), café não torrado (US$ 11,3 bi), milho não moído, exceto milho doce (US$ 8,2 bi) e algodão em bruto (US$ 5,2 bi). De acordo com o *U. S. Department of Agriculture* (USDA), no ano de 2024 o Brasil foi o maior produtor mundial de café, sendo responsável por 38% do total global.

Com os crescentes avanços tecnológicos, o aumento da produtividade das culturas é crucial para a competitividade do setor frente a concorrentes internacionais. Nesse sentido, destacam-se como fatores relevantes para a eficiência da produção agrícola as variáveis meteorológicas. O volume de precipitação, umidade, temperatura mínima e máxima, radiação solar e temperatura média do vento são fatores que podem afetar a produtividade agrícola (Crane-Droesch, 2018).

Devido ao aquecimento global, tem-se conectado a influência antropogênica com o aumento da frequência e intensidade de eventos climáticos extremos. Há evidências do aumento da probabilidade de temperaturas muito altas e da diminuição de temperaturas extremamente baixas em várias partes do mundo (Stott et al., 2016). A variação climática é o principal fator causador de oscilações e perdas de produtividade de café no Brasil (de Camargo, 2010). Dessa forma, faz-se importante investigar como essas alterações podem influenciar a produtividade dessa cultura e prevê-las antes da colheita, de modo a gerenciar os recursos de forma mais estratégica.

A utilização de algoritmos de aprendizado de máquina supervisionados permite a detecção de padrões a partir de dados cuja resposta já é conhecida. Eles podem ser assim utilizados para realizar previsões com base em novas entradas. Alguns desses métodos que podem ser usados para isso são *Random Forest* (Floresta Aleatória), *Gradient Boosting* (Impulsionamento de Gradiente) e *XGBoost* (Impulsionamento de Gradiente Extremo).

No desenvolvimento deste trabalho foi utilizado o modelo de linguagem ChatGPT para revisão e correção do texto bem como para o código. Todas as contribuições foram avaliadas e são de responsabilidade do autor.

O algoritmo *Random Forest* é um método baseado em árvores de decisão que possibilita a resolução de problemas de classificação e regressão (previsão). Ele inicialmente cria amostras aleatórias dos dados de treinamento para treinar as árvores de decisão e em cada nó delas escolhe um subconjunto aleatório das variáveis para evitar vieses. Para os casos de regressão, ele utiliza a média das previsões das árvores.

Por outro lado, o algoritmo *Gradient Boosting*, embora se baseie em árvores de decisão, as constrói de forma independente e sequencial, de modo que a seguinte corrige os erros da anterior até um critério de parada.

Já o algoritmo *XGBoost* é uma versão otimizada do *Gradient Boosting*. Ele utiliza o método *pruning* (poda) para evitar o *overfitting* (sobreajuste) e outras técnicas para encontrar os melhores pontos de divisão.

2 OBJETIVOS

Assim, o objetivo geral do trabalho será criar modelos de previsão da produtividade do café a partir de dados meteorológicos. Para isso serão utilizados dados históricos e públicos de rendimento advindos da pesquisa Produção Agrícola Municipal (PAM) do Instituto Brasileiro de Geografia e Estatística (IBGE) e climatológicos advindos do Instituto Nacional de Meteorologia (INMET). Serão usados os algoritmos de aprendizados supervisionado *Random Forest*, *Gradient Boosting* e *XGBoost*.

Os objetivos específicos do trabalho serão:

- Comparar os modelos obtidos para os diferentes algoritmos e determinar o mais adequado;

- Determinação da variável meteorológica mais importante para o estabelecimento dos modelos de previsão obtidos.

3 FUNDAMENTAÇÃO TEÓRICA

3.1 Aprendizado de máquina

Aprendizado de máquina (AM) é um ramo da inteligência artificial que envolve a construção de modelos matemáticos que expliquem dados. Dessa forma, o aprendizado se dá pela obtenção de parâmetros reguláveis adequados a eles, de modo que é como se o programa aprendesse com as observações obtidas com o mínimo de instruções humanas. A partir disso, é possível determinar padrões, fazer previsões e tomar decisões.

Quanto à presença ou ausência de rótulos, os métodos de AM podem ser classificados em supervisionado, não supervisionado, semissupervisionado e por reforço. São elas:

- Aprendizado de máquina supervisionado: modela o relacionamento entre dados medidos (variáveis explicativas) e rótulos associados a eles (respostas). Uma vez que o modelo é determinado, pode ser usado para prever valores de novos dados. São exemplos de métodos de aprendizado supervisionado análises de regressão, modelos logísticos, modelagem multinível, modelos para dados de contagem e árvores de decisão. Quando são utilizados para prever a classe à que uma observação pertence, resolvem problemas de classificação; já quando são usados para prever um valor contínuo, solucionam problemas de regressão.

- Aprendizado de máquina não supervisionado: modela as características de dados medidos, porém sem rótulos associados. Ele permite fazer análises exploratórias, como identificar e separar grupos distintos de dados e reduzir a sua dimensionalidade. São exemplos de métodos de aprendizado não supervisionado o *clustering* (análise de agrupamento de dados), análise fatorial por componentes principais e análise de correspondência.

- Aprendizado de máquina semissupervisionado: usa dados com e sem rótulos associados para criar o modelo. Os dados rotulados, em menor número, são utilizados para direcionar o aprendizado por meio da geração de rótulos para os não rotulados.

- Aprendizado por reforço: se dá por meio de tentativa e erro, de modo que o agente é recompensado ao acertar a decisão e penalizado ao errar. O aprendizado se dá pelo estabelecimento de uma política/estratégia que maximize a função de recompensa.

Nos métodos de aprendizado de máquina supervisionado, os dados são divididos em conjuntos de treino e teste. O grupo de treino compreende as observações utilizadas para gerar o modelo; já o de teste avalia a a capacidade de generalização do modelo gerado para novos dados.

A divisão nas classes de treino e teste nos algoritmos de aprendizado de máquina supervisionados pode ocorrer das seguintes formas:

- *Hold-out* (separação): técnica em que os dados são divididos em um conjunto de teste e um de treino.

- *Cross-validation* (validação cruzada): técnica em que o conjunto de observações é dividido em subconjuntos (*folds*) que são usados como teste e treino alternadamente, com o processo sendo repetido diversas vezes. Dessa forma, diferentemente do método de *hold*-out, o desempenho do modelo não depende apenas de uma divisão de treino e teste. O modelo é analisado a partir da média das métricas de avaliação de cada iteração.

- *K-fold cross validation* (validação cruzada com k subconjuntos): é um caso específico de *cross-validation* em que os dados são divididos em k subconjuntos e, em cada iteração, k-1 grupos são utilizados como treino e 1 como teste. O processo é repetido k vezes, de modo que cada subconjunto seja usado como treino uma vez. O modelo é avaliado a partir da média das métricas de avaliação de cada iteração.

- *Leave-one-out cross validation* (validação cruzada com um excluído): caso específico do *k-fold cross validation* em que k é o número de observações. O modelo é avaliado a partir da média das métricas de avaliação de cada iteração.

Pode-se também ser utilizado um conjunto de validação que tem a função de, antes do treino do modelo, avaliar a sua eficiência. Essa avaliação tem o intuito de ajustar a escolha dos hiperparâmetros e assim possibilitar uma melhor eficiência do modelo.

Na modelagem por meio de aprendizado de máquina supervisionado, são definidos pelo usuário variáveis que controlam o comportamento do algoritmo, os hiperparâmetros. Eles podem determinar o tempo gasto no treinamento, o uso de recursos computacionais, bem como a complexidade do modelo gerado. Exemplos de hiperparâmetros são a profundidade máxima da árvore, o número mínimo de amostras para dividir um nó e número mínimo de amostras em uma folha, em árvores de decisão; número de árvores na floresta e de variáveis usadas em cada divisão em florestas aleatórias.

Na determinação de modelos usando AM supervisionado, deve-se evitar que ocorram *underfitting* (subajuste) e *overfitting* (sobreajuste). No primeiro caso, o modelo obtido é simples demais para descrever os padrões observados; já no segundo, o sistema torna-se altamente especializado para os dados de treino, porém não consegue generalizar para novas observações. Para impedir esses problemas, podem-se definir hiperparâmetros que ao mesmo tempo evitem que o modelo se torne extremamente complexo mas que não o tornem demasiado simplificado.

