Analytes prediction using Machine Learning algorithms with Near-Infrared reflectance spectra (NIR) data of Cacao almonds and leaves,

Objective

The objectives of this study are:

- Develop and evaluate **Machine Learning (ML)** models with NIR spectroscopy data to predict analytes in cocoa leaves and almonds.
- Comparing predictions between **ML** models and select the best option based on goodness-of-fit metrics.

Data

- There are 2 databases (**almonds and leaves**) that contain the raw spectra collected from the NIRS device.

- The dataset for **almonds** has 626 samples collected, while the dataset for **leaves** has a total of 388 samples collected.

  Variables

  - Both datasets (almonds and leaves) contain 700 columns (or variables) which are the spectra collected for each sample at a given wavelength measured in 2 nm intervals.

  - For almonds, there are 14 analytes (Na, Mg, Al, P, K, Ca, Cr, Mn, Fe, Co, Cu, Zn, Cd, Pb) as target variables , and, for leaves, there are 18 analytes (B, Na, Mg, Al, P, S, K, Ca, Cr, Mn, Fe, Co, Ni, Cu, Zn, As, Mo, and Cd) as target variables.

Methodology

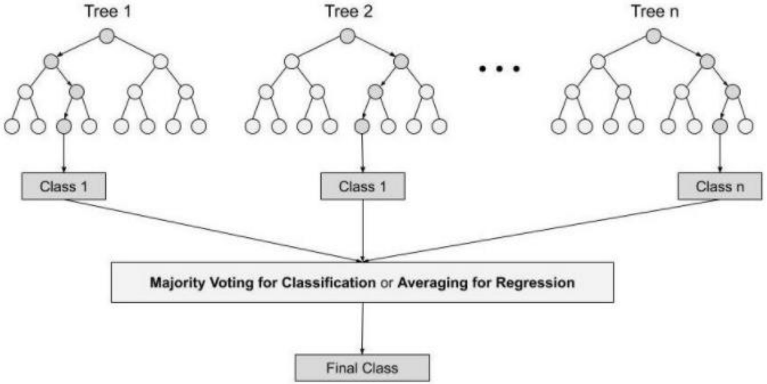
The methodology pipeline has **three phases**:

- Model training: cross-validation and model testing of **ML** algorithms with **raw data**: Random Forest, XGboost and Support Vector Machine (SVM).
    - Data were split into training (70%) and test (30%).
- Algorithm selection: choose the best fit on the test data, according to traditional metrics of goodness-of-fit (RMSE and R-squared).
- **Transform data** into samples with **linear detrending and polynomial** (quadratic or cubic) detrending and estimate, with the best ML algorithm, new models with these data.

Methodology

Random Forest (RF)

- RF is an ensemble algorithm that builds decision trees on different samples and takes their majority vote for classification, and average in case of regression.
    - Ensemble simply means combining multiple models, Thus, a collection of models is used to make predictions rather than an individual model.
- To create the decision trees, this algorithm creates a different training subset from sample training data with replacement (**Bagging**) and the final output is based on **majority voting.**


Tree 1
Tree 2
Tree n
Class 1
Class 1
Class n
Majority Voting for Classification or Averaging for Regression
Final Class

Source: Analytics Vidhia

Methodology

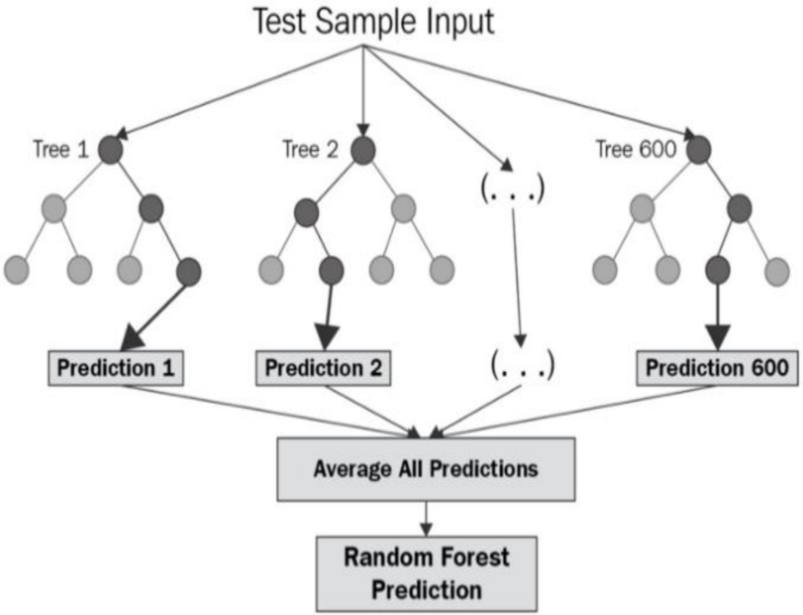
Random Forest (RF)

