agriculture

Article

Identifying Critical Drivers of Mango, Tomato, and Maize Postharvest Losses (PHL) in Low-Income Countries and Predicting Their Impact

Hory Chikez [1,*], Dirk Maier [1], Sigurdur Olafsson [2] and Steve Sonka [1,3]

[1] Department of Agricultural and Biosystems Engineering, Iowa State University, Ames, IA 50011, USA
[2] Department of Industrial and Manufacturing Systems Engineering, Iowa State University, Ames, IA 50011, USA
[3] Ed Snider Center for Enterprise and Markets, University of Maryland, College Park, MD 20742, USA
* Correspondence: horych@iastate.edu

Abstract: Several studies have identified a host of factors to be considered when attempting to reduce food postharvest loss (PHL). However, very few studies have ranked those factors by their importance in driving PHL. This study used the Random Forest model to rank the critical drivers of PHL in maize, mango, and tomato, cultivated in Tanzania, Kenya, and Nigeria, respectively. The study then predicted the maize, mango, and tomato PHLs by changing the levels of the most critical drivers of PHL and the number of farmers at each level. The results indicate that the most critical drivers of PHL consist of pre-harvest and harvest variables in the field, such as the quantity of maize harvested in the maize value chain, the method used to know when to begin mango harvest, and the type of pest that attacks plants in the tomato value chain. Furthermore, changes in the levels of a critical driver and changes in the number of smallholder farmers at a given level both have an impact on PHL, although the impact of the former is much higher than the latter. This study also revealed that the critical drivers of PHL can be categorized as either passive and difficult to manipulate, such as the geographic area within which a smallholder farmer lives, or active and easier to control, such as services provided by the Rockefeller Foundation YieldWise Initiative. Moreover, the affiliation of smallholder farmers to the YieldWise Initiative and a smallholder farmer's geographic location are ubiquitous critical drivers across all three value chains. Finally, an online dashboard was created to allow users to explore further the relationship between several critical drivers, the PHL of each crop, and a desired number of smallholder farmers.

Keywords: variable importance; Random Forest; predictive model; dashboard

check for updates

Citation: Chikez, H.; Maier, D.; Olafsson, S.; Sonka, S. Identifying Critical Drivers of Mango, Tomato, and Maize Postharvest Losses (PHL) in Low-Income Countries and Predicting Their Impact. *Agriculture* **2023**, *13*, 1912. https://doi.org/10.3390/agriculture13101912

Academic Editor: Anna Maria Stellacci

Received: 5 July 2023
Revised: 20 September 2023
Accepted: 22 September 2023
Published: 29 September 2023

1. Introduction

Maize, mango, and tomato are essential crops in Tanzania, Kenya, and Nigeria, respectively. One way to observe this importance is by comparing each crop's annual production to all the other primary crops in each country. Maize in Tanzania, mango in Kenya, and tomato in Nigeria ranked in the 97th, 92nd, and 76th percentile of all primary crops, respectively, in 2019 [1]. Additionally, maize is cultivated by most Tanzanian farmers and occupies 45 percent of Tanzania's cultivated land [2]. Meanwhile, mango production in Kenya, notably the Apple and Ngowe varieties, which are the most prevalent, has increased rapidly over the last decade and is expected to reach an annual production of 1.1 MT in 2022 [3]. Nigeria is the largest producer of tomatoes in Sub-Saharan Africa (SSA), and tomato is a key vegetable consumed throughout the country [4].

Besides production quantities, the nutritional facet of each crop is also valuable in understanding the importance of the crops within the respective country. For example, maize has been reported to provide about 60 percent of Tanzanian's dietary calories and 50 percent of their proteins [5], making it an important food crop in several parts of

Tanzania [6]. Mangos contain high β-carotene content, which is a precursor of vitamin A. Hence, either fresh or dried, mango fruits could reduce vitamin A deficiency in Kenya in vulnerable groups such as women and children [7]. Tomato are a rich source of lycopene, beta-carotene, folate, potassium, vitamin C, flavonoids, and vitamin E, hence, may be considered a valuable component of a cardioprotective diet [8].

Lastly, the economic importance of each crop, particularly for the benefit of smallholder farmers (SHFs), has been well reported in several studies. In Tanzania, for example, SHFs contribute over 80 percent of Tanzania's total maize production [5], while mango farming in Kenya and tomato farming in Nigeria constitute a major income generator for many SHF households [4,9,10].

Yet, despite the evident nutritional and economic value that maize, mango, and tomato crops bring to the populations of SSA, large quantities of these crops are lost during or after harvest [11], thus never reaching the end consumer. For instance, in Tanzania, maize postharvest losses (PHL) of up to 18 percent have been reported along the entire value chain [12]. Similarly, mango and tomato production is usually accompanied by a major PHL, estimated at 40–50% [13,14]. Hence, PHL reduction efforts, especially in SSA, could be a catalyst for increasing profit for food value chain actors while at the same time boosting food availability and ultimately improving food security [5,15]. To this end, several PHL reduction initiatives have emerged over the last decade, predominantly in SSA, which remains the most food-insecure region of the world [16].

Notably, the United Nations Sustainable Development Goals (SDG12.3) aim, by 2030, to reduce food losses along production and supply chains, including postharvest losses [17]. Additionally, The Rockefeller Foundation launched the YieldWise Initiative (YWI) in 2016, which implemented several interventions to help smallholder farmers reduce their PHL in Tanzania, Kenya, and Nigeria [18]. Following the implementation of the YWI, surveys were conducted to collect farm-level data [19].

While the survey instruments used in the YWI captured many explanatory variables related to PHL, identifying critical drivers from such a large number of covariates using standard statistical methods is rather challenging [20]. Therefore, this study used a predictive modeling approach from the field of machine learning to first identify the most critical drivers of PHL from the large numbers of explanatory variables recorded in the datasets and their respective impact on PHL in the maize, mango, and tomato value chains. The advantages of using predictive modeling to this effect include, but are not limited to, their high speed in generating results [21] and their higher predictive accuracy than explanatory statistical models [22]. Moreover, predictive models are well suited for exploring and analyzing diverse data [20,23]. They can capture underlying complex patterns and relationships in the data [22] and quantify relationships between predictors and outcomes [24]. Furthermore, they are well suited for identifying important variables derived from a large dataset. Therefore, this study used the Random Forest predictive model approach to identify the most critical drivers of PHL in the maize, mango, and tomato value chains and predict their impact. Lastly, an online dashboard was created to allow users to predict maize, mango, and tomato PHLs by varying several critical drivers of a value chain and the number of farmers at once. The dashboard is described in Appendix A and can be accessed through the following link: https://phldashboard.shinyapps.io/phldashboard/ (accessed on 8 May 2023).

2. Materials and Methods

2.1. Data Summary

For each value chain, the data collected contained multiple explanatory variables, such as farm demographics, agricultural practices, storage methods, YieldWise affiliation or interventions, and PHL quantity incurred by the farmer between the harvest and point of sale. In each value chain, PHL was expressed as the quantitative and qualitative losses accumulated between harvest and point of sale [19]. After extensively cleaning the YWI survey data, the maize value chain dataset contained 22 explanatory variables and 381 observations (Table 1), the mango value chain dataset contained 21 explanatory variables and 753 observations (Table 2), and the tomato value chain dataset contained 25 explanatory variables and 503 observations (Table 3). Each observation represents a SHF who responded to the survey.

Table 1. Maize value chain data summary.

Variables	Levels (Subset)	Number of Farmers or Observations (n)	21. Average Land Size (ac) per Farmer	22. Average Maize Harvest (kg) per Farmer per Season	Average Maize PHL (%) per Farmer per Season
1. gender	female	161	4	2836	7
	male	220	4	4626	7
2. zone	central	73	8	2410	13
	coastal	24	2	2445	14
	southern highlands	284	3	4365	6
3. sample types	beneficiary	146	3	4627	6
	control	139	6	3210	11
	other	96	3	3673	5
4. YieldWise training	female adults	57	2	3519	6
	female youth	1	2	6000	7
	male adults	84	3	5449	6
	male youth	2	3	2878	2
	other	237	5	3393	8
5. decides on planting	female adults	78	3	2759	9
	male adults	148	4	4163	9
	other	155	4	4148	5
6. decides on harvest	female adults	77	3	2865	8
	male adults	104	5	4002	10
	male youth	1	3	4940	3
	other	199	4	4184	6
7. decides on proceeds	female adults	71	3	2776	9
	male adults	103	4	4007	9
	male youth	1	3	4940	3
	other	206	4	4172	6
8. tarp supplier	agra aggregator	76	3	4399	9
	agro equipment stores	147	5	4572	7
	donated	12	5	3507	10
	other	146	3	2916	6
9. threshing modes	manual	82	3	3388	9
	mechanical	134	6	4472	7
	other	165	3	3619	7
10. point of sale	farm	11	5	3156	12
	homestead	28	5	3199	8
	other	333	4	3924	7
	village market	8	4	5013	7
	warehouse	1	1	3204	1
11. transport mode	other	374	4	3866	8
	sacks on animal cart	4	6	3769	2
	sacks on bicycle	1	5	7371	3
	sacks on wheelbarrow	2	2	2950	5
12. direct client	aggregators	5	2	2266	11
	direct consumers	4	4	5884	4
	other	372	4	3869	7

Table 1. *Cont.*

Variables	Levels (Subset)	Number of Farmers or Observations (n)	21. Average Land Size (ac) per Farmer	22. Average Maize Harvest (kg) per Farmer per Season	Average Maize PHL (%) per Farmer per Season
13. level of dryness	dry	25	5	3923	9
	humid	19	3	4478	10
	other	104	4	3864	5
	very dry	231	4	3829	8
	very humid	2	3	2430	15
14. sun drying practices	field before harvesting	149	4	3572	9
	other	108	4	3818	5
	shallow layer stand	1	1	1336	4
	spread threshed maize	120	4	4328	7
	ventilated crib for cob	3	4	2991	11
15. mc measurement method	farmer experience	110	5	3191	10
	other	255	4	4000	7
	salt test	6	3	5644	4
	smell-based	1	1	2440	2
	use of machines	7	5	9078	3
	weight-based	2	4	1765	5
16. storage method	home storage	32	4	4240	7
	improved granaries	3	3	5060	4
	on ground in house	28	3	5135	5
	other	108	5	3576	6
	pics hermetic bags	13	4	6115	7
	plastic bag	161	4	3676	9
	plastic silos	1	5	3470	3
	traditional granaries	35	4	3393	9
17. received training	no	177	4	3514	9
	yes	204	4	4178	6
18. education level	complete college	2	4	5400	2
	complete primary	293	4	4001	7
	complete secondary	31	5	4126	5
	complete university	1	2	1620	1
	dip certificate	1	5	3798	18
	no formal education	19	3	2600	12
	other	7	6	2585	10
	postgraduate	1	3	5220	0
	some primary	16	3	1882	10
	some secondary	10	4	5501	4
19. employment status	formal sector	21	3	3606	5
	housewife	35	2	3014	6
	informal sector	31	6	3496	11
	not working	21	4	2904	6
	other	21	4	3980	8
	retired	7	4	5301	4
	self employed	232	4	4115	7
	temporarily employed	13	3	3717	9
20. transport mode I	cattle cart	56	4	4102	9
	covered trucks	16	4	4349	4
	open trucks	51	4	6849	5
	other	197	4	2989	8
	sacks on head	16	6	3542	7
	sacks on wheelbarrow	10	2	2656	10
	tractor	28	4	4842	7
	wheelbarrows no sacks	7	3	2586	7

Table 2. Mango value chain data summary.

