soil systems

MDPI

Article

Prediction of Soil Salinity/Sodicity and Salt-Affected Soil Classes from Soluble Salt Ions Using Machine Learning Algorithms

Demis Andrade Foronda [1,2,*] and Gilles Colinet [1]

[1] Water-Soil-Plant Exchanges, TERRA, Gembloux Agro-Bio Tech, University of Liège, Passage des Déportés 2, 5030 Gembloux, Belgium
[2] Facultad de Ciencias Agrícolas y Pecuarias, Universidad Mayor de San Simón, Cochabamba 4926, Bolivia
* Correspondence: dn.andrade@doct.uliege.be

Abstract: Salt-affected soils are related to salinity (high content of soluble salts) and/or sodicity (excess of sodium), which are major leading causes of agricultural land degradation. This study aimed to evaluate the performances of three machine learning (ML) algorithms in predicting the soil exchangeable sodium percentage (ESP), electrical conductivity (EC_e), and salt-affected soil classes, from soluble salt ions. The assessed ML models were Partial Least-Squares (PLS), Support Vector Machines (SVM), and Random Forests (RF). Soil samples were collected from the High Valley of Cochabamba (Bolivia). The explanatory variables were the major soluble ions (Na^+, K^+, Ca^{2+}, Mg^{2+}, HCO_3^-, Cl^-, CO_3^{2-}, SO_4^{2-}). The variables to be explained comprised soil EC_e and ESP, and a categorical variable classified through the US Salinity Lab criteria. According to the model validation, the SVM and RF regressions performed the best for estimating the soil EC_e, as well as the RF model for the soil ESP. The RF algorithm was superior for predicting the salt-affected soil categories. Soluble Na^+ was the most relevant variable for all the predictions, followed by Ca^{2+}, Mg^{2+}, Cl^-, and HCO_3^-. The RF and SVM models can be used to predict soil EC_e and ESP, as well as the salt-affected soil classes, from soluble ions. Additional explanatory features and soil samples might improve the ML models' performance. The obtained models may contribute to the monitoring and management of salt-affected soils in the study area.

Keywords: machine learning; electrical conductivity; exchangeable sodium percentage; salt-affected soil classification

Citation: Andrade Foronda, D.; Colinet, G. Prediction of Soil Salinity/Sodicity and Salt-Affected Soil Classes from Soluble Salt Ions Using Machine Learning Algorithms. *Soil Syst.* **2023**, *7*, 47. https://doi.org/10.3390/soilsystems7020047

Academic Editors: Thomas Baumgartl and Mandana Shaygan

Received: 23 March 2023
Revised: 28 April 2023
Accepted: 5 May 2023
Published: 10 May 2023

1. Introduction

Salt-affected soils are mainly related to arid and semiarid regions and basically comprise saline and/or sodic soils. Saline soils have a significant amount of soluble salts which consist of major ions like sodium (Na^+), potassium (K^+), calcium (Ca^{2+}), magnesium (Mg^{2+}), bicarbonate (HCO_3^-), chloride (Cl^-), carbonate (CO_3^{2-}), and sulfate (SO_4^{2-}). Sodic soils have an excess of exchangeable Na^+ in the cation exchange complex, as well as in the soil solution. Soluble salts and Na^+ normally originate either from natural processes such as weathering (primary salinity/sodicity) or are induced by human activities such as the inappropriate management of land and water resources (secondary salinity/sodicity). Soil salinity negatively affects root growth and crop yield through the osmotic effect caused by the high concentration of soluble salts, and soil sodicity causes adverse effects, such as an increase in soil pH, loss of soil physical structure (clay dispersion, swelling, and plugging of soil pores), and the deterioration of soil–water relations (decrease in infiltration, hydraulic conductivity, retention and drainage), leading to soil erosion, crusting, compaction, runoff, waterlogging, nutrient imbalances, and specific ion effects on plants [1–7].

Salinity levels can be expressed as total soluble salts (TSS) or as soil electrical conductivity (EC) of saturated extract or soil–water suspensions. Sodicity levels are usually

determined as the exchangeable sodium percentage (ESP) through the amount of exchangeable Na^+ as a proportion of either the cation exchange capacity (CEC) or the sum of exchangeable cations [4,8], as well as by the sodium adsorption ratio (SAR) calculated from the soluble Na^+ relative to the soluble Ca^{2+} + Mg^{2+} concentrations in a soil solution using the formula proposed by Richards et al. [9]. The widely used salt-affected soil classification from the US Salinity Lab (USSL)—based on the threshold values of a soil EC_e of 4 dS m^{-1}, ESP of 15%, and pH of 8.5—generates four classes, namely, normal, saline, saline–sodic, and sodic soil. The Australian classification is analogous to the USSL criteria with the exception that it considers a soil ESP threshold value of 6% and takes into account the pH levels [10]. Furthermore, neutral and alkali salts determine the distinction between sodicity and alkalinity, so alkali soils normally have an excess of exchangeable Na^+ and carbonates besides a pH above 8 [11]. Concerning that fact, Chhabra et al. [12] proposed an alternative classification including the ion ratios of $(2CO_3{}^{2-} + HCO_3{}^-)/(Cl^- + 2SO_4{}^{2-})$ and $Na^+/(Cl^- + 2SO_4{}^{2-})$ expressed in mol m^{-3}, besides soil EC_e and ESP, for facilitating the specific management and reclamation of salt-affected soils.

Data mining can be described as the capacity of identifying patterns from data to establish relationships and models through data analysis, and machine learning (ML) is a process of learning from a system's experience for self-improving based on resultant information. Moreover, supervised learning models the relationships and dependencies between the target prediction output and the input data/features to predict the output values for new data. Partial Least-Squares (PLS)—Discriminant Analysis (DA) is a 'supervised' version of principal component analysis (PCA) which achieves dimensionality reduction with complete cognizance of the classes, arriving at a linear transformation that converts the data to a lower dimensional space with as small an error as possible [13]. In addition, PLS regression combines features from PCA and multiple regression, allowing the reduction of the dimensionality while focusing on covariance. Support Vector Machines (SVM) seek to design a decision surface and separate the margin between the different levels, finding this hyperplane using support vectors and margins. Then, the SVM with linear kernel function fits an optimal hyperplane between the classes, making linear and separable small samples [14], while support vector regression fits a line as the hyperplane with the maximum number of points. Breiman and Cutler's Random Forests (RF) algorithm is a tree-based ensemble which generates trees built on resampled subsets of data, with each tree depending on an ensemble of random variables. RF classification combines the trees by unweighted voting and chooses the most voted class over all the tree ensembles at training time if the response is categorical, or combines the resulting trees by unweighted averaging if the response is continuous [15,16].

ML methods have been used to classify soils based on various features such as chemical, physical, and biological variables, as well as on specific criteria. Within the framework of ML algorithms, many methods have been progressively developed to automate the soil classification process, such as Decision Trees, k-Nearest Networks, Artificial Neural Networks, and SVM [17]; in that context, some investigations on various soil type classifications using ML methods were carried out [18–21]. The review on ML and soil sciences by Padarian et al. [22] shows that the modelling of continuous and categorical soil properties is based on their relationships with environmental covariates and is mainly focused on mapping. Some key findings in the compilation by Motia and Reddy [23] were that: the implementation of soil classification uses more ML methods than soil regression; the assessment of soil salinity still shows a low contribution from ML; SVM and RF techniques are widely used in ML predictions of soil parameters and classifications; and the *RMSE* and R^2 are the top metrics used for the performance evaluation of ML prediction models in soil analysis.

