**SUPPLEMENTARY INFORMATION (SI)**

Table SI 1: Datasets compiled from maize yield trials for modelling (modified after Asamoah et al. 2024).

|  |  |
| --- | --- |
| Variable groups | Detailed variables |
| Climate (6) | Rainfall (total for planting season), average temperatures over planting season of the daily minimum and maximum, average relative humidity of the planting season, average evapotranspiration of the planting season, solar radiation |
| Soil (17) | pH, organic carbon, total nitrogen, cation exchange capacity, available phosphorus, exchangeable calcium, exchangeable magnesium, total exchangeable bases, sand, silt, clay, bulk density, electrical conductivity, base saturation, root zone water holding capacity, soil type, total porosity |
| Crop (1) | Genotype |
| Environmental (5) | Normalized difference vegetation index, elevation, terrain wetness index, slope, Agroecological zone |
| Management practices (3) | Application of any organic amendment (e.g. poultry manure, cattle manure), management type, mode of fertilizer application |
| Fertilizer application (3) | Nitrogen, phosphorus, potassium |

Table SI 2: Machine learning models and their hyperparameters.

|  |  |  |  |
| --- | --- | --- | --- |
| Machine learning model, R Package | Hyperparameter | Description | Values |
| SVM, e1071 | Cost, C, | Regularization parameter weighs constraint violations of  the model | 0.1 to 1, by = 0.1 |
| epsilon | Regularization parameter defines ribbon around predictions | 0.1 to 1, by = 0.1 |
| XGBoost, xgboost | nrounds | Number of boosting steps | 50, 100, 200, 300 |
| eta | Learning rate, also called “shrinkage” parameter | 0.01, 0.02, 0.05, 0.1, 0.3 |
| gamma | Number of splits of a tree by assuming a minimal improvement for each split | 0, 0.1, 0.02, 1, 2 |
| subsample | Portion of the observations that is randomly selected in each iteration | 0.6, 0.8, 1 |
| min\_child\_weight | Restriction of the number of splits of each tree | 1, 3, 5 |
| colsample\_bytree | Number of variables that is chosen for the splits of a tree | 1 |
| max\_depth | Maximum depth of a leaf in the decision trees | 3, 6, 8, 10 |
| RF, ranger | num.trees | Number of trees that are combined in the overall ensemble model | 100, 200, 500, 1000 |
| mtry | Number of randomly chosen variables are considered for each split | , 15 %, 25 %, 33.3 % and 40 % of Ev |
| min.node.size | Minimum number of observations in a terminal node | 1, 3, 5 |
| sample.fraction | Number of observations that are randomly drawn for training a specific tree | 0.50, 0.63 and 0.80 |
| KNN, knn | k | Number of neighbours | 2, 3, 4, 5, 6, 7, 8, 9, 10 |

