UNIVERSIDADE FEDERAL DE SÃO CARLOS
CAMPUS SÃO CARLOS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS AMBIENTAIS

ESTOQUE DE CARBONO EM SISTEMAS AGROFLORESTAIS COM CACAU: UMA ABORDAGEM DE CLASSIFICAÇÃO E REGRESSÃO COM RANDOM FOREST

GABRIELA GALETTI RUSCA

Disciplina: Introdução Ao Machine Learning - Cam 413

Departamento de Ciências Ambientais

Email: ggalettirusca@gmail.com | ggrusca@ufscar.estudante.br

Professor: Dr. Marcos Roberto Benso

São Carlos
2025

Resumo executivo

Este relatório apresenta a construção e avaliação de modelos preditivos para estimar a o estoque de carbono (tC/ha) em sistemas agroflorestais tropicais com cultivo de cacau. A modelagem foi realizada com dados do projeto *RESTORE+* ***Productivity and Carbon Estimates for Tree-Based Cropping Systems***, produzido pelo IIASA (*International Institute for Applied Systems Analysis*), utilizando o algoritmo **Random Forest**, reconhecido por sua capacidade de capturar relações não lineares, interações complexas entre variáveis e boa capacidade de generalização, Breiman (2001).

A base de dados é proveniente da Indonésia, composta por informações produtivas, ambientais e de manejo, referentes a consórcios de cacau com espécies como café, dendê e coco. Utilizou-se uma amostragem de 2% da base original para prototipagem, garantindo desempenho computacional e robustez estatística. As variáveis explicativas incluíram: faixa etária dos plantios (age_group), classe de adequação agroecológica (value_class), tipo de consórcio (parameter_name), tipo de solo (soil_type) e carbono estocado (carbon).

Dois modelos principais foram desenvolvidos: **Modelo 1 – Classificação do Estoque de Carbono (Baixo, Médio, Alto):** apresentou acurácia de 97,9% e Kappa = 0,96, com destaque para a idade do plantio e tipo de consórcio como principais preditores e **Modelo 2 – Regressão do Estoque de Carbono**: obteve $R^2 = 0.98$ e RMSE = 0.62 tC/ha, indicando ajuste robusto. A idade do sistema foi a variável mais explicativa (100% de importância relativa), e o tipo de solo não apresentou relevância.

Também foi testado um modelo binário de classificação da produtividade (Alta vs. Baixa), mas os resultados foram insatisfatórios (acurácia de 51% e Kappa < 0), levando à sua exclusão da análise final.

As análises indicaram padrões coerentes do ponto de vista agronômico: o estoque de carbono aumenta com a idade dos sistemas, enquanto a produtividade tende a se estabilizar ou diminuir em estágios mais avançados. O modelo tem potencial aplicação em políticas públicas de agricultura de baixo carbono, planejamento agroecológico e mecanismos de incentivo ambiental. Para aplicação em escala local (como no Brasil), recomenda-se o ajuste do modelo com dados regionais e validação participativa com stakeholders locais.

Palavras-chave: Agrofloresta; Cacau; Machine Learning; Estoque de Carbono; Produtividade.

1. Introdução

O crescimento da agricultura de baixo carbono representa um desafio e uma oportunidade para conciliar produtividade agrícola e mitigação das mudanças climáticas. Em especial, os sistemas agroflorestais (SAFs) com cacau destacam-se por sua capacidade de sequestrar carbono atmosférico ao mesmo tempo em que geram renda para agricultores familiares em regiões tropicais (Kongor et al., 2024). No entanto, fatores que influenciam o estoque de carbono e a produtividade agrícola em SAFs ainda são pouco compreendidos, especialmente quando se consideram interações complexas entre variáveis ambientais e de manejo.