- In a **regression context** the RF algorithm works in the same way.However, each decision tree looks at **MSE** (mean squared error) as its objective or ***cost function***, to be minimized.

- The cost function looks as follows:

$$\min\left(\frac{1}{\mathrm{n}}\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2\right)$$

- The final output is based on **average all predictions.**



Methodology

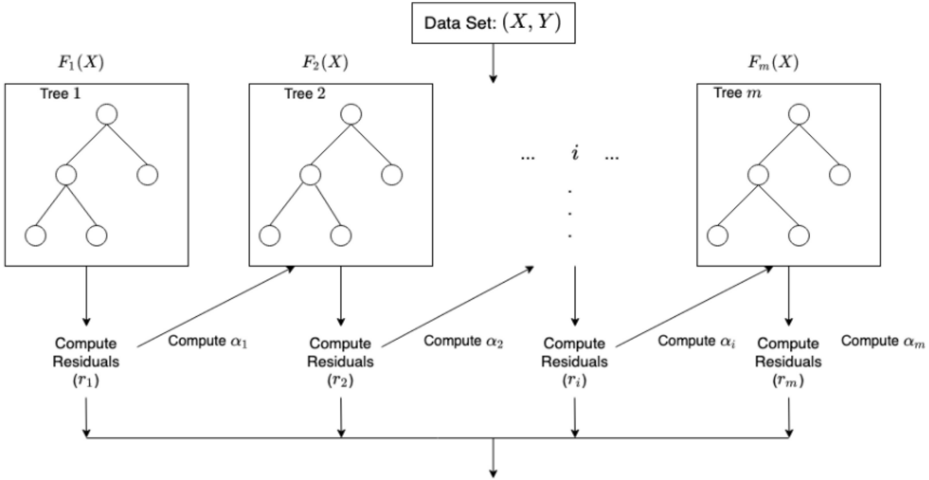
XGboost

- XGboost is an **ensemble learning** algorithm based on decision trees too, but, unlike **RF**, a **boosting** method is used to build the decision trees.
  - Ensemble learning offers a systematic solution to combine the predictive power of multiple learners, The resultant is a single model which gives the aggregated output from several models.
- In **boosting**, the **trees are built sequentially** such that each subsequent tree aims to reduce the errors of the previous tree, Each tree learns from its predecessors and updates the residual errors. Hence, the tree that grows next in the sequence will learn from an updated version of the residuals.

Methodology

XGboost

- In a **regression context,** XGboost minimizes a normalized objective function that combines the loss function (**based on MSE**) and a penalty term for model complexity. The process looks as follow:



$$F_m(X) = F_{m-1}(X) + \alpha_m h_m(X, r_{m-1})$$

- Where Escriba aquí la ecuación., $\mathrm{F}(X)$ represents the vector with the final predictions, so, the ***cost function*** is: $\min\left(\sum_{i=1}^{m} L(Y_i, F(X))\right)$

Source: Amazon Web Services

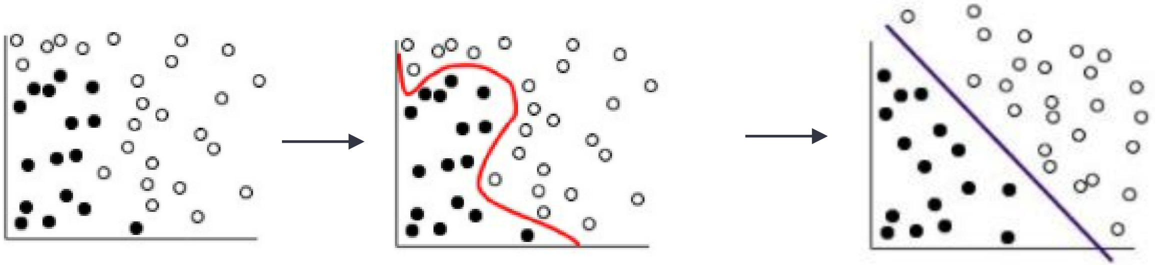
Methodology

Support Vector Machine (SVM)

- **SVM** works by mapping data to a high-dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable.