Ev = number of explanatory variables

Table SI 3: Descriptive statistics of selected variables.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Variable | n | Mean | SD | Min | Max | Variable | n | Mean | SD | Min | Max |
|  | Yield (kg ha-1) | 4496 | 2524 | 1412 | 11.3 | 8815 | AE-P (kg kg-1) | 2858 | 41.2 | 45.1 | -62.7 | 606.7 |
|  | AE-N (kg kg-1) | 3106 | 16.9 | 19.8 | -66.6 | 222.2 | AE-K (kg kg-1) | 2760 | 35.4 | 32.5 | -62.7 | 335.0 |
| Yield | T min PS (°C) | 4496 | 22.1 | 0.8 | 18.0 | 31.9 | pH | 4496 | 5.9 | 0.4 | 4.1 | 7.3 |
| AE-N | 3106 | 22.1 | 0.7 | 18.0 | 26.0 | 3106 | 5.9 | 0.3 | 4.1 | 7.3 |
| AE-P | 2858 | 22.1 | 0.6 | 18.0 | 25.0 | 2858 | 5.9 | 0.4 | 4.3 | 7.3 |
| AE-K | 2760 | 22.1 | 0.6 | 18.0 | 25.0 | 2760 | 5.9 | 0.3 | 4.4 | 7.3 |
| Yield | T max PS (°C) | 4496 | 30.7 | 1.1 | 27.0 | 40.0 | SOC (g kg-1) | 4496 | 0.08 | 0.05 | 0.02 | 0.43 |
| AE-N | 3106 | 30.6 | 1.1 | 27.0 | 40.0 | 3106 | 0.08 | 0.03 | 0.03 | 0.43 |
| AE-P | 2858 | 30.6 | 1.1 | 28.0 | 40.0 | 2858 | 0.07 | 0.03 | 0.03 | 0.18 |
| AE-K | 2760 | 30.7 | 1.1 | 29.0 | 40.0 | 2760 | 0.07 | 0.03 | 0.03 | 0.18 |
| Yield | Sol Rad PS | 4496 | 442.4 | 12.0 | 390.7 | 483.7 | Total N (g kg-1) | 4496 | 0.01 | 0.00 | 0.00 | 0.03 |
| AE-N | 3106 | 442.6 | 10.4 | 390.7 | 483.7 | 3106 | 0.01 | 0.00 | 0.00 | 0.03 |
| AE-P | 2858 | 442.5 | 10.1 | 390.7 | 483.7 | 2858 | 0.01 | 0.00 | 0.00 | 0.03 |
| AE-K | 2760 | 443.0 | 9.4 | 390.7 | 483.7 | 2760 | 0.01 | 0.00 | 0.00 | 0.02 |
| Yield | RH mean (%) | 4496 | 78.8 | 3.0 | 61.9 | 90.0 | CEC (cmol+ kg-1) | 4496 | 7.2 | 6.6 | 0.1 | 82.9 |
| AE-N | 3106 | 78.7 | 2.4 | 61.9 | 90.0 | 3106 | 7.3 | 7.5 | 1.9 | 82.9 |
| AE-P | 2858 | 78.6 | 2.1 | 61.9 | 90.0 | 2858 | 6.5 | 1.9 | 0.1 | 17.5 |
| AE-K | 2760 | 78.6 | 2.2 | 61.9 | 90.0 | 2760 | 7.2 | 7.3 | 1.9 | 82.9 |
| Yield | RA PS (mm) | 4496 | 611.2 | 151.1 | 344.0 | 1049.0 | TEB (cmol+ kg-1) | 4496 | 0.4 | 0.1 | 0.2 | 0.8 |
| AE-N | 3106 | 607.5 | 140.9 | 344.0 | 950.8 | 3106 | 0.4 | 0.1 | 0.2 | 0.8 |
| AE-P | 2858 | 607.0 | 143.5 | 344.0 | 950.8 | 2858 | 0.4 | 0.1 | 0.2 | 0.8 |
| AE-K | 2760 | 605.7 | 143.9 | 344.0 | 950.8 | 2760 | 0.4 | 0.1 | 0.2 | 0.8 |
| Yield | Av ET (mm) | 4496 | 136.2 | 4.2 | 103.9 | 156.1 | Bulk density (g cm-3) | 4496 | 1.3 | 0.1 | 1.1 | 1.7 |
| AE-N | 3106 | 136.1 | 4.0 | 103.9 | 152.0 | 3106 | 1.3 | 0.1 | 1.1 | 1.7 |
| AE-P | 2858 | 136.0 | 4.0 | 103.9 | 145.5 | 2858 | 1.3 | 0.1 | 1.1 | 1.7 |
| AE-K | 2760 | 136.1 | 3.7 | 103.9 | 156.1 | 2760 | 1.3 | 0.1 | 1.1 | 1.7 |
| Yield | NDVI | 4496 | 0.4 | 0.0 | 0.2 | 0.6 | Total Porosity (%) | 4496 | 33.4 | 4.1 | 23.5 | 56.4 |
| AE-N | 3106 | 0.4 | 0.0 | 0.2 | 0.6 | 3106 | 33.4 | 4.5 | 23.5 | 56.4 |
| AE-P | 2858 | 0.4 | 0.0 | 0.2 | 0.6 | 2858 | 33.4 | 4.7 | 23.5 | 56.4 |
| AE-K | 2760 | 0.4 | 0.0 | 0.2 | 0.6 | 2760 | 33.4 | 4.8 | 23.5 | 56.4 |
| Yield | Slope (°) | 4496 | 1.4 | 1.2 | 0.0 | 6.0 | EC (mS m-1) | 4496 | 1.0 | 2.9 | 0.0 | 34.2 |
| AE-N | 3106 | 1.3 | 1.2 | 0.0 | 6.0 | 3106 | 0.9 | 2.6 | 0.0 | 34.2 |
| AE-P | 2858 | 1.3 | 1.2 | 0.0 | 6.0 | 2858 | 1.0 | 2.7 | 0.1 | 34.2 |
| AE-K | 2760 | 1.3 | 1.2 | 0.0 | 6.0 | 2760 | 1.0 | 2.7 | 0.1 | 34.2 |