Neste contexto, este trabalho busca aplicar técnicas de aprendizado de máquina (Random Forest) para modelar e prever o estoque de carbono (tC/ha) e a produtividade do cacau (t/ha/ano) com base em dados reais de campo, apoiando decisões de manejo sustentável e políticas públicas voltadas à agricultura regenerativa. A modelagem foi realizada com dados do projeto *RESTORE+* ***Productivity and Carbon Estimates for Tree-Based Cropping Systems***, produzido pelo IIASA (*International Institute for Applied Systems Analysis*), utilizando o algoritmo **Random Forest**, reconhecido por sua capacidade de capturar relações não lineares, interações complexas entre variáveis e boa capacidade de generalização, Breiman (2001).

2. Descrição da Base de Dados

Os dados utilizados neste estudo são provenientes da Indonésia, coletados no âmbito do projeto do projeto RESTORE+ (2023) - Productivity and Carbon Estimates for Tree-Based Cropping Systems, produzido pelo IIASA (International Institute for Applied Systems Analysis). Trata-se de uma base estruturada em formato tabular, composta por parcelas de SAFs com cacau em diferentes arranjos de consórcio e condições edafoclimáticas.

A base original continha mais de 10 mil observações. Para fins de prototipagem, foi utilizada uma amostra de 2% com balanceamento por faixa etária. Foram considerados apenas os sistemas consorciados com cacau e café, cacau e dendê, ou cacau e coco.

Apesar de sua origem indonésia, os dados representam sistemas agroflorestais tropicais semelhantes aos encontrados em regiões da Amazônia e do sul da Bahia, possibilitando hipóteses comparativas para aplicação futura no Brasil.

Foram utilizadas as seguintes colunas (variáveis):

Tabela 1. Variáveis utilizadas.

Variável	Descrição	Tipo	Unidade
age_group	Faixa etária do sistema agroflorestal	Categórica	anos
value_class	Classe de adequação agroecológica (S1 a N2)	Categórica	—
parameter_name	Tipo de consórcio com cacau (café, dendê ou coco)	Categórica	—
soil_type	Tipo de solo (e.g. peat)	Categórica	—
carbon	Estoque de carbono acumulado	Contínua	tC/ha
yield	Produtividade anual do cacau	Contínua	t/ha/ano

3. Metodologia de Aprendizagem de Máquina

Foi utilizado o algoritmo Random Forest, implementado via pacote caret no R, por sua capacidade de lidar com dados não lineares e variáveis categóricas sem necessidade de normalização.

Etapas metodológicas:

- Pré-processamento: exclusão de NAs, reorganização de variáveis, balanceamento da variável age_group.
- Divisão dos dados: 80% para treino e 20% para teste, com estratificação por tipo de consórcio.
- Modelagem: foi implementada na linguagem R (R CORE TEAM, 2024), utilizando os pacotes *caret* (Kuhn et al., 2024) e *pdp* (Greenwell et al., 2023) para treino dos modelos e interpretação dos resultados.

  - Modelo 1 (Classificação do carbono): a variável carbon foi convertida em uma variável categórica (carbon_class: Baixo, Médio, Alto) a partir dos tercis da distribuição.
  - Modelo 2 (Regressão do carbono): modelagem direta da variável contínua carbon

Hiperparâmetros:

Foi feito ajuste de hiperparâmetros (mtry) por validação cruzada 10-fold, testando valores entre 2 e 6. O critério de seleção foi a maior acurácia (modelo 1) ou menor RMSE (modelo 2).

Avaliação:

- Classificação: acurácia, Kappa, matriz de confusão
- Regressão: RMSE, R^2 e importância das variáveis (varImp())

4. Resultados

4.1. Análise Exploratória dos Dados

Esta subseção oferece uma visão preliminar sobre os padrões observados antes da modelagem preditiva.

4.1.1. Produtividade por tipo de consórcio

Boxplot (Figura 1) mostra que o consórcio com café apresentou mediana de produtividade ligeiramente superior ao dos demais. No entanto, há alta variabilidade e sobreposição entre os sistemas.