Para avaliar a acurácia dos modelos de regressão gerados por modelos de aprendizado de máquina, podem ser usadas métricas que calculam os erros entre as previsões obtidas e os valores obtidas. Algumas das mais comumente usadas são:

- *Mean Error* (ME): representa a média dos erros entre os valores reais (y_i) e os previstos ($\hat{y}_i$) para as *n* observações do conjunto de teste. Pode ser determinada pela seguinte equação:

$$ME = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)$$

A sua utilização possui como desvantagem o fato de que valores positivos de erro podem anular negativos, de modo que isso pode mascarar imprecisões do modelo.

- *Mean Absolute Error* (MAE): representa a média dos erros absolutos entre os valores reais (y_i) e os previstos ($\hat{y}_i$) para as *n* observações do conjunto de teste. Pode ser determinada pela seguinte equação:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

A utilização do MAE, por considerar a média dos módulos dos erros, consegue impedir o problema da anulação dos resíduos positivos pelos negativos.

- *Mean Percentage Error* (MPE): representa a média dos erros percentuais entre os valores reais (y_i) e os previstos ($\hat{y}_i$) para as *n* observações do conjunto de teste. Pode ser determinada pela seguinte equação:

$$MPE = \frac{100}{n}\sum_{i=1}^{n}\frac{(y_i - \hat{y}_i)}{y_i}$$

A utilização do MPE, assim como a do ME, tem como problema a possibilidade de anulação dos resíduos positivos pelos negativos. Além disso, quando algum y_i tende a zero, a sua utilização é inviabilizada. O seu uso tem como vantagem a capacidade de comparação entre diferentes datasets, por se tratar de uma medida relativa.

- *Mean Absolute Percentage Error* (MAPE): representa a média dos erros percentuais absolutos entre os valores reais (y_i) e os previstos ($\hat{y}_i$) para as *n* observações do conjunto de teste. Pode ser determinada pela seguinte equação:

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

A utilização do MAPE, assim como a do MPE, tem como vantagem o fato de permitir comparações entre diferentes *datasets*, por se tratar de uma medida relativa. Além disso, ele

evita a anulação de resíduos positivos pelos negativos por ser calculado de forma absoluta. Porém, sua utilização é inviabilizada se algum y_i tender a zero.

- *Root Mean Square Error* (RMSE): mede a raiz da média do quadrado dos erros entre os valores reais (y_i) e os previstos ($\hat{y}_i$) para as *n* observações do conjunto de teste. Pode ser determinada pela seguinte equação:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- *Mean Square Error* (MSE): mede a média do quadrado dos erros entre os valores reais (y_i) e os previstos ($\hat{y}_i$) para as *n* observações do conjunto de teste. Pode ser determinada pela seguinte equação:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- Coeficiente de Determinação (R^2): mede o quanto a variabilidade dos erros é explicada pelo modelo gerado. Pode ser determinado pela seguinte equação:

$$R^2 = 1 - \frac{SQ_{res}}{SQ_{tot}}$$

Em que:

$$SQ_{res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

E:

$$SQ_{tot} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

Onde y_i são os valores reais, $\hat{y}_i$ os valores previstos e $\bar{y}$ a média dos valores observados. O SQ_{res} (Soma dos Quadrados dos Resíduos) representa os erros que não são explicados pelo modelo, ao passo que o SQ_{tot} (Soma Total dos Quadrados) mede a variabilidade dos dados ao redor da média. O coeficiente de determinação pode estar entre 0 e 1 e, de maneira geral, quanto mais próximo de 1, mais assertivo o modelo.

3.2 Árvore de decisão

Árvore de decisão é um algoritmo de AM supervisionado que pode ser utilizado para regressão ou classificação. Ela funciona como uma sequência de partições binárias, hierárquicas

e sucessivas (semelhante a uma árvore). Em cada divisão é feita uma pergunta ou imposta uma condição de modo que, dependendo se o dado satisfaz aquele requisito ou não, ele é dividido em um subconjunto. Essas divisões são feitas até que se atinja um critério de parada. Dessa forma, se visa atingir homogeneidade dos dados em relação à variável resposta.

Seus principais elementos são:

- Raiz (*Root Node)*: primeiro nó da árvore, onde há a primeira partição dos dados.

- Nó interno (*decision node*): nós intermediários da árvore, em que são incluídas as condições para dividir os dados.

- Ramo (*branch*): conexões entre os nós, que ligam a saída de um teste ao próximo nó ou folha.

- Folha (*leaf node*): nó no final do ramo, que representam o resultado final da classificação ou regressão. No primeiro caso, representam uma classe e no segundo valores contínuos.

- Critério de divisão: regra que determina a melhor variável e ponto de corte a ser utilizada para segmentar os dados. Em uma árvore de decisão regressiva, pode-se usar o MSE como parâmetro, por exemplo.

3.3 Floresta Aleatória

A floresta aleatória (*Random Forest*) é um método de AM supervisionado que combina várias Árvores de Decisão individuais para fazer as previsões de classificação ou regressão. Pelo fato de ela juntar mais de um algoritmo para gerar a previsão, ela é do tipo *ensemble* (conjunto). Seu objetivo é de, dessa forma, evitar a dependência de uma única árvore para gerar a previsão e assim diminuir a probabilidade de *overfitting*.

Para cada árvore, é selecionado um conjunto de observações aleatórias com reposição do conjunto de teste. Dessa forma, valores podem ser excluídos e outros serem usados mais de uma vez. Em seguida, é feito o treinamento de cada árvore de forma paralela e isolada, porém apenas algumas variáveis são selecionadas de forma aleatória. Em problemas de classificação, cada árvore retorna uma classe; em problemas de regressão, é gerado um número real. No primeiro caso, para novos valores, é prevista a classe mais votada; já no segundo, é a média das previsões individuais.

Devido ao fato de que cada árvore é treinada isoladamente e de modo paralelo e que as observações são selecionadas ao acaso com reposição, as florestas aleatórias são um exemplo de algoritmo *bagging* (*bootstrap aggregating*).

3.4 *Gradient Boosting*

O *Gradient Boosting* (Impulsionamento de Gradiente) é um método de AM supervisionado do tipo *ensemble* que combina os algoritmos (geralmente árvores de decisão) de modo sequencial. Cada novo modelo é treinado para corrigir os erros do anterior. Pode ser utilizado para problemas de classificação ou de regressão.

Neste método, inicialmente é feita uma previsão inicial e os resíduos decorrentes dele são calculados. A seguir, uma árvore de decisão é treinada para prever os erros e, assim, tentar corrigi-los. A nova previsão é dada pela anterior somada a um fator de correção. Associado a este fator está a taxa de aprendizado (η), que determina o peso deste fator de correção. A realização dessas previsões se repete até que seja atingido o número de árvores indicado ou os resíduos sejam suficientemente pequenos.

A taxa de aprendizado determina a celeridade de convergência do método. Se for elevada, faz com que haja convergência rápida, porém aumenta a possibilidade de *overfitting*; se for baixa, o modelo aprende mais lentamente mas generaliza melhor.

3.5 *XGBoost*

O *XGBoost* (Impulsionamento de Gradiente Extremo) é um aprimoramento do *Gradient Boosting*, sendo mais preciso e rápido. Além de considerar os resíduos na montagem das árvores, ele também leva em conta a variação dos erros. Ademais, o *XGBoost* inclui penalizações para controlar a complexidade das árvores e dessa forma evitar o *overfitting*.

3.6 Revisão Bibliográfica

Neste item é apresentada uma breve revisão bibliográfica de estudos que analisaram a influência de fatores climáticos sobre a produtividade do café.

Kouadio et al. (2021) utilizaram um modelo baseado em processos biofísicos para prever a produtividade da espécie Café Robusta (*Coffea canephora*) em função de variações climáticas nas prinicipais provínciais produtoras do Vietnã. Essa abordagem incluiu os diferentes processos do ciclo de vida da planta (crescimento ativo, partição de biomassa e crescimento passivo) usando equações que representam mecanismos que influenciam o crescimento das plantas, como fotossíntese e alocação de biomassa. Foram utilizados no modelo

dados de temperatura máxima e mínima, radiação solar e pluviosidade com granularidade diária advindos de fontes oficiais de 2001 a 2014 e coletados em campo no período de 2008 a 2017. Os valores de raiz do erro quadrático médio (RMSE) entre os valores de produtividade prevista e estimada variaram entre 0,24 a 0,33 t/ha e os erros percentuais absolutos médios (MAPE) entre 9 e 14%.

