AgriEngineering

MDPI

Article

Smart Coffee: Machine Learning Techniques for Estimating Arabica Coffee Yield

Cleverson Henrique de Freitas [1,*], Rubens Duarte Coelho [1], Jéfferson de Oliveira Costa [2] and Paulo Cesar Sentelhas [1,†]

[1] Biosystems Engineering Department, Luiz de Queiroz College of Agriculture, University of Sao Paulo (USP), Piracicaba 13418-900, SP, Brazil; rdcoelho@usp.br (R.D.C.); pcsentel.esalq@usp.br (P.C.S.)
[2] Experimental Field of Gorutuba, Minas Gerais Agricultural Research Agency (EPAMIG), Nova Porteirinha 39525-000, MG, Brazil; costajo@alumni.usp.br
* Correspondence: chfreitas@alumni.usp.br
† In memoriam.

Abstract: Coffee is a global commodity, with Brazil being a major producer, particularly in the Minas Gerais state. This study applied machine learning to predict the Arabica coffee yield in the region, analyzing two groups of cultivars (G1 and G2) using data from 1993 to 2020. The Factor Analysis of Mixed Data (FAMD) was employed to explore the relationships between climatic factors, management practices, and the coffee yield. Four machine learning models, such as Multiple Linear Regression (MLR), Random Forest (RF), XGBoost (XGB), and Support Vector Machines (SVM) were calibrated and evaluated for yield prediction. The FAMD revealed complex interactions among variables, requiring four principal components to explain approximately 64.6% of the total variance. Management practices, such as the planting density and pruning, had a stronger influence on G1 cultivars, while G2 cultivars were more sensitive to climatic conditions, particularly the air temperature. Among the machine learning models, RF and XGB performed best in the yield estimation, whereas MLR and SVM were less effective, particularly for values above 60 bags ha^{-1} (1 bag = 60 kg). These findings underscore the variability in the yield across cultivars and demonstrate the potential of machine learning to guide tailored management strategies for different coffee cultivars.

Keywords: *Coffea arabica*; exploratory analysis; productivity estimation; agricultural management; climatic conditions; crop modeling

Citation: Freitas, C.H.d.; Coelho, R.D.; Costa, J.d.O.; Sentelhas, P.C. Smart Coffee: Machine Learning Techniques for Estimating Arabica Coffee Yield. *AgriEngineering* **2024**, *6*, 4925–4942. https://doi.org/10.3390/agriengineering6040281

Academic Editors: Sotirios K. Goudos, Shaohua Wan and Achilles Boursianis

Received: 14 November 2024
Revised: 10 December 2024
Accepted: 17 December 2024
Published: 20 December 2024

1. Introduction

Coffee stands as one of the world's most important commodities, holding significant economic and cultural value [1]. Its influence transcends daily consumption, serving as a livelihood for millions of families, particularly in developing nations like Brazil, renowned as the foremost producer and exporter worldwide [1,2].

Statistical projections for the 2023/24 harvest pinpoint Brazil as the chief contributor, responsible for approximately 31.4% of the global coffee production. This is closely followed by Vietnam (18.0%) and Colombia (6.6%), encompassing both the *Coffea arabica* and *Coffea canephora* species [3]. Notably, within Brazil, the Minas Gerais state takes the lead as a primary coffee producer, representing a substantial share of the country's total production, contributing 43.1% in the 2022/23 season [3].

The coffee yield is intricately linked to a complex combination of environmental factors like temperature, precipitation, and solar radiation, alongside various management techniques such as irrigation, pruning, planting density, fertilization, and pest and disease control [4–8]. Particularly, meteorological factors like temperature and precipitation significantly impact all developmental stages of coffee, be it vegetative or reproductive [4,9–11].

Accurate estimation of the coffee yield holds paramount importance for various stakeholders including farmers, researchers, and policymakers, facilitating planning and

decision making within coffee cultivation [12]. This necessitates the utilization of diverse simulation models to anticipate and forecast the coffee yield across different temporal and spatial scales, including mechanistic or biophysical models [13–16], physiological–mathematical models [17–20], statistical models [11,12,21–25], as well as those incorporating remote-sensing data [26–29].

While biophysical models offer a comprehensive analysis about physical and biological factors shaping crop growth and yield, their intricate nature and requisite parameterization often pose challenges, especially in data-limited environments [18,20]. With technological advancements, more sophisticated methods for estimating the yield are emerging, such as machine learning (ML) techniques, serving as pivotal tools for enhancing the coffee production chain's development, implementation, and management [26,30–32].

These models are adept at processing large datasets, identifying complex patterns, capturing non-linear relationships, and dynamically adapting to new information [23,26,30,32–35]. Moreover, recent studies underscore the growing relevance of integrating ML in addressing challenges in the coffee yield estimation [12,21,26,28,36], quality assessment [32,37,38], and disease diagnosis [39–41]. By integrating meteorological, agronomic, and remote-sensing data, ML enhances the precision and scalability in yield forecasting.

Thus, this study introduces the concept of "Smart Coffee", a metaphor for applying intelligent technologies to optimize the coffee yield prediction and management within the broader framework of smart agriculture [35]. By applying four ML techniques, such as Multiple Linear Regression (MLR), Random Forest (RF), XGBoost (XGB), and Support Vector Machines (SVM), this research aims to estimate the yield of Arabica coffee cultivars across three key locations in Minas Gerais. Unlike conventional approaches, this study seeks to integrate meteorological and management variables, addressing key gaps in the current methodologies and contributing to the advancement of precision agriculture for coffee production.

2. Materials and Methods

2.1. Study Area and Arabica Coffee Data

This study analyzed coffee yield data collected from 1993 to 2020 and provided by partner companies involved in the project. The data encompass plantings established between 1970 and 2016 for nine cultivars: Acaiá (ACA), Bourbon (BRB), Catuaí (CTI), Catucaí (CTC), Icatú (ICT), Mundo Novo (MNV), Rubi (RUB), Topázio (TPZ), and the gene bank (BNL). These records represent three diverse locations within Minas Gerais, Brazil, the country's leading coffee-producing state (Table 1).

Table 1. Coffee-producing locations and their geographical and climatic characteristics.

Location	State	Latitude	Longitude	Altitude	Mean Temperature [1]	Mean Accumulated Precipitation [1]
		(Degree)	(Degree)	(m)	(°C)	(mm)
Alfenas	MG	−21.54	−45.93	886.3	21.1	1535.0
Alfenas	MG	−21.30	−45.93	793.0	21.2	1442.3
Conceição do Rio Verde	MG	−21.90	−45.21	929.0	20.5	1511.7

[1] Mean meteorological data considering the period from January 1991 to December 2020.

According to the Köppen climate classification system, the local climate was categorized as Cwb, denoting a humid subtropical climate with temperate summers [42]. This classification indicates favorable climatic conditions conducive for the growth of Arabica coffee, characterized by temperatures ranging between 20.5 °C and 21.2 °C, annual rainfall exceeding 1400 mm, and elevations surpassing 700 m above sea level [43]. The soil type of the studied locations ranges from *Latossolo Vermelho distrófico* (Oxisols) to *Cambissolo Háplico Tb distrófico* (Inceptisols) [44].

Variables concerning coffee trees were considered, including the mean yield (Ymean, 60 kg bags ha^{-1}), mean yield under irrigation (Yirr, 60 kg bags ha^{-1}) and rainfed conditions (Yrain, 60 kg bags ha^{-1}), plant population (Pop, plants ha^{-1}), mean age (Age, years), and area per plant (Apl, m^2 plants), which is calculated by multiplying the spacing between rows and between plants. The yield data were stratified into years of low and high yields, mirroring the biennial nature common in coffee crops [45]. Factors such as irrigation usage (irrigated or rainfed) and pruning practices (yes or no) were also taken into consideration in the analytical models.

Given the extensive dataset comprising approximately 3079 yield observations and the variability stemming from management practices influencing coffee cultivation, a cluster analysis was conducted to explore the interplay between different Arabica coffee cultivars based on these management variables. Subsequently, two distinct cultivar groups were identified and utilized for the calibration and evaluation of the yield estimation models. The mean values for each variable within the respective cultivar groups are presented in Table 2. Group 1 (G1) exhibited higher yield levels and plant densities, while Group 2 (G2) demonstrated wider spacing (>Apl) between plants and older planting ages.

Table 2. Mean yield (Ymean, 60 kg bags ha^{-1}), irrigated yield (Yirr, 60 kg bags ha^{-1}) and rainfed yield (Yrain, 60 kg bags ha^{-1}), plant population (Pop plants ha^{-1}), mean age (Age, years), and area per plant (Apl, m^2 pl) for the two Arabica coffee cultivar-formed groups.

Group	Cultivars	Ymean	Yirr	Yrain	Pop	Age	Apl
		60 kg Bags ha^{-1}			Plants ha^{-1}	Years	m^2 pl
G1	ACA, BRB, BNL, CTC, RUB and TPZ	44.4	50.5	38.3	4642.6	10.3	2.2
G2	CTI, ICT and MNV	40.1	41.8	38.4	3550.2	16.1	3.1

2.2. Agrometeorological Variables

Daily meteorological data were used, including the minimum (Tmin, °C), mean (Tmean, °C), and maximum (Tmax, °C) air temperature, precipitation (Prec, mm), wind speed at 2 m (U2, m s^{-1}), relative humidity (RH, %), and global solar radiation (Qg, MJ m^{-2} day^{-1}), which were subsequently transformed to a 10-day scale, provided by partner companies to the project. Missing data were filled in using the databases from the Brazilian National Institute of Meteorology (INMET), the National Water and Sanitation Agency (ANA), and the BR-DWGD data (Brazilian Daily Weather Gridded Data) described by Xavier et al. [46].

The use of gridded data for filling gaps is well-supported by several studies demonstrating their reliability [47–51]. Specifically, the BR-DWGD dataset has shown high accuracy for meteorological variables across Brazil, performing comparably to observed weather data. It has proven effective for gap-filling, crop simulation models, and climate change projections [47–51].