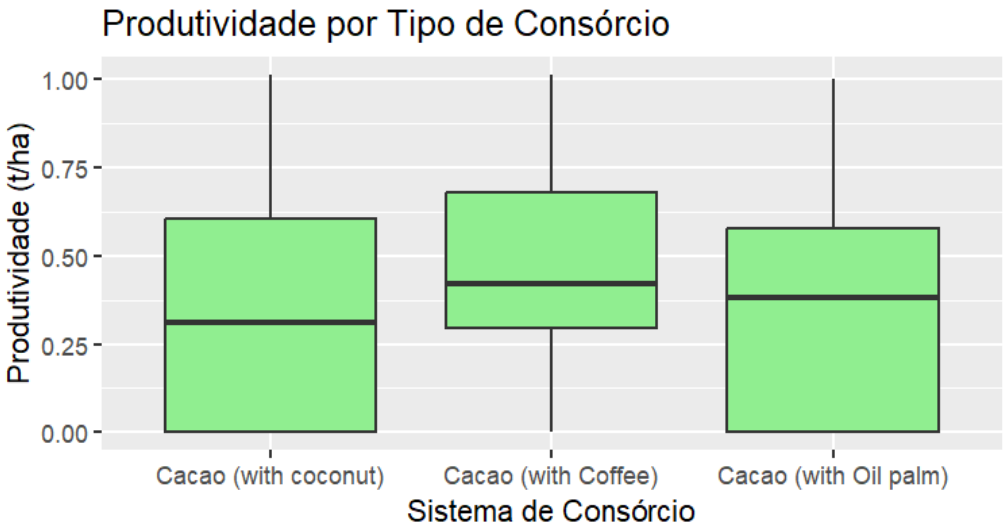

Figura 1 – Produtividade por tipo de consórcio agroflorestal.

4.1.2. Carbono vs Produtividade por consórcio

O gráfico de dispersão com regressão por grupo (Figura 2) mostra que, para sistemas com dendê (Oil Palm), o aumento do carbono tende a se associar a queda de produtividade, enquanto sistemas com café mantêm produtividade estável ou crescente com o carbono.

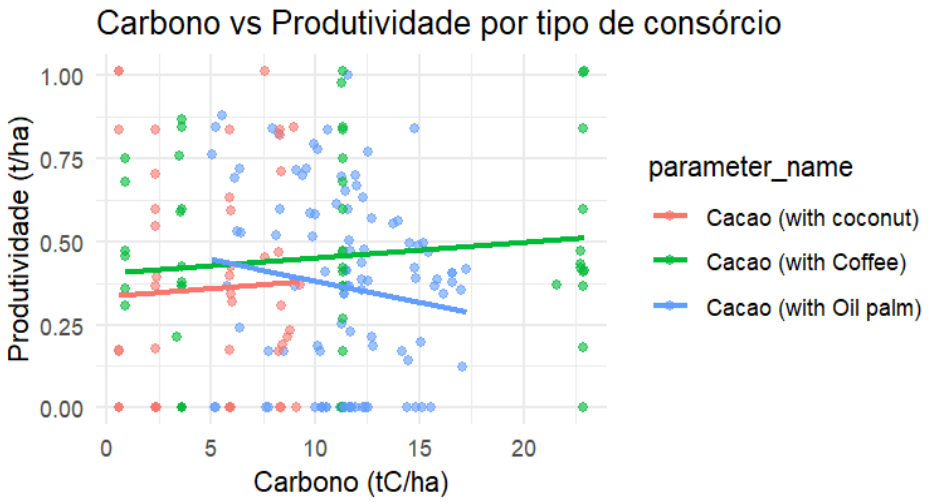

Figura 2 – Relação entre carbono e produtividade por tipo de consórcio.

4.1.3. Carbono e produtividade por faixa etária

O gráfico da média da produtividade e carbono por faixa etária evidencia que o carbono aumenta com a idade do sistema, ao passo que a produtividade tem pico na faixa 9–25 anos, caindo depois — possivelmente por competição ou sombreamento excessivo.