Barbosa et al. (2021) utilizaram veículos aéreos não tripulados para obter imagens de cafeeiros e a partir delas prever a produtividade do café utilizando *feature selection* e *deep learning*. Foram utilizados no modelo de previsão a estimativa da altura das árvores e dos diâmetros das copas, o índice de área foliar (LAI) e os valores individuais das bandas RGB (vermelho, verde e azul) obtidas. Com base no método *feature selection*, foram determinados como parâmetros mais importantes o LAI e o diâmetro das copas. Para desenvolver os modelos de produtividade, foram usados os algoritmos de máquinas de vetores de suporte linear (SVM), regressão por *gradient boosting* (GBR), regressão por floresta aleatória (RFR), regressão por mínimos quadrados parciais (PLSR) e neuroevolução de topologias aumentadas (NEAT). Considerando-se o MAPE como critério de comparação, o algoritmo mais assertivo foi o NEAT utilizando apenas o LAI e o diâmetro das copas, com um erro percentual médio absoluto de 31,75%.

Legesse (2019) analisou como as mudanças climáticas impactam na produtividade do café, além de propor soluções para mitigar problemas. Os efeitos observados das variações meteorológicas foram redução do rendimento e qualidade, aumento dos custos de produção e de doenças e diminuição de áreas propícias para cultivo. O aumento da temperatura leva a uma diminuição da fotossíntese da planta, além de a tornar mais suscetível ao ataque de fungos, como o *Hemileia vastatrix,* e insetos, como o *Hypothenemus hampei*. A disponibilidade e o momento das chuvas também afetam nas taxas de fotossíntese e qualidade dos grãos.

Dinh et al. (2022) estudaram a sensibilidade da produtividade da espécie *Coffea canephora* em relação a variações climáticas a nível distrital e provincial no Vietnã, determinaram os momentos em que o clima é mais influente no rendimento e a partir de quando é possível estimar o rendimento da safra. Para isso, utilizaram séries temporais da produtividade entre 2000 e 2018 e regressão linear múltipla com análise de componentes principais (PCA) e validação cruzada *leave-one-out* (LOO). Foi concluído que a sua produtividade é dependente da estação chuvosa do ano anterior à colheita, sendo que alta pluviosidade no período de crescimento favorece o aumento do rendimento ao passo que pouca chuva durante a formação dos grãos o diminui. Foi possível, dependendo do local, prever a produtividade do café com uma antecedência de 3 a 6 meses.

Jayakumar et al. (2016) estudaram a influência de variáveis climáticas sobre a produtividade das espécies *Coffea arabica* e *Coffea canephora* no distrito Wayanad do estado Kerala, na Índia. Foram analisados os parâmetros precipitação, temperatura máxima e mínima e umidade relativa média ao longo do período de 1980 a 2009. Para se fazer a correlação, foi utilizada regressão múltipla *stepwise*, sendo que o teste t de Student foi aplicado para avaliar a significância estatística do efeito dos fatores. Foi observada uma correlação positiva entre a produtividade da *C. arabica* com a temperatura máxima no mês de janeiro e a umidade relativa em julho. Em relação à espécie *C. canephora* encontrou-se uma relação entre o rendimento e a temperatura máxima de fevereiro e uma relação negativa com a umidade relativa de fevereiro.

Pham et al. (2019) estudaram o impacto das mudanças climáticas na produção de café. Segundo os autores, pode ocorrer diminuição de áreas adequadas do cultivo, redução na produtividade, aumento da disseminação de pragas e doenças, perda de áreas ideais de plantação principalmente no Brasil e Vietnã e necessidade de transferência para locais de maior altitude. Por outro lado, a maior concentração de dióxido de carbono na atmosfera pode compensar um pouco os efeitos negativos e áreas não propícias podem se tornar adequadas para plantação no futuro.

Kittichotsatsawat et al. (2022) investigaram a influência de fatores climáticos sobre a produtividade da espécia *Coffea arabica* bem como fizeram um modelo para sua previsão. Para isso, utilizaram dados de umidade relativa, temperatura máxima e mínima, pluviosidade, área cultivada e zona produtora no período de 2004 a 2018. Os algoritmos utilizados no estudo foram regressão linear múltipla (MLR) e redes neurais artificiais (ANN). Dentre estes, o ANN teve mais sucesso, com coeficiente de determinação de 0,9524, ao passo que o do MLR foi de 0,9235. Os fatores mais relevantes para a produtividade foram a temperatura mínima e máxima mensal para o MLR e a temperatura máxima e precipitação total para a ANN. Foi possível observar que a produtividade é máxima quando a temperatura é inferior a 29 ºC e a precipitação mensal é menor do que 100 mm.

Com base nestes estudos, foi possível perceber que variáveis climáticas afetam a produtividade do café. Além disso, modelos de previsão (regressão linear múltipla, redes neurais artificiais e aprendizado profundo) que utilizam estes fatores demonstraram eficiência. Variáveis como temperatura e pluviosidade possuem papel crítico no desenvolvimento das plantas e em seu rendimento.

4 METODOLOGIA

4.1 Preparação dos Dados

Os dados de precipitação total, temperatura do ar – bulbo seco e umidade relativa do ar dos municípios estudados foram obtidos por meio de bases abertas disponibilizadas pelo INMET – Instituto Nacional de Meteorologia. Já os de produtividade de café foram adquiridos a partir do repositório público da pesquisa PAM do IBGE. A produtividade de cada município foi obtida de forma anual, ao passo que os dados meteorológicos tiveram granularidade horária. Foram estudadas as safras dos anos de 2010 a 2019.

Foram considerados os municípios do estado de São Paulo que tiveram produção de café e que possuem dados meteorológicos disponíveis em todo o período de estudo. Foram eles: Ariranha, Avaré, Barra do Turvo, Bauru, Casa Branca, Franca, Ibitinga, Itapira, Ituverava, Jales, José Bonifácio, Ourinhos e São Carlos.

Inicialmente, foi realizado pré-processamento dos dados. Isso incluiu análise exploratória por meio de estatística descritiva, verificação de outliers e identificação de falhas, como ausência ou inconsistência. Essa etapa visou garantir que os dados estivessem prontos para uso e para se obter percepções que sejam úteis nas etapas de análise posteriores.

4.2 Aprendizado de Máquina

4.2.1 *Random Forest*

Para a construção do modelo preditivo, foram consideradas como variáveis explicativas as médias anuais das variáveis temperatura do ar – bulbo seco e umidade relativa do ar e a média diária da precipitação total; já o atributo resposta foi o rendimento médio da produção no ano. As observações foram divididas em conjunto de treino (70% do total) e teste (30%). Foi utilizada validação cruzada com 5 subconjuntos para estimativa dos hiperparâmetros.

Para a atribuição dos hiperparâmetros número de árvores, profundidade máxima de cada árvore, número mínimo de amostras necessárias para dividir um nó e número de amostras que uma folha deve conter foi utilizada a classe *GridSearchCV* da biblioteca *scikit-learn* do Python. Foram empregados os seguintes hiperparâmetros:

- Número de árvores: 200
- Profundidade máxima de cada árvore: 3

- Número mínimo de amostras necessárias para dividir um nó: 2

- Número de amostras que uma folha deve conter: 2

O desempenho do modelo foi avaliado com base no coeficiente de determinação (R^2) e a raiz do erro quadrático médio (RMSE). Para a determinação do modelo foi usada a classe *RandomForestRegressor* da biblioteca *scikit-learn* do Python.

Para a determinação da importância de cada variável foi utilizado o atributo *feature_importances_* do *scikit-learn*.