Variables	Levels (Subset)	Number of Farmers or Observations (n)	20. Average Number of Mango Trees per Farmer	21. Average Mango Price (KES) per Fruit	Average Mango PHL (%) per Farmer per Season
1. county	central	13	59	5	25
	coast	345	19	5	32
	eastern	389	71	7	25
	northeastern	6	9	7	18
2. treatment control	non beneficiary	282	42	6	31
	YieldWise beneficiary	471	49	6	27
3. farm ownership	no	135	33	6	28
	yes	618	49	6	28
4. labor costs	no	468	33	6	30
	yes	285	69	7	25
5. who harvested mango	buyer	411	58	6	27
	farmer	181	24	6	33
	other	161	43	7	27
6. inform when to harvest	days after blooming	5	51	8	14
	fruit color	165	70	7	30
	fruit size or shape	49	22	6	43
	other	521	41	6	27
	test for maturity	13	44	6	29
7. frequency of harvest	daily	53	100	7	31
	fortnightly	231	36	6	29
	monthly	52	44	6	34
	other	109	51	5	28
	weekly	308	44	7	26
8. methods of harvest	harvesting tools	49	57	9	25
	other	160	35	5	25
	traditional practices	544	49	6	30
9. how farmer identified buyer	brokers	407	48	6	29
	farmer-based organization	12	169	12	12
	other	81	48	4	33
	own effort	253	38	7	26
10. harvested mango graded	no	346	49	5	30
	yes	407	44	7	27
11. market destination	export	106	88	8	19
	local market	362	45	6	29
	other	239	34	5	31
	processing	41	24	6	34
	supermarket	5	63	14	25
12. storage after harvesting	cold store	18	75	6	25
	other	49	55	6	29
	traditional practices	686	45	6	28
13. package for sale	other	314	46	6	31
	plastic crates	320	50	7	24
	traditional practices	119	39	6	34
14. receive production training	no	534	41	6	31
	yes	219	60	7	21
15. have bank account	no	374	35	6	31
	yes	379	58	7	26

Table 2. *Cont.*

Variables	Levels (Subset)	Number of Farmers or Observations (n)	20. Average Number of Mango Trees per Farmer	21. Average Mango Price (KES) per Fruit	Average Mango PHL (%) per Farmer per Season
16. have mobile money account	no	95	29	5	32
	yes	658	49	6	28
17. receive remittances	no	467	47	6	28
	yes	286	46	7	29
18. taken loan for farm	no	695	44	6	29
	yes	58	71	8	25
19. production PHL practices	fruit fly traps	125	76	6	25
	other	203	53	8	27
	tarp	115	17	5	33
	traditional practices	310	41	6	29

Table 3. Tomato value chain data summary.

Variables	Levels (Subset)	Number of Farmers or Observations (n)	21. Average Income (NGN) per Farmer	22. Average Labor Cost (NGN) per Season	23. Average Frequency of Pesticide Applications per Season	24. Average Distance (km) to Market	25. Average Harvest (kg) per Farmer per Season	Average Tomato PHL (%) per Farmer per Season
1. treatment or control	control	166	235,630	46,796	7	18	214	20
	intervention	337	271,738	46,151	6	19	253	15
2. state	jigawa	134	309,038	43,330	6	21	185	22
	kano	248	195,535	34,452	5	20	266	13
	katsina	121	337,077	74,137	9	12	247	19
3. gender	female	5	316,200	11,000	6	4	280	14
	male	498	259,255	46,719	6	18	239	17
4. dry month	january	43	244,118	42,531	7	20	260	17
	february	6	641,667	110,000	6	4	188	5
	august	6	185,000	57,583	6	7	66	14
	september	52	398,327	70,133	8	19	219	26
	october	229	215,823	37,574	7	21	256	18
	november	106	289,778	63,218	5	15	214	15
	december	61	235,737	25,149	5	15	249	7
5. tomato varieties	chibli	20	235,880	36,019	5	13	126	4
	other	236	275,188	59,668	7	17	233	15
	roma	65	296,261	25,218	5	21	229	17
	uc82b	182	229,512	37,801	6	19	265	20
6. intercropped with tomatoes	no	343	265,509	50,783	7	20	250	16
	yes	160	247,628	36,890	6	15	217	19
7. main fertilizers	npk	63	158,847	29,177	5	11	196	24
	other	402	284,891	52,018	7	18	239	15
	ssp	21	144,647	9266	5	39	404	15
	urea	17	183,476	22,176	5	31	208	17
8. major pests attacks	aphids	50	278,472	35,752	5	7	154	10
	other	287	279,685	54,281	7	16	236	14
	thrips bugs	18	203,922	21,406	5	9	228	6
	tuta	148	221,799	37,630	6	27	278	26
9. pesticides usage	combine	329	254,836	53,248	7	18	252	17
	other	12	152,592	62,167	1	4	222	10
	single	162	277,887	31,212	6	21	217	16
10. diseases attack	leaf blight	41	177,815	29,810	4	11	238	10
	leaf virus	92	180,828	26,667	6	32	326	18
	nematodes	43	382,616	62,744	6	14	216	18
	other	327	276,181	51,827	7	16	219	17
11. herbicides	glyphosate	77	193,075	33,197	6	32	240	15
	manual	337	300,756	54,205	6	14	233	15
	other	67	166,036	26,795	6	19	291	20
	primextra	22	152,004	31,927	12	29	189	27

Table 3. *Cont.*

Variables	Levels (Subset)	Number of Farmers or Observations (n)	21. Average Income (NGN) per Farmer	22. Average Labor Cost (NGN) per Season	23. Average Frequency of Pesticide Applications per Season	24. Average Distance (km) to Market	25. Average Harvest (kg) per Farmer per Season	Average Tomato PHL (%) per Farmer per Season
12. irrigation type	drip	104	226,746	36,039	6	26	250	17
	flood	265	284,821	46,744	7	15	236	18
	other	109	266,513	61,075	5	20	258	12
	sprinkler	25	103,240	21,140	8	8	158	18
13. harvesting containers	other	9	307,444	49,889	4	18	223	13
	plastic crates	1	8500	23,000	6	1	170	2
	raffia baskets	474	244,296	45,798	6	19	246	17
	sacks	19	637,816	60,031	5	10	98	11
14. harvest destination	agg center	9	243,889	52,333	7	8	99	34
	buyer picks up	164	264,128	54,764	7	29	296	17
	market	197	216,168	30,184	6	11	202	17
	other	133	320,247	59,567	7	16	235	14
15. transportation method	1-ton truck	198	273,308	39,966	6	16	254	15
	2-ton truck	48	267,115	51,281	9	17	245	25
	30-ton truck	2	850,000	25,000	10	237	400	23
	5-ton truck	1	115,000	40,000	14	10	234	4
	motorcycle	134	225,802	35,320	5	10	168	17
	other	81	276,929	78,418	7	38	345	14
	tricycle	39	237,174	45,420	6	10	182	19
16. YieldWise services	credit	23	143,978	22,039	4	52	324	15
	inputs	19	238,474	55,237	6	33	239	25
	market access	73	151,985	37,607	5	10	234	15
	none	177	326,968	63,824	7	18	208	17
	other	186	254,787	33,542	6	17	259	15
	training	25	259,562	59,340	6	11	267	24
17. information channel	friend	15	234,557	26,273	7	25	240	27
	market	5	239,200	45,000	11	6	280	30
	neighbor	3	96,667	20,167	6	1	197	14
	other	331	287,468	54,454	7	19	249	17
	phone	3	186,000	33,333	4	35	227	27
	posters	1	50,000	5000	3	2	50	33
	radio	145	206,384	31,118	5	17	220	13
18. farmer group member	no	180	283,348	47,792	5	14	208	16
	yes	323	246,710	45,567	7	21	258	17
19. inputs on credit	no	473	262,040	46,149	6	18	239	16
	yes	30	224,833	49,753	8	16	248	24
20. contract with Dangote	do not know	18	136,544	39,617	4	10	214	17
	no	451	267,166	40,408	6	19	236	17
	yes	34	227,653	128,929	9	15	301	7

2.2. Random Forest

Predictive models range from linear models to decision trees, neural networks, support vector machines, cluster models, expert systems, etc. Each type has its strengths and weaknesses. However, the Random Forest predictive model was deemed more appropriate for this study's objectives since it does not require any distributional assumptions [25] to analyze the complex YWI data. Furthermore, Random Forests are more effective when predictors are categorical and are not converted into dummy variables [24], as is the case in the YWI datasets. They are robust to a noisy response [24], fast to execute, require minimal storage [26], and, overall, are an effective tool in prediction [27]. Lastly, a Random Forest analysis can create several predictive models called decision trees and then combine them [21] to produce more precise predictions [22].

Random Forest models can be divided into two main categories: regression and classification. Regression Forests are used to predict a quantitative or continuous response, whereas classification Forests are used to predict a qualitative or categorical response [28]. Because the response variables in this study are maize, mango, and tomato PHL (quantitative), the regression Random Forest was used. Fundamentally, the Random Forest regression model consists of recording the predictions of each regression tree, T_b, for a new observation and then taking the average over B trees [23]:

$$\hat{f}_B(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \quad (1)$$

where

$T_b(x)$ = prediction outcome of each regression tree b for a new observation x

B = total number of regression trees in the Random Forest model

$\hat{f}_B(x)$ = sum of T_b divided by B the for a new observation x

The Random Forest predictive model was built using R software (RStudio, Boston, MA, USA). The overarching process used to attain the objectives of this study is shown in Figure 1. Further, the specific procedures used in this study are detailed within the following subsections. In addition, it should be noted that a first step in this study's procedures was to conduct an extensive screening and cleaning process, as described by [19].