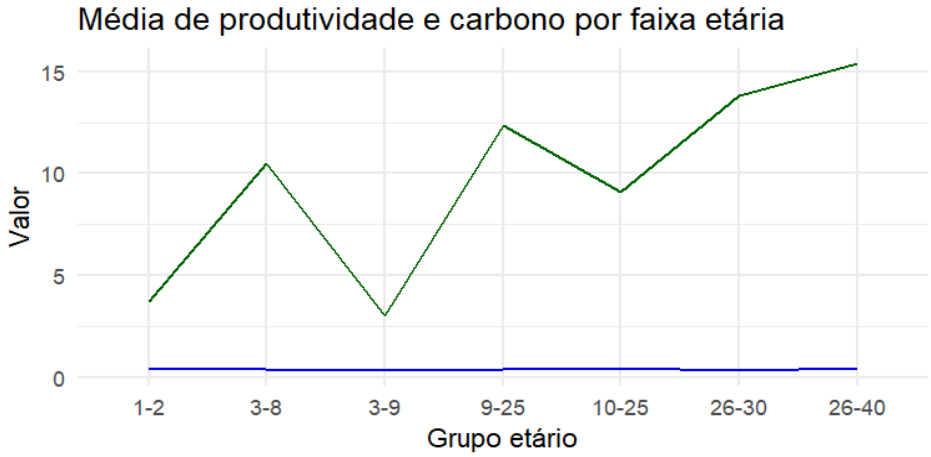

Figura 3 – Média de produtividade e carbono por faixa etária.

4.2. Modelo 1 - Classificação do Estoque de Carbono

4.2.1. Desempenho do modelo de classificação

O modelo Random Forest foi utilizado para classificar o estoque de carbono em três categorias: Baixo, Médio e Alto, com base em variáveis ambientais e agronômicas (tipo de solo, idade do plantio, tipo de consórcio, classe de adequação agroecológica e produtividade).

O processo de tuning do modelo Random Forest testou diferentes valores de mtry (número de preditores considerados por árvore) com validação cruzada 10-fold. A validação cruzada (10-fold) apresentou acurácia média crescente conforme o número de variáveis preditoras (mtry) aumentava. A acurácia média aumentou com o número de preditores até o valor máximo testado. O melhor desempenho foi obtido com mtry = 6 (Figura 4).

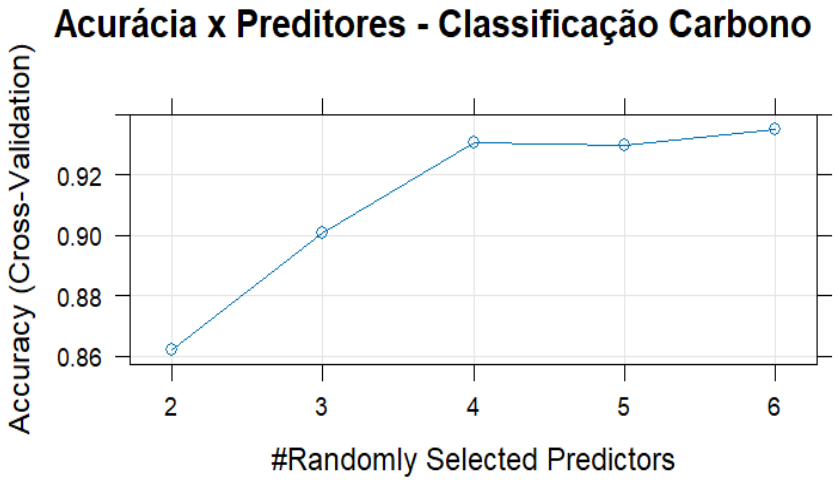

Figura 4 – Acurácia média por número de preditores (mtry) - Classificação do Estoque de Carbono.

A **Acurácia média:** 93,5% e **Kappa:** 0,90 e esses valores indicam forte capacidade preditiva, estabilidade do modelo e baixo risco de sobreajuste.