Variables from the water balance were also considered, such as an accumulated water deficit (WDac, mm), accumulated water surplus (SURac, mm), relative evapotranspiration (rET = ETR/ETP, dimensionless), and the number of days with storage of less than 50% (NDSto50), considering the phenological phase of the coffee plant between the beginning of flowering and physiological maturation. Potential evapotranspiration (ETP, mm) was estimated using the method proposed by Thornthwaite [52], being a suitable method for estimating evapotranspiration in tropical regions, and the other agrometeorological variables were calculated through the water balance, considering a water holding capacity of 100 mm [53,54].

The beginning of flowering was calculated using the model proposed by Zacharias et al. [55], which was later corrected using observed data on flowering dates. Thus, the beginning of flowering corresponded to the first 10-day period in which there was an accumulation of 1740 degree days (Tb = 10 °C) after a rainfall exceeding 7 mm, counted from

the first 10-day period of April each year. The physiological maturity was determined after an accumulation of 3000 degree days (Tb = 10 °C) following the beginning of flowering [19].

2.3. Exploratory Analysis of Arabica Coffee Yield

To understand how the Arabica coffee yield relates to agrometeorological and management variables, a Factorial Analysis of Mixed Data (FAMD) was conducted. This allows for a combined analysis of data containing quantitative variables (yield, plant age, agrometeorological variables, etc.) and qualitative variables (use of irrigation, pruning, bienniality, etc.) [56]. The analysis was carried out using the "FactoMineR" and "factoextra" packages in R language [57]. FAMD is a technique that combines elements of Principal Component Analysis (PCA) and Multiple Correspondence Analysis (MCA), effectively handling both quantitative and qualitative variables by normalizing them during the analysis to ensure both have an equal influence on the outcome [56].

To facilitate the interpretation of the results, a biplot graph was generated, which allows a clear visualization of the most significant principal components. This graph is useful for understanding the similarity among individuals based on a mixed set of variables and for exploring the associations between all variables involved in the study.

2.4. Agricultural Modeling of Arabica Coffee Yield

Based on the two cultivars groups, four different machine learning techniques were applied, including Multiple Linear Regression (MLR), Random Forest (RF), eXtreme Gradient Boosting or XGBoost (XGB), and a Support Vector Machine (SVM), for estimating the coffee yield based on agrometeorological conditions, crop characteristics such as age and bienniality, and management practices such as irrigation, spacing, and pruning.

Multiple Linear Regression (MLR, Equation (1)) is characterized by a linear relationship between a dependent variable (Y) and multiple independent variables (X_1, X_2, …, X_n) [58].

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon \quad (1)$$

where β represents the coefficients and ϵ is the error term.

This technique is often applied when the predictors have a straightforward, additive influence on the response variable. However, it is sensitive to multicollinearity, which can distort the interpretation of coefficients, and assumes independence between predictors, limiting its use in datasets with highly correlated variables [58].

Random Forest (RF) is an ensemble learning method that constructs multiple decision trees during training and aggregates their predictions for improved accuracy [59]. Each tree is trained on a bootstrapped sample of the data, and at each node, a random subset of predictors is chosen to split the data, which reduces overfitting and increases the model diversity [59]. RF is robust to outliers, multicollinearity, and irrelevant variables, making it particularly suitable for complex datasets. On the other hand, the method is computationally expensive and can lack interpretability due to the large number of trees in the ensemble.

Instead, eXtreme Gradient Boosting (XGB) is a powerful boosting algorithm that iteratively combines weak learners, typically decision trees, to minimize prediction errors [60]. Unlike traditional boosting, XGB incorporates regularization techniques to avoid overfitting and uses the stochastic sampling of data and predictors at each iteration to enhance generalization [60]. It is able to process the high-dimensional data and model complex, nonlinear associations, but is computationally expensive and demands the cautious tuning of hyperparameters.

Finally, a Support Vector Machine (SVM), developed by Boser et al. [61], classifies data by identifying hyperplanes that optimally separate classes in a multidimensional feature space. For nonlinear problems, a SVM leverages kernel functions such as a linear, polynomial, and radial basis function (RBF), transforming the data into higher-dimensional spaces where a hyperplane can effectively separate classes [61]. Although SVM is a powerful tool

for classification, it requires the proper scaling of data and is less effective when applied to very large datasets due to computational limitations.

For the application of machine learning techniques to the two groups of Arabica coffee cultivars, during the testing phase (about 80% of the original data), the models' hyperparameters were calibrated using cross-validation and the Caret package (Classification and Regression Training) available in R software, a technique that involves dividing the training dataset into k equally sized subsets, such that one of these is used for testing and the rest for training [62]. To address the temporal aspect of the data, the biennial variable (Bien) was included as a predictor. This variable captures the characteristic biennial production cycle of coffee, effectively controlling temporal variations in the yield across years [45,63].

Before modeling, outliers were identified and addressed through a visual inspection of the data. Furthermore, predictors were normalized to ensure that variables with differing scales contributed equally to the model's performance. Multicollinearity among predictor variables was also evaluated, where it was pointed out that the variables minimum air temperature (Tmin), accumulated water surplus (SURac), and relative evapotranspiration (rET) had a high degree of collinearity, high correlation with other predictor variables, and were then removed from the analyses.

Both in the calibration phase and the evaluation phase, the performance of the models was verified through the following statistical indices: mean absolute error (MAE), root mean square error (RMSE), percent BIAS index (PBIAS%), correlation coefficient (r), coefficient of determination (R^2), index of agreement (d) [64], performance index (c) [49], and modeling efficiency index (NSE) [65]. The performance index (c) is obtained by the product between the coefficient (r) and (d) and interpreted as: "excellent" (c > 0.85); "very good" (c between 0.76 and 0.85); "good" (c between 0.66 and 0.75); "medium" (c between 0.61 and 0.65); "weak" (c between 0.51 and 0.60); "very weak" (c between 0.41 and 0.50); and "poor" (c < 0.40).

3. Results and Discussion

3.1. Exploratory Analysis Results

The Factorial Analysis of Mixed Data (FAMD) and the contribution of each variable per principal component are presented in Figure 1 and Table 3. For this study, four principal components were necessary to explain most of the total variance of the data. These components presented eigenvalues equal to 4.8, 3.0, 2.4, and 2.1 and their respective percentages of variance explanation, being 25.2%, 16.0%, 12.4%, and 11.0%. Considering the four components, about 64.6% of the total variance of the data is explained. Moreover, the number of necessary principal components (four) highlights the complexity of the relationships between the studied variables.

When considering each principal component separately, the first component (PC1) seems to capture the variation related to the water balance (WDac, SURac, rET, and NDSto50) and environmental conditions, suggesting that factors such as precipitation (Prec) and relative humidity (RH) are crucial for differentiating between the analyzed groups. Additionally, the accumulated water deficit (WDac) has the largest contribution to the first component (PC1) with 17.08% (Table 3). This suggests that this variable is one of the main factors differentiating the groups in their analysis along the first component.

Meanwhile, the second component (PC2) stands out for its emphasis on management variables (Apl and Population) and temperature (Tmean and Tmin). The mean temperature (Tmean) is the most significant variable, contributing 18.61% (Table 3). This indicates that management practices related to the planting density and plant population are important factors in this component, as well as the thermal conditions of the locations.