Apart from simple/multivariate regression-based models, most of the studies based on ML methods in predicting and mapping salinity use variables from remote sensing (spectral bands and derived indices) [24–29], and combined with other environmental covariates (elevation, geology, hydrology, morphometry, and climate) [30–34]. Field-measured data

(physical and chemical soil–water properties), which are used to a lesser extent, may improve the prediction performances for soil salinity, even more if alternative salt-term parameters are considered. Moreover, the determination of the content of exchangeable cations—and thus the soil ESP—is usually less cost-effective and more time-consuming than that of soluble ion concentrations, which are often used for estimating salinity/sodicity indirectly. Therefore, this study aimed to evaluate and compare the prediction performances of three ML regression and classification algorithms (PLS, SVM, and RF) for estimating the soil EC_e and ESP, and classifying salt-affected soils from soluble salt ions. Then, the results may contribute alternative covariates for modelling as well as to the characterization and management of salt-affected soils in the study area.

2. Materials and Methods

2.1. Study Area and Data

The observations (135 soil samples) were collected at a depth of ~25 cm from the agricultural lands of the High Valley of Cochabamba-Bolivia (Figure 1), under the framework of the survey by Weber [35]. The area is located between the latitude boundaries of −17°29′47.7″ to −17°39′48.6″ and longitude of −66°5′16.8″ to −65°45′13.0″, at an elevation of ~2750 m. The climate of the valley is semiarid with a mean annual temperature and rainfall of 15–16 °C and 450–550 mm, respectively. Regarding the geomorphic characterization of this area [36] (Metternicht and Zinck, 2010), most of the salt-affected soils are in the landscape of a valley with a relief type consisting of lagunary depressions, aluvio-lagunary/lagunary facies, a landform consisting of lagunary flats, and soil associations consisting of Ustalfic Haplargids/Ustochreptic Camborthids and Typic Salorthids/Natric Camborthids. The soil textural classes consisting of loam, silty loam, and silty clay loam were predominant among the samples.

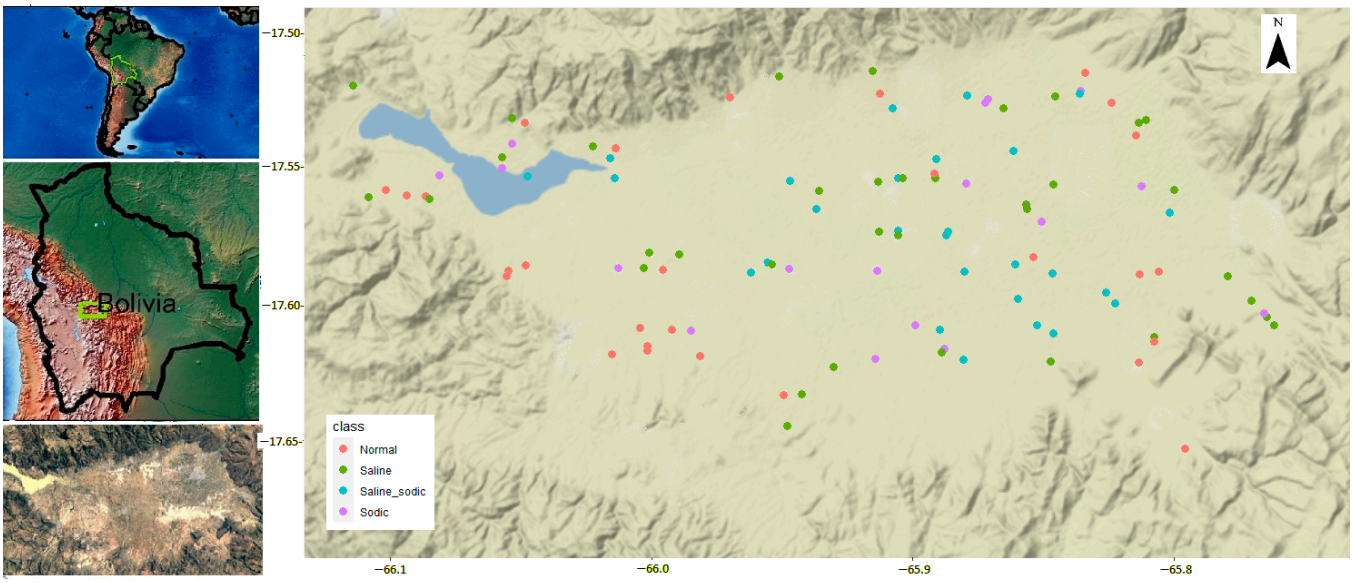


Figure 1. Soil sampling points and their salt-affected classes (USSL criteria) in the High Valley of Cochabamba, Bolivia.

2.2. Variables

As the explanatory variables, concentrations of soluble cations (Na^+, K^+, Ca^{2+}, Mg^{2+}) and anions (HCO_3^-, Cl^-, CO_3^{2-}, SO_4^{2-}) were determined from a paste extract, following the standard procedures of Richards et al. [9] at the Soil-Water Lab, Faculty of Agricultural and Livestock Sciences, *Universidad Mayor de San Simón* (Bolivia).

The continuous variables to be predicted were the soil EC_e and the soil ESP calculated using the formula (Equation (1)) [4,8] with the exchangeable cation values obtained through a derived ISO 22171 at a pH of 7 and atomic adsorption spectroscopy at the *Station Provin-*

ciale d'analyses agricoles Lab (Belgium), taking into account the assessment by So et al. [37] for overcoming their overestimation as total extractable cations. The categorical variable to be explained comprises four categories classified using the USSL criteria [9], namely: normal (ESP < 15%, EC_e < 4 dSm^{-1}, pH < 8.5), saline (ESP < 15%, EC_e > 4 dSm^{-1}, pH < 8.5), saline–sodic (ESP > 15%, EC_e > 4 dSm^{-1}, pH < > 8.5), and sodic (ESP > 15%, EC_e < 4 dSm^{-1}, pH > 8.5). For practical purposes, the alkali soil was classified as sodic.

$$ESP = \left(\frac{Na^+}{Ca^{2+} + Mg^{2+} + Na^+ + K^+} \right) 100 \quad (1)$$

where cations are expressed as a concentration in $cmol_c\ kg^{-1}$.

2.3. Data Preparation and Model Implementation

The flow process of the modelling is described in Figure 2. Extreme values in the dataset were checked by applying a threshold value using the *Mahalanobis* distance from the PCA, and then 10 observations were discarded. In overcoming the possibility of hidden dependencies of the cross-validation (CV) and for testing purposes, the models were evaluated through an internal validation by partitioning the dataset into two sets, calibration (75%) and validation (25%), for both regression and classification models. The data were scaled into each calibration process. A subsequent performance evaluation showed that a min–max normalization was not needed.