AE-N = agronomic efficiency of nitrogen, AE-P = agronomic efficiency of phosphorus, AE-K = agronomic efficiency of potassium, n = number of treatment plots, Mean = average, SD = standard deviation, Min = minimum, Max = maximum, T min PS = minimum temperature during planting season, T max PS = maximum temperature during planting season, Sol Rad PS = solar radiation during planting season, RH mean = mean relative humidity, RA PS = total rainfall during planting season, Av ET = average evapotranspiration, NDVI = normalized difference vegetation index, OC = soil organic carbon, Total N = total nitrogen, CEC = cation exchange capacity, TEB = total exchangeable bases, EC = electrical conductivity

Table SI 4: Distribution of maize trials sites and observations for yield modelling.

|  |  |  |
| --- | --- | --- |
| Agroecological zone | Sites | Observations |
| Coastal Savanna | 3 | 11 |
| Forest-Savanna Transition | 93 | 1036 |
| Guinea Savanna | 421 | 2426 |
| Semi-Deciduous Forest | 92 | 1007 |
| Sudan Savanna | 2 | 16 |
| Total | 611 | 4496 |

Table SI 5: Distribution of maize trials sites and observations for AE-N modelling.

|  |  |  |
| --- | --- | --- |
| Agroecological zone | Sites | Observations |
| Coastal Savanna | 1 | 2 |
| Forest-Savanna Transition | 73 | 729 |
| Guinea Savanna | 395 | 1699 |
| Semi-Deciduous Forest | 83 | 669 |
| Sudan Savanna | 1 | 7 |
| Total | 553 | 3106 |

Table SI 6: Distribution of maize trials sites and observations for AE-P modelling.

|  |  |  |
| --- | --- | --- |
| Agroecological zone | Sites | Observations |
| Forest-Savanna Transition | 61 | 680 |
| Guinea Savanna | 375 | 1651 |
| Semi-Deciduous Forest | 61 | 520 |
| Sudan Savanna | 1 | 7 |
| Total | 498 | 2858 |

Table SI 7: Distribution of maize trials sites and observations for AE-K modelling.

|  |  |  |
| --- | --- | --- |
| Agroecological zone | Sites | Observations |
| Forest-Savanna Transition | 50 | 580 |
| Guinea Savanna | 375 | 1651 |
| Semi-Deciduous Forest | 60 | 522 |
| Sudan Savanna | 1 | 7 |
| Total | 486 | 2760 |

A screenshot of a graph

AI-generated content may be incorrect.

Figure SI 1*:* Distribution of maize yield and agronomic efficiencies. Panels A – D each show, on the left (i, iii, v, vii), histograms of frequency distribution, and on the right (ii, iv, vi, viii) corresponding density plots for a) maize yield measured in kg ha-1 (i-ii), b) AE-N measured in kg kg-1 (iii-iv), c) AE-P measured in kg kg-1 (v-vi), and d) AE-K measured in kg kg-1 (vii-viii).

A group of colorful boxes

AI-generated content may be incorrect.

Figure SI 2*:* Variation in maize yield and nutrient agronomic efficiency across agroecological zones in Ghana. Boxplots illustrate the distribution of (a) maize yield measured in kg ha-1, (b) AE-N measured in kg kg-1, (c) AE-P measured in kg kg-1, and (d) AE-K measured in kg kg-1 across five major agroecological zones: Coastal Savanna, Forest-Savanna Transition, Guinea Savanna, Semi-Deciduous Forest, and Sudan Savanna.

A diagram of a triangle

AI-generated content may be incorrect.

Figure SI 3*:* Pairwise correlation matrix of agronomic, environmental, and management variables associated with maize yield and nutrient agronomic efficiency across Ghana.

*A diagram of different colors

Description automatically generated*

Figure SI 4*:* Scatter density plots comparing AE-N predictions across different ML models.

A group of colored graphs

Description automatically generated with medium confidence

Figure SI 5*:* Scatter density plots comparing AE-P predictions across different ML models.

A group of colored graphs

Description automatically generated with medium confidence

Figure SI 6*:* Scatter density plots comparing AE-K predictions across different ML models.