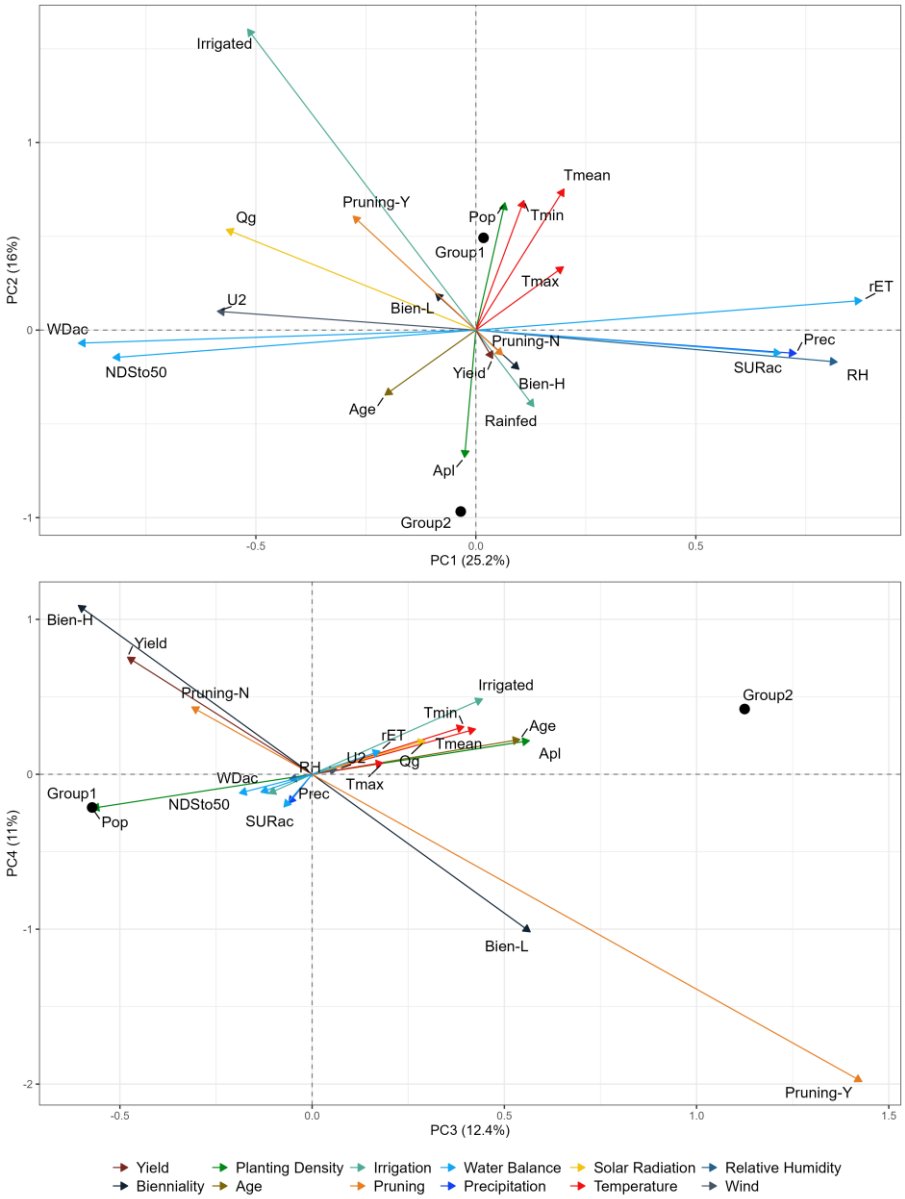

Figure 1. Factorial Analysis of Mixed Data (FAMD) of Arabica coffee yield (YIELD, 60 kg bags ha^{-1}) based on agrometeorological variables and management techniques in Minas Gerais, Brazil. Variables: Group 1 (ACA, BNL, BRB, CTC, RUB, and TPZ), Group 2 (CTI, ICT, and MNV), bienniality (Bien-H and Bien-L), planting density (area per plant: Apl—m^2 pl and plant population: Pop—plants ha^{-1}), mean age (Age, years), irrigation (Irrigated and Rainfed), pruning (Pruning-Y and Pruning-N), water balance (accumulated water deficit: WDac—mm; accumulated water surplus: SURac—mm; relative evapotranspiration: rET—dimensionless; and number of days with STO < 50%: NDSto50—days), precipitation (Prec, mm), solar radiation (Qg, MJ m^{-2} 10-day^{-1}), temperature (maximum air temperature: Tmax—°C; mean air temperature: Tmean—°C; and minimum air temperature: Tmin—°C), relative humidity (RH, %), and wind speed at 2m (U2, m s^{-1}).

Table 3. Contribution of each variable by principal component in the Factorial Analysis of Mixed Data (FAMD).

Variables	Acronym (Unit)	Contribution of Variables by Principal Component (%)			
		PC1 25.2%	**PC2 16.0%**	**PC3 12.4%**	**PC4 11.0%**
Group	G1 or G2	0.00	5.14	11.57	2.06
Yield	Yield (bags ha^{-1})	0.03	0.75	9.78	27.29
Bienniality	Bien (H or L)	0.04	0.44	6.20	25.29
Area per plant	Apl (m^2 plants)	0.01	15.13	13.55	2.25
Population	Pop (plants ha^{-1})	0.09	15.17	13.88	2.30
Age	Age (years)	0.89	3.94	12.39	2.41
Irrigation	Irrigated or Rainfed	0.30	7.07	0.90	1.37
Pruning	Pruning (Y or N)	0.07	0.87	8.05	19.65
Accumulated water deficit	WDac (mm)	17.08	0.16	0.75	0.63
Accumulated water surplus	SURac (mm)	10.05	0.49	0.23	2.12
Relative evapotranspiration	rET (--)	16.11	0.82	1.32	1.07
Number of days with Storage < 50%	NDSto50 (days)	14.24	0.71	1.52	0.71
Precipitation	Prec (mm)	11.10	0.51	0.16	1.67
Maximum temperature	Tmax (°C)	0.83	3.74	1.44	0.26
Mean temperature	Tmean (°C)	0.84	18.61	7.66	4.06
Minimum temperature	Tmin (°C)	0.25	15.70	6.61	4.52
Global solar radiation	Qg (MJ m^{-2} 10-day^{-1})	6.73	9.46	3.67	2.24
Relative air humidity	RH (%)	14.11	0.95	0.15	0.06
Wind speed at 2m	U2 (m s^{-1})	7.22	0.33	0.18	0.05

The third component (PC3) is notable for its diversity of influential variables, including those related to the group and management, such as the planting density (Apl and Pop) and the age of the plants, as well as the yield, which contributes 9.78% (Table 3). This component suggests that the yield is affected by a combination of management factors and genetic factors, given the presence of 11.57% from the Groups (Group 1 and Group 2).

The fourth component (PC4), in turn, is dominated by the Yield, contributing 27.29%, followed by bienniality (25.29%) and management practices such as pruning, which contributes 19.65% (Table 3). This component indicates that the yield is a variable strongly affected by specific management practices and intrinsic factors of the crop, such as the bienniality [45,63,66].

When considering these four components together, a picture of significant complexity emerges. The first two components seem to focus more on climatic and management variables, while components 3 and 4 reveal the importance of the yield and specific management practices. Meanwhile, variables such as irrigation, the maximum temperature, global solar radiation, and wind speed had a low percentage of contribution in the four components, suggesting that these variables may have complex and interactive effects that are not easily captured by a single metric or component (Table 3). It should be emphasized that deficit irrigation is commonly practiced in these regions, as water availability is the most important limiting factor. Consequently, efforts are made to irrigate the largest possible area by applying an irrigation depth below the optimal requirement for coffee plants.

3.2. Model Calibration and Importance of Predictor Variables

The calibrated hyperparameter values for each technique and each cultivar group are presented in Table 4. The components of the Factorial Analysis of Mixed Data (FAMD) and the differences in hyperparameters demonstrate that the two groups have different levels of yield and are influenced differently by climatic and management variables.

Table 4. Calibration of machine learning technique parameters used for estimating coffee yield for the two cultivar groups in Minas Gerais, Brazil.

Technique	Method	Hyperparameters	Range	Group 1		Group 2	
				Final Values [1]		Final Values [1]	
Multiple Linear Regression	--	b0	--	64.047	.	15.190	
		b1—Age		−0.0439		−0.0124	
		b2—Pop		0.0015		−0.0015	
		b3—Apl		3.0162	*	−1.8851	.
		b4—Irrigation_NI		−5.4107	***	1.4111	
		b5—Prn_Y		−25.493	***	−22.566	***
		b6—BienL		−24.244	***	−23.290	***
		b7—Tmax		−0.1608		0.8149	.
		b8—Tmean		−0.5570		−2.2604	**
		b9—Prec		0.0011		−0.0039	.
		b10—U2		2.9692		3.1374	
		b11—RH		0.0380		0.6643	
		b12—Qg		−0.6223		1.3001	
		b13—NDSto50		0.0180		0.0159	
		b14—WDac		−0.0155		−0.0279	
Random Forest	--	maxnodes	10–1000	500		200	
		ntree	500–1000	1000		1000	
		nodesize	15–25	15		25	
		mtry	2–8	4		2	
XGBoosting	--	nrounds	100–10,000	1900		200	
		max_depth	2–6	4		2	
		eta	0.01–0.3	0.015		0.1	
		gamma	0–1	0.05		0	
		colsample_bytree	0.4–1	0.6		0.6	
		min_child_weight	1–3	1		1	
		subsample	0.5–1	0.75		0.75	
Support Vector Machine	Linear	C	0–2	0.1053		1.5789	
	Polynomial	degree	1–3	3		3	
		scale	0.001–1	0.1		0.1	
		C	0.25–2	0.5		0.25	
	Radial [2]	C	0.25–128	8		2	
		sigma	Maximized	0.0563		0.0602	

[1] Significance of coefficients: <0.10% (***), <10% (**), <50% (*), and <10% (.). [2] Radial is the best method for the SVM technique.

Considering multiple linear regression (MLR), Group 1, in addition to having a higher yield, was shown to be more sensitive to management factors such as irrigation and pruning, while for Group 2, some agrometeorological variables also need to be considered, such as the temperature and precipitation. Rainfed conditions (IrrigationNI), pruning (PrnY), and years of low yield (BienL) negatively impacted the yield in both groups, being more pronounced in G1. Moreover, the planting density (Apl) and plant population (Pop) have opposite impacts on the two groups. While XGBoosting (XGB) suggests that a more complex model may be necessary for Group 1, with more rounds and greater depth, the Random Forest (RF) and Support Vector Machine (SVM) reveal that the complexity and fitting of the models vary between groups.

After the calibration and training of the models, it was possible to identify the importance of each predictor variable, with the top ten shown in Figure 2, for the four machine learning techniques used in this study. For both groups, the bienniality (Bien) and pruning (Prn) variables had high percentages compared to other predictor variables. For Group 1, the third most important predictor variable differed among the techniques, being irrigation for MLR, the mean age for RF, the plant population for XGB, and the mean temperature for

SVM. For Group 2, except for XGB, which indicated the plant population as the third most important variable, the other techniques pointed to the mean air temperature as the third most important predictor variable. Moreover, for both groups, the RF technique showed a greater balance in the importance percentage of the variables in comparison to the other techniques, with values greater than 25% for the top ten variables.