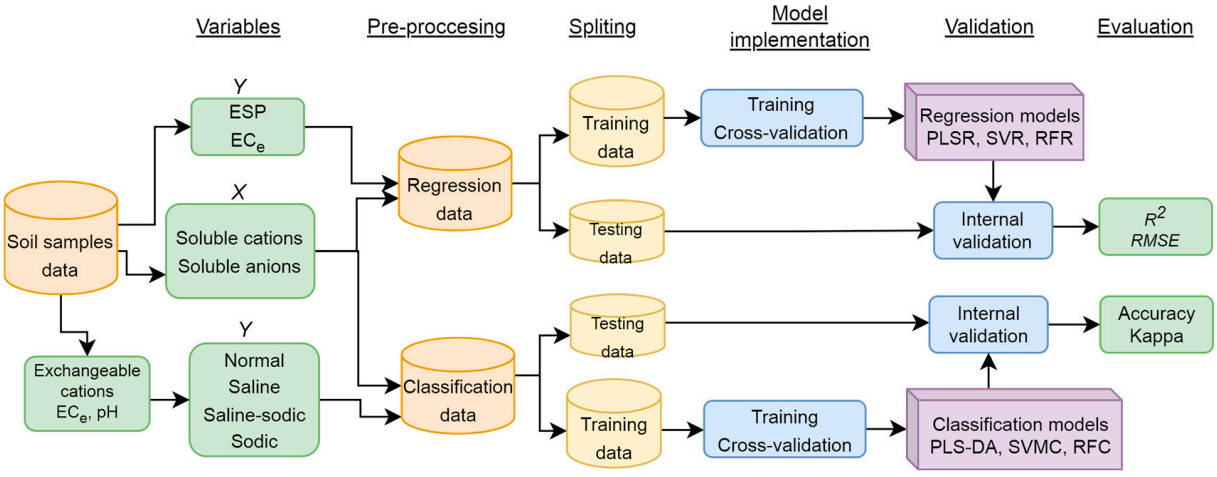


Figure 2. Flow chart of the methodological path of this study.

Three supervised ML algorithms were used: Partial Least-Squares (PLS) and Support Vector Machines (SVM) with linear kernel function as discriminating methods, and Random Forests (RF) as a tree-based method, for the respective regression (PLS-R, SV-R, RF-R) and classification (PLS-DA, SVM-C, RF-C) algorithms. A multivariate linear regression (ML-R) model was added for comparison purposes. The models were trained with tenfold groups, and CV was repeated five times. The specific tuning of the parameters for the training and CV of regression and classification models is shown in Table A1.

2.4. Model Performance Evaluation

The prediction was performed for the three regression/classification methods by using the obtained models from the training process on the testing datasets; then, the performances were compared. The metrics to evaluate the effectiveness of the regression techniques were the determination coefficient R^2 (Equation (2)) and the root mean square error *RMSE* (Equation (3)) as the standard deviation of the error. For classification models, the metrics were the overall accuracy (Equation (4)) as the correct classification of the data obtained by executing the model, and Cohen's kappa statistics (Equation (5)) like the

strength of the agreement as the extent to which the data are correct representations of the variables measured [38]. Additionally, the measures of sensitivity and specificity, as the proportions of true positives and true negatives correctly predicted, respectively, were calculated for classification.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (p_i - o_i)^2}{\sum_{i=1}^{n} \left(\bar{o} - o_i\right)^2} \quad (2)$$

$$RMSE = \left[n^{-1} \sum_{i=1}^{n} (p_i - o_i)^2\right]^{1/2} \quad (3)$$

where n is the number of observations, p_i is the predicted values, o_i is the observed data, and $\bar{o}$ is the mean for o_i.

$$\text{Accuracy} = \sum_{i=1}^{n} \frac{True\ classification}{Total\ cases} \quad (4)$$

$$Kappa = \frac{P_o - P_e}{1 - P_e} \quad (5)$$

where n is the number of classes, P_o is the total agreement probability, and P_e is the agreement probability due to chance.

2.5. Other Assessments

The relative importance of the variables was assessed through the RF measures of Mean Decrease Accuracy/Gini for classification and the percent increase in MSE and increase in node purity for regressions. For overcoming the imbalance caused by the sodic category, a resampling technique was applied. The stability of the models was assessed in function to three different data partitions as an indicator of the change in the level of performance; then, the dataset was split for obtaining a validation dataset proportion over (30%) and below (20%) the referential of 25%. Finally, the models were assessed with additional explanatory variables, namely, soil pH and EC_e determined from the same solution in which the soluble ions were measured, total organic carbon (TOC), and soil texture (clay, silt, and sand).

2.6. Software

Statistical analysis and ML modelling/evaluation were performed by using the R software (v.4.1.3) [39] and RStudio (v.1.31093) [40]. The regression and classification models (PLS, SVM, RF) were trained and evaluated through the package *caret* (classification and regression training) [41], and complementary packages for data preparation, analysis and visualization such as *randomForest* [42] and *FactomineR*, among others, were used.

3. Results and Discussion

3.1. Statistical Overview

Some descriptive statistics of the dataset are shown in Table A2. The distribution of samples according to the salt-affected soil classes was relatively balanced, except for the sodic soil category (Figure 3a). Among the explanatory variables, soluble Ca^{2+} with Mg^{2+} (r of 0.87) and Na^{+} with the anions were relatively highly correlated, as well as the soil EC_e and ESP with Na^{+} and soluble anions (Figure A1). Despite these relatively high relationships, it should be considered that ML algorithms deal with multicollinearity through regularizations and by focusing the prediction and accuracy instead of the influence among variables; moreover, all soluble salt ions are part of the dominant composition and balance in the soil solution of each site-specific sample. Correlations between the contents of cations in the soil sorption complex and those in the soil–water solution are relatively low (Table A3) in contrast to the findings of Porębska and Ostrowska [43]. According to

the PCA, around 98% of the variance was explained by seven out of eight components. The components are not so good for discriminating the clusters (Figure 3b); consequently, for a complete separation of the soil categories, the PLS-DA, SVM, and RF classification algorithms were performed.

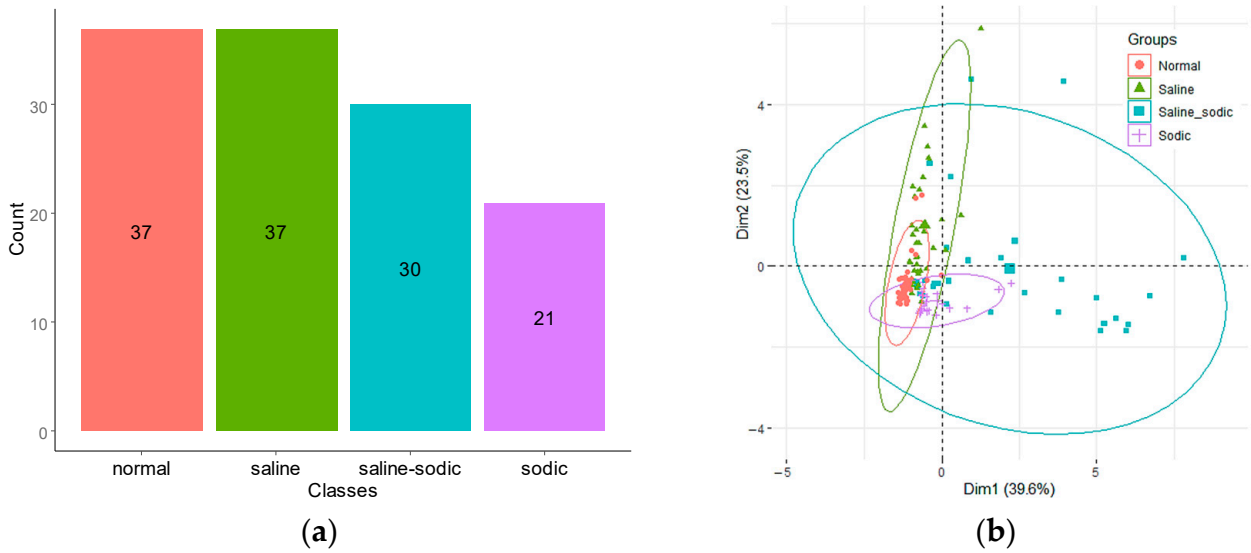


Figure 3. Distribution of the observations according to the salt-affected categories (**a**). PCA plot of observations grouped by soil categories (**b**).