- A separator between the categories is found, then the data are transformed in such a way that the separator could be drawn as a hyperplane.

- Characteristics of new data can be used to predict the group to which a new record should belong.



Original Dataset,

The two categories can be separated with a curve,

After the transformation, the boundary between the two categories can be defined by a **hyperplane**,

- The mathematical function used for the transformation is known as the kernel function and can be linear, polynomial, sigmoid and others functional forms.

Source: IBM

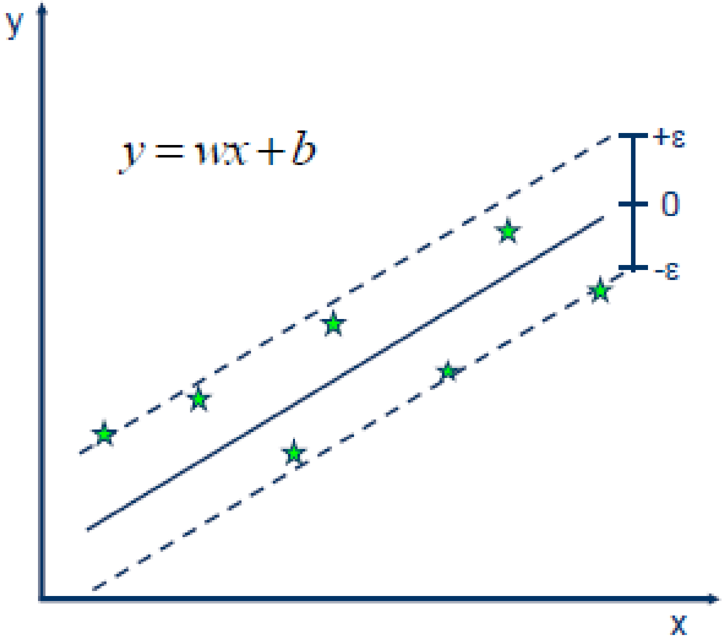
Methodology

Support Vector Machine (SVM)

- **In regression, SVM** algorithm uses the same principles as the SVM for classification, with only few differences:
    - A margin of tolerance (**epsilon**) is set for the estimation.
    - With the epsilon set, a **decision boundary** is established.
    - The objective is to basically consider the points tha are within the decision boundary line, The best fit line is the hyperplane that has a maximum numbe of points, i,e,, it minimizes the error, keep in mind that part of the error is tolerated.

The **cost function** looks as follows:

$$\min\left(\frac{1}{2}\,||w||^2 + c\,\sum_{i=1}^{N}(\xi_i + \xi_i^*)\right)$$



Source: IBM

Results

Almonds Raw Data

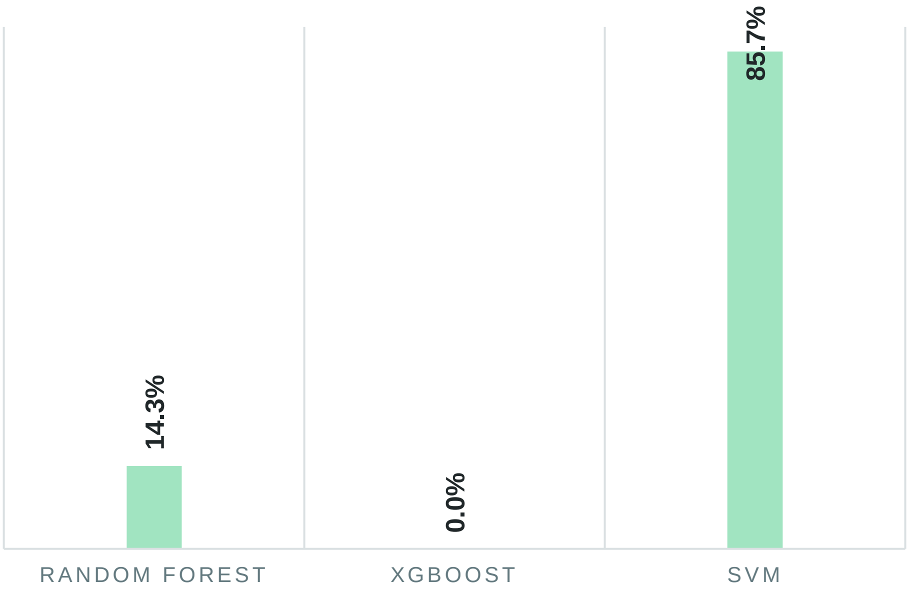
Table 1: Metrics performance by ML algorithms with test data