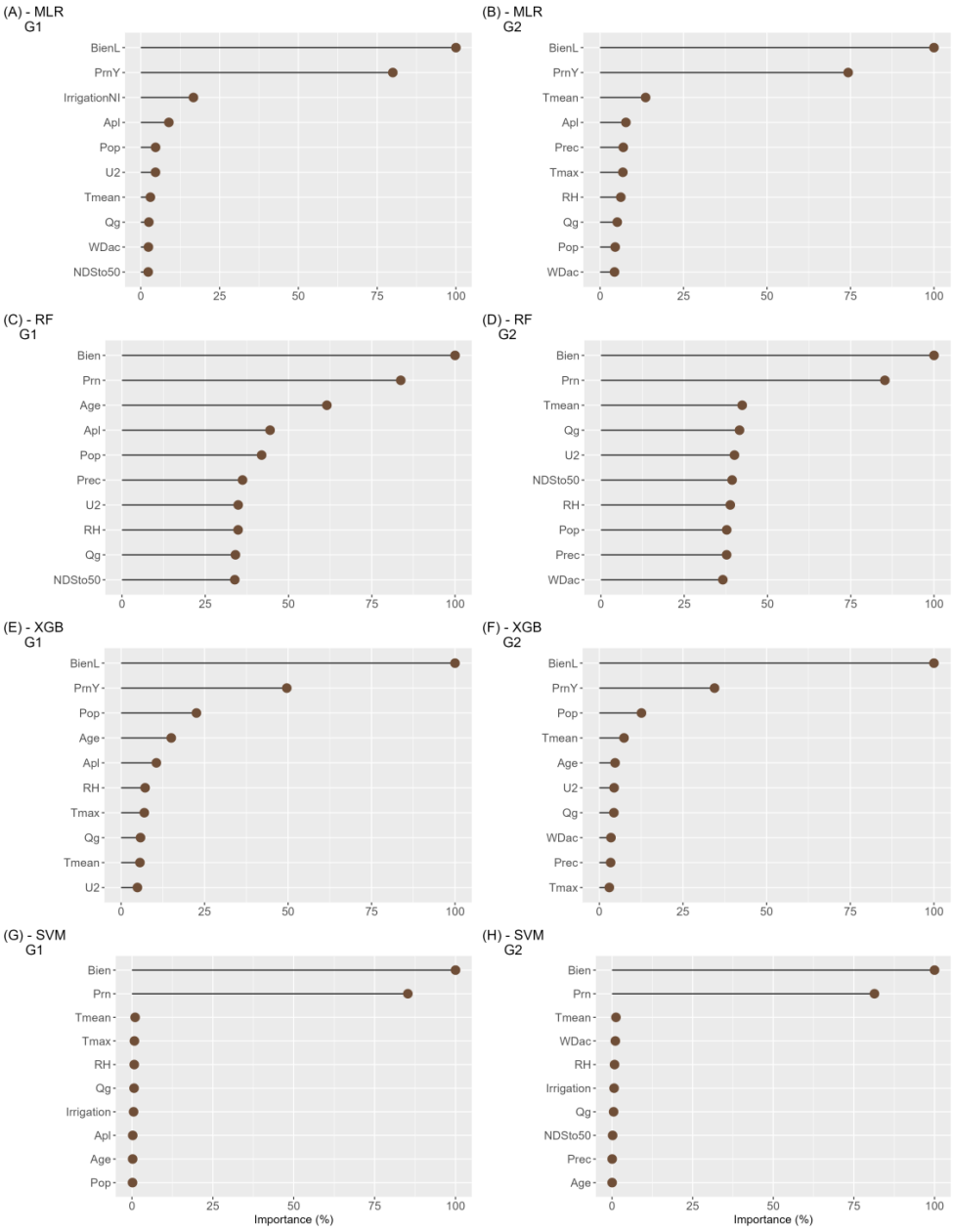

Figure 2. Classification of the top ten determinant variables of Arabica coffee yield for G1 (G1, ACA, BNL, BRB, CTC, RUB, and TPZ) and G2 (CTI, ICT, and MNV), using the method of Multiple Linear Regression (MLR, (**A**,**B**)), Random Forest (RF, (**C**,**D**)), XGBoosting (XGB, (**E**,**F**)), and Support Vector Machine (SVM, (**G**,**H**)). Variables include area per plant (Apl, m^2 pl), bienniality (high and low), irrigation (irrigated and rainfed), pruning (yes or no), plant population (Pop, plants ha^{-1}), mean age (Age, years), accumulated water deficit (WDac, mm), number of days with STO < 50% (NDSto50, days), precipitation (Prec, mm), wind speed (U2, m s^{-1}), maximum air temperature (Tmax, °C), mean air temperature (Tmean, °C), global solar radiation (Qg, MJ m^{-2} 10-day^{-1}), and relative humidity (RH, %).

These results indicate that, for both groups, the "Bien" and "Prn" variables stand out in terms of importance, suggesting that these management factors are critical for the coffee yield, regardless of the group. Additionally, the variation in the third most important predictor variable among the different techniques for Group 1 suggests that this group's sensitivity to different factors can vary depending on the model used. For Group 2, the consistency in identifying the "mean air temperature" as an important variable (except in the case of XGB) reinforces the idea that agrometeorological factors are especially relevant for this group. These observations can be extremely useful for the development of more effective and customized management strategies for each group of coffee cultivars.

The importance of agrometeorological variables in estimating coffee yields, such as the temperature, water deficit, and precipitation, in addition to information about cultivation, such as bienniality and management techniques, has also been mentioned in various studies with different simulation models [11,12,18,21,24–26].

Valeriano et al. [18], in addition to climatic conditions, showed the importance of the geographical location in estimating the coffee yield using the physiological–mathematical model described by Santos and Camargo [17]. This study also suggested that gridded data could be a viable alternative for yield estimation, which could be an interesting direction for future research, in addition to the use of physiological or dynamic models.

In Figure 3, Venn diagrams are presented to graphically represent the agreement of the top ten predictor variables for each of the four machine learning techniques. For Group 1 (Figure 3A), all techniques identified bienniality, pruning, the area per plant (Apl), the plant population (Pop), and global solar radiation (Qg) as the main variables in estimating the yield of these coffee cultivars, indicating once again the importance of management for this group. On the other hand, for Group 2 (Figure 3B), in addition to bienniality and pruning, the mean air temperature (Tmean), precipitation (Prec), global solar radiation (Qg), and accumulated water deficit (WDac) were the main variables for all techniques. An interesting factor is that for Group 1, except for Qg, the models indicated biennially (Bien) and management techniques (pruning, spacing, and plant population) as the main predictor variables, while for Group 2, environmental variables (mean temperature, rainfall, radiation, and water deficit) were highlighted as more important, also including bienniality and pruning, showing that, for Group 2, both management and environmental conditions are critical factors in determining the yield.

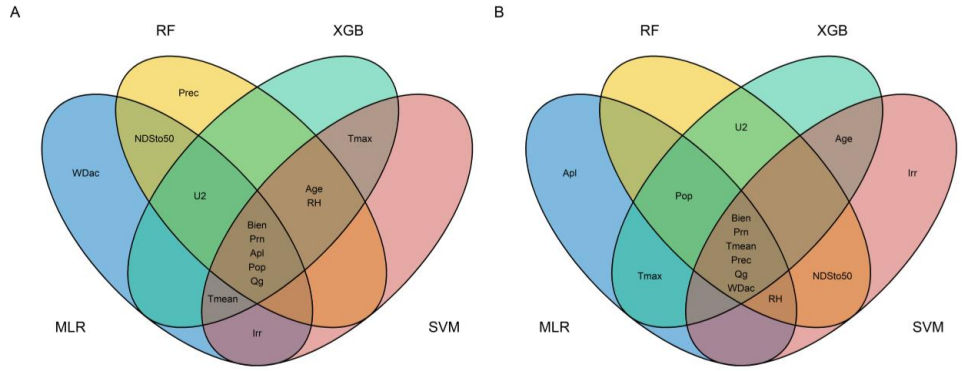

Figure 3. Venn diagram to indicate the agreement among four machine learning techniques (Blue: Multiple Linear Regression—MLR; Red: Support Vector Machine—SVM; Green: XGBoosting; Yellow: Random Forest—RF) in identifying the main predictor variables of Arabica coffee yield for Group 1 ((**A**)—ACA, BNL, BRB, CTC, RUB, and TPZ) and Group 2 ((**B**)—CTI, ICT, and MNV). Variables include area per plant (Apl, m^2 pl), bienniality (high and low), irrigation (irrigated and rainfed), pruning (yes or no), plant population (Pop, plants ha^{-1}), mean age (Age, years), accumulated water deficit (WDac, mm), number of days with STO < 50% (NDSto50, days), precipitation (Prec, mm), wind speed (U2, m s^{-1}), maximum air temperature (Tmax, °C), mean air temperature (Tmean, °C), global solar radiation (Qg, MJ m^{-2} 10-day^{-1}), and relative humidity (RH, %).

3.3. Estimation of Arabica Coffee Yield

The performance of the models in estimating the coffee yield cultivars of Group 1 is presented in Figure 4, for the four machine learning techniques. During the training phase, 1632 data points were used, representing about 80% of the total data, these being 2041, while in the model-testing phase, 409 independent data points were used. In general, all techniques underestimated the coffee yield values above 60.0 bags ha^{-1} (1 bag = 60 kg), resulting in negative PBIAS values.