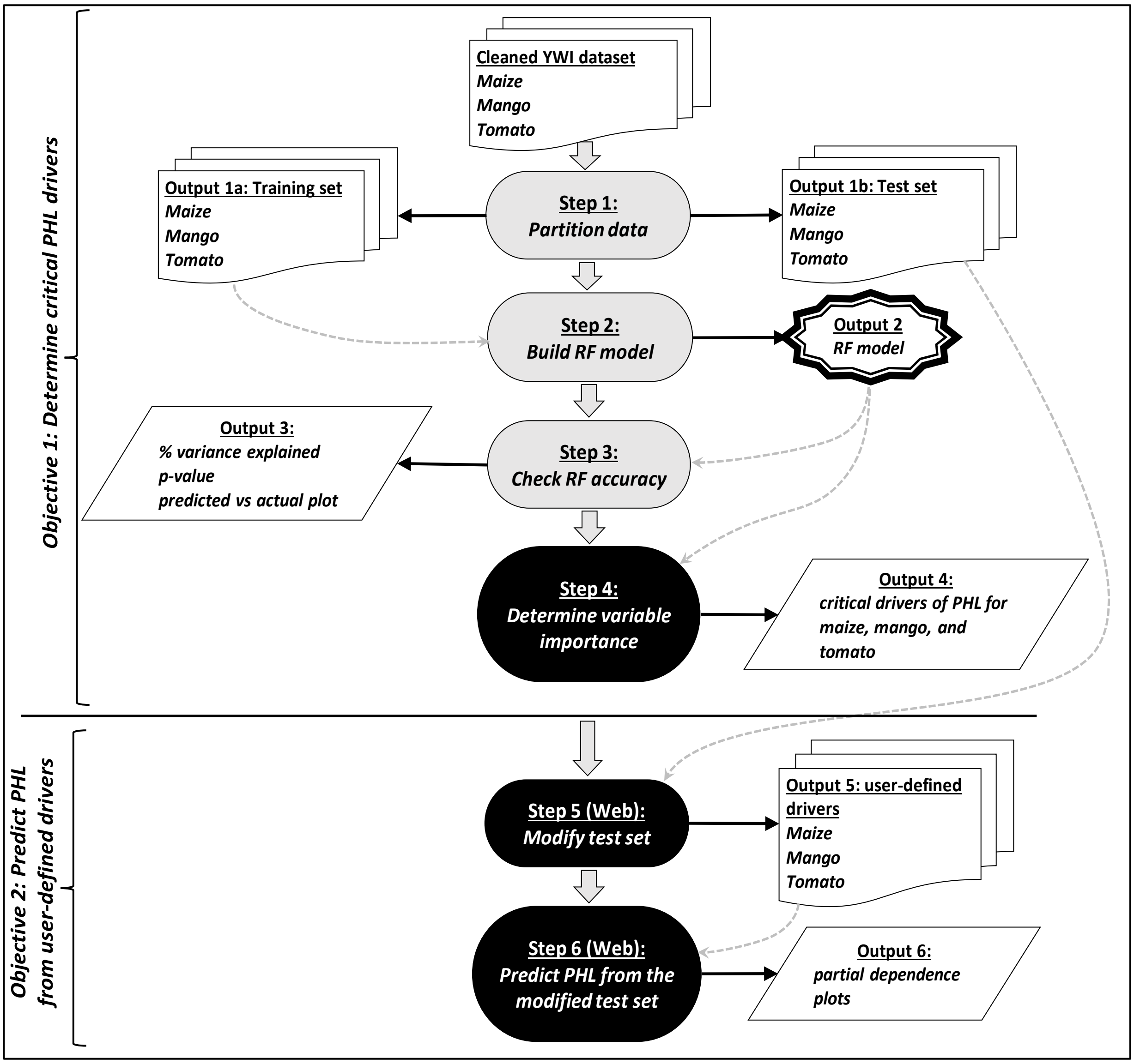

Figure 1. Regression Random Forest model and predictions. Web refers to the online dashboard.

2.3. Random Forest

To prevent the Random Forest model from overfitting and to accurately evaluate it, it is important to split the data into a training set and a testing set [29]. The training set was used to develop the model, while the test or validation set was used to evaluate the model's performance [24]. Several splitting rules can be used to partition data; however, empirical studies show that the best results are obtained when 20–30% of the data are used for testing and the remaining 70–80% are used for training [29]. Within this study, partitioning the maize, mango, and tomato datasets was accomplished through the DataPartition() function [30] and resulted in training sets of 269, 530, and 354 observations and test sets of 112, 223, and 149 observations in the maize, mango, and tomato value chain datasets, respectively (Figure 1 step 1).

2.4. Tuning the Random Forest Model

The Random Forest package and the Classification and Regression Training (caret) package were first loaded to simplify the tuning process. Then, three key parameters were considered to improve the model's accuracy, namely: the random seed, the number of trees to be built in the predictive model, and the number of predictors randomly sampled at each split, also referred to as "mtry" in R.

Setting a random seed allows the results of the predictive model to be reproduced [28], since the Random Forest predictive model is built by selecting predictors randomly. The random seed was set as set.seed(1234) by using the set.seed() function [28]. The number of trees specifies how many trees will be built to populate the Random Forest. The default value is generally set at 500 [31], since a larger number of trees in a Forest only increases its computational cost, has no significant performance gain [32], and could yield overfitting [33]. Hence, the number of trees in this study was left to R's default setting of 500 trees. As for the number of predictors randomly sampled at each split, the RandomForest() function calculates this value by dividing the total number of predictors found in the dataset by three for the regression Forest [28]. Since the maize dataset contained 22 predictors, the mango dataset, 21, and the tomato dataset, 25, the resulting mtry values were seven for the maize and mango value chain predictive models and eight for the tomato model. Three Random Forest predictive models were built for each value chain's dataset. Each Random Forest predictive model's significance was computed using the rfUtilities package and the rf.significance() function. In addition to the significance of the predictive models, the proportion of variance explained and predicted vs. actual plots was also generated to assess further the accuracy of each Random Forest predictive model.

2.5. Variable Importance

The variable importance calculation identifies important predictors or variables highly related to the response variable [34]. Hence, computing the variable importance was essential in achieving this study's first objective, which consists of identifying the most critical drivers of maize, mango, and tomato PHL. The variable importance was computed along with the corresponding p-values by using the rfPermute package and the rfPermute() and rp.importance() functions. The variable importance was expressed in units of "percent increase in mean squared error (%IncMSE)", which represents the mean decrease in accuracy in predictions when a given predictor or independent variable is excluded from the predictive model [28]. Hence, the higher the %IncMSE, the greater the importance of the corresponding variable in the predictive model. After computing the variable importance, the critical drivers of PHL were identified as those variables whose %IncMSE were significant (p-value < 0.05) (see Appendix B). Critical drivers with the highest %IncMSE were referred to as "most critical drivers".

2.6. Predictions

Partial dependence plots were generated for each model as they are useful in interpreting the complexity of the Random Forest model [35] by displaying the relationship between the predicted outcome (PHL) and predictors of interest (critical drivers) [36]. Partial dependence plots were first generated by using the plotmo package, plotmo() function, and method = "partdep" argument [37] to have an overall view of the changes in the predicted PHL as a function of several variables contained in the dataset (Appendix C).

Additional partial dependence plots were also created to explore predicted PHL changes by varying the levels or subsets of a given critical driver. This process entailed altering a critical driver of interest in the test set while leaving all other critical drivers of PHL unchanged and subsequently running the predict() function on the modified test sets. The process of altering a critical driver of interest varied depending on whether the critical driver was categorical or numerical. When a critical driver of interest was categorical, all the levels of that critical driver were changed to a single level of interest (Figure 1 step 5), then the predict() function was run to predict PHL as a function of the changes made. The predicted PHL values were subsequently averaged over the total number of observations (Figure 1 step 6) and plotted on a bar graph, with the predicted average on the vertical axis and the chosen subset of the critical drivers on the horizontal axis. However, when the critical driver was numerical, it was first altered by adding a constant to each observation (Figure 1 step 5), then the predict() function was run to predict PHL as a function of the changes made (Figure 1 step 6). The predicted PHL values were subsequently plotted in a line graph, and placed on the vertical axis, while the altered values of the numerical driver were placed on the horizontal axis. While creating partial dependence plots helps understand the marginal effect that one or two critical drivers have on the predicted PHL, adding more than two critical drivers to the plot makes it difficult to visualize. Therefore, an online dashboard was created using the Shiny package to assess PHL by varying the subsets of one or more critical drivers and the number of farmers.

3. Results

3.1. Random Forest Models

The proportion of variance explained (R-squared) was low for the predictive models of the three value chains (Table 4). The low proportion of variance explained is also reflected in the predicted vs. actual value plots of each crop chain (Figure 2).

Table 4. Summary of the Random Forest predictive model for each value chain.

Random Forest Predictive Model Value Chain	*p*-Value	% Var Explained (R-Squared)	Mean Squared Residual
Maize	0.001 *	20	67
Mango	0.001 *	13	447
Tomato	0.001 *	21	327

* Indicates that the model is significant at $p < 0.05$.

Obtaining such low R-squared values in a Random Forest regression model is not uncommon, especially when using a large dataset with irregular patterns [25], such as the YWI dataset. The fact that the PHL values were estimated by farmers and not measured likely led to errors, which in turn contributed to the lower R-squared values, as is sometimes the case in social science studies [38]. One way to minimize the errors caused by estimating would be to measure PHL by quantifying each crop's total weight loss and subsequently expressing it as a percentage of the total harvested weight [39]. The Random Forest model summary also produced the mean squared residual values listed in Table 4, which can be used to compare alternative predictive models to the Random Forest predictive models used in this study.

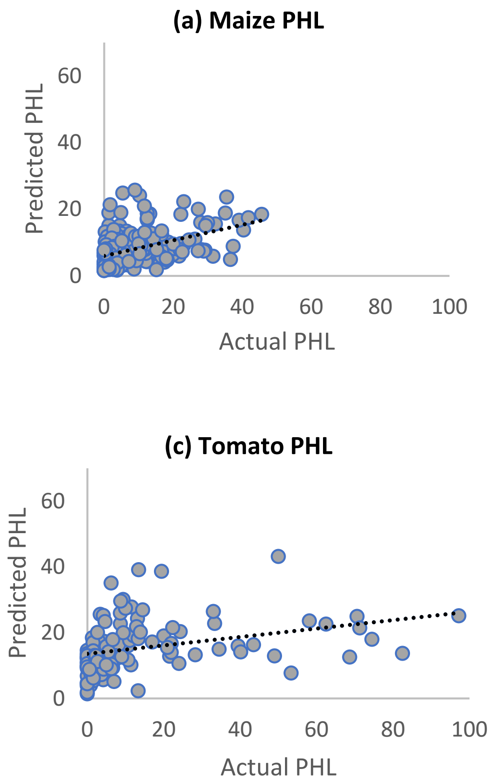

Figure 2. Predicted PHL vs. actual PHL.

3.2. Critical Drivers of PHL

Overall, the number of critical drivers of PHL in the mango and tomato value chains shown in Figures 3b and 3c, respectively, is approximately three times larger than those in the maize value chain shown in Figure 3a. Hence, these results suggest that perishable crops have more critical drivers of PHL than nonperishable crops. Kiaya [40] supports this notion, as PHL in nonperishable crops is primarily due to exogenous factors such as moisture, insects, or rodents, while PHL in perishable crops is usually due to exogenous factors and endogenous factors such as respiration, transpiration, and germination.