A group of graphs with different colored dots

Description automatically generated with medium confidence

Figure SI 7: Scatter density plot of predicted against actual AE-P with a. RF, b. SVM, c. KNN, d. XGBoost, ML models with LOOCV, LSOCV and LZOCV.

A screenshot of a graph

Description automatically generated

Figure SI 8: Scatter density plot of predicted against actual AE-K with a. RF, b. SVM, c. KNN, d. XGBoost, ML models with LOOCV, LSOCV and LZOCV.

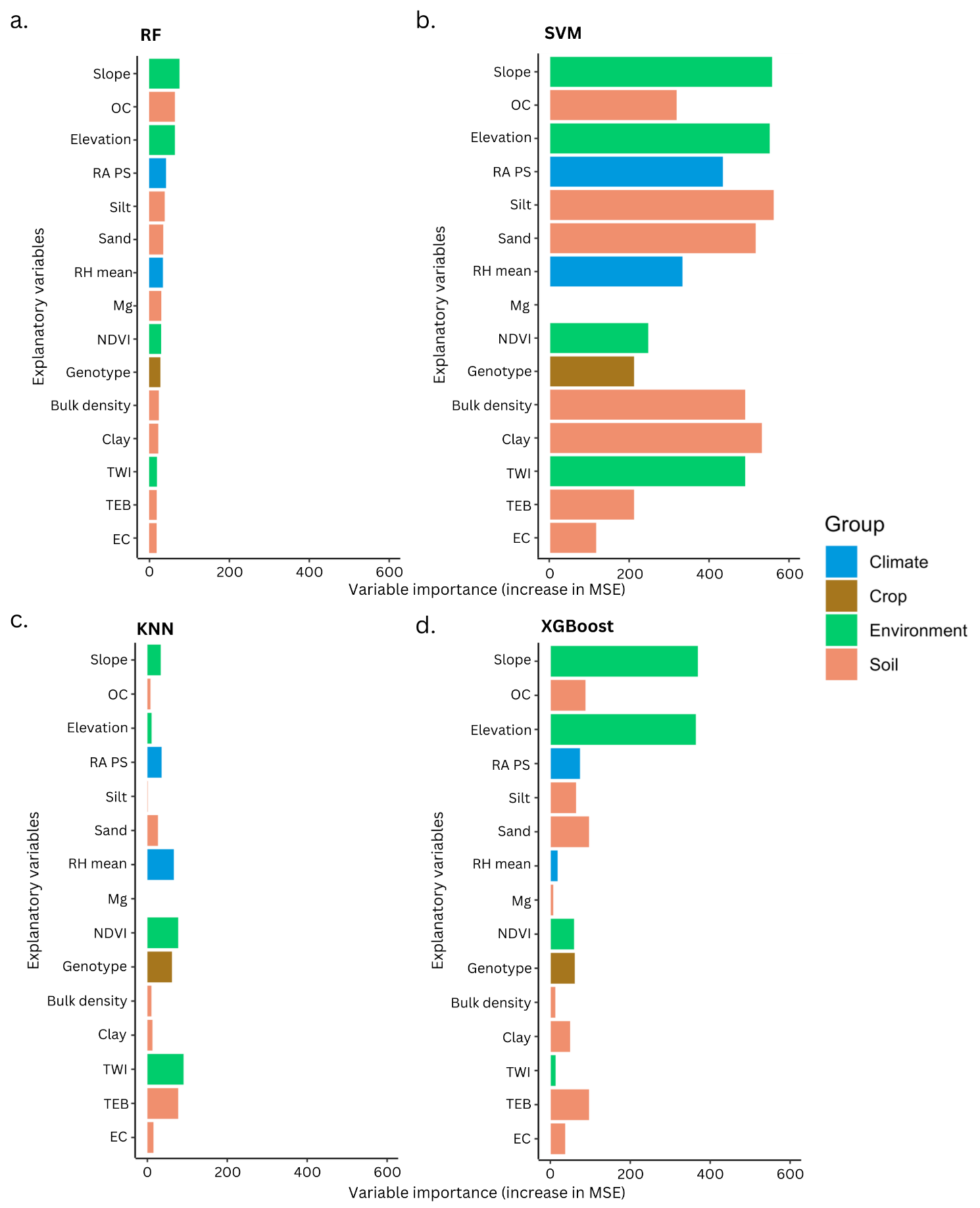


Figure SI 9: Variable importance rankings for different machine learning models used to predict AE-P. The models include (a) RF, (b) SVM, (c) KNN, and (d) XGBoost. The variable importances are arranged in the order of the RF model.