4.2.2 *Gradient Boosting*

Da mesma forma que para o *Random Forest*, foram consideradas como variáveis explicativas as médias anuais das variáveis temperatura do ar – bulbo seco e umidade relativa do ar e a média diária da precipitação total; já o atributo resposta foi o rendimento médio da produção no ano. As observações foram divididas em conjunto de treino (70% do total) e teste (30%). Foi utilizada validação cruzada com cinco subconjuntos para a estimativa dos hiperparâmetros.

Para a atribuição dos hiperparâmetros taxa de aprendizado, profundidade máxima de cada árvore e número de árvores foi utilizada a classe *GridSearchCV* da biblioteca *scikit-learn* do Python. Foram empregados os seguintes hiperparâmetros:

- Taxa de aprendizado: 0,01

- Profundidade máxima de cada árvore: 3

- Número de árvores: 100

O desempenho do modelo foi avaliado com base no coeficiente de determinação (R^2) e a raiz do erro quadrático médio (RMSE). Para a determinação do modelo foi usada a classe *GradientBoostingRegressor* da biblioteca *scikit-learn* do Python. Para a determinação da importância de cada variável foi utilizado o atributo *feature_importances_* do *scikit-learn*.

4.2.3 *XGBoost*

Da mesma forma que para o *Random Forest* e o *Gradient Boosting*, foram consideradas como variáveis explicativas as médias anuais das variáveis temperatura do ar – bulbo seco e umidade relativa do ar e a média diária da precipitação total; já o atributo resposta foi o rendimento médio da produção no ano. As observações foram divididas em conjunto de

treino (70% do total) e teste (30%). Foi utilizada validação cruzada com cinco subconjuntos para a estimativa dos hiperparâmetros.

Para a atribuição dos hiperparâmetros taxa de aprendizado, profundidade máxima de cada árvore e número de árvores foi utilizada a classe *GridSearchCV* da biblioteca *scikit-learn* do Python. Foram empregados os seguintes hiperparâmetros:

- Taxa de aprendizado: 0,01
- Profundidade máxima de cada árvore: 5
- Número de árvores: 100

O desempenho do modelo foi avaliado com base no coeficiente de determinação (R^2) e a raiz do erro quadrático médio (RMSE). Para a determinação do modelo foi usada a classe *XGBRegressor* da biblioteca *XGBoost* do Python. Para a determinação da importância de cada variável foi utilizado o atributo *feature_importances_* do *scikit-learn*.

5 RESULTADOS E DISCUSSÃO

Neste capítulo são apresentados os resultados obtidos - construção de modelos de previsão da produtividade do café usando *Random Forest*, *Gradient Boosting* e *XGBoost* - e é feita a discussão deles.

Os valores de precipitação, temperatura, umidade média e produtividade utilizado para para cada município e ano estudado são apresentados na Tabela 1 a seguir:

Tabela 1 - Precipitação, Temperatura, Umidade e Produtividade.

Município	Ano	Precipitação (mm/dia)	Temperatura (ºC)	Umidade (%)	Produtividade (kg/ha)
Ariranha	2010	0,67	23,0	67,6	500
Ariranha	2011	3,1	22,2	70,1	500
Ariranha	2012	1,6	22,7	68,2	667
Ariranha	2013	3,4	22,2	71,3	667
Ariranha	2014	3,0	22,9	66,8	667
Ariranha	2015	3,2	23,7	73,4	500
Ariranha	2016	2,8	22,6	71,0	500
Ariranha	2017	3,2	23,5	70,2	500
Ariranha	2018	2,3	22,7	69,9	500
Ariranha	2019	1,7	22,9	68,5	500
Avaré	2010	3,3	20,2	72,2	900
Avaré	2011	3,3	20,0	73,3	960
Avaré	2012	4,3	20,6	73,4	960
Avaré	2013	4,8	20,5	73,7	960
Avaré	2014	3,6	22,0	67,1	960
Avaré	2015	5,3	20,9	75,4	760
Avaré	2016	3,7	20,3	73,3	960
Avaré	2017	5,4	20,6	72,6	960
Avaré	2018	4,4	21,1	72,5	1000
Avaré	2019	3,0	21,2	70,8	1057
Barra do Turvo	2010	5,6	18,0	87,6	1666
Barra do Turvo	2011	6,1	17,3	88,0	1000
Barra do Turvo	2012	4,0	18,0	86,9	1333
Barra do Turvo	2013	4,3	17,0	87,3	1333
Barra do Turvo	2014	4,9	18,5	86,5	1333
Barra do Turvo	2015	5,4	18,8	88,7	667
Barra do Turvo	2016	4,7	17,6	88,2	667
Barra do Turvo	2017	5,5	18,1	89,2	667
Barra do Turvo	2018	3,3	18,0	89,6	667
Barra do Turvo	2019	3,6	18,3	90,4	1000
Bauru	2010	2,3	21,1	69,5	553
Bauru	2011	3,9	21,2	72,8	714
Bauru	2012	3,5	21,8	73,5	760
Bauru	2013	4,6	21,7	75,2	760
Bauru	2014	2,6	22,0	68,4	760
Bauru	2015	3,8	22,1	74,5	1100
Bauru	2016	3,2	21,5	72,3	1100
Bauru	2017	3,8	21,8	71,8	1000
Bauru	2018	3,3	22,4	71,3	1000
Bauru	2019	2,3	24,5	76,6	1000

Tabela 1 - Precipitação, Temperatura, Umidade e Produtividade (continuação).

Município	Ano	Precipitação (mm/dia)	Temperatura (ºC)	Umidade (%)	Produtividade (kg/ha)
Casa Branca	2010	0,72	21,8	65,8	1333
Casa Branca	2011	2,3	21,7	65,3	1460
Casa Branca	2012	3,1	21,7	65,0	1388
Casa Branca	2013	5,0	21,4	69,7	1510
Casa Branca	2014	2,1	22,5	60,8	1510
Casa Branca	2015	3,0	22,7	69,4	1073
Casa Branca	2016	1,9	22,0	68,0	1168
Casa Branca	2017	1,4	22,2	64,2	1197
Casa Branca	2018	1,7	22,1	65,5	1241
Casa Branca	2019	2,7	22,4	65,5	1353
Franca	2010	4,9	21,5	64,5	2100
Franca	2011	4,7	21,0	67,1	840
Franca	2012	4,1	21,3	65,4	842
Franca	2013	4,5	21,0	68,3	1080
Franca	2014	2,4	21,9	60,0	1680
Franca	2015	3,7	22,0	65,2	1020
Franca	2016	3,5	21,7	63,9	2400
Franca	2017	4,2	21,6	62,3	840
Franca	2018	4,8	21,6	64,2	2400
Franca	2019	5,3	22,5	68,2	1080
Ibitinga	2010	2,8	21,6	66,4	900
Ibitinga	2011	4,1	22,0	70,2	900
Ibitinga	2012	2,9	22,5	69,7	1200
Ibitinga	2013	3,4	21,8	71,7	3000
Ibitinga	2014	2,6	23,0	68,1	1000
Ibitinga	2015	4,2	22,9	71,7	500
Ibitinga	2016	3,9	22,3	69,4	2000
Ibitinga	2017	2,1	22,4	67,6	1182
Ibitinga	2018	1,4	22,6	68,9	1154
Ibitinga	2019	1,7	23,0	65,8	1308

Tabela 1 - Precipitação, Temperatura, Umidade e Produtividade (continuação).

Município	Ano	Precipitação (mm/dia)	Temperatura (ºC)	Umidade (%)	Produtividade (kg/ha)
Itapira	2010	3,3	21,7	69,2	960
Itapira	2011	4,5	21,4	71,0	960
Itapira	2012	4,3	22,4	69,6	960
Itapira	2013	4,5	21,3	71,3	960
Itapira	2014	2,5	22,4	63,3	768
Itapira	2015	4,3	22,4	70,9	780
Itapira	2016	3,7	21,6	69,9	900
Itapira	2017	3,6	21,8	67,6	900
Itapira	2018	3,2	21,7	69,0	2068
Itapira	2019	3,1	21,9	67,8	1589
Ituverava	2010	1,5	22,3	67,8	1500
Ituverava	2011	4,3	22,0	69,5	2280
Ituverava	2012	3,6	22,4	68,7	1500
Ituverava	2013	2,6	22,1	71,6	1632
Ituverava	2014	2,9	22,8	65,0	1671
Ituverava	2015	4,7	23,0	69,4	1671
Ituverava	2016	3,8	22,5	69,0	2096
Ituverava	2017	3,5	22,8	66,9	2096
Ituverava	2018	4,3	22,5	69,8	2096
Ituverava	2019	3,4	23,0	68,8	2099
Jales	2010	0,67	23,0	67,6	886
Jales	2011	6,0	24,5	66,2	718
Jales	2012	3,3	24,2	63,4	1600
Jales	2013	3,4	23,6	65,9	933
Jales	2014	3,0	24,2	62,1	1800
Jales	2015	4,5	24,4	66,7	1200
Jales	2016	2,9	24,5	61,9	1250
Jales	2017	5,0	24,0	64,6	1250
Jales	2018	4,0	24,5	64,2	867
Jales	2019	2,4	24,4	61,3	867

Tabela 1 - Precipitação, Temperatura, Umidade e Produtividade (continuação).