Additionally, the most critical drivers of PHL, labeled "the quantity (kg) of maize harvested by a smallholder farmer" (Figure 3a), "the method used to know when to begin mango harvest" (Figure 3b), and "the type of pest that attacked the tomato plant" (Figure 3c), are all related to either pre-harvest or harvest activities. Hence, these results suggest that PHL reduction efforts should start before a harvest and continue during the harvest. Incidentally, several studies have attempted to understand how pre-harvest factors affect PHL. For example, [41] looked at how physiological processes and field management strategies influenced the ultimate quality of perishable crops. Similarly, [42] established that the prevention of pre-harvest infection of maize by toxigenic *A. flavus* strains should be a critical control point to reduce PHL due to aflatoxin contamination.

The per cent increase in mean squared error (%IncMSE) represents the mean decrease of accuracy in predictions when a given predictor or independent variable is excluded from the predictive model. YWI refers to critical drivers related to the YieldWise Initiative.

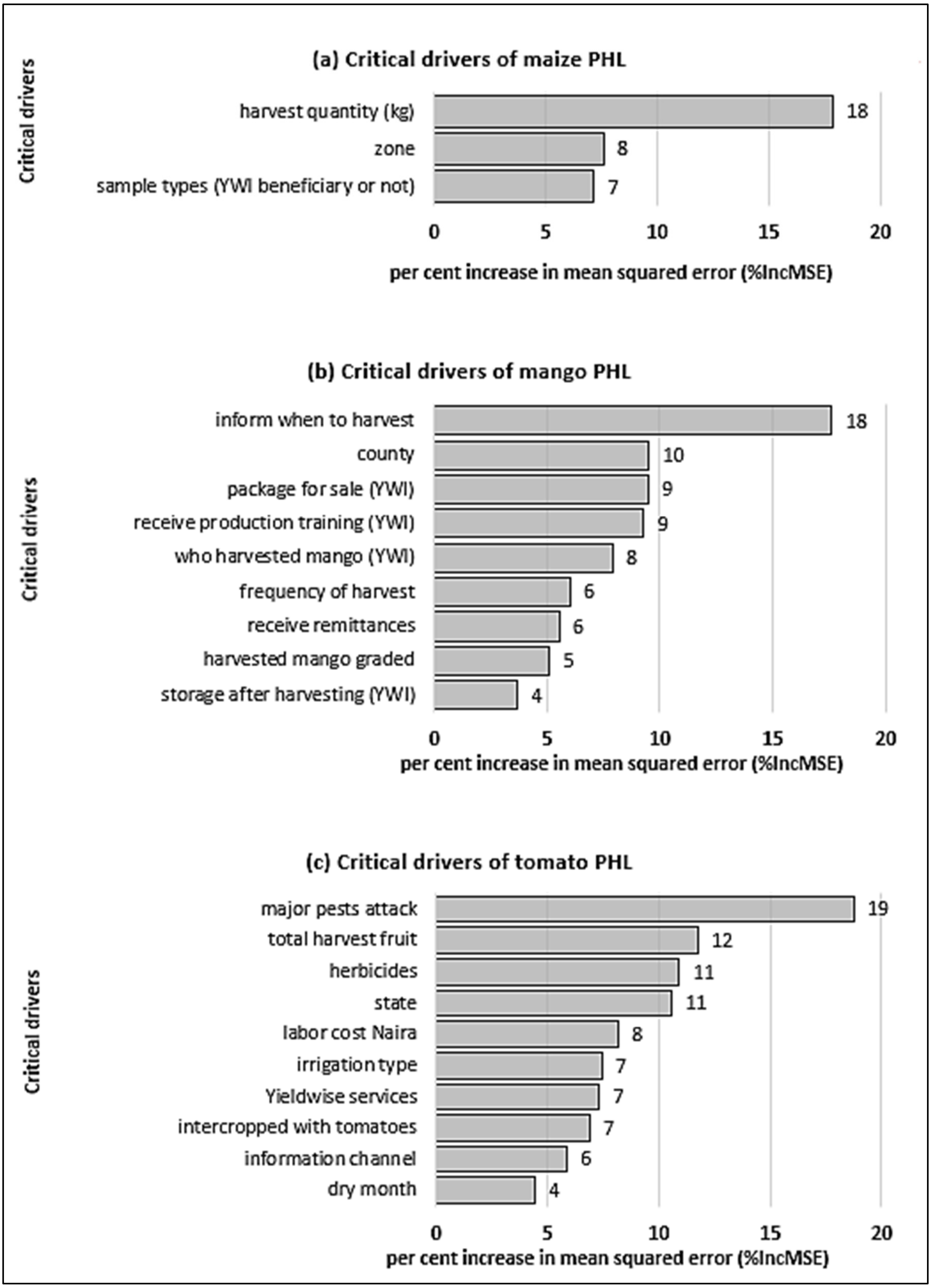

Figure 3. Critical drivers of PHL.

Lastly, two types of critical drivers were ubiquitous across all three crop value chains (Figure 3). First, the geographic location of the smallholder farmer (SHF) was labeled as "zone" in the maize value chain, "province" in the mango value chain, and "state" in the tomato value chain. Second, the affiliation of a SHF with the YWI defined as the "sample types" in the maize value chain, the "type of packaging" material in the mango value chain, and the "YWI services" in the tomato value chain. These results also indicate that the YWI played an important role as a driver of PHL reduction in the three value chains [43]. While the geographic location of a SHF can be difficult to modify [44], the various YWI services could be further explored to identify combinations that best reduce PHL. For this purpose, an online dashboard was created and made accessible at https://phldashboard.shinyapps.io/phldashboard/ (accessed on 8 May 2023). Screenshots of the dashboard can be found in Appendix A.

3.3. Assessing the Critical Drivers of PHL

The assessment of the critical drivers of PHL was carried out through partial dependence plots (Figure 4). These show the relationship between the most critical driver in each value chain, the PHL of each crop, and the number of SHFs. In the maize value chain, the results reveal that as the quantity of harvested maize increases, typically to more than 1000 kg, the maize PHL decreases, regardless of the change in the number of farmers.

In the mango value chain, counting the number of days after blooming to know when to begin harvest was associated with the least PHL. Incidentally, several studies have used the number of days after blooming to either determine the optimal mango harvest date to mitigate PHL during storage [45] or know when mango fruit develops the best organoleptic characteristics during ripening [46,47]. The partial dependence plot in Figure 4b also reveals that the mango PHL increases as the levels change from "days after blooming" to "fruit size or shape". Moreover, the PHL increases monotonically as the number of farmers at each level changes from 10 to 263. However, the PHL increase due to changes in the levels (from "days after blooming" to "fruit size or shape") is larger than the increase due to changes in the number of farmers (from 10 to 223) at each level. These results suggest that an optimum PHL mitigation practice or technology should be identified first before increasing its adoption among farmers.

In the tomato value chain, *Thysanoptera*, "thrips bugs", and *Emitteri*, "aphids", are equally associated with less PHL than the tomato leaf miner *Tuta absoluta*, "tuta". Incidentally, other studies have reported the lepidoptera "tuta" as a destroyer of tomato plants in seven northern states in Nigeria, mainly due to SHFs lacking knowledge of integrated pest management practices [48,49]. Like the maize and mango value chains, as the levels of the "types of pest attack" change from *Thysanoptera*, "thrips bugs", to *Emitteri*, "aphids", and then to *Tuta absoluta*, "tuta" (Figure 4c), the shift in tomato PHL is more pronounced than when changing the number of farmers at each level.

In addition to the established relationship between the levels of various critical drivers and PHLs, it should be noted that PHL is mainly impacted by the change in the levels of a critical driver rather than the change in the number of SHFs at each level.

Since PHL is a multifaceted issue that requires considering multiple critical drivers at a time, an online dashboard was created to explore the relationship between critical drivers, the PHL of each crop, and the desired number of SHFs. The ability to insert a selected number of farmers in the dashboard allows the user to predict PHLs based on a new, arbitrary number of farmers that are not part of the YWI dataset. The online dashboard is accessible at https://phldashboard.shinyapps.io/phldashboard/ (accessed on 8 May 2023) and is described in Appendix A.

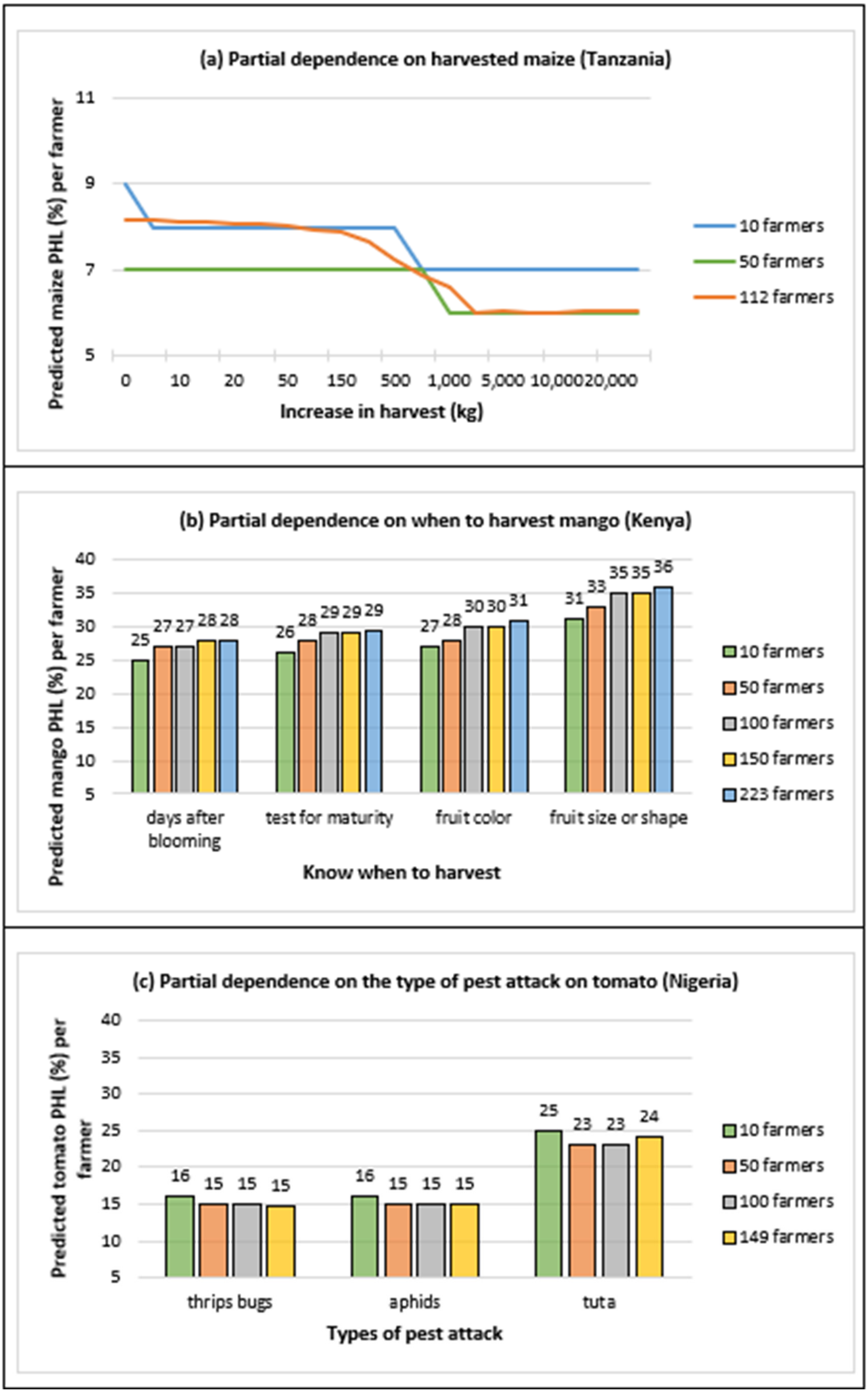

Figure 4. Partial dependence plots for the most critical drivers of PHL.