Component	Random Forest		XGboost		SVM	
	RMSE	R-squared	RMSE	R-squared	RMSE	R-squared
	Test	Test	Test	Test	Test	Test
Na	114,4	0,17	114,5	0,16	116,4	0,13
Mg	445,9	0,07	441,5	0,08	439,5	0,09
K	1520,2	0,34	1519,4	0,33	1281,1	0,52
Ca	268,9	0,35	288,8	0,25	266	0,36
Al	9,5	0,49	9,7	0,46	9,09	0,53
P	877,5	-0,01	892	-0,04	701,7	0,35
Cr	0,15	0,23	0,13	0,38	0,12	0,48
Mn	10,8	0,07	11	0,03	10,3	0,14
Fe	8,5	0,2	8,3	0,23	7,5	0,36
Co	0,61	-0,03	0,6	0,01	0,6	0,02
Cu	7,8	0,04	7,5	0,1	7,09	0,21
Zn	11,6	0,18	11,8	0,14	11,16	0,23
Cd	1,32	-0,07	1,2	-0,02	1,2	0,01
Pb	0,67	0,31	0,68	0,3	0,69	0,27

Results

Almonds Raw Data

Figure 1: Best algorithm to predict analytes



- The best algorithm to predict almonds analytes is the SVM.
- 85,7% of the analytes have a better prediction fit with the SVM, while the remaining 14,3% corresponds to the Random Forest algorithm.

Results

Leaves Raw Data

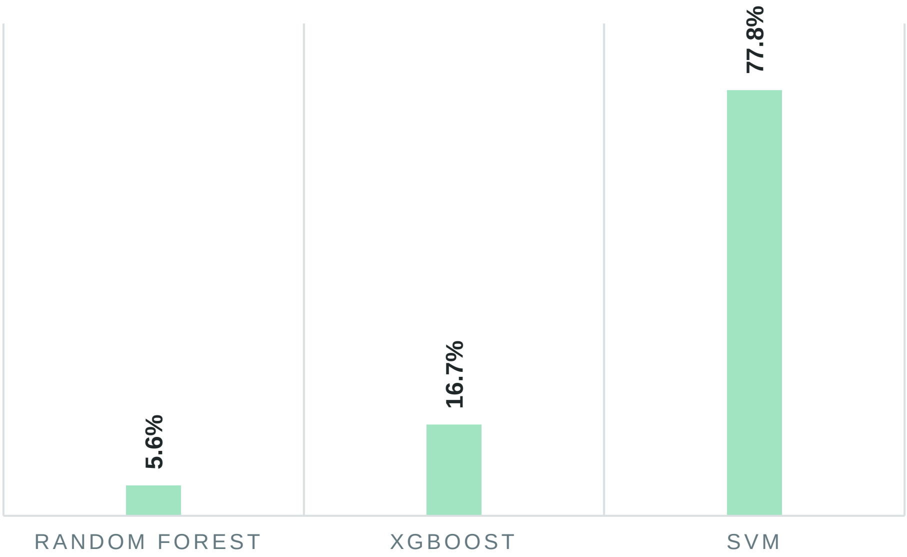
Table 2: Metrics performance by ML algorithms with test data