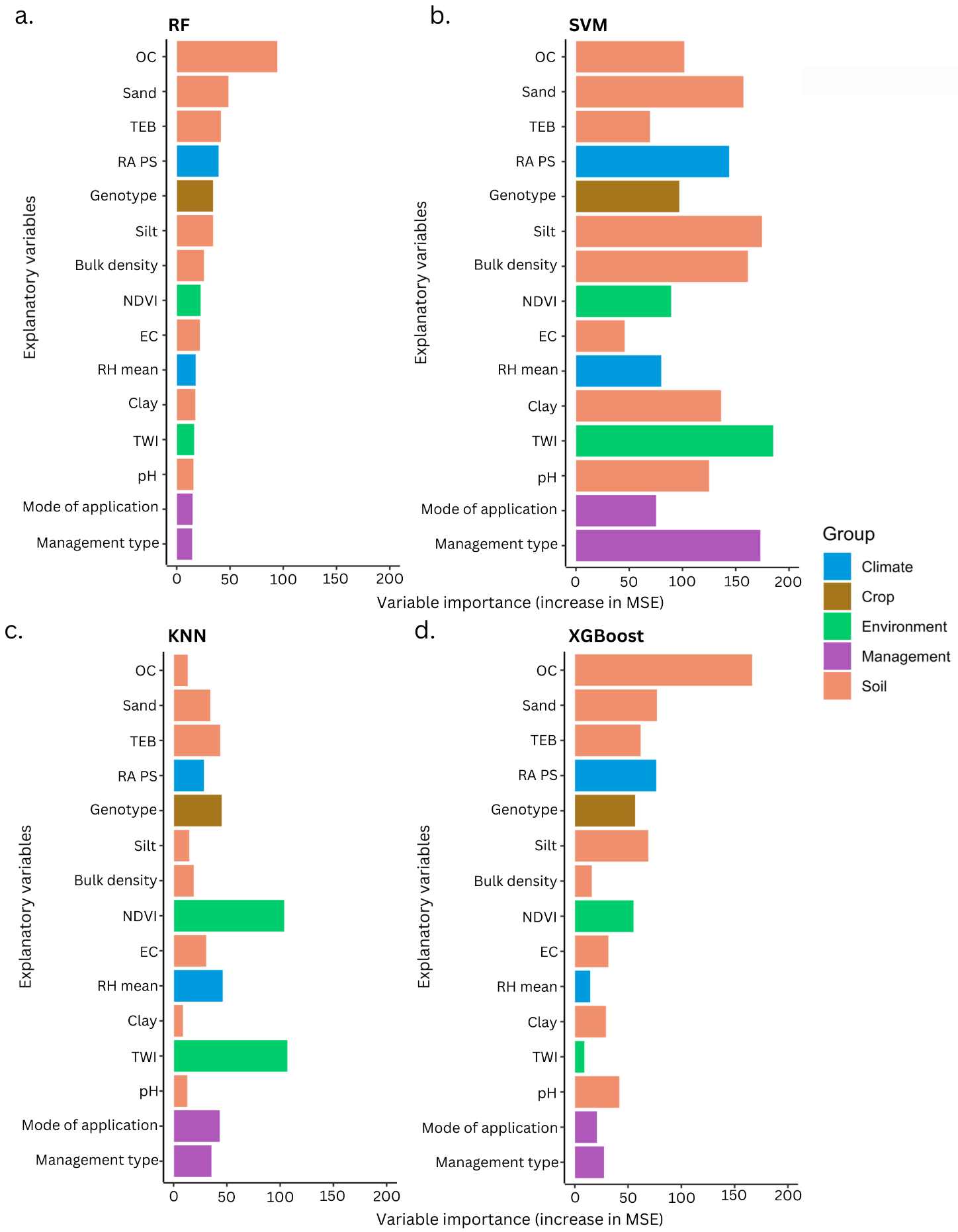


Figure SI 10: Variable importance rankings for different machine learning models used to predict AE-K. The models include (a) RF, (b) SVM, (c) KNN, and (d) XGBoost. The variable importances are arranged in the order of the RF model.