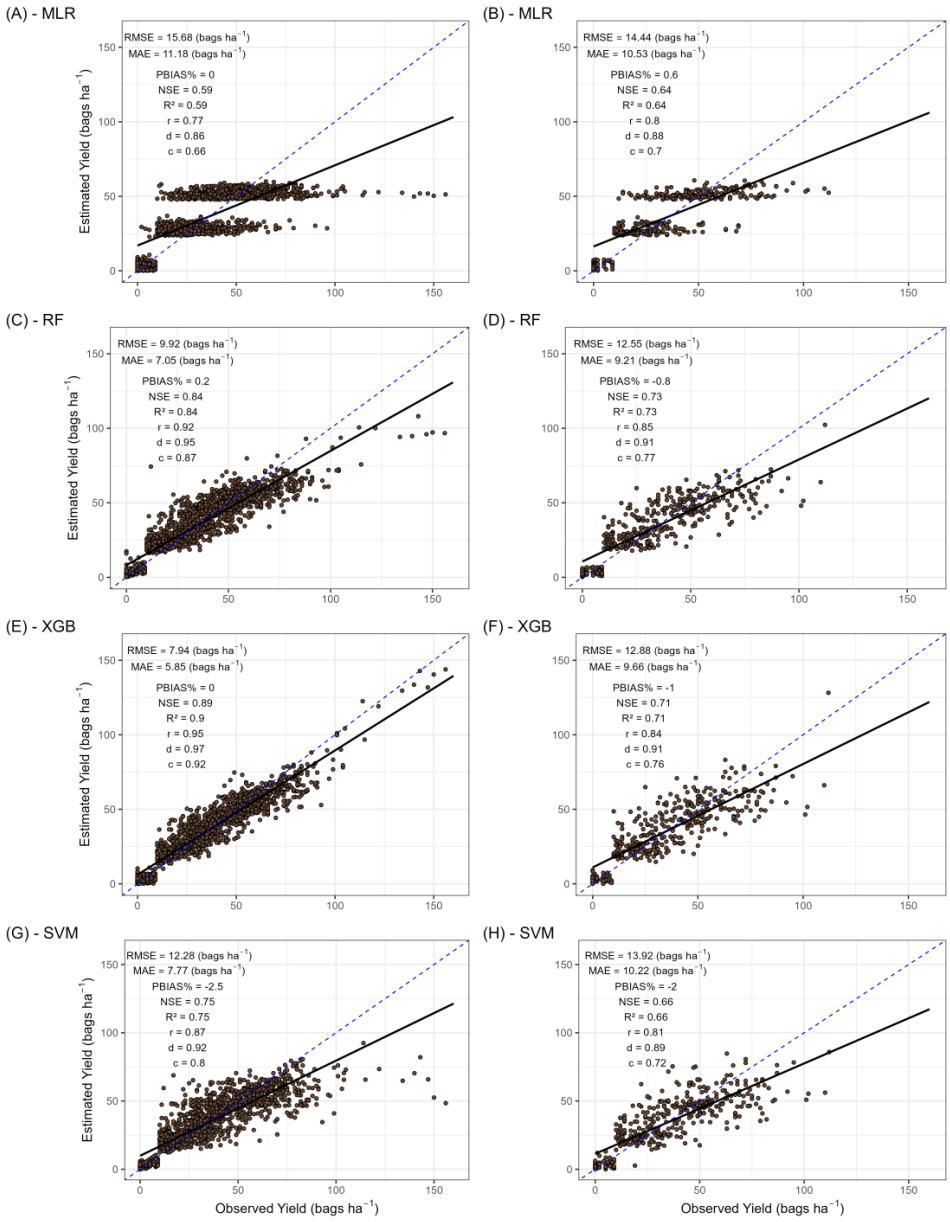

Figure 4. Observed and estimated Arabica coffee yield for the cultivar group G1 (ACA, BNL, BRB, CTC, RUB, and TPZ) in the training phase (**A**,**C**,**E**,**G**) and testing phase (**B**,**D**,**F**,**H**) using the method of Multiple Linear Regression (MLR, (**A**,**B**)), Random Forest (RF, (**C**,**D**)), XGBoosting (XGB, (**E**,**F**)), and Support Vector Machine (SVM, (**G**,**H**)). Statistical indices include Root Mean Square Error (RMSE, 60 kg bags ha^{-1}), Mean Absolute Error (MAE, 60 kg bags ha^{-1}), Percent Bias (PBIAS, %), Nash–Sutcliffe Efficiency Index (NSE), Coefficient of Determination (R^2), Correlation Coefficient (r), Index of Agreement (d), and Performance Index (c). The dashed blue line indicates the 1:1 line.

The MLR model during the training and testing phases was the least performing model compared to the others, with an R^2 accuracy of 0.59 and RMSE of 15.68 60 kg bags

ha^{-1} in training, and R^2 of 0.64 and RMSE of 14.44 60 kg bags ha^{-1} in the testing phase, being classified as "good" according to the c index for both phases. Furthermore, the MLR model showed stratification in the yield estimation, with the estimated values restricted to the ranges 0–10, 20–40, and 40–60 bags ha^{-1} (1 bag = 60 kg) (Figure 4A,B).

The RF (Figure 4C,D) and XGB (Figure 4E,F) techniques showed the best performances in estimating the coffee yield for Group 1, with both classified as "excellent" (RF: c = 0.87 and XGB: c = 0.92) during the training phase and "very good" (RF: c = 0.77 and XGB: c = 0.76) during the testing phase. They also showed high accuracy (RF: R^2 = 0.84 and XGB: R^2 = 0.90) and low error values (RF: RMSE = 9.92 60 kg bags ha^{-1} and XGB: RMSE = 7.94 60 kg bags ha^{-1}), especially during the training phase, with a highlight for XGB. In the testing phase, the RF model was slightly superior to XGB, with an RMSE of 12.55 60 kg bags ha^{-1} and an accuracy of 0.73, compared to XGB, which showed an R^2 of 0.71 and RMSE of 12.88 60 kg bags ha^{-1}.

Regarding the performance of the other models, the SVM technique was the second lowest performing (Figure 4G,H), being classified as "very good" (c = 0.80) in the training phase and "good" (c = 0.72) in the testing phase with an R^2 accuracy of 0.75 and 0.66, and errors ranging between 12.28 60 kg bags ha^{-1} and 13.92 60 kg bags ha^{-1} during the training and testing phases, respectively.

In Figure 5, the performances of the models in estimating the coffee yield cultivars of Group 2 are presented for the four machine learning techniques. During the training phase, 830 data points were used, representing about 80% of the total data, these being 1038, while in the model testing phase, 208 independent data points were used. Generally, all techniques, as with Group 1, underestimated the yield values above 60 bags ha^{-1} (1 bag = 60 kg). Moreover, the yield estimation models for Group 2 were slightly inferior to those calibrated for Group 1.

As with Group 1, for Group 2, the MLR model showed an inferior performance, also stratifying yield estimates, both in the training phase (R^2 = 0.64 and RMSE = 13.33 60 kg bags ha^{-1}) and in the testing phase (R^2 = 0.61 and RMSE = 13.94 60 kg bags ha^{-1}) and being classified according to the c index as "good" (Figure 5A,B).

The other techniques were, in the training phase, classified as "very good," with the accuracy varying between R^2 = 0.77 and 0.81 for the SVM (Figure 5G) and RF (Figure 5C) models, respectively, and errors between RMSE = 10.09 bags ha^{-1} and 10.68 bags ha^{-1} for the XGB (Figure 5E) and SVM models, respectively. In the testing phase, all models were classified as "good," with the c index being between 0.71 and 0.73, for the SVM (Figure 5H) and XGB (Figure 5F) models, respectively, with accuracy between 0.65 (SVM) and 0.68 (RF, Figure 5D) and errors between 12.72 bags ha^{-1} (XGB) and 13.13 bags ha^{-1} (SVM). Even though the RF model (Figure 5C,D) was slightly more accurate in both phases, that is, with less dispersion (higher R^2), the XGB model was the most accurate, with the lowest errors in both the training and testing phases (Figure 5E,F).

Regarding the coffee yield estimation using ML techniques, the MLR model demonstrated the weakest performance compared to the others, both in the training and testing phases for the two groups. This could be due to the model's simplicity, which may not be capable of capturing the data's complexity. Similarly, while SVM performed better than MLR, it was still inferior by tree-based models, possibly due to its sensitivity to hyperparameters and the need for fine-tuning.

In contrast, RF and XGB achieved the best performances, with XGB slightly outperforming RF in the training phase, but being slightly inferior in the testing phase. This performance trend could be explained by the algorithmic concepts of RF and XGB, both of which are tree-based methods [67]. These algorithms perform better when the dataset's features exhibit pronounced patterns, as in our study. Furthermore, such characteristics are effectively captured by tree-based models, enabling RF and XGB to achieve comparable levels of accuracy and corroborate the effectiveness of more sophisticated methods compared to MLR [67].

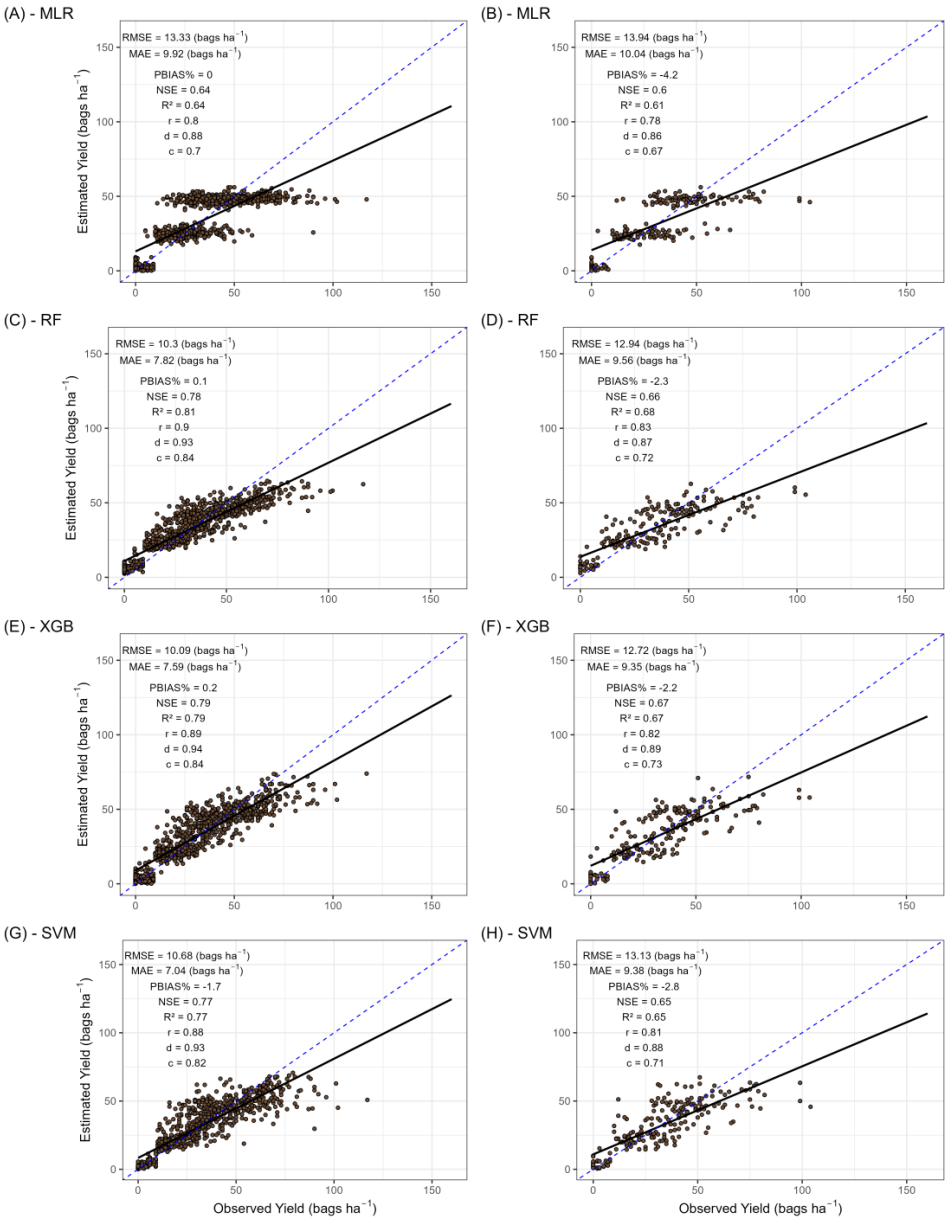

Figure 5. Observed and estimated Arabica coffee yield for the cultivar group G2 (CTI, ICT, and MNV) in the training phase (**A**,**C**,**E**,**G**) and testing phase (**B**,**D**,**F**,**H**) using the method of Multiple Linear Regression (MLR, (**A**,**B**)), Random Forest (RF, (**C**,**D**)), XGBoosting (XGB, (**E**,**F**)), and Support Vector Machine (SVM, (**G**,**H**)). Statistical indices include Root Mean Square Error (RMSE, 60 kg bags ha^{-1}), Mean Absolute Error (MAE, 60 kg bags ha^{-1}), Percent Bias (PBIAS, %), Nash–Sutcliffe Efficiency Index (NSE), Coefficient of Determination (R^2), Correlation Coefficient (r), Index of Agreement (d), and Performance Index (c). The dashed blue line indicates the 1:1 line.