3.2. Evaluation of the Regression Models

Among the assessed ML regression models for predicting soil EC_e, the SV-R and RF-R algorithms performed the best with relatively similar values of R^2 and *RMSE*, followed by ML-R and PLS-R models, which, in contrast, showed good cross-validation performances (Table 1). The overall high proportions of soil EC_e variance explained by the soluble ions agree with the fact that the soluble major ions complex is normally a good predictor for the soil EC and vice versa, and also coincide with the high correlations between soil EC_e and soluble ions as total dissolved salts [44,45]. Furthermore, the low performance of the PLS-R model agrees with the fact that it is better in cases where the number of explanatory variables is high or where multicollinearity is an issue. As a partially related study, Wang et al. [46] found that RF regression performed comprehensively better than SVM among other ML models in predicting salinity from field-measured spectral and salinity parameter data.

Table 1. Prediction performances of the regression models for estimating soil EC_e and ESP.

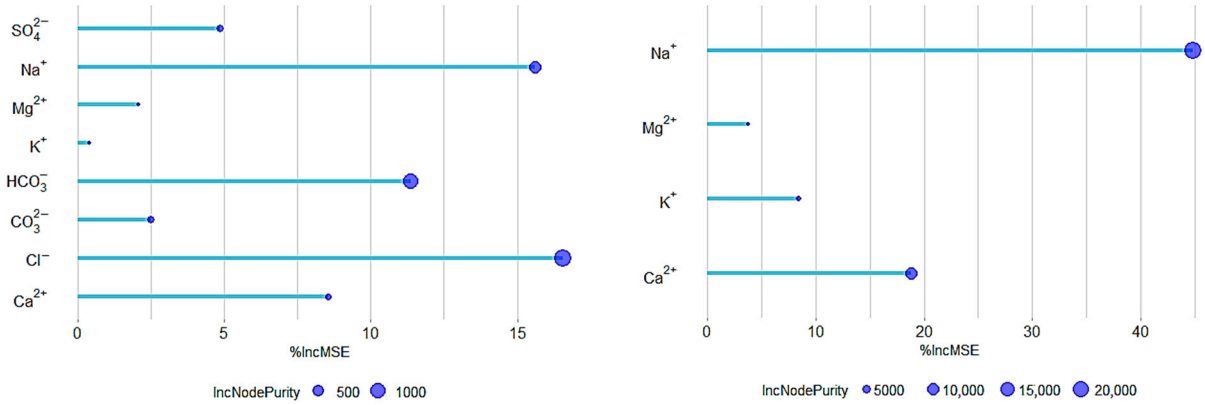
Method	EC_e		ESP	
	RMSE	R^2	*RMSE*	R^2
PLS-R	2.9 (3.3)	0.82 (0.72)	19.0 (13.6)	0.41 (0.63)
SV-R	1.9 (3.5)	0.92 (0.74)	18.4 (14.0)	0.40 (0.65)
RF-R	2.1 (3.7)	0.91 (0.66)	12.6 (12.4)	0.71 (0.60)
ML-R	2.4 (2.8)	0.88 (0.81)	19.1 (13.6)	0.40 (0.54)

RMSE stands for root mean square error. Values in parentheses mean the CV performances.

For estimating the soil ESP, the RF-R obtained the best prediction performance (R^2 of 0.71 and *RMSE* of 12.6), followed by the rest of the models with similar results; even so, they obtained relatively good cross-validation performances (Table 1). The relatively high performance of the RF-R model for predicting soil ESP is partly related to the relationships between SAR and ESP or exchangeable sodium ratio (ESR) (Table A3), and has some correspondence to the results obtained to predict ESP from SAR by using simple

regression [47–50], and there is also correspondence with those to estimate the ESR from SAR [51,52].

Through the variable importance analysis by using the RF-R algorithm, two measures were obtained: percent increase in mean square error (MSE) as the prediction error of each variable if omitted from the analysis, and the increase in node purity as how much the model error increases when a particular variable is randomly permuted or shuffled. According to these metrics, Na^+ is the most important variable for predicting both soil ESP and EC_e, besides Cl^- and HCO_3^- which are indispensable for estimating soil EC_e, as well as Ca^{2+} for ESP (Figure 4a,b). In addition, despite the relatively low importance of K^+ in predicting soil ESP (Figure 4b), it might be important to keep this cation for modelling because it influences soil dispersion, as demonstrated through the exchangeable cation ratio (ECR) [53] and the cation ratio of soil structural stability (CROSS) [54] as alternative indicators for soil ESP and SAR, respectively.



(**a**) (**b**)

Figure 4. Variable importance as the percent increase in mean square error (%IncMSE) and the increase in node purity (IncNodePurity) from the RF model for the soil EC_e (**a**) and ESP (**b**).

3.3. Evaluation of the Classification Models

According to the internal validation, the RF-C model obtained the best performance with the highest prediction accuracy (87%) indicating a good classification with a significant strength of agreement beyond chance (kappa of 82%), followed by the SVM-C and PLS-DA models, both with a regular classification and moderate agreement. Additionally, according to the CV analysis, the RF-C and SVM-C algorithms performed better than the PLS-DA model with relatively similar results (Table 2).

Table 2. Accuracy and kappa values of the model training and model testing.

Method	Calibration/CV *		Validation	
	Accuracy	Kappa	Accuracy	Kappa
PLS-DA	0.55	0.37	0.67	0.52
SVM-C	0.63	0.49	0.70	0.58
RF-C	0.61	0.47	0.87	0.82

* CV stands for cross-validation.

The overall Out of Bag (OOB) error of the RF bootstrapping was 37.9%, and the error classes were 0.29, 0.38, 0.26, and 0.68 for normal, saline, saline–sodic, and sodic soil, respectively. The misclassifications of sodic soil were mainly due to its imbalance in contrast to the other categories. The soil pH used to classify the soil may decrease the quality of the classification models because it is not directly related to the soluble/exchangeable cations, as the soil EC_e and ESP are. Based on the predictions in the confusion matrixes (Table A4), the measures of sensitivity and specificity were calculated. Overall, the sensitivity as the

true positive rate was regular to good for predicting the normal, saline, and saline–sodic classes but poor for the sodic class; in addition, the RF-C model generated higher values of sensitivity than those of the SVM-C and PLS-DA (Table A5).