Component	Random Forest		Xgboost		SVM	
	RMSE	R-squared	RMSE	R-squared	RMSE	R-squared
	Test	Test	Test	Test	Test	Test
Al	79,1	0,11	76,9	0,15	68,9	0,32
As	0,06	-0,5	0,04	-0,03	0,05	-0,2
B	12,3	-0,03	11,9	0,02	10,05	0,31
Ca	4619,5	0,35	4388,7	0,42	3940	0,53
Cd	4,11	-0,23	3,7	-0,02	3,43	0,14
Co	6,6	0,05	6,7	0,02	6,35	0,12
Cr	0,32	0,31	0,29	0,44	0,25	0,58
Cu	13,3	-0,16	11,9	0,05	12,5	-0,03
Fe	126,9	0,25	121,5	0,31	118,9	0,34
K	3988,9	0,36	3650,3	0,47	1565	0,9
Mg	1729,9	0,19	1521	0,37	682	0,87
Mn	341,5	0,09	337	0,12	292	0,33
Mo	1,06	-0,05	1,03	-0,00	0,96	0,13
Na	146,7	-0,03	140,7	0,05	135,4	0,12
Ni	18,22	0,04	18,3	0,02	19,3	-0,07
P	442,3	0,22	434,4	0,25	418	0,3
S	515,2	-0,01	507,5	0,01	416,2	0,34
Zn	84,11	-0,01	80	0,08	81,6	0,04

Results

Leaves Raw Data

Figure 2: Best algorithm to predict analytes



- The best algorithm to predict leaves analytes is the SVM.
- 77,8% of the analytes have a better prediction fit with the SVM, while the remaining 22,2% corresponds to the Random Forest (5,6%) and Xgboost (16,7%) algorithms.

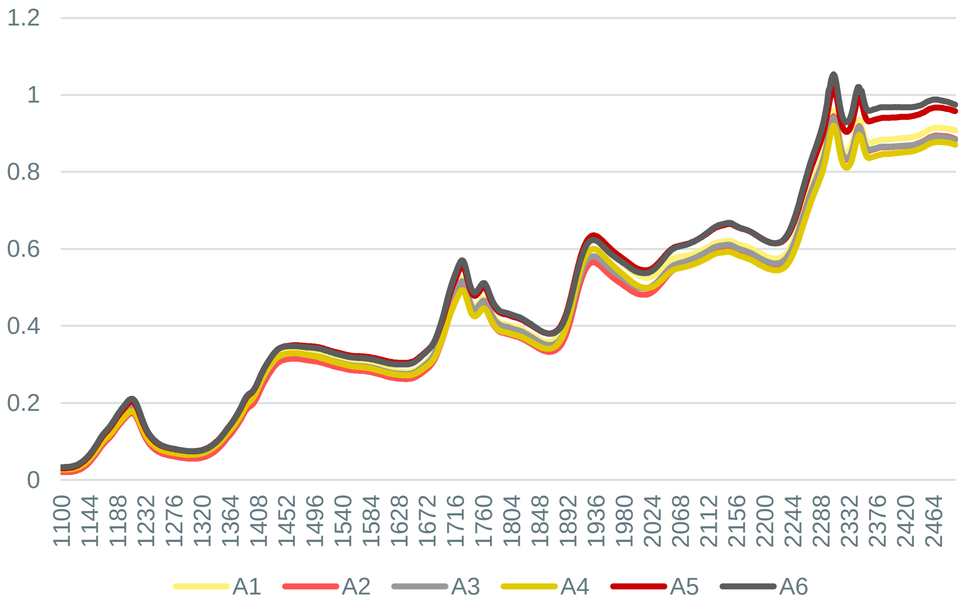
Results

Data transformation: polynomial detrending

- Due to the structure of the **wavelength measured data** for the analytes, we perform a polynomial detrending.

- We tested two versions: **linear** and **cubic** detrending transformation.

Figure 3: Example of wavelength measured data (NA in almonds)