Município	Ano	Precipitação (mm/dia)	Temperatura (ºC)	Umidade (%)	Produtividade (kg/ha)
José Bonifácio	2010	0,42	23,3	65,8	1000
José Bonifácio	2011	3,90	23,0	65,1	1000
José Bonifácio	2012	3,23	23,5	62,1	1091
José Bonifácio	2013	3,82	23,1	58,4	800
José Bonifácio	2014	2,41	23,6	54,9	1200
José Bonifácio	2015	4,18	24,0	73,8	1200
José Bonifácio	2016	3,41	23,2	71,0	1200
José Bonifácio	2017	3,17	23,4	68,2	1200
José Bonifácio	2018	2,72	23,4	68,0	1200
José Bonifácio	2019	3,04	24,8	68,0	1800
Ourinhos	2010	3,19	21,2	73,4	700
Ourinhos	2011	3,31	21,5	71,2	720
Ourinhos	2012	3,33	22,4	72,5	938
Ourinhos	2013	3,84	21,6	73,2	1000
Ourinhos	2014	4,24	22,4	70,9	950
Ourinhos	2015	4,98	22,2	75,2	588
Ourinhos	2016	4,48	21,7	73,0	917
Ourinhos	2017	5,33	23,1	70,6	1593
Ourinhos	2018	3,63	22,1	71,2	750
Ourinhos	2019	2,97	22,8	71,1	708
São Carlos	2010	3,32	20,8	68,0	1200
São Carlos	2011	4,80	20,6	68,8	998
São Carlos	2012	3,80	20,6	67,9	840
São Carlos	2013	4,29	20,4	71,6	844
São Carlos	2014	2,57	21,5	64,5	675
São Carlos	2015	3,88	21,4	71,2	900
São Carlos	2016	4,87	20,8	68,9	840
São Carlos	2017	4,32	21,1	68,2	900
São Carlos	2018	4,07	21,6	69,2	900
São Carlos	2019	3,86	21,6	68,6	900

Os valores de coeficiente de determinação (R^2) para os conjuntos de treino e de teste, raiz do erro quadrático médio (RMSE) para os dados de treinamento e teste e o tempo de execução de cada modelo obtido são apresentados na Tabela 2 a seguir:

Tabela 2: Desempenho dos algoritmos.

Modelo	R^2 de treino	R^2 de teste	RMSE de treino (kg/ha)	RMSE de teste (kg/ha)	Tempo de execução (s)
Random Forest	0,43	0,11	355,3	415,1	9,9
Gradient Boosting	0,41	0,10	359,9	416,6	1,6
XGBoost	0,39	0,12	367,8	412,6	4,9

Por meio dos valores de R^2 de treino obtidos, é possível perceber que os modelos obtidos pelos algoritmos *Random Forest*, *Gradient Boosting* e *XGBoost* apresentaram limitações na explicação da variabilidade da produtividade frente às variáveis precipitação, temperatura e umidade. Além disso, a capacidade de predição dos três métodos utilizados foi moderada. Vale ressaltar que nenhum dos três algoritmos obteve desempenho significativamente superior aos outros dois.

Isto pode ser devido ao fato de que existem outras variáveis que exercem influência sobre a produtividade, como por exemplo o tipo de solo, se há a utilização de adubação e defensores agrícolas, se a plantação é irrigada ou não, intensidade da radiação solar, etc.

Além disso, a distância entre as plantações e as estações meteorológicas que coletam os dados pode acarretar, devido a presença de microclimas, divergências entre as variáveis obtidas e as condições que estão efetivamente sobre a cultura estudada.

Outro fator que pode ter originado isso pode ter sido o fato de que os valores de rendimento médio apontados pela pesquisa PAM do IBGE são obtidos por meio de estimativas de agentes em contato com técnicos do setor agrícola, produtores e entidades específicas de controle. Isso pode gerar ruídos e incertezas nesta variável, fazendo com que haja diminuição da capacidade preditiva do modelo.

Ademais, o fato de que foram usados dados de diversos municípios para a geração de um modelo geral pode ter diminuído as assertividades de previsão obtidas, visto que cada um deles possui particularidades que podem fazer com que seja difícil ajustar bem a todas as distribuições ao mesmo tempo.

As Figuras 1 e 2 a seguir apresentam os valores de R^2 e RMSE obtidos para os conjuntos de treino e teste dos três modelos utilizados:

Figura 1 - Comparação do R^2 entre treino e teste.

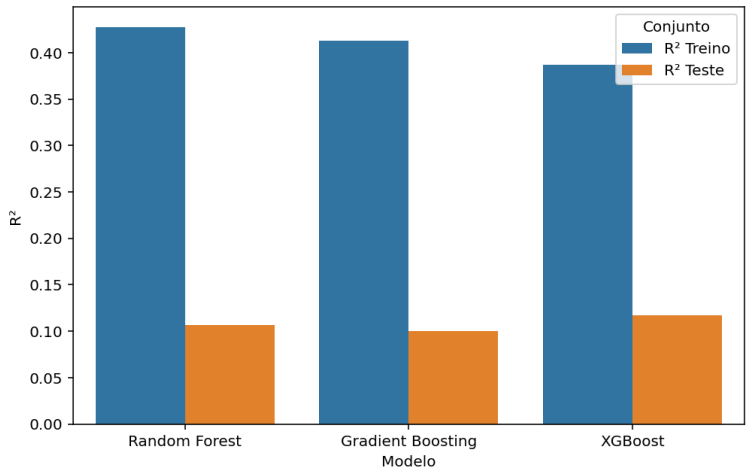
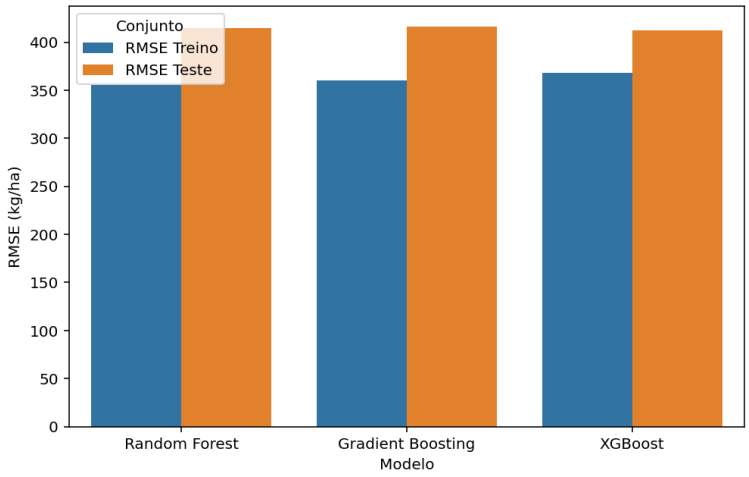


Figura 2 - Comparação do RMSE entre treino e teste.



A partir das Figuras 1 e 2, é possível perceber que os modelos tiveram R^2 e RMSE maiores para os conjuntos de treino em comparação com os de teste. Isso indica que os algoritmos tiveram moderada capacidade de generalização. Além disso, é possível perceber que tanto o *Random Forest*, o *Gradient Boosting* e o *XGBoost* possuem resultados de coeficiente de determinação e de raiz do erro quadrático médio semelhantes, de modo que nenhum deles é mais assertivo que os outros dois.