According to the estimation of the variables' relative importance using the Mean Decrease Accuracy and Mean Decrease Gini calculations, the soluble Na^+ was the most relevant parameter for classifying the salt-affected soils, followed by Ca^{2+}, Mg^{2+}, and Cl^- (Figure 5a,b). These rankings coincide with the variable selection through RF backward elimination and become important for eventually discarding the less important variables if and when the performance of the model is improved. The importance estimations have some correspondence with the ratio of soluble Na^+ to the base cations expressed by the SAR and also with the relevance of neutral salts over alkali salts for these soils.

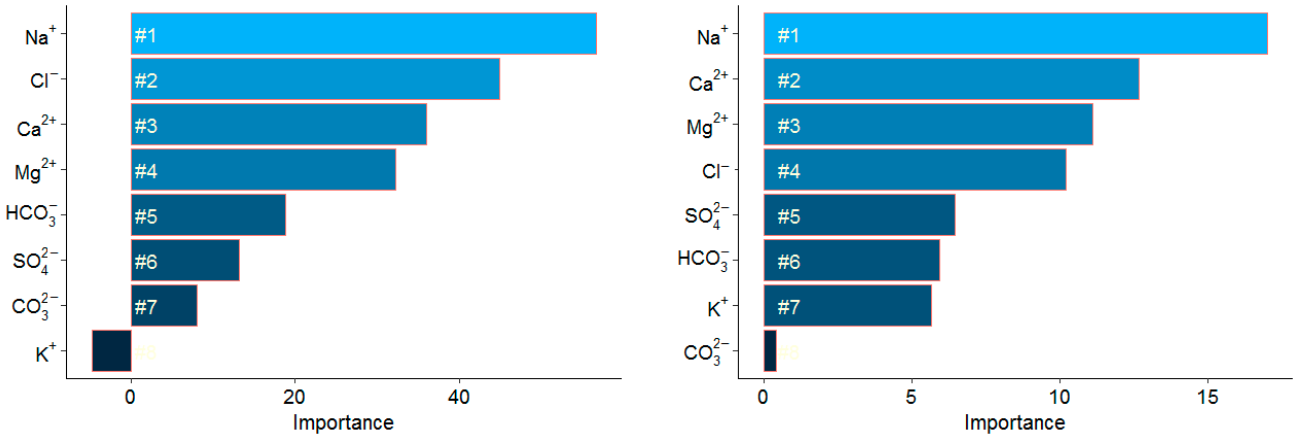


Figure 5. RF relative importance estimation of the explanatory variables according to the measures of *MeanDecreaseAccuracy* (**a**) and *MeanDecreaseGini* (**b**).

3.4. Resampling

For overcoming the imbalance generated by the minority class (sodic), the models were trained a second time by applying the resampling method 'Synthetic Minority Over-Sampling Technique' through the Smote function [55]; then, the results from the performance validation showed a slight improvement for the SVM-C model, but a significant decrease for the RF-C model in accuracy and kappa values (Table 3), compared to those without resampling (Table 2).

Table 3. Accuracy and kappa values of the model training and testing with the Smote function.

Method	Calibration/CV *		Validation	
	Accuracy	Kappa	Accuracy	Kappa
PLS-DA	0.55	0.39	0.60	0.48
SVM-C	0.61	0.46	0.73	0.62
RF C	0.60	0.45	0.77	0.68

* CV stands for cross-validation.

3.5. Stability Analysis

The stability was evaluated by performing a new validation of the regression and classification models based on three different partitions (percent calibration datasets of 70, 75, and 80). The RF regression models for predicting soil EC_e and ESP obtained lower differences between performances of the three calibration data amounts than those of SV-R and PLS-R, whereas, for the classification, the PLS-DA followed by the SVM-C technique were more stable than the RF-C model in predicting soil categories (Table 4).

Table 4. Stability assessment for the performance validations of the models.

Model and Metrics	Method	Percent of Calibration Dataset			Difference *
		70%	75%	80%	
EC_e Regression $RMSE/R^2$	PLS-R	3.5/0.68	2.9/0.82	2.3/0.92	1.2/0.24
	SV-R	3.4/0.71	2.0/0.92	1.9/0.95	1.5/0.24
	RF-R	2.9/0.79	2.1/0.91	3.0/0.88	1.7/0.15
ESP Regression $RMSE/R^2$	PLS-R	15.1/0.52	18.9/0.41	14.9/0.57	7.8/0.27
	SV-R	15.5/0.54	18.4/0.40	15.5/0.58	5.8/0.32
	RF-R	12.6/0.65	12.6/0.71	11.1/0.78	1.5/0.13
Classification Accuracy/Kappa	PLS-DA	0.65/0.51	0.67/0.52	0.71/0.57	0.06/0.06
	SVM-C	0.70/0.58	0.70/0.58	0.79/0.69	0.09/0.11
	RF-C	0.78/0.70	0.87/0.82	0.79/0.71	0.17/0.23

* Difference = sum of absolute differences among the metric values of the three partitions.

3.6. Additional Variables

By adding the soil pH, EC_e, TOC, clay, silt, and sand to the matrix of predictor variables, only the performances of PLS and SVM regressions to predict soil ESP showed a significant improvement (Table 5) compared to those in Table 1. These results partly contrast with those of Keshavarzi et al. [56] who obtained R^2/*MSE* values of 0.84/5.36 and 0.90/5.09 for the AI-based models Multi-Layer Perceptron and Adaptive Neuro-Fuzzy Inference System, respectively, for predicting ESP from EC_e, pH, and clay. Although the RF classification model obtained a slight increase in effectiveness (Table 5), should be noted the redundancy caused by the soil EC_e and pH as explanatory variables and as classifiers of the explained categories at the same time; however, their further inclusion might be pertinent if more easily obtained parameters are used, such as EC and pH measured in soil–water suspensions.

Table 5. Obtained model performances by adding features to the matrix of explanatory variables.

Method	Regression—EC_e		Regression—ESP		Classification	
	RMSE	R^2	*RMSE*	R^2	Accuracy	Kappa
PLS	7.6	0.89	12.5	0.62	0.61	0.45
SVM	4.4	0.96	12.1	0.63	0.61	0.47
RF	12.1	0.55	12.7	0.62	0.90	0.87

3.7. Some Remarks

Overall, RF and SVM regression models performed the best for predicting soil EC_e from soluble ions, as well as the RF model for estimating the soil ESP from soluble cations; and the RF followed by the SVM classification algorithm outperformed the PLS-DA in predicting salt-affected soil classes from soluble salt ions. Considering that it is important to apply tailored reclamation techniques based on modelling and predictive tools calibrated and validated for site-specific salt-affected soils [57], the obtained models become important tools for the monitoring and management of salt-affected soils for the study area, and also as source of alternative covariates for further modelling.

As tentative limitations, all the models still need an optimization of their prediction effectiveness; therefore, additional observations might be included in the dataset for improving the performance and stability of the classification/regression models, as well as for overcoming class imbalances and reinforcing the selection of variables. Additionally, the input of additional features such as remote sensing data and field-measured soil properties can also be useful for improving the modelling and predictions. Further classification modelling could consider alternative classification systems such as that of Chhabra et al. [12] which generates only three soil classes (normal, saline, and alkali).