4.2.2. Avaliação no conjunto de teste

O modelo treinado com Random Forest foi avaliado em dados não vistos previamente (conjunto de teste), correspondentes a 20% da amostra original (49 observações) e todos os consórcios foram classificados corretamente, com **acurácia elevada (97,96%)** e **Kappa = 0,97**, o que sugere que as classes estavam bem separadas no espaço amostral utilizado. Isso pode refletir estruturas distintas entre os consórcios ou forte contribuição das variáveis preditoras.

- **Intervalo de Confiança 95%:** [0,89 ; 0,999]
- **P-valor (Acc > NIR):** < 2.2e-16

A matriz de confusão mostrou que as três classes de carbono foram previstas corretamente, com apenas um erro de classificação (Figura 5).

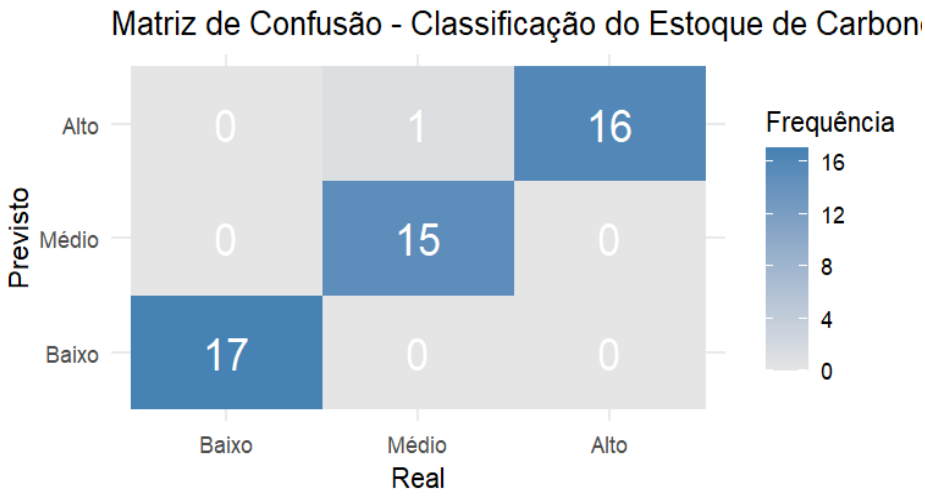

Figura 5 – Matriz de Confusão (Conjunto de Teste).

4.2.3. Importância das variáveis

No algoritmo Random Forest, a importância de uma variável representa quanto ela contribui para reduzir o erro das previsões. Quanto maior o valor, mais essencial ela é para o modelo aprender padrões úteis nos dados.

Esses valores são relativos: a variável mais importante é padronizada para 100%, e as demais são escaladas proporcionalmente. Sendo assim, ela permite identificar quais atributos ambientais e de manejo têm maior influência sobre o estoque de carbono em sistemas agroflorestais com cacau.

A variável mais importante para a classificação do carbono foi a faixa etária do plantio (age_group), seguida pelo tipo de consórcio (parameter_name) e a classe de adequação agroecológica (value_class). A produtividade do cacau (yield) teve contribuição pequena, enquanto o tipo de solo (soil_type) teve importância nula. Esta análise de importância das variáveis no modelo de classificação do estoque de carbono revelou padrões coerentes com o conhecimento agronômico sobre sistemas agroflorestais com cacau:

- **Faixa etária do plantio (age_group)** foi a variável mais influente no modelo. Diversas transformações ortogonais da variável (e.g., age_group.L, age_group^5, age_group^6, etc.) apareceram entre os termos mais importantes, o que indica que a idade do sistema agroflorestal tem forte associação com o acúmulo de carbono. Isso é esperado, já que sistemas

mais antigos tendem a acumular biomassa ao longo do tempo, especialmente em arranjos com espécies lenhosas.

- **Tipo de consórcio (parameter_name)** também teve peso elevado no modelo, especialmente as categorias relacionadas ao cultivo de cacau com dendê e café. Esses consórcios diferem quanto à composição de espécies, densidade e arquitetura vegetal, o que impacta diretamente o potencial de sequestro de carbono. Notou-se que sistemas com dendê apresentaram, em média, estoques mais elevados.