Results

Almonds Detrending Data

Table 3: SVM metrics performance with test detrending data

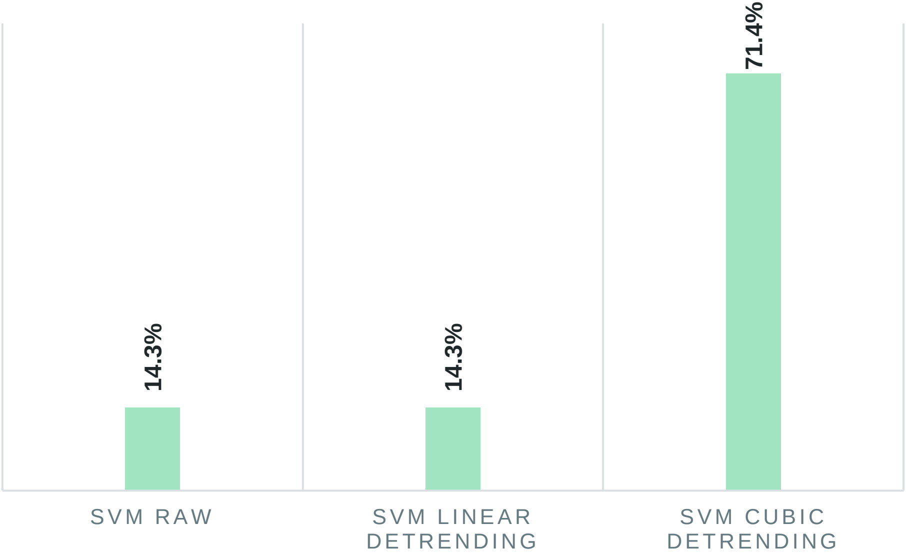
Component	SVM raw data		SVM Linear Dt		SVM Cubic Dt	
	RMSE	R-squared	RMSE	R-squared	RMSE	R-squared
	Test	Test	Test	Test	Test	Test
Na	116,40	0,10	110,20	0,22	107,36	0,27
Mg	439,50	0,10	406,80	0,22	404,08	0,24
K	1281,10	0,50	1248,50	0,55	1247,24	0,55
Ca	266,00	0,40	265,20	0,36	264,11	0,37
Al	9,10	0,50	8,60	0,57	8,64	0,58
P	701,70	0,40	696,50	0,36	691,69	0,37
Cr	0,10	0,50	0,12	0,48	0,13	0,47
Mn	10,30	0,10	10,30	0,16	10,43	0,14
Fe	7,50	0,40	7,50	0,37	7,46	0,38
Co	0,60	0,00	0,59	0,03	0,60	0,02
Cu	7,10	0,20	6,90	0,24	6,86	0,26
Zn	11,20	0,20	10,90	0,26	10,89	0,27
Cd	1,20	0,00	1,26	0,01	1,26	0,02
Pb	0,70	0,30	0,69	0,28	0,68	0,30

Results

Almonds Detrending Data

- The results in Figure 2 compare the fit of the models with detrending data using the best algorithm using raw data, i,e,, the results with the SVM algorithm.

Figure 2: Best algorithm to predict analytes with detrending data



- The best algorithm is the SVM with cubic detrended data.
- 71,4% of the analytes have a better prediction fit with the cubic detrended data, while the remaining 28,6 % corresponds to raw data (14,3%) and linear detrend (14,3%).

Results

Leaves Detrending Data

Table 4: SVM metrics performance with test detrending data

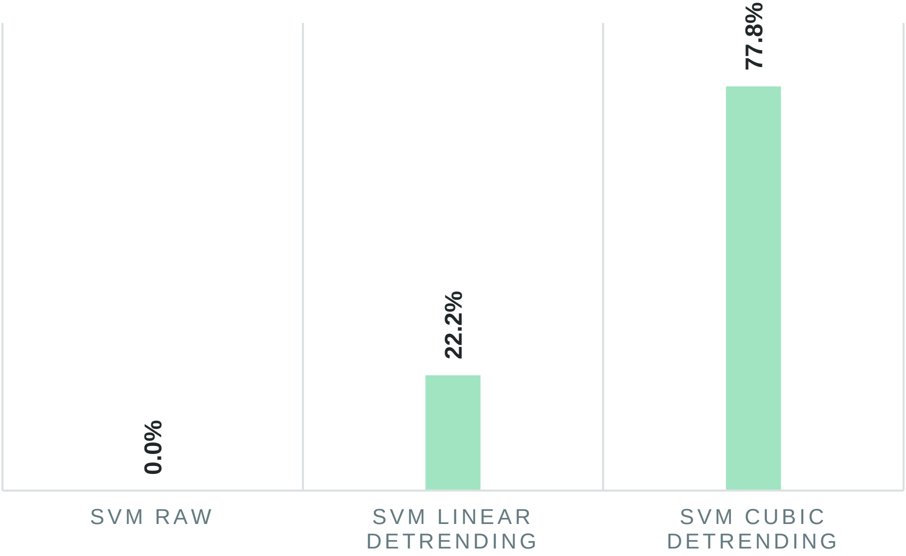
Component	SVM Raw Data		SVM Linear Dt		SVM Cubic Dt	
	RMSE	R-squared	RMSE	R-squared	RMSE	R-squared
	Test	Test	Test	Test	Test	Test
Al	68,9	0,32	64,7	0,4	62,8	0,44
As	0,05	-0,2	0,05	-0,36	0,05	-0,28
B	10,05	0,31	9,24	0,41	9,21	0,42
Ca	3940	0,53	1332	0,95	1301,65	0,95
Cd	3,43	0,14	3,41	0,15	3,42	0,15
Co	6,35	0,12	6,15	0,18	6,25	0,16
Cr	0,25	0,58	0,25	0,57	0,25	0,58
Cu	12,5	-0,03	12,5	-0,02	12,47	-0,02
Fe	118,9	0,34	110,7	0,43	108,52	0,45
K	1565	0,9	1502,45	0,91	1650,54	0,89
Mg	682	0,87	676,66	0,88	670,81	0,88
Mn	292	0,33	247,29	0,53	245,56	0,53
Mo	0,96	0,13	0,9497	0,16	0,9533	0,15
Na	135,4	0,12	121	0,3	118,87	0,32
Ni	19,3	-0,07	17,52	0,12	17,02	0,17
P	418	0,3	395,18	0,38	393,74	0,39
S	416,2	0,34	378	0,46	376,71	0,46
Zn	81,6	0,04	82,04	0,04	81,83	0,04