4. Conclusions

The performances of ML classification and regression algorithms (PLS, SVM, and RF) in predicting soil EC_e, ESP, and salt-affected soil classes were evaluated and compared. Among the assessed ML regressions, SVM and RF obtained the best performances for predicting the soil EC_e, whereas the RF model was superior for estimating the soil ESP. The RF classification algorithm showed the best prediction accuracy (87%) with a kappa value of 82%, followed by SVM and PLS-DA. Soluble Na^+ was the most important explanatory variable for all the prediction models, followed by Ca^{2+}, Mg^{2+}, Cl^-, and HCO_3^- which were important for classification, as well as for regression. The sodic class was poorly predicted, and the applied resampling for overcoming its imbalance did not significantly improve the classification performances. The stability analysis showed that the amount of training data generated less impact on the RF regression models, whereas the SVM and PLS-DA were more stable than RF for classification. Additional explanatory variables somewhat improved the PLS and SVM regressions to predict ESP and the RF classification effectiveness. It can be concluded that the RF or SVM and the RF regression can be suitable to estimate the soil EC_e and ESP, respectively. In addition, the RF and SVM classification models can be appropriate in predicting salt-affected soil classes from soluble salt ions. Additional samples and explanatory features can be included in the dataset for improving the prediction performances. The assessed models might contribute significantly to the monitoring, mapping, and management of salt-affected soils in the study area.

Author Contributions: Conceptualization, D.A.F.; methodology, D.A.F.; validation, D.A.F.; formal analysis, D.A.F. and G.C.; investigation, D.A.F.; writing—original draft preparation, D.A.F.; writing—review and editing, D.A.F. and G.C. All authors have read and agreed to the published version of the manuscript.

Funding: Funded by the 'Académie de recherche et d'enseignement supérieur'—ARES (Belgium) and provided by the 'Universidad Mayor de San Simón'—UMSS (Bolivia). Code: AI-11420/UMSR1.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data are available on request from the corresponding author.

Acknowledgments: The team from Soil-Water-Plant Exchanges—TERRA—GABT (University of Liège) and the team from the Soil-Water Lab and CISTEL (UMSS).

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

Appendix A

Table A1. The setting of parameters for model training and cross-validation analysis.

Model	Algorithms	Parameters/Values
EC_e and ESP Regression	PLS-R	Number of components: 1 (EC_e), 3 (ESP)
	SV-R	CF grid: 0.01, 0.1, 0.25, 0.5, 1
	RF-R	NT of 3000, MTRY of 5 (EC_e), 2 (ESP)
Multiple classification	PLS-DA	Number of components: 2
	SVM-C	CF grid: 0.05, 0.1, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2
	RF-C	NT of 3000, NS of 10, MTRY of 2

R = regression; C = classification; NT = number of trees; NS = minimum node size; MTRY = number of randomly selected predictors; CF = capacity factor for SVM.

Table A2. Descriptive statistics of explanatory variables (soluble salt ions), ESP and EC_e.

Item	Mean	SD	CV	Min	Max	Median	Count
Ca^{2+}	3.7	4.5	1.2	0.1	26.2	2.2	125
Mg^{2+}	1.7	1.9	1.1	0.09	9.4	1.0	125
Na^{+}	27.4	54.9	2.0	0.02	326.1	5.6	125
K^{+}	0.5	0.5	1.0	0.02	2.2	0.4	125
Cl^{-}	17.4	35.3	2.0	0	205.0	5	125
SO_4^{2-}	14.2	29.6	2.1	1.2	153.4	3.7	125
HCO_3^{-}	5.4	6.6	1.2	0.5	34.0	3.0	125
CO_3^{2-}	6.3	22.2	3.5	0.0	134.0	0.0	125
ESP	16.3	20.4	1.2	0.1	77.0	4.9	125
EC_e	6.1	6.5	1.1	0.3	33.4	4.1	125

SD = standard deviation; CV = coefficient of variation.

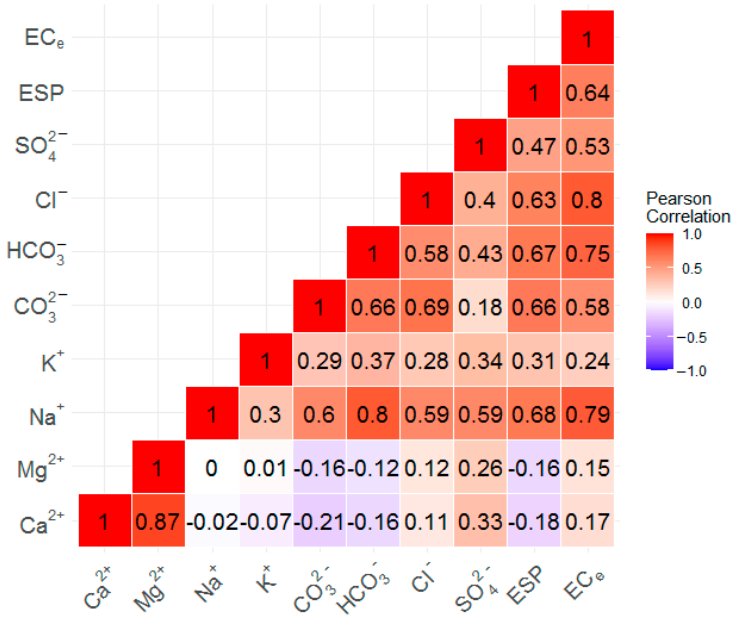


Figure A1. Correlation matrix among the explanatory variables, ESP and EC_e.

Table A3. Correlation matrix among sums of soluble and exchangeable cations, sodicity parameters, and EC_e.

	Sum-Sol Cations	Sum-Sol-Anions	Sum-Exc-Cations	SAR	ESR	ESP	EC_e
Sum-Sol Cations	1						
Sum-Sol-Anions	0.78	1					
Sum-Exc-Cations	0.32	0.42	1				
SAR	0.90	0.75	0.33	1			
ESR	0.57	0.77	0.45	0.61	1		
ESP	0.66	0.75	0.50	0.66	0.93	1	
EC_e	0.81	0.84	0.30	0.73	0.64	0.64	1

Sum-Sol = Sum of soluble; Sum-Exc = Sum of exchangeable; SAR = sodium adsorption ratio; ESR = exchangeable sodium ratio (ESP/100-ESP).

Table A4. Confusion matrixes of the predictions for the three ML classification algorithms.

Class	PLS-DA				SVM-C				RF-C			
	NO	SA	SS	SO	NO	SA	SS	SO	NO	SA	SS	SO
Normal	9	2	1	5	9	2	1	4	9	0	0	1
Saline	1	6	1	0	1	6	0	0	1	8	0	1
Saline–sodic	0	0	5	0	0	0	5	0	0	0	7	1
Sodic	0	0	0	0	0	0	1	1	0	0	0	2

NO = normal; SA = saline; SS = saline–sodic; SO = sodic.

Table A5. Sensitivity and specificity for the three classification models.