- **Classe de adequação agroecológica (value_class)**, embora com menor influência que idade e consórcio, também se destacou. Como essa variável sintetiza aspectos edáficos, topográficos e climáticos, sua importância sugere que ambientes mais adequados (S1, S2) favorecem o acúmulo de carbono, possivelmente por proporcionarem melhores condições de crescimento e densidade vegetal.

- **Produtividade do cacau (yield)** teve importância marginal, indicando que a produção de frutos não está fortemente correlacionada ao estoque de carbono. Essa separação é comum em SAFs maduros, onde a produtividade pode cair mesmo quando o estoque de biomassa total segue elevado, por competição ou sombreamento excessivo.

- **Tipo de solo (soil_type)** teve importância nula, o que pode ser atribuído à pouca variabilidade de tipos de solo na amostra analisada (predominância de um único tipo, como “peat” ou solos orgânicos mal representados), ou ao fato de que o tipo de consórcio e idade do sistema encapsulam, de forma indireta, os efeitos do solo sobre o carbono.

Imagem 1. Importância das Variáveis no Modelo de Classificação do Estoque de Carbono (Random Forest).

variables are sorted by maximum importance across the classes

	Baixo	Médio	Alto
parameter_nameCacao (with Coffee)	54.757	100.00	72.626
value_class	36.610	77.22	65.788
age_group^6	63.820	60.86	71.560
age_group^5	50.366	53.96	44.628
age_group.L	46.104	46.40	43.349
parameter_nameCacao (with Oil palm)	43.303	28.92	26.474
age_group.Q	27.902	35.37	34.259
age_group^4	32.143	32.85	33.201
age_group.C	28.121	32.53	33.123
soil_typepeat	22.840	14.40	0.000
yield	6.709	21.28	5.779

Fonte: Elaborado pela autora no software R, com base na função varImp() aplicada ao modelo Random Forest de classificação do estoque de carbono (modelo_class_carbon). 2025.

4.2.4. Modelo de Classificação da Produtividade – Não incluído

Um modelo de classificação binária da produtividade (Alta vs Baixa) também foi testado, com base nas mesmas variáveis explicativas utilizadas na modelagem do carbono (faixa etária do plantio, tipo de consórcio, classe de adequação agroecológica, produtividade e tipo de solo).

No entanto, os resultados foram insatisfatórios sendo a Acurácia 51% e o Kappa: -0,01 (valor negativo, indicando desempenho inferior ao acaso).

A elevada sobreposição entre as classes de produtividade (Alta vs Baixa) dificultou a capacidade do modelo de identificar padrões discriminativos consistentes, especialmente considerando o pequeno número de observações e a variabilidade dentro de cada classe.

Por esse motivo, optou-se por não incluir o modelo de classificação da produtividade nos resultados principais. A abordagem por regressão contínua foi considerada mais robusta e informativa para representar a influência das variáveis sobre a produtividade do cacau nos sistemas agroflorestais analisados.

4.3. Modelo de Regressão do Estoque de Carbono

4.3.1. Desempenho do modelo

Este modelo teve como objetivo prever o estoque de carbono acumulado (t C/ha) com base em variáveis ambientais e agronômicas, utilizando a abordagem de Random Forest. O desempenho do modelo foi excelente, com $R^2 = 0,98$, RMSE = 0,624 tC/ha e MAE = 0,468 tC/ha, indicando elevada capacidade preditiva do modelo, explicando aproximadamente 98,7% da variabilidade do estoque de carbono no conjunto de treinamento.

4.3.2. Importância das Variáveis

A variável mais importante para o modelo de regressão do carbono foi age_group.L (faixa etária linear da plantação), com 100% de importância relativa, evidenciando que a idade do sistema agroflorestal é o principal determinante do estoque de carbono. Dois níveis da variável

parameter_name, que representam diferentes tipos de consórcios agroflorestais, também se destacaram com importância entre 41% e 66%, reforçando a influência do arranjo de espécies no acúmulo de biomassa.