4. Conclusions

This study used Random Forest modeling to analyze the YWI data collected by the Rockefeller Foundation YieldWise Initiative in three value chains, namely maize in Tanzania, mango in Kenya, and tomato in Nigeria. The following conclusions emerged from this analysis.

1. Three critical drivers of PHL were identified in the maize value chain, nine in the mango value chain, and ten in the tomato value chain. Hence, perishable crops such as tomato and mango have more critical drivers to consider when attempting to reduce PHL than nonperishable crops such as maize.
2. The most critical drivers of PHL were the quantity of maize harvested by a smallholder farmer in the maize value chain, the method used to know when to begin mango harvest in the mango value chain, and the type of pest that attacked the tomato plant in the tomato value chain. It was then noted that the most critical drivers are all related to pre-harvest and harvest activities in the field. Hence, PHL reduction efforts should begin in the field before harvest and continue during harvest.
3. The critical drivers of PHL fall into two categories: passive critical drivers that are difficult to manipulate, such as the geographic area within which a smallholder farmer lives, and active critical drivers that are easier to manage, such as the services provided by the YieldWise Initiative. Moreover, the geographic location of a smallholder farmer and the smallholder farmers' affiliation with the YieldWise Initiative were both ubiquitous drivers across all three value chains.
4. PHL is impacted by changes in the levels of a critical driver as well as changes in the number of smallholder farmers at each level, although the former has a much higher impact. Hence, the optimum PHL mitigation practices or technologies should be identified first before attempting to increase their adoption among smallholder farmers.
5. An online dashboard was created to (a) visually display maize, mango, and tomato PHLs in bar graphs for the numerous variables found in the YieldWise Initiative dataset, (b) rank the critical drivers of maize, mango, and tomato PHL reduction, and (c) explore the relationship between several critical drivers in each value chain, the PHL of each crop, and a desired number of smallholder farmers.

While the data that led to the above conclusions were estimated by farmers and not measured, applying the Random Forest regression algorithm to assess effects across the three different agricultural commodity types is a strength of this research.

Author Contributions: Conceptualization, H.C., D.M. and S.O.; data curation, H.C.; formal analysis, H.C. and S.O.; investigation, H.C., D.M., S.O. and S.S.; methodology, H.C., D.M. and S.O.; software, H.C. and S.O.; visualization, H.C. and D.M.; roles/writing—original draft, H.C.; writing—review and editing, H.C. and D.M.; funding acquisition, D.M. and S.S.; project administration, D.M. and S.S.; resources, D.M. and S.S.; supervision, D.M., S.O. and S.S.; validation, D.M., S.O. and S.S. All authors have read and agreed to the published version of the manuscript.

Funding: Funding for this study was provided by The Rockefeller Foundation (2018 FOD 004), the Foundation for Food and Agriculture Research (DFs-18-0000000008), and the Iowa Agriculture and Home Economics Experiment Station (DD14481).

Data Availability Statement: An online interactive mango PHL dashboard was created at https://phldashboard.shinyapps.io/phldashboard/ (accessed on 8 May 2023) to support this study's results.

Acknowledgments: The authors would like to thank The Rockefeller Foundation, the Foundation for Food and Agriculture Research, and the Iowa Agriculture and Home Economics Experiment Station for providing funding which made this research possible.

Conflicts of Interest: The funders had no role in the design of this study, analyses, or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

Appendix A. YieldWise Initiative PHL Data Dashboard

This online dashboard serves as a quantitative information management tool that

- Provides a visual display of the maize, mango, and tomato postharvest losses (PHLs) in the form of bar graphs as a function of the numerous variables found in the Rockefeller YieldWise Initiative datasets.
- Ranks the critical drivers of the maize, mango, and tomato PHLs in order of importance.
- Predicts the maize, mango, and tomato PHLs as a function of their three most critical drivers as well as the number of smallholder farmers of interest.

The dashboard can be accessed by copying the following link into a web browser: https://phldashboard.shinyapps.io/phldashboard/ (accessed on 8 May 2023). This will open the dashboard, which can be used as explained in Figures A1–A7 below.

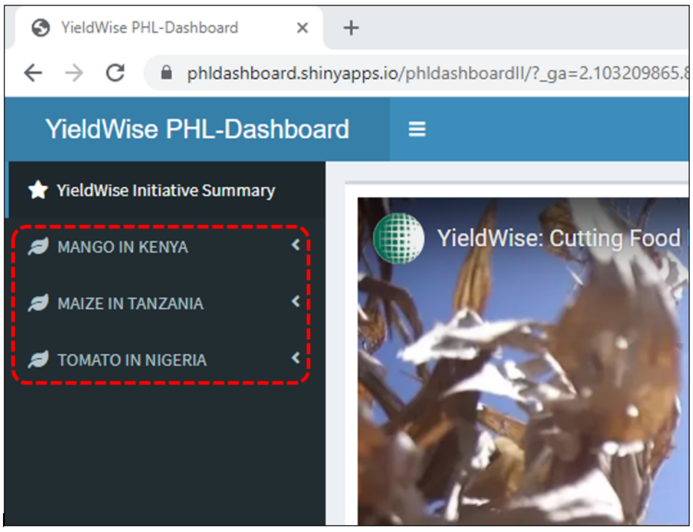

Figure A1. On the left-hand side of the screen is a menu item with the three value chains of the YieldWise Initiative.

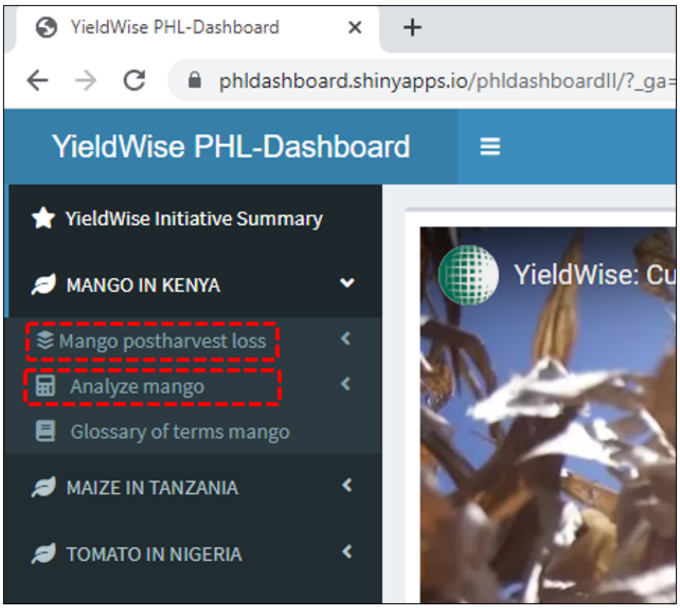

Figure A2. The user can expand a desired value chain by clicking on it. This will allow them to either visualize PHL or analyze PHL.

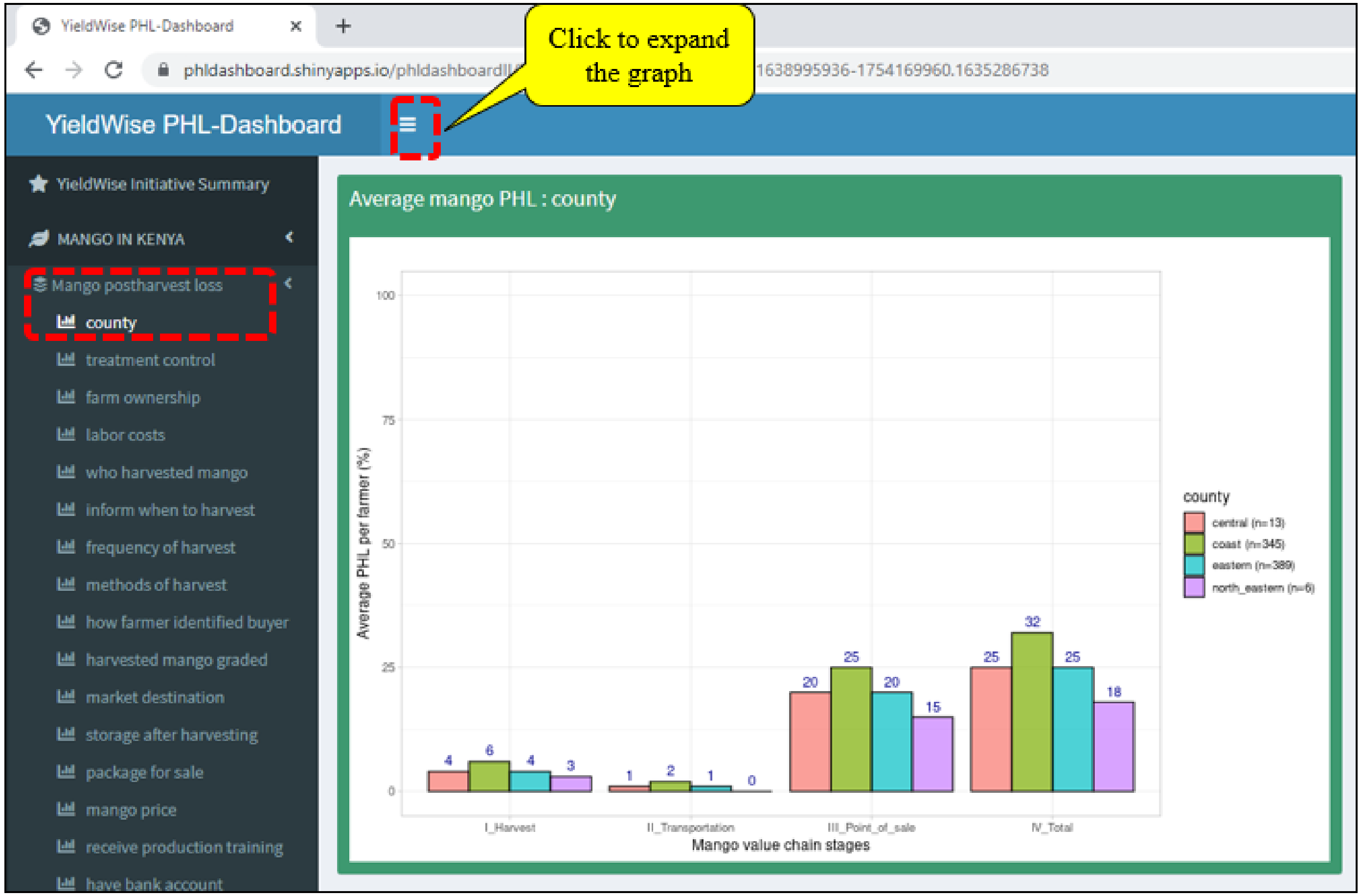

Figure A3. By selecting "Mango postharvest loss" as shown in the figure above, the user will access a list of variables. The user can then select a desired variable such as the "county" variable selected in the figure above as an example. The user will then automatically see a bar graph of various PHL along the value chain, categorized by counties. The number of farmers in each category of the county is also specified in the legend. Finally, the user can collapse the list of variables by clicking on "Mango postharvest loss" one more time.