This finding is consistent with previous studies such as those by Aparecido et al. [12] and Miranda et al. [24], which, although they obtained promising results with MLR, with R^2 values over 0.80, also showed limitations, especially when the dataset is small or the variables are complex. Furthermore, the study by Alves et al. [26], which employed techniques such as Decision Trees (CART) and RF, also highlighted the efficacy of these more advanced methods, especially when dealing with large datasets and multiple variables, and revealed significant importance for variables related to coffee tree nutrients, mainly Mg, Fe, and Ca.

The performance indices observed in this study are also comparable to those reported in studies addressing agrometeorological models. For example, Victorino et al. [68] an-

alyzed the influence of a water deficit and reported correlation indices (r) ranging from 0.58 to 0.96, while Rosa et al. [27] found R^2 values between 0.54 and 0.89 when evaluating agrometeorological–spectral models. Similarly, Valeriano et al. [18] reported RMSE values ranging from 7.75 bags ha^{-1} to 7.97 bags ha^{-1} depending on the meteorological data sources, results that align closely with those observed here.

Notably, the ML models in this study performed comparably to those in Freitas et al. [20], who utilized an agrometeorological model achieving an RMSE of 8.65 bags ha^{-1} and MAE of 6.77 bags ha^{-1}, with an R^2 of 0.65. Furthermore, this study's results outperform more complex mechanistic models like DynACof (Dynamic Agroforestry Coffee crop model) [15] and CAF2014 [69], which were applied to simulate the growth and development of coffee plants under Central America's conditions. Specifically, the CAF2014 model showed R^2 ranging from 0.54 to 0.64 [69], whereas the DynACof model exhibited a model efficiency coefficient (NSE) ranging from −1.14 to 0.14 [15].

Recent advancements, such as integrating satellite-derived data and remote sensing indices with ML, have further improved yield forecasting [36]. Studies like Abreu Júnior et al. [28] demonstrated the use of Sentinel-2 imagery in combination with models such as Neural Networks (NN), Linear Regression (LR), RF, and SVM. Their findings revealed that NN and RF achieved lower RMSE% values (23% and 27%, respectively), indicating higher accuracy compared to LR and SVM, which exhibited higher RMSE% values of 39% and 36%. Regarding R^2, the NN model showed the strongest performance with an R^2 of 0.85, while LR displayed the weakest performance with an R^2 of 0.67.

Despite their advantages, ML models also have limitations in relation to mechanistic or biophysical models [14–16] and physiological–mathematical models [17,19,20]. The dependency on specific datasets restricts their generalizability to regions with distinct climatic conditions. However, many coffee-growing regions in Brazil share similar climates, which could allow for the broader applicability of these models. Another limitation is that ML models do not enable the direct calculation of yield gaps, which are important for identifying areas where management practices could be optimized [20,70–72].

Another important factor was the consideration of the biennial cycle of the coffee plant. Previous studies, such as that by Soares et al. [11] and Freitas et al. [20], also highlighted the importance of the biennial cycle, in addition to the main flowering and different agrometeorological variables according to phenology, in the variability of the coffee yield. The inclusion of this variable may have contributed to the overall efficacy of the more advanced models, such as RF and XGB, as well as to the good performance of MLR. Furthermore, the studies by Aparecido et al. [21] and Aparecido and Rolim [22] emphasized the importance of the air temperature and water deficit at different phenological stages of coffee, especially during fruit formation, as a significant factor affecting the yield, as well as the models for the G2 cultivar group.

3.4. Future Research Directions

Despite the promising results, this study faced significant challenges in the data acquisition, particularly concerning detailed soil characteristics such as the Available Water Capacity (AWC) and root depth, along with the amounts of irrigation applied. The scarcity of accessible and high-quality data on these aspects imposes limitations on the prediction models. The accurate modeling of the coffee yield relies heavily on the comprehensive understanding of these soil and water variables, which are instrumental in reflecting the true agronomic conditions.

Future research should focus on integrating comprehensive soil- and water-related variables to enhance model precision and reliability. Detailed data on soil properties, including the composition, texture, organic matter content, and water retention capacity, combined with precise irrigation records, could significantly improve the model performance under varying management and environmental conditions. Additionally, leveraging geospatial data and remote-sensing technologies offers a promising avenue to address data scarcity, enabling broader and more detailed datasets for analysis.

Expanding the dataset to include additional agronomic factors, such as pest and disease pressures, nutrient availability, and plant physiology metrics, could further validate the robustness of the present models. These improvements are particularly relevant given the complexities that advanced platforms such as DSSAT [73], APSIM [74], or AquaCrop [75] have not yet fully addressed. Although these platforms are powerful, they currently do not account for coffee cultivation simulation, emphasizing the relevance and potential of the models developed in this study and those proposed by Freitas et al. [20].

Moreover, adopting advanced machine learning techniques, such as Deep Neural Networks (DNNs), or hybrid approaches that integrate ML with process-based models could offer significant opportunities for improving yield predictions. These strategies have the potential to bridge the gap between empirical and mechanistic approaches, enabling more precise and generalizable coffee yield estimations and addressing data and methodological limitations.

4. Conclusions

The Factor Analysis of Mixed Data (FAMD) highlighted the complexity of the relationships among climatic factors, management practices, and the Arabica coffee yield. Four principal components were required to explain approximately 64.6% of the variance. Climatic and management factors, such as the water balance and air temperature, dominated the first two components, while the yield and specific management practices characterized the third and fourth components. This analysis emphasized that the Arabica coffee yield is shaped by both climatic and management factors.

The variables "Bien" (bienniality) and "Prn" (pruning) were identified as significant contributors to the yield in both groups, indicating that the yield is influenced by a combination of management and genetic factors. This study also identified differences between the coffee cultivar groups. Group 1 was more sensitive to management factors such as the planting density and pruning, while Group 2, in addition to management, was also influenced by climatic variables, mainly by the mean air temperature, indicating that different cultivar groups may require distinct management strategies.

Regarding the performance of machine learning models, the Random Forest (RF) and XGBoost (XGB) techniques proved to be more effective in estimating the yield for both groups, while Multiple Linear Regression (MLR) and Support Vector Machines (SVM) had inferior performances. Moreover, the yield estimation models for Group 2 were slightly inferior to the models calibrated for Group 1, and for both groups of cultivars, all models faced difficulties in estimating the yield values above 60 bags ha^{-1} (1 bag = 60 kg).

This study underscores the value of machine learning in understanding the intricate factors influencing the Arabica coffee yield and providing a basis for tailored management strategies. Incorporating additional variables, such as soil properties, detailed irrigation practices, and coffee plant conditions such as stem nutrient accumulation, hormonal levels, and the number of fruitful branches could further enhance the model performance and support the development of more robust predictive tools. Future research could also investigate advanced machine learning techniques, such as Deep Neural Networks, to improve the prediction accuracy and address the limitations identified in this study.