A classe de adequação agroecológica (value_class) também teve contribuição moderada (16%). Em contraste, a variável yield (produtividade do cacau) teve importância muito baixa (7%) para explicar o carbono, o que indica uma dissociação entre produtividade e estoque de carbono nos SAFs analisados, e o tipo de solo (soil_typepeat) não apresentou qualquer influência no modelo (0%).

Esses resultados sugerem que o carbono está fortemente associado à idade das plantações e ao tipo de consórcio adotado, mas não diretamente à produtividade agrícola. Isso reforça a interpretação de que o estoque de carbono representa uma característica estrutural e de longo prazo do sistema agroflorestal, mais sensível a fatores temporais e de manejo do que ao desempenho produtivo pontual.

Imagem 2 – Importância relativa das variáveis no modelo de regressão para o estoque de carbono (Random Forest).

	Overall
age_group.L	100.00
parameter_nameCacao (with Coffee)	66.66
age_group^5	50.19
age_group^6	41.92
parameter_nameCacao (with Oil palm)	26.69
age_group^4	24.15
age_group.C	22.93
value_class	16.35
age_group.Q	14.18
yield	7.23
soil_typepeat	0.00

Fonte: Elaborado pela autora no software R, com base na função varImp() aplicada ao modelo Random Forest de regressão para o estoque de carbono (modelo_carbon_rf). 2025.

4.4. Testes Estatísticos Adicionais

4.4.1. Diferenças de carbono entre consórcios (ANOVA + Tukey)

A ANOVA indicou diferença estatisticamente significativa no estoque de carbono entre os tipos de consórcio (p = 0,0001).

O teste de Tukey revelou: cacau com café e com dendê têm estoques significativamente maiores que o sistema com coco e não houve diferença significativa entre café e dendê.

5. Recomendações

O modelo desenvolvido pode contribuir com:

- Políticas públicas de pagamento por serviços ambientais (PSA);
- Programas de incentivo à agricultura de baixo carbono (Plano ABC+);
- Planejamento de manejo agroflorestal, com foco em produtividade e acúmulo de carbono ao longo do tempo;
- Zoneamento agroecológico, considerando a idade dos sistemas e tipo de consórcio.

Recomenda-se o uso do modelo como ferramenta de apoio à decisão, mas com validação adicional em escalas regionais. É importante comunicar que a produtividade não está diretamente associada ao estoque de carbono, o que evidencia *trade-offs* em SAFs maduros.

Além disso, a variável solo mostrou-se não significativa, sugerindo necessidade de maior diversidade amostral para capturar esse fator.

6. Referências Bibliográficas

BREIMAN, L. Random Forests. *Machine Learning*, v. 45, n. 1, p. 5–32, 2001. DOI: https://doi.org/10.1023/A:1010933404324. Acesso em: 15 jun. 2025.

GREENWELL, B. M. et al. *pdp: Partial Dependence Plots*. R package version 0.7.0, 2023. Disponível em: https://cran.r-project.org/package=pdp. Acesso em: 15 jun. 2025.

KONGOR, J. E.; GOCKOWSKI, J.; WESSEL, M.; ASARE, R.; BESSEAU, P.; WEISE, S. Cocoa agroforestry systems for sustainable development: A global review. *Sustainability*, v. 16, n. 2, p. 750, 2024. DOI: https://doi.org/10.3390/su16020750. Acesso em: 15 jun. 2025.

KUHN, M. et al. *caret: Classification and Regression Training*. R package version 6.0-94, 2024. Disponível em: https://cran.r-project.org/package=caret. Acesso em: 15 jun. 2025.

RESTORE+. *Tree Productivity and Carbon in Agroforestry Systems – RESTORE+ Dataset Technical Note*. IIASA, 2023. Disponível em: https://www.restoreplus.org. Acesso em: 15 jun. 2025.