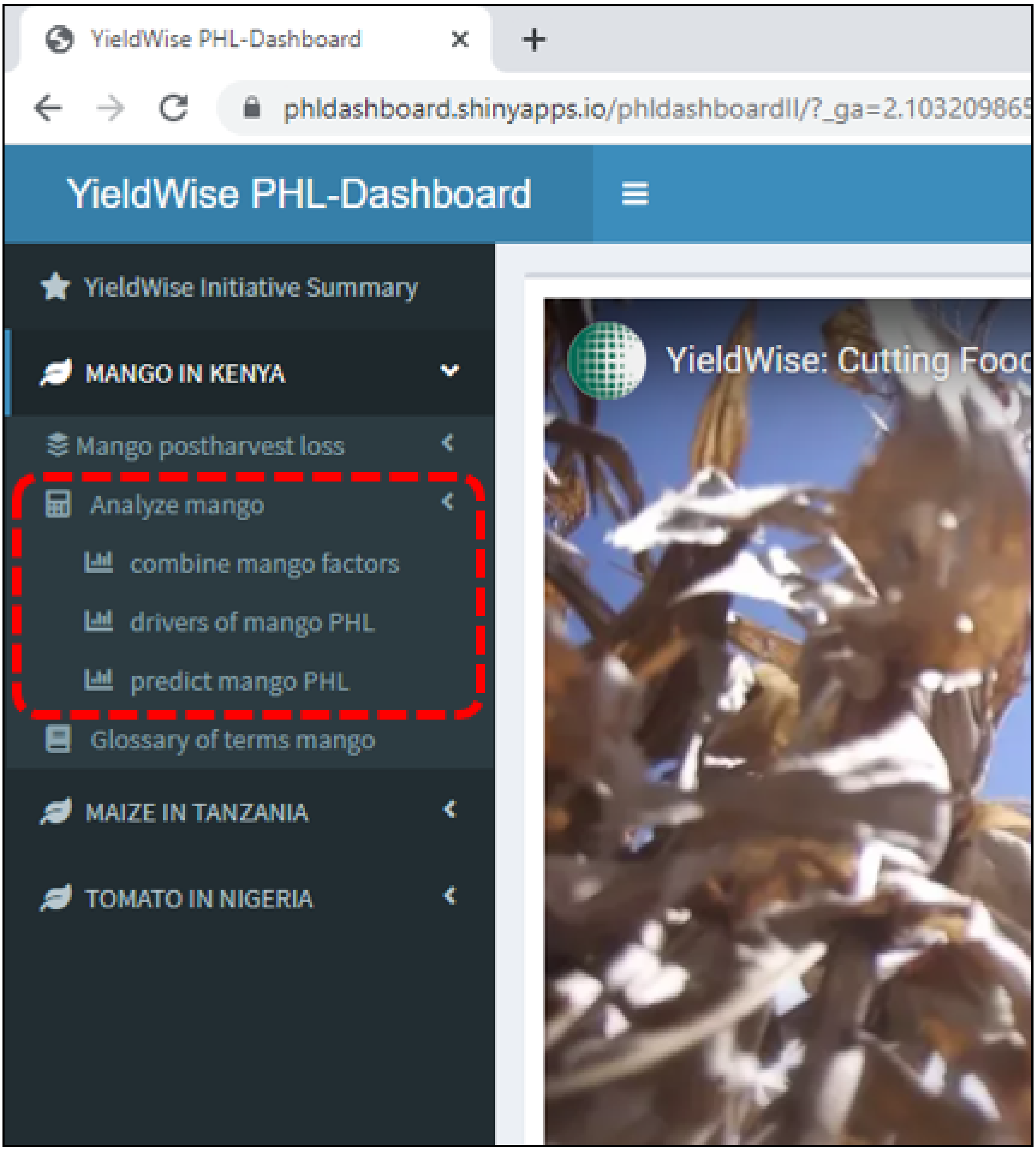

Figure A4. By clicking on "Analyze mango", as shown in the figure above, the user can "combine factors", which allows them to view PHL along the value chain as a function of combining multiple desired variables as opposed to looking at only one variable. Or the user could also select "drivers of mango PHL" to look at the critical drivers of mango PHL ranked in their order of importance, that is, from the most critical driver or variable to the least. Finally, under "Analyze mango", the user can predict a PHL by selecting a given driver of PHL and a desired number of farmers.

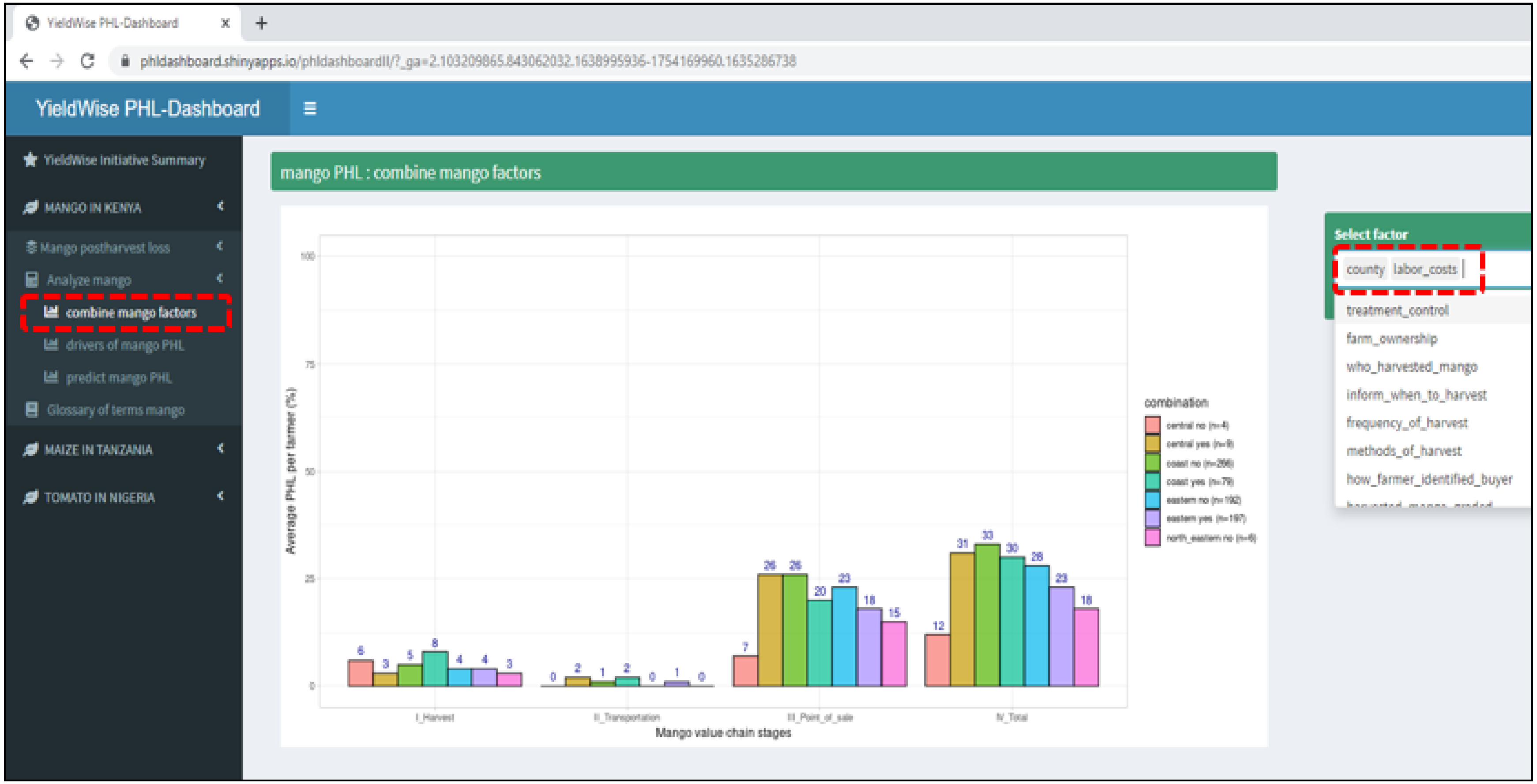

Figure A5. The figure illustrates an example of how the "combine mango factors" is used to visualize PHL along the value chain as a function of combining the "county" AND "labor cost" variables.

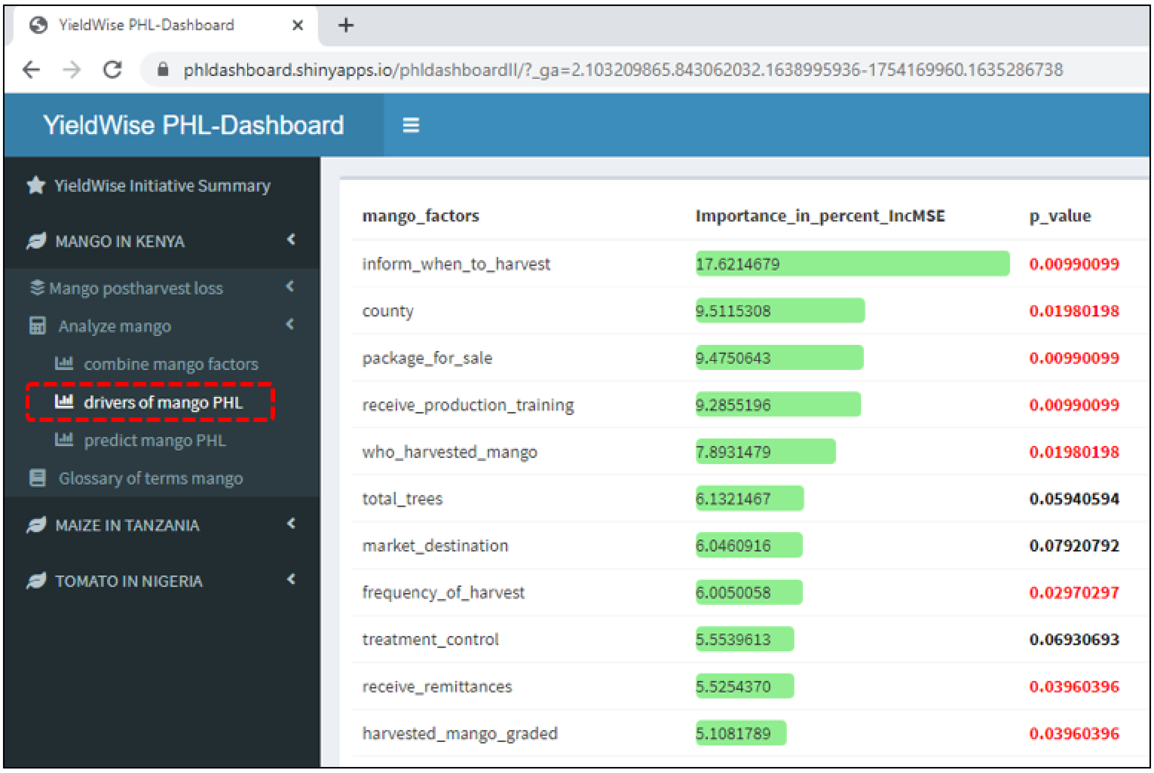

Figure A6. This figure shows the ranking of the drivers of mango PHL, i.e., "mango_factors" in the first column, followed by their relative importance (expressed in percent in mean squared error) in driving PHL, and the p-values in the third column. The critical drivers are factors with $p < 0.05$ (red text in column 3).