As Figuras 3, 4 e 5 a seguir apresentam os valores dos resíduos dos valores previstos, respectivamente, pelos modelos *Random Forest*, *Gradient Boosting* e *XGBoost*:

Figura 3 - Resíduos obtidos pelo modelo *Random Forest*.

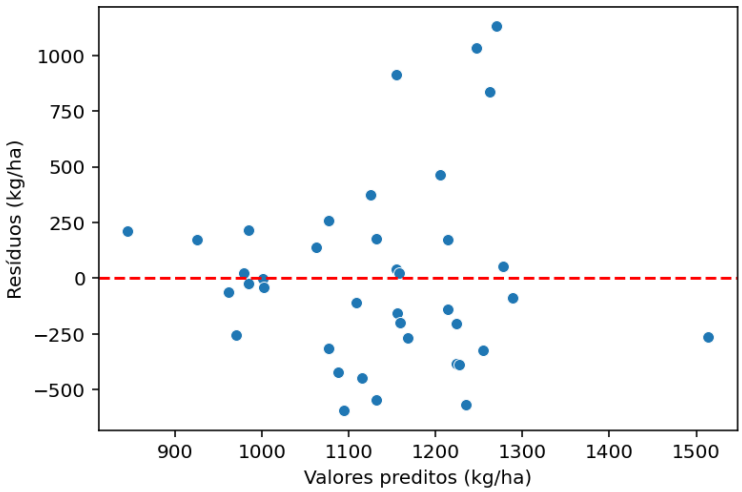


Figura 4 - Resíduos obtidos pelo modelo *Gradient Boosting*.

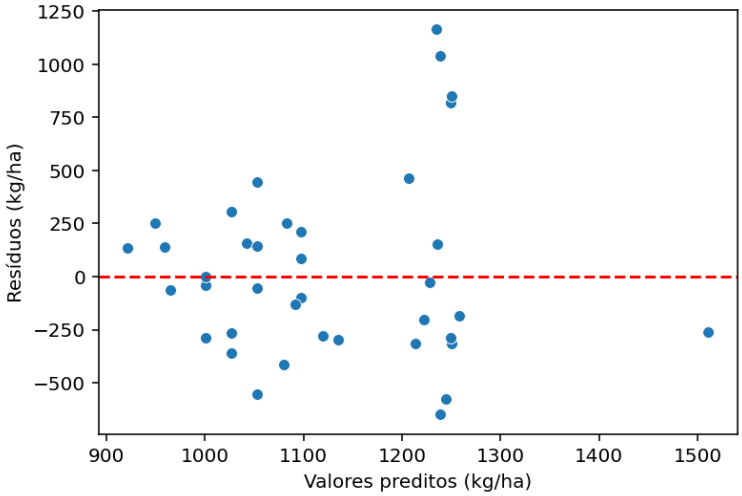
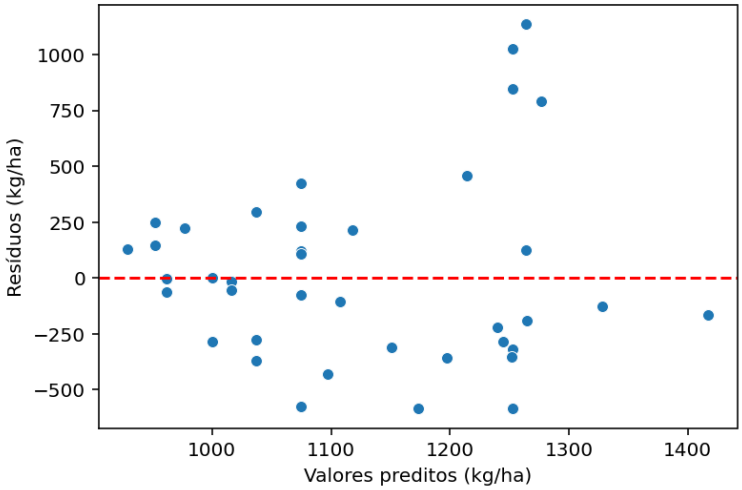


Figura 5 - Resíduos obtidos pelo modelo *XGBoost*.



Por meio dos gráficos das Figuras 3, 4 e 5 é possível notar que, para os três modelos, os resíduos estão dispersos ao redor do eixo das abscissas (linha tracejada). Isso indica que não há viés sistemático nas previsões. Além disso, é possível notar que na faixa de previsão entre 1200 e 1300 kg/ha os erros são maiores, o que indica que os modelos têm maior dificuldade de prever com assertividade neste intervalo.

As Figuras 6, 7 e 8 a seguir apresentam os histogramas da distribuição dos resíduos obtidos pelos modelos *Random Forest*, *Gradient Boosting* e *XGBoost*:

Figura 6 - Histograma dos resíduos obtidos pelo *Random Forest*.

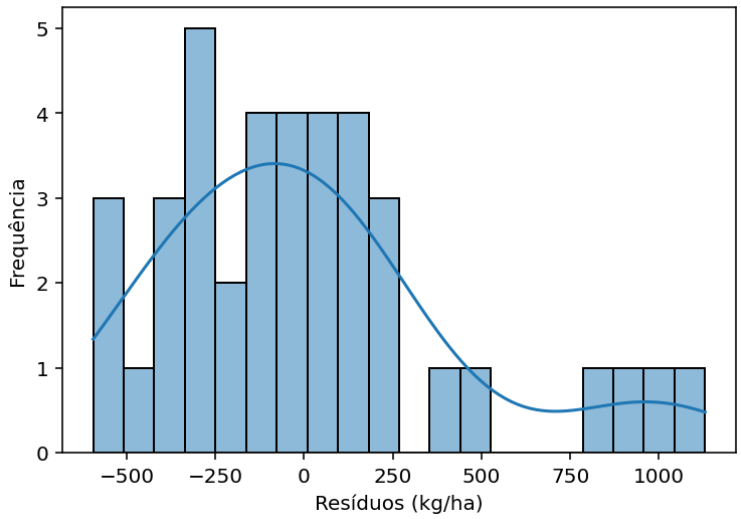


Figura 7 - Histograma dos resíduos obtidos pelo *Gradient Boosting*.

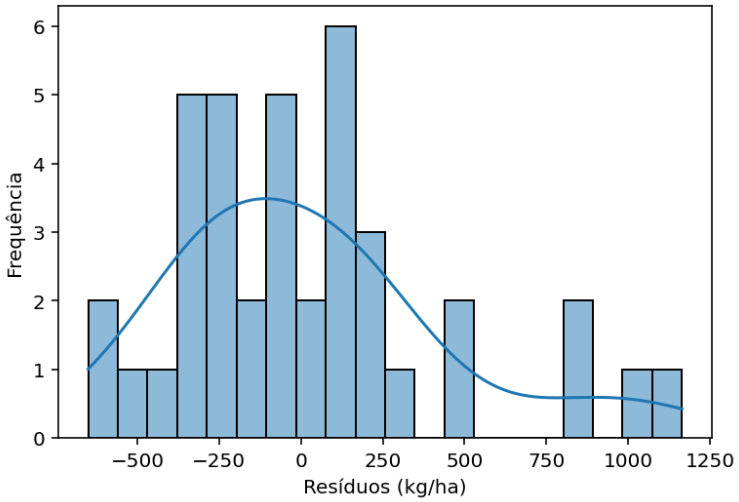
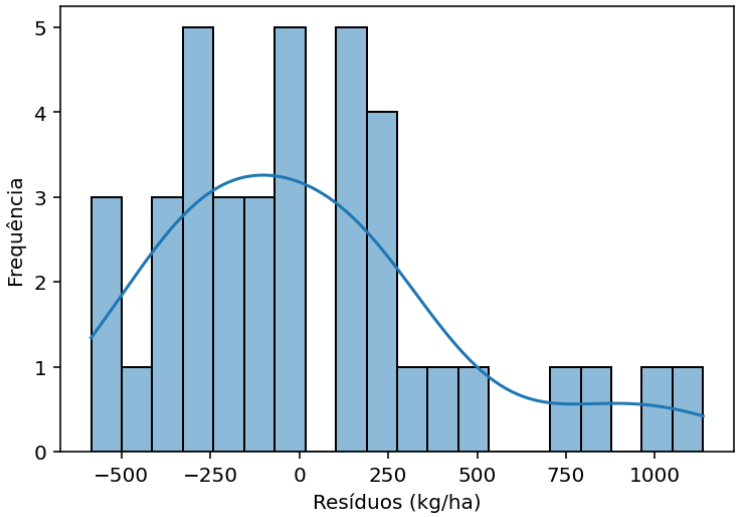


Figura 8 - Histograma dos resíduos obtidos pelo *XGBoost*.