Class	Sensitivity			Specificity		
	PLS-DA	SVM-C	RF-C	PLS-DA	SVM-C	RF-C
Normal	0.90	0.90	0.90	0.60	0.65	0.95
Saline	0.75	0.75	1.00	0.91	0.95	0.90
Saline–sodic	0.71	0.71	1.00	1.00	1.00	0.96
Sodic	0.00	0.20	0.40	1.00	0.96	1.00

References

1. Qadir, M.; Schubert, S.; Ghafoor, A.; Murtaza, G. Amelioration Strategies for Sodic Soils: A review. *Land Degrad. Dev.* **2001**, *12*, 357–386. [CrossRef]
2. Qadir, M.; Schubert, S. Degradation Processes and Nutrient Constraints in Sodic Soils. *Land Degrad. Dev.* **2002**, *13*, 275–294. [CrossRef]
3. Levy, G.J.; Shainberg, I. Sodic Soils. In *Encyclopedia of Soils in the Environment*; Hillel, D., Ed.; Elsevier: Amsterdam, The Netherlands, 2005; pp. 504–513, ISBN 9780123485304. [CrossRef]
4. Qadir, M.; Oster, J.D.; Schubert, S.; Noble, A.D.; Sahrawat, K.L. Phytoremediation of Sodic and Saline-Sodic Soils. In *Advances in Agronomy*; Elsevier: Amsterdam, The Netherlands, 2007; Volume 96, pp. 197–247, ISBN 978-0-12-374206-3.
5. Keren, R. Salt-Affected Soils Reclamation. In *Encyclopedia of Soils in the Environment*; Elsevier: Amsterdam, The Netherlands, 2005; pp. 454–461, ISBN 978-0-12-348530-4.
6. Lin, Z.-Q.; Bañuelos, G.S. Soil Salination Indicators. In *Environmental Indicators*; Armon, R.H., Hänninen, O., Eds.; Springer: Dordrecht, The Netherlands, 2015; pp. 319–330, ISBN 978-94-017-9498-5.
7. Andrade Foronda, D.; Colinet, G. Combined Application of Organic Amendments and Gypsum to Reclaim Saline–Alkali Soil. *Agriculture* **2022**, *12*, 1049. [CrossRef]
8. Sumner, M.E.; Rengasamy, P.; Naidu, R. Sodic soils: A reappraisal. In *Sodic Soil: Distribution, Management and Environmental Consequences*; Sumner, M.E., Naidu, R., Eds.; Oxford University Press: New York, NY, USA, 1998; pp. 3–17.
9. Richards, L.; Allison, L.; Bernstein, C.; Bower, J.; Brown, M.; Fireman, J.; Richards, W. *Diagnosis and Improvement of Saline Alkali Soils*; United States Salinity Laboratory Staff—Department of Agriculture; Agricultural Research Service: Washington, DC, USA, 1954; 169p.
10. Rengasamy, P. Soil Processes Affecting Crop Production in Salt-Affected Soils. *Funct. Plant Biol.* **2010**, *37*, 613–620. [CrossRef]
11. Gupta, R.K.; Abrol, I.P. Salt-Affected Soils: Their Reclamation and Management for Crop Production. In *Advances in Soil Science*; Lal, R., Stewart, B.A., Eds.; Springer: New York, NY, USA, 1990; Volume 11, pp. 227–229.
12. Chhabra, R. Classification of Salt-Affected Soils. *Arid Land Res. Manag.* **2004**, *19*, 61–79. [CrossRef]
13. Ruiz-Perez, D.; Guan, H.; Madhivanan, P.; Mathee, K.; Narasimhan, G. So You Think You Can PLS-DA? *BMC Bioinform.* **2020**, *21*, 2. [CrossRef] [PubMed]
14. Mohan, L.; Pant, J.; Suyal, P.; Kumar, A. Support Vector Machine Accuracy Improvement with Classification. In Proceedings of the 2020 12th International Conference on Computational Intelligence and Communication Networks, Bhimtal, India, 25–26 September 2020; IEEE: Piscataway, NJ, USA, 2010; pp. 477–481.
15. Cutler, A.; Cutler, D.R.; Stevens, J.R. Random Forests. In *Ensemble Machine Learning*; Zhang, C., Ma, Y., Eds.; Springer: Boston, MA, USA, 2012; pp. 157–175. ISBN 978-1-4419-9325-0.
16. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
17. Chandan, T.R. Recent Trends of Machine Learning in Soil Classification: A Review. *Int. J. Comput. Eng.* **2018**, *8*, 25–32.
18. Kovačević, M.; Bajat, B.; Gajić, B. Soil Type Classification and Estimation of Soil Properties Using Support Vector Machines. *Geoderma* **2010**, *154*, 340–347. [CrossRef]
19. Harlianto, P.A.; Adji, T.B.; Setiawan, N.A. Comparison of Machine Learning Algorithms for Soil Type Classification. In Proceedings of the 2017 3rd International Conference on Science and Technology-Computer (ICST), Yogyakarta, Indonesia, 11–12 July 2017; pp. 7–10. [CrossRef]