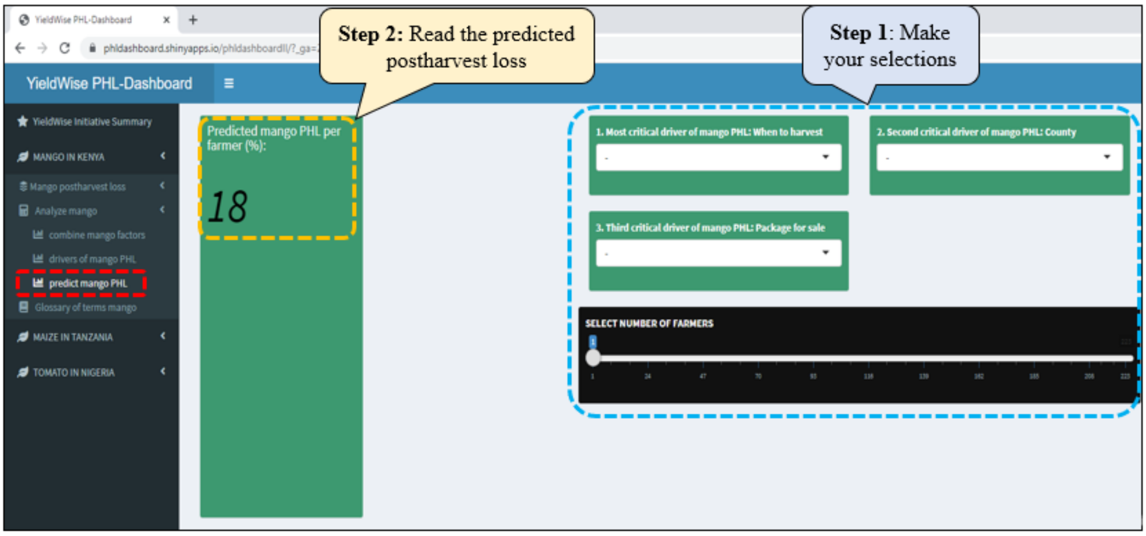

Figure A7. This figure illustrates how to predict PHL. The user will first have to click on the "predict mango PHL" menu item circled in red. Then the user can select the desired subset or level of a given driver, as well as a desired number of farmers circled in blue. Lastly the user will be able to read the predicted PHL that will be displayed in the box circled in yellow based on the previously made selections.

The mango value chain was chosen as an example to illustrate how the dashboard works. The user can conduct the same operations in the maize and tomato value chains on the dashboard.

Appendix B. Variable Importance

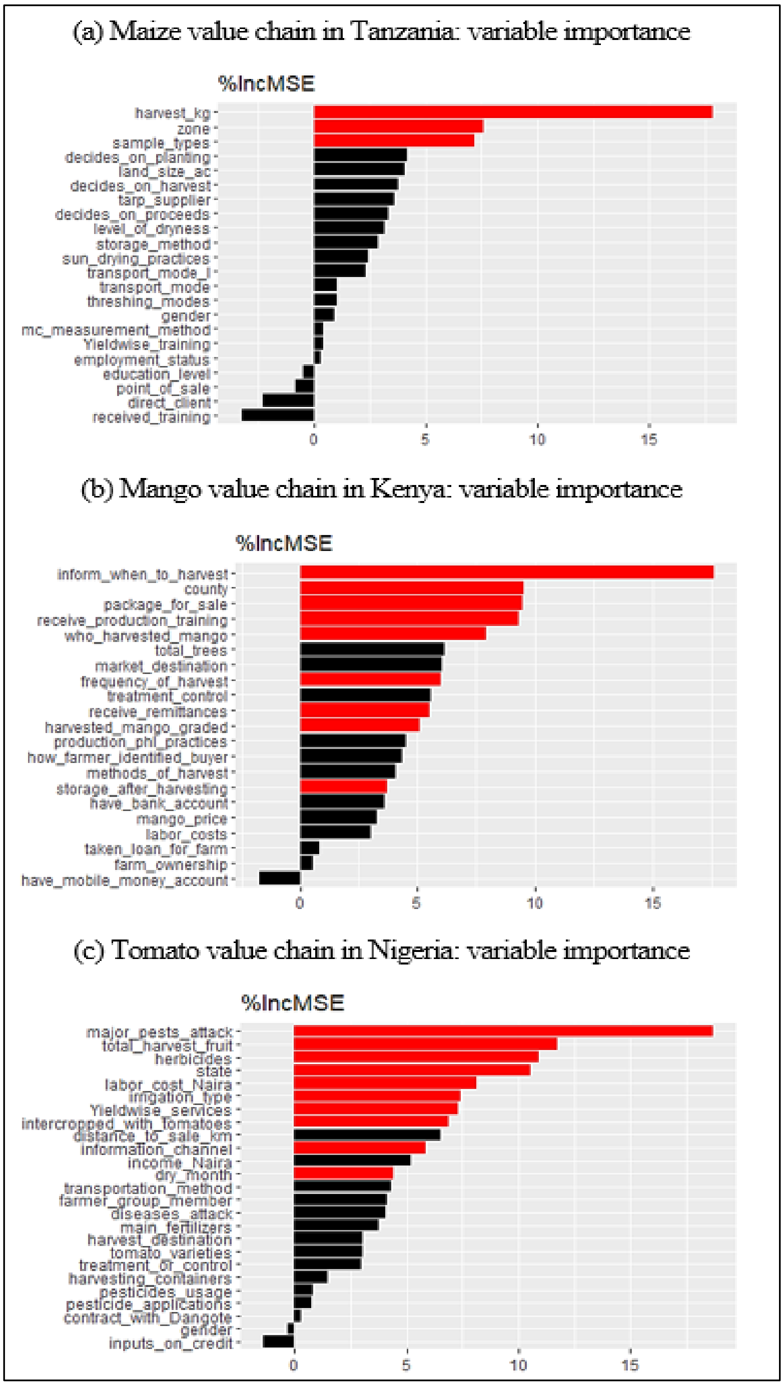

Figure A8. The vertical axis lists all explanatory variables found in each value chain dataset. The horizontal axis shows the %IncMSE, which represents the importance of each explanatory variable. The red bars represent the critical drivers of PHL or the explanatory variables with statistically significant importance ($p < 0.05$).

Appendix C. Partial Dependence Plots

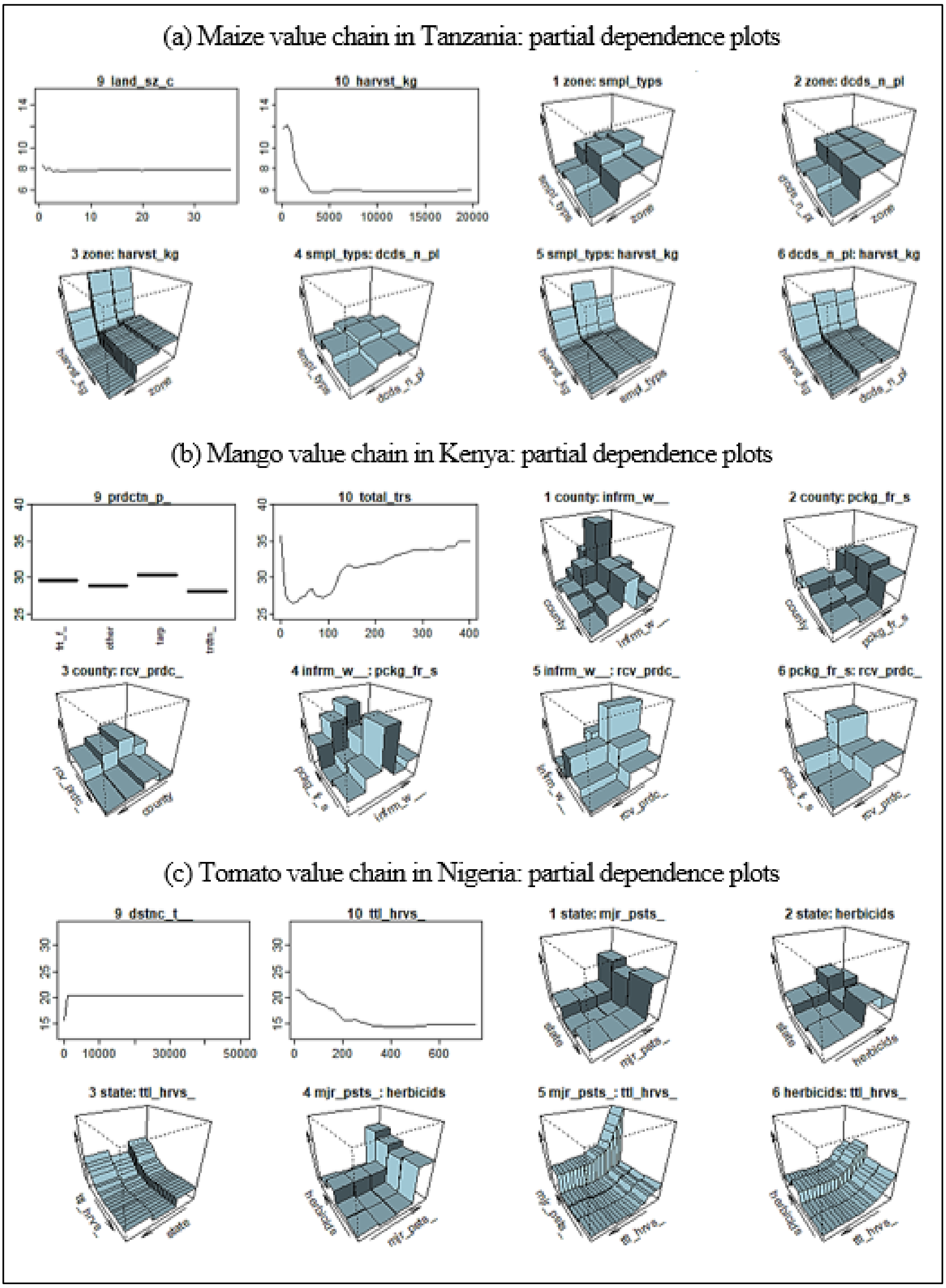

Figure A9. Partial dependence plots (PDP). PDPs illustrate the change in the response variables for given subsets of a critical driver of the highest importance in each value chain. The response variables are averages of the predicted PHLs and are located on the vertical axis. Each exhibit shows two 2D plots of the partial dependence and four 3D plots of the combined relationships between multiple critical drivers of PHL.