Por meio das Figuras 6, 7 e 8 é possível notar que há, para os três modelos, uma maior frequência dos resíduos negativos. Isso indica que há uma leve tendência deles de superestimarem os valores de produtividade. Além disso, é possível ver que, por haver uma cauda longa à direita, em alguns casos em que há subestimação do rendimento os resíduos são elevados. O fato de que as distribuições dos resíduos não seguem a distribuição normal indica que os modelos possuem limitação de capacidade de generalização.

As Figuras 9, 10 e 11 a seguir apresentam os valores da contribuição das variáveis precipitação, temperatura e umidade para os modelos *Random Forest*, *Gradient Boosting* e *XGBoost* respectivamente:

Figura 9 - Importância das variáveis para o modelo *Random Forest*.

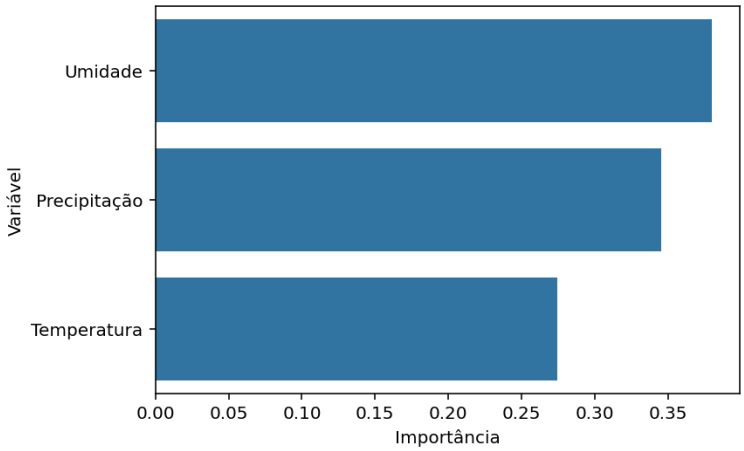


Figura 10 - Importância das variáveis para o modelo *Gradient Boosting*.

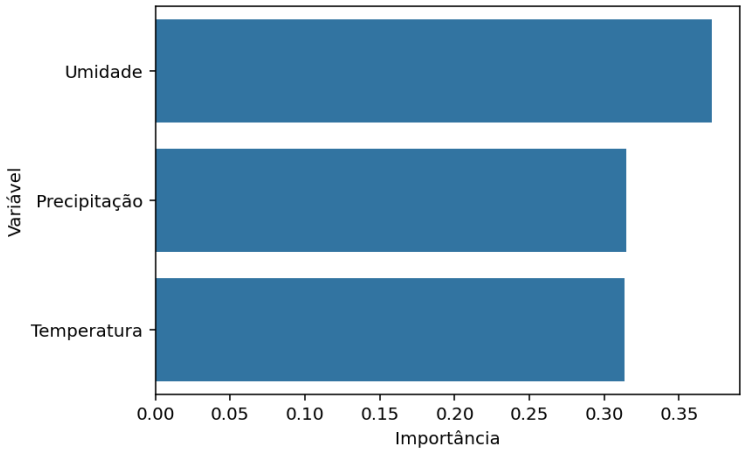
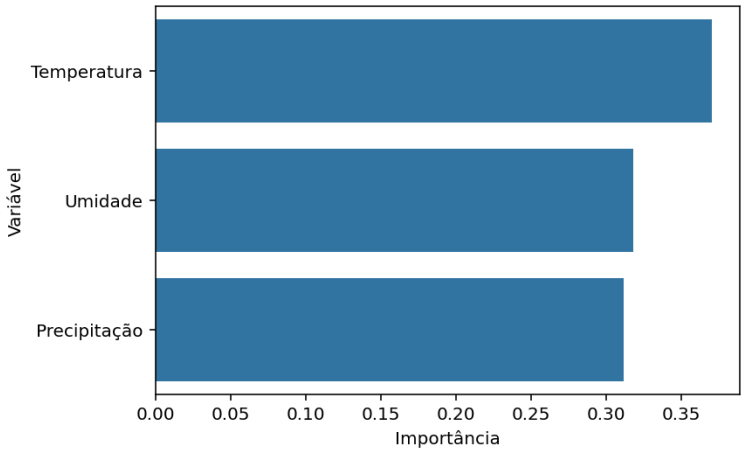


Figura 11 - Importância das variáveis para o modelo *XGBoost*.



Pelas Figuras 9, 10 e 11 é possível perceber que as três variáveis possuem importância semelhante para a construção dos modelos, sendo que a umidade foi a mais importante para o *Random Forest* e o *Gradient Boosting* e a temperatura para o *XGBoost*. Essa diferença entre os

modelos se deve ao fato de que cada um deles calcula essa importância de forma diferente, utilizando métricas diversas.

6 CONCLUSÃO

Este trabalho teve como objetivo geral a criação de modelos de previsão da produtividade do café a partir de variáveis meteorológicas, utilizando para isso algoritmos de aprendizado de máquina: *Random Forest*, *Gradient Boosting* e *XGBoost*.

Os modelos obtidos tiveram desempenho semelhante para os três algoritmos, porém o *XGBoost* apresentou assertividade levemente superior, com maior coeficiente de determinação e menor raiz do erro quadrático médio. Foram obtidos valores de R^2 de 0,12 (*XGBoost*), 0,11 (*Random Forest*) e 0,10 (*Gradient Boosting*); já os de RMSE foram de 412,6 (*XGBoost*), 415,1 (*Random Forest*) e 416,6 kg/ha (*Gradient Boosting*).

A análise da importância das variáveis para a previsão da produtividade indicou que, embora as três possuam semelhante significância, a umidade aparece como a mais relevante quando usados os métodos *Random Forest* e *Gradient Boosting*. Isso reforça a importância da irrigação para o rendimento agrícola.

Já ao utilizar o algoritmo *XGBoost* obteve-se a temperatura como variável mais relevante. Essa diferença é devido à diferença de cálculo da importância nestes métodos de aprendizado de máquina.

As limitações dos modelos de previsão obtidos podem ser explicadas pela ausência de variáveis chave no modelo, como radiação solar, tipo de solo e condições de cultivo; a heterogeneidade dos municípios de estudo e a forma de obtenção dos dados meteorológicos e de produtividade.

Conclui-se que, embora os modelos apresentados não tenham apresentado elevado poder preditivo, foi possível obter entendimentos relevantes acerca da influência das variáveis meteorológicas sobre a produtividade do café. Além disso, percebeu-se o potencial destes algoritmos para a análise agrícola desde que sejam feitas abordagens metodológicas adequadas.

7 SUGESTÕES DE TRABALHOS FUTUROS

Como possibilidades de futuros trabalhos, apresentam-se as seguintes alternativas:

- Incorporação de novas variáveis explicativas no modelo de predição da produtividade, como tipo de solo de cultivo, radiação solar, temperatura máxima e mínima, etc.
- Construção dos modelos de forma individual para cada município, incorporando uma faixa de tempo maior de modo a que haja mais dados de treinamento.
- Uso de outros algoritmos de aprendizado de máquina para os modelos de previsão, como redes neurais artificiais, *Support Vector Machines*, *Deep Learning*, regressão linear, etc.
- Utilização das variáveis explicativas em granularidade menor (mensal ou trimestral, por exemplo).
- Utilização de dados meteorológicos e de produtividade coletados *in loco* a nível de fazenda ou plantação, de modo que seja possível ter maior precisão e controle. Além disso, dessa forma é possível investigar a influência das variáveis meteorológicas nas fases de crescimento da planta.

REFERÊNCIAS

BARBOSA, B. D. S.; FERRAZ, G. A. S.; COSTA, L.; AMPATZIDIS, Y.; VIJAYAKUMAR, V.; DOS SANTOS, L. M. 2021. UAV-based coffee yield prediction utilizing feature selection and deep learning. **Smart Agricultural Technology** 1: 100010.