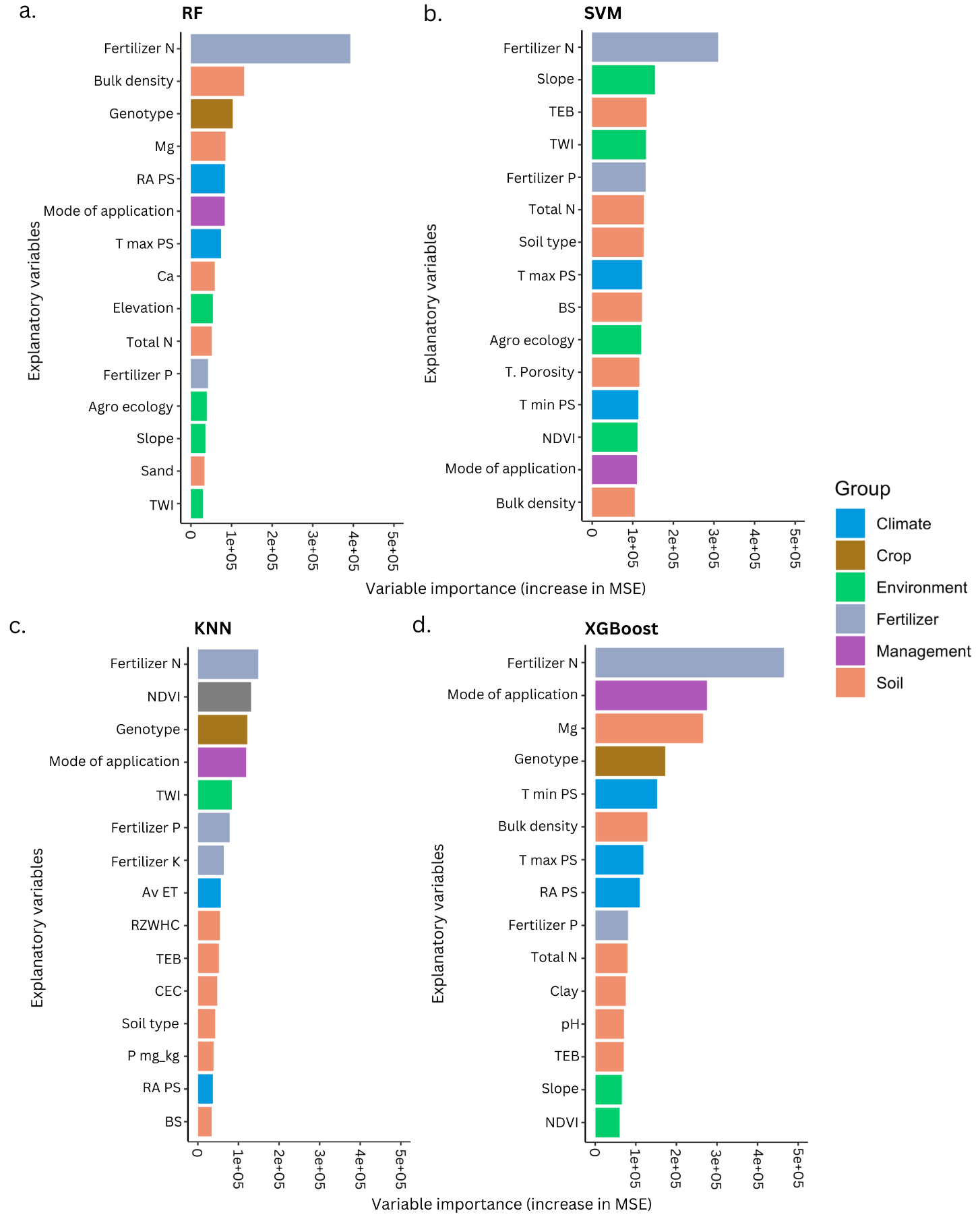


Figure SI 11: Variable importance rankings for different machine learning models used to predict maize yield. The models include (a) RF, (b) SVM, (c) KNN, and (d) XGBoost. The variable importances are arranged in descending order, with the most influential variables appearing at the top.



Figure SI 12: Variable importance rankings for different machine learning models used to predict AE-N. The models include (a) RF, (b) SVM, (c) KNN, and (d) XGBoost. The variable importances are arranged in descending order, with the most influential variables appearing at the top.