Author Contributions: Conceptualization, C.H.d.F. and P.C.S.; data curation, C.H.d.F. and P.C.S.; methodology, C.H.d.F. and P.C.S.; formal analysis, C.H.d.F.; software, C.H.d.F.; validation, C.H.d.F., R.D.C., J.d.O.C. and P.C.S.; visualization, C.H.d.F., R.D.C. and P.C.S.; writing—original draft preparation, C.H.d.F.; writing—review and editing, C.H.d.F., R.D.C. and J.d.O.C.; supervision, R.D.C. and P.C.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by Brazilian research agencies: the 'Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP)', grant number 2020/11465-8, and the 'Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)', grant number 140143/2019-0.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: We extend our gratitude to 'Ipanema Coffees' for providing the data essential for the realization of this study.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. INTERNATIONAL COFFEE ORGANIZATION Trade Statistics Tables—Total Production by All Exporting Countries. Available online: https://ico.org/coffee-development-report-2/ (accessed on 6 July 2023).
2. Ubilava, D. El Niño, La Niña, and World Coffee Price Dynamics. *Agric. Econ.* **2012**, *43*, 17–26. [CrossRef]
3. CONAB—Companhia Nacional de Abastecimento Séries Históricas Das Safras—Café Arábica. Available online: https://www.conab.gov.br/info-agro/safras/serie-historica-das-safras/itemlist/category/894-cafe-arabica (accessed on 10 September 2023).
4. De Camargo, M.B.P. The Impact of Climatic Variability and Climate Change on Arabic Coffee Crop in Brazil. *Bragantia* **2010**, *69*, 239–247. [CrossRef]
5. da Mota, R.P.; Ferraz-Almeida, R.; de Camargo, R.; Franco, M.H.R.; Delvaux, J.C.; Lana, R.M.Q. Organomineral Fertilizer in Coffee Plant (Coffea Arabica L.): Fertilizer Levels and Application Times. *Coffee Sci.* **2023**, *18*, e182098. [CrossRef]
6. DaMatta, F.M.; Ramalho, J.D.C. Impacts of Drought and Temperature Stress on Coffee Physiology and Production: A Review. *Braz. J. Plant Physiol.* **2006**, *18*, 55–81. [CrossRef]
7. de Souza, T.L.; de Oliveira, D.P.; Santos, C.F.; Reis, T.H.P.; Cabral, J.P.C.; da Silva Resende, É.R.; Fernandes, T.J.; de Souza, T.R.; Builes, V.R.; Guelfi, D. Nitrogen Fertilizer Technologies: Opportunities to Improve Nutrient Use Efficiency towards Sustainable Coffee Production Systems. *Agric. Ecosyst. Environ.* **2023**, *345*, 108317. [CrossRef]
8. Verdin Filho, A.C.; Volpi, P.S.; Ferrão, M.A.G.; Ferrão, R.G.; Mauri, A.L.; da Fonseca, A.F.A.; Tristão, F.A.; de Andrade Júnior, S. New Management Technology for Arabica Coffee: The Cyclic Pruning Program for Arabica Coffee. *Coffee Sci.* **2016**, *4*, 475–483.
9. Bongase, E.D. Impacts of Climate Change on Global Coffee Production Industry: Review. *Afr. J. Agric. Res.* **2017**, *12*, 1607–1611. [CrossRef]
10. Nunes, F.L.; de Camargo, M.B.P.; Fazuoli, L.C.; Rolim, G.d.S.; Pezzopane, J.R.M. Modelos Agrometeorológicos de Estimativa Da Duração Do Estádio Floração-Maturação Para Três Cultivares de Café Arábica. *Bragantia* **2010**, *69*, 1011–1018. [CrossRef]
11. Soares, L.d.S.; Rezende, T.T.; Beijo, L.A.; Franco Júnior, K.S. Interaction between Climate, Flowering and Production of Dry Coffee (Coffea Arabica L.) in Minas Gerais. *Coffee Sci.* **2021**, *16*, 1–10. [CrossRef]
12. Aparecido, L.E.d.O.; Rolim, G.d.S.; Lamparelli, R.A.C.; de Souza, P.S.; dos Santos, E.R.; de Souza Rolim, G.; Camargo Lamparelli, R.A.; de Souza, P.S.; dos Santos, E.R.; Rolim, G.d.S. Agrometeorological Models for Forecasting Coffee Yield. *Agron. J.* **2017**, *109*, 249–258. [CrossRef]
13. Rodríguez, D.; Cure, J.R.; Gutierrez, A.P.; Cotes, J.M.; Cantor, F. A Coffee Agroecosystem Model: II. Dynamics of Coffee Berry Borer. *Ecol. Modell.* **2013**, *248*, 203–214. [CrossRef]
14. van Oijen, M.; Dauzat, J.; Harmand, J.M.; Lawson, G.; Vaast, P. Coffee Agroforestry Systems in Central America: II. Development of a Simple Process-Based Model and Preliminary Results. *Agrofor. Syst.* **2010**, *80*, 361–378. [CrossRef]
15. Vezy, R.; le Maire, G.; Christina, M.; Georgiou, S.; Imbach, P.; Hidalgo, H.G.; Alfaro, E.J.; Blitz-Frayret, C.; Charbonnier, F.; Lehner, P.; et al. DynACof: A Process-Based Model to Study Growth, Yield and Ecosystem Services of Coffee Agroforestry Systems. *Environ. Model. Softw.* **2020**, *124*, 104609. [CrossRef]
16. Kouadio, L.; Tixier, P.; Byrareddy, V.; Marcussen, T.; Mushtaq, S.; Rapidel, B.; Stone, R. Performance of a Process-Based Model for Predicting Robusta Coffee Yield at the Regional Scale in Vietnam. *Ecol. Model.* **2021**, *443*, 109469. [CrossRef]
17. Santos, M.A.; de Camargo, M.B.P. Parametrização de Modelo Agrometeorológico de Estimativa de Produtividade Do Cafeeiro Nas Condições Do Estado de São Paulo. *Bragantia* **2006**, *65*, 173–183. [CrossRef]
18. Valeriano, T.T.B.; Rolim, G.d.S.; Aparecido, L.E.d.O.; Moraes, J.R.d.S.C.d. Estimation of Coffee Yield from Gridded Weather Data. *Agron. J.* **2018**, *110*, 2462–2477. [CrossRef]
19. Verhage, F.Y.F.; Anten, N.P.R.; Sentelhas, P.C. Carbon Dioxide Fertilization Offsets Negative Impacts of Climate Change on Arabica Coffee Yield in Brazil. *Clim. Change* **2017**, *144*, 671–685. [CrossRef]
20. Freitas, C.H.d.; Coelho, R.D.; de Oliveira Costa, J.; Sentelhas, P.C. Equationing Arabica Coffee: Adaptation, Calibration, and Application of an Agrometeorological Model for Yield Estimation. *Agric. Syst.* **2025**, *223*, 104181. [CrossRef]
21. de Oliveira Aparecido, L.E.; Lorençone, J.A.; Lorençone, P.A.; Torsoni, G.B.; Lima, R.F.; dade Silva CabralMoraes, J.R. Predicting Coffee Yield Based on Agroclimatic Data and Machine Learning. *Theor. Appl. Climatol.* **2022**, *148*, 899–914. [CrossRef]
22. Aparecido, L.E.d.O.; Rolim, G.d.S. Forecasting of the Annual Yield of Arabic Coffee Using Water Deficiency. *Pesqui. Agropecuária Bras.* **2018**, *53*, 1299–1310. [CrossRef]
23. Kouadio, L.; Byrareddy, V.M.; Sawadogo, A.; Newlands, N.K. Probabilistic Yield Forecasting of Robusta Coffee at the Farm Scale Using Agroclimatic and Remote Sensing Derived Indices. *Agric. For. Meteorol.* **2021**, *306*, 108449. [CrossRef]
24. Miranda, J.M.; Reinato, R.A.O.; Silva, A.B. Modelo Matemático Para Previsão Da Produtividade Do Cafeeiro. *Rev. Bras. Eng. Agrícola e Ambient.* **2014**, *18*, 353–361. [CrossRef]
25. Freitas, C.H.d. Adaptation, Calibration, and Application of Coffee Crop Simulation Models for Assessing the Impact of Climate Change in Brazilian Conditions. Ph.D. Thesis, Universidade de São Paulo, Piracicaba, Brazil, 2024. [CrossRef]