CRANE-DROESCH, A. Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. 2018. **Environmental Research Letters** 13: 114003.

DE CAMARGO, M. B. P. 2010. The impact of climatic variability and climate change on arabic coffee crop in Brazil. **Bragantia**, Campinas 69: 239-247.

DINH, T. L. A.; AIRES, F.; RAHN, E. 2022. Statistical Analysis of the Weather Impact on Robusta Coffee Yield in Vietnam. **Frontiers in Environmental Science** 10: 820916.

JAYAKUMAR, M.; RAJAVEL, M.; SURENDRAN, U. 2016. **Int J Biometeorol** 60: 1943-1952.

KITTICHOTSATSAWAT, Y.; TIPPAYAWONG, N.; TIPPAYAWONG, K. Y. 2022. Prediction of arabica coffee production using artificial neural network and multiple linear regression techniques. **Nature** 12: 14488.

KOUADIO, L.; TIXIER, P.; BYRAREDDY, V.; MARCUSSEN, T.; MUSHTAQ, S.; RAPIDEL, B.; VAUTARD, R.; STONE, R. 2021. Performance of a process-based model for predicting robusta coffee yield at the regional scale in Vietnam. **Ecological Modelling** 443: 109469.

LEGESSE, A. 2019. Climate Change Effect on Coffee Yield and Quality: A Review. **International Journal of Forestry and Horticulture** 5: 1-9.

Ministério do Desenvolvimento, Indústria, Comércio e Serviços [MDIC]. 2024. **ComexVis**. Disponível em: <http://comexstat.mdic.gov.br/pt/comex-vis>. Acesso em: 03 de março de 2025.

PHAM, Y.; REARDON-SMITH, K.; MUSHTAQ, S.; COCKFIELD, G. 2019. The impact of climate change on coffee production: a systematic review. **Climatic Change** 156: 609-630.

STOTT, P. A.; CHRISTIDIS, N.; OTTO, F. E. L.; SUN, Y.; VANDERLINDEN, J.; VAN OLDENBORGH, G. J.; VAUTARD, R.; VON STORCH, H.; WALTON, P.; YIOU, P.; ZWIERS, F. W. 2016. Attribution of extreme weather and climate-related events. **WIREs Climate Change** 7: 23-41.

U. S. Department of Agriculture [USDA]. 2024. **Production – Coffee**. Disponível em: <https://www.fas.usda.gov/data/production/commodity/0711100>. Acesso em: 03 de março de 2025.

Apêndice A – *Script*

A seguir é apresentado o *script* usado no trabalho.

```
# %% Importação de bibliotecas
import pandas as pd
import numpy as np
import time
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.model_selection import train_test_split, GridSearchCV, KFold
from sklearn.ensemble import RandomForestRegressor, GradientBoostingRegressor
from sklearn.metrics import mean_squared_error, r2_score
from xgboost import XGBRegressor

# %% Leitura dos dados
df = pd.read_excel("Todos.xlsx")
X = df[["Precipitação", "Temperatura", "Umidade"]]
y = df["Produtividade"]

# Divisão treino/teste
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.3, random_state=42
)

# Configuração validação cruzada
cv = KFold(n_splits=5, shuffle=True, random_state=42)

# Função auxiliar para treinar, avaliar e medir tempo
```

```
def treinar_modelo(nome, modelo, param_grid):
    inicio = time.time()
    grid = GridSearchCV(modelo, param_grid, cv=cv, scoring='r2', n_jobs=-1)
    grid.fit(X_train, y_train)
    melhor_modelo = grid.best_estimator_
    fim = time.time()

    # Previsões treino e teste
    y_train_pred = melhor_modelo.predict(X_train)
    y_test_pred = melhor_modelo.predict(X_test)
    residuos = y_test - y_test_pred

    # Métricas treino
    rmse_train = np.sqrt(mean_squared_error(y_train, y_train_pred))
    r2_train = r2_score(y_train, y_train_pred)

    # Métricas teste
    rmse_test = np.sqrt(mean_squared_error(y_test, y_test_pred))
    r2_test = r2_score(y_test, y_test_pred)
    tempo = fim - inicio

    # Importância das features
    importancia = pd.DataFrame({
        "Variável": X.columns,
        "Importância": melhor_modelo.feature_importances_
    }).sort_values(by="Importância", ascending=False)

    return {
        "nome": nome,
```

```
        "modelo": melhor_modelo,
        "params": grid.best_params_,
        "rmse_train": rmse_train,
        "rmse_test": rmse_test,
        "r2_train": r2_train,
        "r2_test": r2_test,
        "tempo": tempo,
        "y_pred": y_test_pred,
        "residuos": residuos,
        "importancia": importancia
    }

# %% Modelos e grids de hiperparâmetros
param_rf = {
    "n_estimators": [100, 200, 300],
    "max_depth": [None, 3, 5, 10],
    "min_samples_split": [2, 5],
    "min_samples_leaf": [1, 2]
}

param_gbr = {
    "n_estimators": [100, 200, 300],
    "learning_rate": [0.01, 0.05, 0.1],
    "max_depth": [2, 3, 4, 5]
}

param_xgb = {
    "n_estimators": [100, 200, 300],
    "learning_rate": [0.01, 0.05, 0.1],
```

```
    "max_depth": [2, 3, 4, 5]
}

# %% Treinamento
resultados = []
resultados.append(treinar_modelo("Random Forest", RandomForestRegressor(random_state=42), param_rf))
resultados.append(treinar_modelo("Gradient Boosting", GradientBoostingRegressor(random_state=42), param_gbr))
resultados.append(treinar_modelo("XGBoost", XGBRegressor(random_state=42, objective="reg:squarederror"), param_xgb))

# %% Tabela comparativa
comparacao = pd.DataFrame([{
    "Modelo": r["nome"],
    "R² Treino": r["r2_train"],
    "R² Teste": r["r2_test"],
    "RMSE Treino": r["rmse_train"],
    "RMSE Teste": r["rmse_test"],
    "Tempo (s)": r["tempo"],
    "Melhores parâmetros": r["params"]
} for r in resultados])

print(comparacao)

# %% Gráficos
for r in resultados:
    nome = r["nome"]
    y_pred = r["y_pred"]
    residuos = r["residuos"]
```

```
        # Gráfico de resíduos
        plt.figure(figsize=(6,4))
        sns.scatterplot(x=y_pred, y=residuos)
        plt.axhline(0, color="red", linestyle="--")
        plt.title(f"Resíduos vs Preditos - {nome}")
        plt.xlabel("Valores preditos (kg/ha)")
        plt.ylabel("Resíduos (kg/ha)")
        plt.show()

        # Histograma dos resíduos
        plt.figure(figsize=(6,4))
        sns.histplot(residuos, bins=20, kde=True)
        plt.title(f"Distribuição dos Resíduos - {nome}")
        plt.xlabel("Resíduos (kg/ha)")
        plt.ylabel("Frequência")
        plt.show()

        # Importância das variáveis
        plt.figure(figsize=(6,4))
        sns.barplot(data=r["importancia"], x="Importância", y="Variável")
        plt.title(f"Importância das variáveis - {nome}")
        plt.show()

    # %% Gráfico comparativo de desempenho
    # R² Treino vs Teste
    plt.figure(figsize=(8,5))
    comparacao_melt_r2 = comparacao.melt(id_vars="Modelo", value_vars=["R² Treino",
"R² Teste"],
                                         var_name="Conjunto", value_name="R²")
    sns.barplot(data=comparacao_melt_r2, x="Modelo", y="R²", hue="Conjunto")
```

```
plt.title("Comparação R² - Treino vs Teste")
plt.ylabel("R²")
plt.show()


# RMSE Treino vs Teste
plt.figure(figsize=(8,5))
comparacao_melt_rmse = comparacao.melt(id_vars="Modelo", value_vars=["RMSE Treino", "RMSE Teste"],
                                       var_name="Conjunto", value_name="RMSE")
sns.barplot(data=comparacao_melt_rmse, x="Modelo", y="RMSE", hue="Conjunto")
plt.title("Comparação RMSE - Treino vs Teste")
plt.ylabel("RMSE (kg/ha)")
plt.show()
```