References

1. FAO. FAOSTAT. 2021. Available online: https://www.fao.org/faostat/en/#data/QCL (accessed on 24 November 2021).
2. Flanagan, K.; Robertson, K.; Hanson, C. *Reducing Food Loss Setting a Global Action Agenda*; World Resources Institute: Washington, DC, USA, 2019.
3. Owuor, T.O. Guide to Export of Fresh and Processed Mango from Kenya: A Manual for Exporters: August 2015. Available online: https://www.scribd.com/document/432862956/Mango-Export-Guide-Final (accessed on 7 April 2023).
4. The Rockefeller Foundation. Saving Tomatoes for the Sauce. 2021. Available online: https://www.rockefellerfoundation.org/wp-content/uploads/2021/04/YieldWise-Tomato-Overview-V4.pdf (accessed on 7 April 2023).
5. Wilson, R.T.; Lewis, J. The Maize Value Chain in Tanzania. A report from the Southern Highlands Food Systems Programme 2015. Available online: https://www.fao.org/fileadmin/user_upload/ivc/PDF/SFVC/Tanzania_maize.pdf (accessed on 7 April 2023).
6. Mboya, R.; Tongoona, P.; Derera, J.; Mudhara, M.; Langyintuo, A. The dietary importance of maize in Katumba ward, Rungwe district, Tanzania, and its contribution to household food security. *Afr. J. Agric. Res.* **2011**, *6*, 2617–2626.
7. Muoki, P.N.; Makokha, A.O.; Onyango, C.A.; Ojijo, N.K.O. Potential contribution of mangoes to reduction of vitamin A deficiency in Kenya. *Ecol. Food Nutr.* **2009**, *48*, 482–498. [CrossRef] [PubMed]
8. Willcox, J.K.; Catignani, G.L.; Lazarus, S. Tomatoes and cardiovascular health. *Crit. Rev. Food Sci. Nutr.* **2003**, *43*, 1–18. [CrossRef] [PubMed]
9. Grant, W.; Kadondi, E.; Mbaka, M.; Ochieng, S. *Opportunities for Financing the Mango Value Chain: A Case Study of Lower Eastern Kenya*; FSD Kenya: Nairobi, Kenya, 2015; p. 52.
10. Wangithi, C.M.; Muriithi, B.W.; Belmin, R. Adoption and dis-adoption of sustainable agriculture: A case of farmers' innovations and integrated fruit fly management in kenya. *Agriculture* **2021**, *11*, 338. [CrossRef]
11. Suri, T.; Udry, C. Agricultural Technology in Africa. *J. Econ. Perspect.* **2022**, *36*, 33–56. [CrossRef]
12. APHLIS. APHLIS+. 2021. Available online: https://www.aphlis.net/en (accessed on 4 May 2021).
13. Engineering for Change. Landscape analysis of Post-Harvest Technologies for Mango Production in East Africa. 2020. Available online: https://www.engineeringforchange.org/wp-content/uploads/2020/10/landscape_analysis_mango_postharvest_tech.pdf (accessed on 7 April 2023).
14. Ugonna, C.; Jolaoso, M.; Onwualu, A. Tomato Value Chain in Nigeria: Issues, Challenges and Strategies. *J. Sci. Res. Rep.* **2015**, *7*, 501–515. [CrossRef]
15. Sheahan, M.; Barrett, C.B. Food loss and waste in Sub-Saharan Africa: A critical review. *Food Policy* **2017**, *70*, 1–12. [CrossRef]
16. Xie, H.; Perez, N.; Anderson, W.; Ringler, C.; You, L. Can Sub-Saharan Africa feed itself? The role of irrigation development in the region's drylands for food security. *Water Int.* **2018**, *43*, 796–814. [CrossRef]
17. Flanagan, K.; Lipinski, B.; Goodwin, L. SDG Target 12.3 on Food Loss and Waste: 2019 Progress Report. 2019, pp. 1–16. Available online: https://champions123.org/sites/default/files/2020-09/champions-12-3-2019-progress-report.pdf (accessed on 7 April 2023).
18. The Rockefeller Foundation. YieldWise—The Rockefeller Foundation. 2022. Available online: https://www.rockefellerfoundation.org/initiative/yieldwise/ (accessed on 8 January 2022).
19. Chikez, H.; Maier, D.; Sonka, S. Mango postharvest technologies: An observational study of the yieldwise initiative in Kenya. *Agriculture* **2021**, *11*, 623. [CrossRef]
20. Best, K.B.; Gilligan, J.M.; Baroud, H.; Carrico, A.R.; Donato, K.M.; Ackerly, B.A.; Mallick, B. Random forest analysis of two household surveys can identify important predictors of migration in Bangladesh. *J. Comput. Soc. Sci.* **2020**, *4*, 77–100. [CrossRef]
21. Finlay, S. *Predictive Analytics, Data Mining and Big Data*; Springer: Berlin/Heidelberg, Germany, 2014. [CrossRef]
22. Shmueli, G. To explain or to predict? *Stat. Sci.* **2010**, *25*, 289–310. [CrossRef]
23. Kern, C.; Klausch, T.; Kreuter, F. Tree-based machine learning methods for survey research. *Surv. Res. Methods* **2019**, *13*, 73–93. [PubMed]
24. Kuhn, M.; Johnson, K. *Applied Predictive Modeling*; Springer Science Business Media: New York, NY, USA, 2013. [CrossRef]
25. Hengsdijk, H.; de Boer, W.J. Post-harvest management and post-harvest losses of cereals in Ethiopia. *Food Secur.* **2017**, *9*, 945–958. [CrossRef]
26. Criminisi, A.; Shotton, J. *Decision Forests for Computer Vision and Medical Image Analysis*; Springer Science Business Media: New York, NY, USA, 2013.
27. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
28. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning*; 8th Printing 2017; Springer: New York, NY, USA, 2013; Volume 102. [CrossRef]
29. Gholamy, A.; Kreinovich, V.; Kosheleva, O. Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation. *Comput. Sci.* **2018**, 1–6.
30. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **2008**, *28*, 1–26. [CrossRef]
31. Williams, G. *Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery*; Springer: New York, NY, USA, 2011. [CrossRef]
32. Perner, P. Machine Learning and Data Mining in Pattern Recognition. In Proceedings of the 7th International Conference, MLDM 2011, New York, NY, USA, 30 August–3 September 2011; Volume 2961.

33. Couronné, R.; Probst, P.; Boulesteix, A.-L. Random forest versus logistic regression: A large-scale benchmark experiment. *BMC Bioinform.* **2018**, *19*, 27. [CrossRef]
34. Grömping, U. Variable importance assessment in regression: Linear regression versus random forest. *Am. Stat.* **2009**, *63*, 308–319. [CrossRef]
35. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
36. Greenwell, B.M. pdp: An R package for constructing partial dependence plots. *R J.* **2017**, *9*, 421–436. [CrossRef]
37. Milborrow, S. Plotting Regression Surfaces with Plotmo Stephen Milborrow. 2019. Available online: http://www.milbo.org/doc/plotmo-notes.pdf (accessed on 7 April 2023).
38. Itaoka, K. Regression and Interpretation of Low R-squared. In Proceedings of the Social Research Network 3rd Meeting, Noosa, Australia, 12–13 April 2012.
39. Hurburgh, C.R., Jr.; Bern, C.J.; Wilcke, W.F.; Anderson, M.E. Shrinkage and Corn Quality Changes in on-Farm Handling Operations. *Trans. ASAE* **1983**, *26*, 1854–1857. [CrossRef]
40. Kiaya, V. Post-Harvest Losses and Strategies to Reduce Them. 2014. Available online: https://www.actioncontrelafaim.org/wp-content/uploads/2018/01/technical_paper_phl__.pdf (accessed on 7 April 2023).
41. Weston, L.A.; Barth, M.M. Preharvest factors affecting postharvest quality of vegetables. *HortScience* **1997**, *32*, 812–816. [CrossRef]
42. Mahuku, G.; Nzioki, H.S.; Mutegi, C.; Kanampiu, F.; Narrod, C.; Makumbi, D. Pre-harvest management is a critical practice for minimizing aflatoxin contamination of maize. *Food Control.* **2019**, *96*, 219–226. [CrossRef]
43. Sonka, S.; Lee, H.; Shah, S. The YieldWise Approach to Post-Harvest Loss Reduction: Creating Market-Driven Supply Chains to Support Sustained Technology Adoption. *Agriculture* **2023**, *13*, 910. [CrossRef]
44. Delgado, L.; Schuster, M.; Torero, M. *The Reality of Food Losses: A New Measurement Methodology*; no. IFPRI Discussion Paper 01686; International Food Policy Research Institute: Washington, DC, USA, 2017; p. 40.
45. Mathieu, L.; Jacques, J. Quality and maturation of mango fruits of cv. Cogshall in relation to harvest date and carbon supply. *Aust. J. Agric. Res.* **2006**, *57*, 419–426.
46. Dick, E.; N'DaAdopo, A.; Camara, B.; Moudioh, E. Influence of maturity stage of mango at harvest on its ripening quality. *Fruits* **2009**, *64*, 13–18. [CrossRef]
47. Stathers, T.; Holcroft, D.; Kitinoja, L.; Mvumi, B.M.; English, A.; Omotilewa, O.; Kocher, M.; Ault, J.; Torero, M. A scoping review of interventions for crop postharvest loss reduction in sub-Saharan Africa and South Asia. *Nat. Sustain.* **2020**, *3*, 821–835. [CrossRef]
48. Borisade, O.A.; Kolawole, A.O.; Adebo, G.M.; Uwaidem, Y.I. The tomato leafminer (*Tuta absoluta*) (Lepidoptera: Gelechiidae) attack in Nigeria: Effect of climate change on over-sighted pest or agro-bioterrorism? *J. Agric. Ext. Rural Dev.* **2017**, *9*, 163–171. [CrossRef]
49. Bala, I.; Mukhtar, M.M.; Saka, H.K.; Abdullahi, N.; Ibrahim, S.S. Determination of insecticide susceptibility of field populations of tomato leaf miner (*Tuta absoluta*) in northern Nigeria. *Agriculture* **2019**, *9*, 7. [CrossRef]