26. Alves, M.d.C.; Sanches, L.; Pozza, E.A.; Pozza, A.A.A.; da Silva, F.M. The Role of Machine Learning on Arabica Coffee Crop Yield Based on Remote Sensing and Mineral Nutrition Monitoring. *Biosyst. Eng.* **2022**, *221*, 81–104. [CrossRef]
27. Rosa, V.G.C.; Moreira, M.A.; Rudorff, B.F.T.; Adami, M. Estimativa Da Produtividade de Café Com Base Em Um Modelo Agrometeorologico-Espectral. *Pesqui. Agropecu. Bras.* **2010**, *45*, 1478–1488. [CrossRef]
28. Abreu Júnior, C.A.M.d.; Martins, G.D.; Xavier, L.C.M.; Vieira, B.S.; Gallis, R.B.d.A.; Junior, E.F.F.; Martins, R.S.; Paes, A.P.B.; Mendonça, R.C.P.; Lima, J.V.D.N. Estimating Coffee Plant Yield Based on Multispectral Images and Machine Learning Models. *Agronomy* **2022**, *12*, 3195. [CrossRef]
29. Zanella, M.A.; Nogueira Martins, R.; Moreira da Silva, F.; Carvalho, L.C.C.; de Carvalho Alves, M.; Fim Rosas, J.T. Coffee Yield Prediction Using High-Resolution Satellite Imagery and Crop Nutritional Status in Southeast Brazil. *Remote Sens. Appl. Soc. Environ.* **2024**, *33*, 101092. [CrossRef]
30. Liakos, K.G.; Busato, P.; Moshou, D.; Pearson, S.; Bochtis, D. Machine Learning in Agriculture: A Review. *Sensors* **2018**, *18*, 2674. [CrossRef] [PubMed]
31. Zanetti, W.A.L.; Marques, M.d.S.; do Amaral, A.M.S.; da Silva, A.B.; Barcelos, J.P.d.Q.; Goés, B.C.; Putti, F.F. Analysis of the Technological Evolution of Coffee Production in Brazil. *J. Agric. Stud.* **2021**, *9*, 352–362. [CrossRef]
32. Motta, I.V.C.; Vuillerme, N.; Pham, H.H.; de Figueiredo, F.A.P. Machine Learning Techniques for Coffee Classification: A Comprehensive Review of Scientific Research. *Artif. Intell. Rev.* **2025**, *58*, 15. [CrossRef]
33. Bunn, C.; Läderach, P.; Ovalle Rivera, O.; Kirschke, D. A Bitter Cup: Climate Change Profile of Global Production of Arabica and Robusta Coffee. *Clim. Chang.* **2015**, *129*, 89–101. [CrossRef]
34. Johnston, D.B.; Pembleton, K.G.; Huth, N.I.; Deo, R.C. Comparison of Machine Learning Methods Emulating Process Driven Crop Models. *Environ. Model. Softw.* **2023**, *162*, 105634. [CrossRef]
35. Vidhya, K.; George, S.; Suresh, P.; Brindha, D.; Jebaseeli, T.J. Agricultural Farm Production Model for Smart Crop Yield Recommendations Using Machine Learning Techniques. *Eng. Proc.* **2023**, *59*, 20. [CrossRef]
36. Sanya, R.; Nabiryo, A.L.; Tusubira, J.F.; Murindanyi, S.; Katumba, A.; Nakatumba-Nabende, J. Coffee and Cashew Nut Dataset: A Dataset for Detection, Classification, and Yield Estimation for Machine Learning Applications. *Data Br.* **2024**, *52*, 109952. [CrossRef] [PubMed]
37. Arwatchananukul, S.; Xu, D.; Charoenkwan, P.; Aung Moon, S.; Saengrayap, R. Implementing a Deep Learning Model for Defect Classification in Thai Arabica Green Coffee Beans. *Smart Agric. Technol.* **2024**, *9*, 100680. [CrossRef]
38. Przybył, K.; Gawrysiak-Witulska, M.; Bielska, P.; Rusinek, R.; Gancarz, M.; Dobrzański, B.; Siger, A. Application of Machine Learning to Assess the Quality of Food Products—Case Study: Coffee Bean. *Appl. Sci.* **2023**, *13*, 10786. [CrossRef]
39. Martínez, F.; Montiel, H.; Martínez, F. A Machine Learning Model for the Diagnosis of Coffee Diseases. *Int. J. Adv. Comput. Sci. Appl.* **2022**, *13*, 968–974. [CrossRef]
40. Parraga-Alava, J.; Cusme, K.; Loor, A.; Santander, E. RoCoLe: A Robusta Coffee Leaf Images Dataset for Evaluation of Machine Learning Based Methods in Plant Diseases Recognition. *Data Br.* **2019**, *25*, 104414. [CrossRef] [PubMed]
41. de Oliveira Aparecido, L.E.; Lorençone, J.A.; Lorençone, P.A.; de Souza Rolim, G.; de Meneses, K.C.; da Silva Cabral de Moraes, J.R.; Torsoni, G.B. Can Nonlinear Agrometeorological Models Estimate Coffee Foliation? *J. Sci. Food Agric.* **2022**, *102*, 584–596. [CrossRef]
42. Alvares, C.A.; Stape, J.L.; Sentelhas, P.C.; De Moraes Gonçalves, J.L.; Sparovek, G. Köppen's Climate Classification Map for Brazil. *Meteorol. Zeitschrift* **2013**, *22*, 711–728. [CrossRef]
43. de Camargo, A.P.; de Camargo, M.B.P. Definição e Esquematização Das Fases Fenológicas Do Cafeeiro Arábica Nas Condições Tropicais Do Brasil. *Bragantia* **2001**, *60*, 65–68. [CrossRef]
44. dos Santos, H.G.; Jacomine, P.K.T.; dos Anjos, L.H.C.; de Oliveira, V.A.; Lumbreras, J.F.; Coelho, M.R.; de Almeida, J.A.; Araújo de Filho, J.C.; de Oliveira, J.B.; Cunha, T.J.F. *Sistema Brasileiro de Classificação de Solos*, 5th ed.; Embrapa Solos: Brasília, DF, Brazil, 2018; ISBN 978-85-7035-800-4.
45. DaMatta, F.M.; Ronchi, C.P.; Maestri, M.; Barros, R.S. Coffee: Environment and Crop Physiology. In *Ecophysiology of Tropical Tree Crops*; DaMatta, F., Ed.; Nova Science Publishers, Inc.: New York, NY, USA, 2010; pp. 181–216, ISBN 978-1-60876-392-4.
46. Xavier, A.C.; Scanlon, B.R.; King, C.W.; Alves, A.I. New Improved Brazilian Daily Weather Gridded Data (1961–2020). *Int. J. Climatol.* **2022**, *42*, 8390–8404. [CrossRef]
47. Dias, H.B.; Sentelhas, P.C. Assessing the Performance of Two Gridded Weather Data for Sugarcane Crop Simulations with a Process-Based Model in Center-South Brazil. *Int. J. Biometeorol.* **2021**, *65*, 1881–1893. [CrossRef] [PubMed]
48. Monteiro, A.F.M.; Martins, F.B.; Torres, R.R.; de Almeida, V.H.M.; Abreu, M.C.; Mattos, E.V. Intercomparison and Uncertainty Assessment of Methods for Estimating Evapotranspiration Using a High-Resolution Gridded Weather Dataset over Brazil. *Theor. Appl. Climatol.* **2021**, *146*, 583–597. [CrossRef]
49. Duarte, Y.C.N.; Sentelhas, P.C. NASA/POWER and DailyGridded Weather Datasets-How Good They Are for Estimating Maize Yields in Brazil? *Int. J. Biometeorol.* **2020**, *64*, 319–329. [CrossRef] [PubMed]
50. Bender, F.D.; Sentelhas, P.C. Solar Radiation Models and Gridded Databases to Fill Gaps in Weather Series and to Project Climate Change in Brazil. *Adv. Meteorol.* **2018**, *2018*, 1–15. [CrossRef]
51. Battisti, R.; Bender, F.D.; Sentelhas, P.C. Assessment of Different Gridded Weather Data for Soybean Yield Simulations in Brazil. *Theor. Appl. Climatol.* **2019**, *135*, 237–247. [CrossRef]
52. Thornthwaite, C.W. An Approach toward a Rational Classification of Climate. *Geogr. Rev.* **1948**, *38*, 55. [CrossRef]

53. Pereira, A.R. Symplifying the Thornthwaite-Mather Water Balance. *Bragantia* **2005**, *64*, 311–313. [CrossRef]
54. Thornthwaite, C.W.; Mather, J.R. (Eds.) *The Water Balance*, 1st ed.; Laboratory of Climatology: Centerton, AR, USA, 1955.
55. Zacharias, A.O.; de Camargo, M.B.P.; Fazuoli, L.C. Modelo Agrometeorológico de Estimativa Do Início Da Florada Plena Do Cafeeiro. *Bragantia* **2008**, *67*, 249–256. [CrossRef]
56. Pagès, J. *Multiple Factor Analysis by Example Using R*; CRC Press: New York, NY, USA, 2014; ISBN 9781482205480.
57. Le, S.; Josse, J.; Husson, F. FactoMineR: An R Package for Multivariate Analysis. *J. Stat. Softw.* **2008**, *25*, 1–18. [CrossRef]
58. Hu, Y.; Yu, S.; Qi, X.; Zheng, W.; Wang, Q.; Yao, H. An Overview of Multiple Linear Regression Model and Its Application. *Chin. J. Prev. Med.* **2019**, *53*, 653–656. [CrossRef]
59. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
60. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16), San Francisco, CA, USA, 13–17 August 2016; Association for Computing Machinery: New York, NY, USA, 2016; Volume 13, pp. 785–794. [CrossRef]
61. Boser, B.E.; Guyon, I.M.; Vapnik, V.N. A training algorithm for optimal margin classifiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, Pittsburgh, PA, USA, 27–29 July 1992; pp. 144–152. [CrossRef]
62. Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *Int. Jt. Conf. Artif. Intell.* **1995**, *2*, 1137–1143.
63. Carvalho, H.F.; Galli, G.; Ventorim Ferrão, L.F.; Vieira Almeida Nonato, J.; Padilha, L.; Perez Maluf, M.; Ribeiro de Resende, M.F.; Guerreiro Filho, O.; Fritsche-Neto, R. The Effect of Bienniality on Genomic Prediction of Yield in Arabica Coffee. *Euphytica* **2020**, *216*, 1–16. [CrossRef]
64. Willmott, C.J.; Robeson, S.M.; Matsuura, K. A Refined Index of Model Performance. *Int. J. Climatol.* **2012**, *32*, 2088–2094. [CrossRef]
65. Nash, J.E.; Sutcliffe, J.V. River Flow Forecasting through Conceptual Models Part I—A Discussion of Principles. *J. Hydrol.* **1970**, *10*, 282–290. [CrossRef]
66. Pereira, S.P.; Bartholo, G.F.; Baliza, D.P.; Sobreira, F.M.; Guimarães, R.J. Crescimento, Produtividade e Bienalidade Do Cafeeiro Em Função Do Espaçamento de Cultivo. *Pesqui. Agropecuária Bras.* **2011**, *46*, 152–160. [CrossRef]
67. Li, X.; Su, X.; Li, J.; Anwar, S.; Zhu, X.; Ma, Q.; Wang, W.; Liu, J. Coupling Image-Fusion Techniques with Machine Learning to Enhance Dynamic Monitoring of Nitrogen Content in Winter Wheat from UAV Multi-Source. *Agriculture* **2024**, *14*, 1797. [CrossRef]
68. Victorino, E.C.; Carvalho, L.G.; Ferreira, D.F. Modelagem Agrometeorológica Para a Previsão de Produtividade de Cafeeiros Na Região Sul Do Estado de Minas Gerais. *Coffee Sci.* **2016**, *11*, 211–220.
69. Ovalle-Rivera, O.; Van Oijen, M.; Läderach, P.; Roupsard, O.; de Melo Virginio Filho, E.; Barrios, M.; Rapidel, B. Assessing the Accuracy and Robustness of a Process-Based Model for Coffee Agroforestry Systems in Central America. *Agrofor. Syst.* **2020**, *94*, 2033–2051. [CrossRef]
70. Freitas, C.H.d.; Elli, E.F.; Sentelhas, P.C. On-Farm Assessment of Eucalypt Yield Gaps—A Case Study for the Producing Areas of the State of Minas Gerais, Brazil. *Int. J. Biometeorol.* **2021**, *65*, 1659–1673. [CrossRef] [PubMed]
71. Sentelhas, P.C.; Battisti, R.; Monteiro, L.A.; Duarte, T.C.N.; Visses, F.A.V. Yield Gap: Concepts, Definitions and Examples (in Portuguese). *Int. Plant Nutr. Inst.* **2016**, *155*, 9–12.
72. Van Ittersum, M.K.; Cassman, K.G.; Grassini, P.; Wolf, J.; Tittonell, P.; Hochman, Z. Yield Gap Analysis with Local to Global Relevance-A Review. *F. Crop. Res.* **2013**, *143*, 4–17. [CrossRef]
73. Jones, J.W.; Hoogenboom, G.; Porter, C.H.; Boote, K.J.; Batchelor, W.D.; Hunt, L.A.; Wilkens, P.W.; Singh, U.; Gijsman, A.J.; Ritchie, J.T. The DSSAT Cropping System Model. *Eur. J. Agron.* **2003**, *18*, 235–265. [CrossRef]
74. Holzworth, D.P.; Huth, N.I.; deVoil, P.G.; Zurcher, E.J.; Herrmann, N.I.; McLean, G.; Chenu, K.; van Oosterom, E.J.; Snow, V.; Murphy, C.; et al. APSIM—Evolution towards a New Generation of Agricultural Systems Simulation. *Environ. Model. Softw.* **2014**, *62*, 327–350. [CrossRef]
75. Raes, D.; Steduto, P.; Hsiao, T.C.; Fereres, E. Aquacrop-The FAO Crop Model to Simulate Yield Response to Water: II. Main Algorithms and Software Description. *Agron. J.* **2009**, *101*, 438–447. [CrossRef]