20. Bhargavi, P.; Jyothi, D.S. Soil Classification Using Data Mining Techniques: A Comparative Study. *Int. J. Eng. Technol.* **2011**, *2*, 55–59.
21. Raza Ansari, S. Application of Machine Learning Techniques for Soil Type Classification of Karanataka. Master's Thesis, National College of Ireland, Dublin, Ireland, 2018. Available online: https://norma.ncirl.ie/id/eprint/3443 (accessed on 25 January 2023).
22. Padarian, J.; Minasny, B.; McBratney, A.B. Machine Learning and Soil Sciences: A Review Aided by Machine Learning Tools. *Soil* **2020**, *6*, 35–52. [CrossRef]
23. Motia, S.; Reddy, S. Exploration of Machine Learning Methods for Prediction and Assessment of Soil Properties for Agricultural Soil Management: A Quantitative Evaluation. *J. Phys. Conf. Ser.* **2021**, *1950*, 012037. [CrossRef]
24. Allbed, A.; Kumar, L. Soil Salinity Mapping and Monitoring in Arid and Semi-Arid Regions Using Remote Sensing Technology: A Review. *Adv. Remote Sens.* **2013**, *2*, 373–385. [CrossRef]
25. Kaplan, G.; Gasparovic, M.; Alqasemi, A.; Aldhaher, A.; Abuelgasim, A.; Ibrahim, M. Soil salinity prediction using Machine Learning and Sentinel—2 Remote Sensing Data in Hyper-Arid areas. *Phys. Chem. Earth Parts A/B/C* **2023**, *130*, 103400. [CrossRef]
26. Wang, J.; Peng, J.; Li, H.; Yin, C.; Liu, W.; Wang, T.; Zhang, H. Soil Salinity Mapping Using Machine Learning Algorithms with the Sentinel-2 MSI in Arid Areas, China. *Remote Sens.* **2021**, *13*, 305. [CrossRef]
27. Wu, W.; Zucca, C.; Muhaimeed, A.S.; Al-Shafie, W.M.; Fadhil Al-Quraishi, A.M.; Nangia, V.; Zhu, M.; Liu, G. Soil Salinity Prediction and Mapping by Machine Learning Regression in Central Mesopotamia, Iraq. *Land Degrad. Dev.* **2018**, *29*, 4005–4014. [CrossRef]
28. Zarei, A.; Hasanlou, M.; Mahdianpari, M. A Comparison of Machine Learning Models for Soil Salinity Estimation Using Multi-Spectral Earth Observation Data. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2021**, *3*, 257–263. [CrossRef]
29. Zurqani, H.; Mikhailova, E.; Post, C.; Schlautman, M.; Sharp, J. Predicting the Classes and Distribution of Salt-Affected Soils in Northwest Libya. *Commun. Soil Sci. Plant Anal.* **2018**, *49*, 689–700. [CrossRef]
30. Boudibi, S.; Sakaa, B.; Benguega, Z.; Fadlaoui, H.; Othman, T.S.; Bouzidi, N. Spatial Prediction and Modeling of Soil Salinity Using Simple Cokriging, Artificial Neural Networks, and Support Vector Machines in El Outaya Plain, Biskra, Southeastern Algeria. *Acta Geochim.* **2021**, *40*, 390–408. [CrossRef]
31. Merembayev, T.; Amirgaliyev, Y.; Saurov, S.; Wójcik, W. Soil Salinity Classification Using Machine Learning Algorithms and Radar Data in the Case from the South of Kazakhstan. *J. Ecol. Eng.* **2022**, *23*, 61–67. [CrossRef]
32. Nabiollahi, K.; Taghizadeh-Mehrjardi, R.; Shahabi, A.; Heung, B.; Amirian-Chakan, A.; Davari, M.; Scholten, T. Assessing Agricultural Salt-Affected Land Using Digital Soil Mapping and Hybridized Random Forests. *Geoderma* **2021**, *385*, 114858. [CrossRef]
33. Vermeulen, D.; Van Niekerk, A. Machine Learning Performance for Predicting Soil Salinity Using Different Combinations of Geomorphometric Covariates. *Geoderma* **2017**, *299*, 1–12. [CrossRef]
34. Wang, F.; Shi, Z.; Biswas, A.; Yang, S.; Ding, J. Multi-Algorithm Comparison for Predicting Soil Salinity. *Geoderma* **2020**, *365*, 114211. [CrossRef]
35. Weber, A. Identification des Échelles Spatiales et des Facteurs de Variations des Sols et de Leurs Propriétés au Sein de la Valle Alto de Cochabamba (Bolivie). Master's Thesis, Gembloux Agro-Bio Tech-Université de Liège, Liège, Belgique, 2018. Available online: https://matheo.uliege.be/handle/2268.2/5035 (accessed on 12 January 2023).
36. Metternicht, G.; Zinck, J.A. Spatial Discrimination of Salt- and Sodium-Affected Soil Surfaces. *Int. J. Remote Sens.* **1997**, *18*, 2571–2586. [CrossRef]
37. So, H.B.; Menzies, N.W.; Bigwood, R.; Kopittke, P.M. Examination into the Accuracy of Exchangeable Cation Measurement in Saline Soils. *Commun. Soil Sci. Plant Anal.* **2006**, *37*, 1819–1832. [CrossRef]
38. McHugh, M.L. Interrater Reliability: The Kappa Statistic. *Biochem. Med.* **2012**, *22*, 276–282. [CrossRef]
39. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2013. Available online: http://www.R-project.org/ (accessed on 12 December 2022).
40. RStudio Team. *RStudio: Integrated Development for R*; RStudio. PBC: Boston, MA, USA, 2020; Available online: http://www.rstudio.com/ (accessed on 12 December 2022).
41. Kuhn, M. *Caret: Classification and Regression Training, R Package Version 6.0-93*; The R Project for Statistical Computing: Vienna, Austria, 2022. Available online: https://CRAN.R-project.org/package=caret (accessed on 12 December 2022).
42. Liaw, A.; Wiener, M. Classification and Regression by randomForest. *R News* **2002**, *2*, 18–22.
43. Porębska, G.; Ostrowska, A. Relationships between Exchangeable and Water-Soluble Cations in the Forest Soil. *Ochr. Srodowiska Zasobów Nat.* **2016**, *27*, 1–7. [CrossRef]
44. Simón, M.; García, I. Physico-Chemical Properties of the Soil-Saturation Extracts: Estimation from Electrical Conductivity. *Geoderma* **1999**, *90*, 99–109. [CrossRef]
45. Chang, C.; Sommerfeldt, T.G.; Carefoot, J.M.; Schaalje, G.B. Relationships of Electrical Conductivity with Total Dissolved Salts and Cation Concentration of Sulfate-Dominant Soil Extracts. *Can. J. Soil Sci.* **1983**, *63*, 79–86. [CrossRef]
46. Wang, S.; Chen, Y.; Wang, M.; Li, J. Performance Comparison of Machine Learning Algorithms for Estimating the Soil Salinity of Salt-Affected Soil Using Field Spectral Data. *Remote Sens.* **2019**, *11*, 2605. [CrossRef]
47. Chi, C.-M.; Zhao, C.-W.; Sun, X.-J.; Wang, Z.-C. Estimating Exchangeable Sodium Percentage from Sodium Adsorption Ratio of Salt-Affected Soil in the Songnen Plain of Northeast China. *Pedosphere* **2011**, *21*, 271–276. [CrossRef]

48. Elbashier, M.; Xiaohou, S.; Ali, A.; Osman, B. Modeling of Soil Exchangeable Sodium Percentage Function to Soil Adsorption Ratio on Sandy Clay Loam Soil, Khartoum-Sudan. *Int. J. Plant Soil Sci.* **2016**, *10*, 1–6. [CrossRef] [PubMed]
49. Seilsepour, M.; Rashidi, M.; Khabbaz, B.G. Prediction of Soil Exchangeable Sodium Percentage Based on Soil Sodium Adsorption Ratio. *Am.-Eurasian J. Agric. Environ. Sci.* **2009**, *5*, 1–4.
50. Andrade Foronda, D.; Rodríguez, E.G.; Colinet, G. Estimación del Porcentaje de Sodio Intercambiable en Función de la Relación de Adsorción de Sodio para Suelos Afectados por Sales en el Valle Alto de Cochabamba. *Rev. Agric.* **2020**, *62*, 31–36.
51. Harron, W.R.A.; Webster, G.R.; Cairns, R.R. Relationship between Exchangeable Sodium and Sodium Adsorption Ratio in a Solonetzic Soil Association. *Can. J. Soil. Sci.* **1983**, *63*, 461–467. [CrossRef]
52. Shirmohamm, Z.; Heydari, S. Modeling of Exchangeable Sodium Ratio on the Saline Soil. *Pak. J. Biol. Sci.* **2020**, *23*, 159–165. [CrossRef] [PubMed]
53. Marchuk, A.; Marchuk, S.; Bennett, J.; Eyres, M.; Scott, E. An Alternative Index to ESP to Explain Dispersion Occurring in Australian Soils When Na Content Is Low. In Proceedings of the National Soil Science Conference (NSS 2014), Melbourne, Australia, 23–27 November 2014; Patti, A., Tang, C., Wong, V., Eds.; Australian Society of Soil Science Incorporated: Warragul, Australia, 2014.
54. Rengasamy, P.; Marchuk, A. Cation Ratio of Soil Structural Stability (CROSS). *Soil Res.* **2011**, *49*, 280. [CrossRef]
55. Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-Sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
56. Keshavarzi, A.; Bagherzadeh, A.; Omran, E.S.E.; Iqbal, M. Modeling of Soil Exchangeable Sodium Percentage using Easily Obtained Indices and Artificial Intelligence-Based Models. *Model. Earth Syst. Environ.* **2016**, *2*, 130. [CrossRef]
57. Shaygan, M.; Baumgartl, T. Reclamation of Salt-Affected Land: A Review. *Soil Syst.* **2022**, *6*, 61. [CrossRef]