Results

Leaves Detrending Data

- The results in Figure 4 compare the fit of the algorithms with detrended data using the best choice, i,e,, the results with the SVM algorithm.

Figure 4: Best algorithm to predict analytes with detrending data



- The best model is the SVM with cubic detrend data,
- 77,8% of the analytes have a better prediction fit with the cubic detrend data, while the remaining 22,2 % corresponds to SVM with linear detrend data,

Results

PLS technique

- Following Castillo et al, (2020), Yuming Guo et al, (2021) and Salehi et al, (2021), we compare the predictions from ML with the algorithms developed with Principal Component Analysis (PCA) technique (traditional statistical method) combined with **Partial Least Squares (PLS)** methods to predict the analytes.

- The following tables show the results with the PLS and SVM techniques with detrended data of the almonds and leaves datasets.

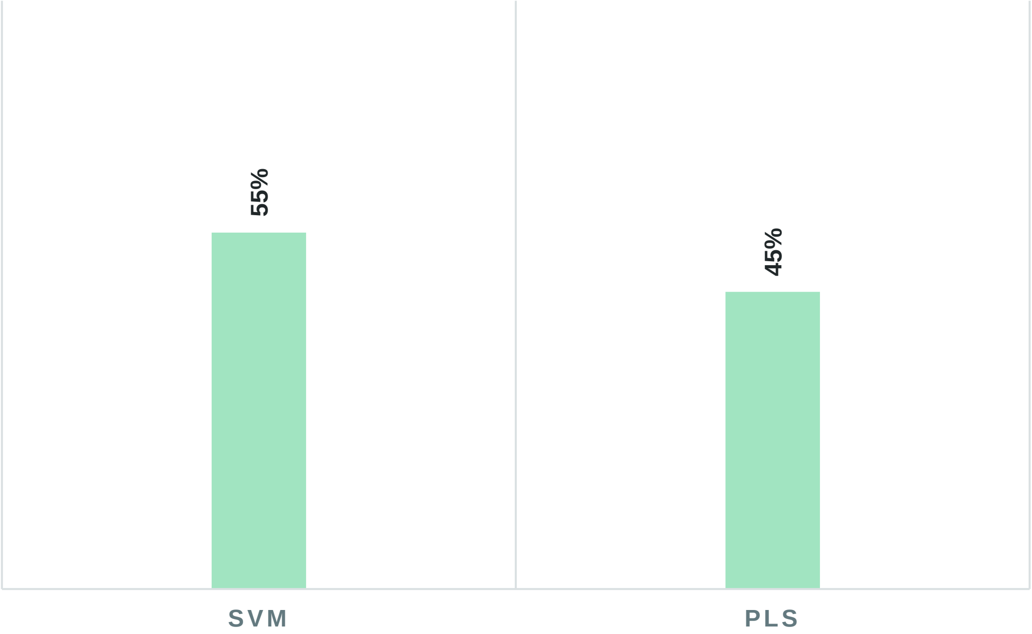
Results

Almonds PLS and SVM technique

Table 5: Metrics performance with test detrended data

Component	R- Squared	
	PCA	SVM
Al	0,45	0,58
Ca	0,39	0,37
Cu	0,20	0,26
Fe	0,43	0,38
K	0,45	0,55
Mg	0,23	0,24
Mn	0,20	0,14
Na	0,41	0,27
P	0,40	0,37
Pb	0,17	0,30
Zn	0,20	0,27