Figure SI 13: Variable importance rankings for different machine learning models used to predict AE-P. The models include (a) RF, (b) SVM, (c) KNN, and (d) XGBoost. The variable importances are arranged in descending order, with the most influential variables appearing at the top.

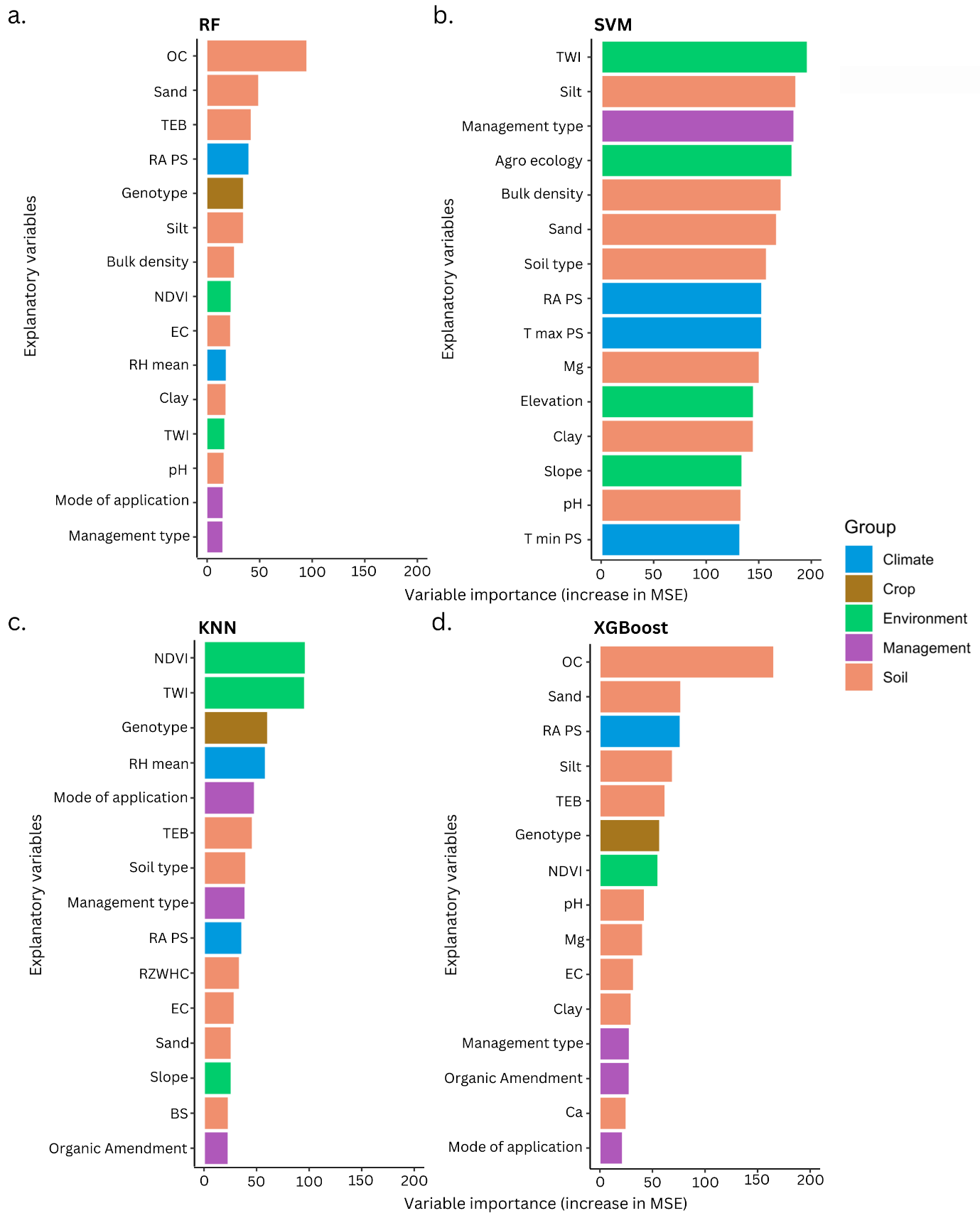


Figure SI 14: Variable importance rankings for different machine learning models used to predict AE-K. The models include (a) RF, (b) SVM, (c) KNN, and (d) XGBoost. The variable importances are arranged in descending order, with the most influential variables appearing at the top.