Figure 5: Best technique to predict analytes with detrended data



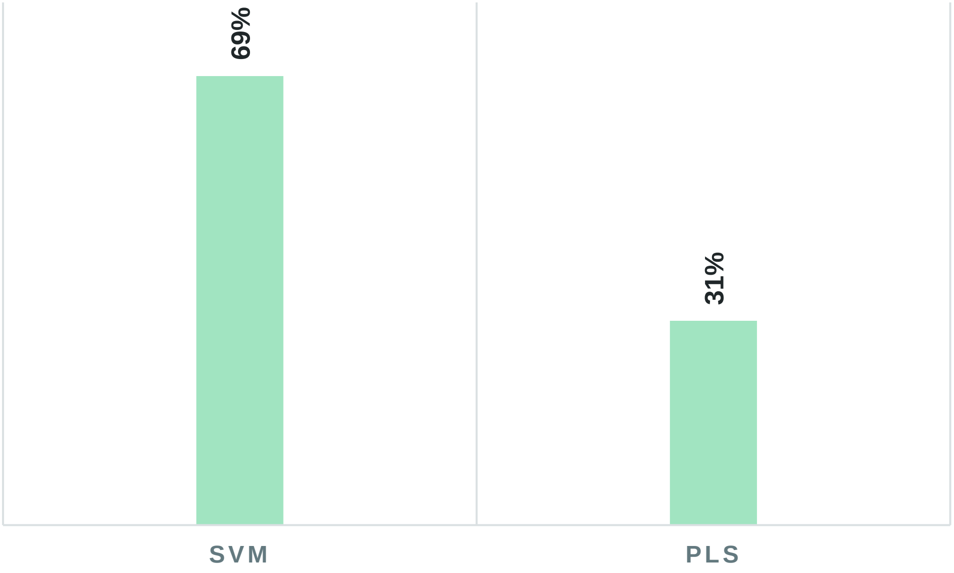
Results

Leaves PLS and SVM technique

Table 6: Metrics performance with test detrended data

Component	R- Squared	
	PCA	SVM
Al	0,34	0,44
B	0,35	0,42
Ca	0,93	0,95
Cd	0,20	0,15
Co	0,18	0,16
Cr	0,48	0,58
Fe	0,35	0,45
K	0,90	0,89
Mg	0,90	0,88
Mn	0,53	0,53
Mo	0,07	0,15
Na	0,39	0,32
Ni	0,06	0,17
P	0,30	0,39
S	0,37	0,46
Zn	0,03	0,04

Figure 6: Best technique to predict analytes with detrended data



Results

PLS and SVM : ML Forecast dominance with detrended data

- The ML dominance ratio compares the results from the SVM and PLS R-squared on the test data se. The ratio looks as follows:

$$\frac{\sum_{R2:\,SVM>PLS}(R^2_{svm}-R^2_{pls})}{\sum_{R2:\,SVM<PLS}(|R^2_{svm}-R^2_{pls}|)}$$

- In **Almonds**, the ML algorithm has a better fit 1,74 of the time, while in **Leaves** ML algorithm has a better fit 4,79 of the time.

Table 7: ML dominance - Conditional bias ratio

Dataset	Test
Almonds	1,74
Leaves	4,79

Conclussions

- Three ML algorithms were used to predict the analytes of cocoa almonds and leaves: Random Forest. XGboost and Support Vector Machine (SVM).
    - SVM was the best algorithm to predict almonds and leaves analytes. In almonds, 85,7% of the analytes have a better prediction fit with SVM, while in leaves 77,8% of analytes have a better prediction with this algorithm.

- Better results are obtained by using a **cubic detrending** in the dataset. 71,4% of almonds, and, 77,8% of leaves analytes have a better prediction fit with the SVM cubic detrended data.

- After comparing the predictions between **SVM** and **PLS** algorithms with the detrended dataset, a better job by the SVM (ML algorithm) is observed.
    - In almonds, the ML algorithm has a dominance ratio of 1,74, while in leaves the ML algorithm has a dominance ratio of 4,79.

Thanks