Journal of Forestry Research (2026) 37:49
https://doi.org/10.1007/s11676-026-01996-2

SHORT COMMUNICATION

How performance metric choice influences individual tree mortality model selection

Aitor Vázquez-Veloso[1] · Andrés Núñez-Bravo[2,3] · Astor Toraño-Caicoya[4] · Hans Pretzsch[1,4] · Felipe Bravo[1]

Received: 31 July 2025 / Accepted: 25 September 2025

Abstract Understanding tree mortality is crucial to understand forest dynamics and is essential for growth models and simulators. Although factors such as competition, drought, and pathogens drive mortality, their underlying mechanisms remain difficult to model. While substantial attention has focused on selecting appropriate algorithms and covariates, evaluating individual tree mortality models also requires careful selection of performance criteria. This study compares seven different metrics to assess their impact on model evaluation and selection. Results show that candidate models exhibited varying performances across metrics and that the choice of metric significantly influences the selection of the best model. When no confusion matrix was available, the area under the precision-recall curve (AUCPR) emerged as a more reliable alternative to the area under the ROC curve (AUC), offering a more informative assessment for imbalanced datasets. When a confusion matrix was available, Cohen's Kappa coefficient (K) and Matthews correlation coefficient (MCC) outperformed accuracy-based metrics, providing a fairer evaluation of both live and dead tree classifications. These findings emphasize the importance of choosing appropriate evaluation standards to enhance mortality model assessment and ensure reliable predictions in forestry applications.

Keywords Forest modeling · Survival · Binary classification · Area under the precision-recall curve (AUCPR) · Mathews correlation coefficient (MCC)

Abbreviations

ACC	Overall accuracy
ACCa	Accuracy in predicting live trees
ACCd	Accuracy in predicting death trees
AUC	Area Under the ROC Curve
AUCPR	Area Under the Precision-Recall Curve
bai	Basal area increment
BAL	Basal area larger than the subject tree
d	Tree diameter at breast height (1.30 m)
dMI	De Martonne Aridity Index
DT	Decision Trees
FP	False Positive values (from confusion matrix)
FPR	False Positive Rate
FN	False Negative values (from confusion matrix)
FNR	False Negative Rate

Project funding: This study was supported by the European Union and Junta de Castilla y León Education Council (ORDEN EDU/842/2022) and the IMFLEX Grant PID2021-126275OB-C22 funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe".

The online version is available at https://link.springer.com/.

Corresponding editor: Tao Xu.

✉ Aitor Vázquez-Veloso
aitor.vazquez.veloso@uva.es

[1] SMART Ecosystems Group. Departamento de Producción Vegetal y Recursos Forestales. Instituto Universitario de Investigación en Gestión Forestal Sostenible (iuFOR), ETS Ingenierías Agrarias, Universidad de Valladolid, 34004 Palencia, Spain

[2] Departamento de Estadística E Investigación Operativa, ETS Ingenierías Agrarias, Universidad de Valladolid, 34004 Palencia, Spain

[3] Agresta S. Coop, C/Duque de Fernán Núñez, 2, 1º, 28012 Madrid, Spain

[4] Tree Growth & Wood Physiology, TUM School of Life Sciences Weihenstephan, Technical University of Munich, Hans-Carl-Von-Carlowitz-Platz 2, 85354 Freising, Germany

G	Stand basal area
h	Tree total height
K	Cohen's Kappa coefficient
KNN	K-Nearest Neighbour
LR	Logistic binomial Regression
MCC	Matthews Correlation Coefficient
ML	Machine Learning
NB	Naive Bayes
RF	Random Forest
ROC Curve	Receiver Operating Characteristic Curve
SDI	Stand Density Index
SI	Site Index
SPEI	Standardized Precipitation Evapotranspiration Index
SVM	Support Vector Machine
TI	Time elapsed
TP	True Positive values (from confusion matrix)
TPR	True Positive Rate
TN	True Negative values (from confusion matrix)
TNR	True Negative Rate

Introduction

Forest modeling has advanced significantly in assessing forest dynamics, yet many challenges remain. One of the most persistent difficulties is accurately predicting individual tree mortality (Bugmann et al. 2019). Although factors such as competition (Bravo-Oviedo et al. 2006; Pretzsch et al. 2023) and drought (Senf et al. 2020) can partially explain mortality events, the predictive performance of current models often is lower than desired. In recent years, widespread tree mortality following heat and drought episodes has been observed globally, underscoring the need for more reliable mortality predictions under increasingly challenging and changing conditions (Hartmann et al. 2022). Accurate estimation of tree mortality is essential not only for evaluating forest health, but also for improving deadwood quantification and decay rate estimation (Paletto et al. 2014; Fravolini et al. 2018), both of which are critical for understanding forest dynamics and modeling carbon stock decomposition.

Proposed solutions for improving mortality models address different aspects of the problem. The covariates analyzed typically cover a range of factors, including those mentioned earlier, and are usually selected and compared carefully in each study based on their availability and explanatory power. Modeling approaches have traditionally relied on logistic regression (LR) (Hülsmann et al. 2017; Shifley et al. 2017; Pretzsch et al. 2020), although alternative algorithms have also been explored (Shearman et al. 2019; McNellis et al. 2021; Venturas et al. 2021).

Regardless of the algorithm used, candidate mortality models must be evaluated to identify the most accurate one. Binary classification model outputs range from 0 to 1, with intermediate values representing the gradual transition between the negative class (0, alive tree) and the positive class (1, dead tree). To evaluate these probabilistic outputs, performance metrics can be calculated at different classification thresholds and plotted against each other, forming curves. One such curve is the receiver operating characteristic (ROC) curve, which plots the true positive rate (TPR) against the false positive rate (FPR). In this context, TPR represents the proportion of correctly predicted dead trees among all actual dead trees, while FPR indicates the proportion of live trees incorrectly classified as dead. The area under the ROC curve (AUC) has become the standard metric for assessing mortality model classification performance (Bravo-Oviedo et al. 2006; Adame et al. 2010; Hülsmann et al. 2018; Salas-Eljatib and Weiskittel 2020; McNellis et al. 2021). However, its suitability is debatable, especially given that single-tree mortality datasets are often highly imbalanced (that is, datasets in which one class is significantly underrepresented due to the low probability of an event; in our case, dead trees resulting from natural mortality). For example, annual mortality rates have been reported as low as 0.35% in Pyrenean oak monospecific stands of north-west Spain (Adame et al. 2010), around 8% in Scots pine and Mediterranean pine monospecific stands in Spain (Bravo-Oviedo et al. 2006), and ranging from 5% to 20% in young stands to 1%–2% in mature stands of unmanaged monospecific plots of common tree species across Europe (Pretzsch et al. 2023). Since AUC is sensitive to class imbalance, alternative metrics may be more appropriate. The area under the precision-recall curve (AUCPR) (Saito and Rehmsmeier 2015) follows a similar approach and is more suitable for imbalanced datasets, yet it remains largely unknown in assessing individual tree mortality. In this context, precision refers to the proportion of correctly predicted dead trees among all trees predicted as dead, while recall (also known as the TPR) refers to the proportion of actual dead trees that were correctly identified. AUCPR emphasizes performance on the minority class (in our case, dead trees), making it a valuable tool for evaluating models under class imbalance conditions.

Alternatively, when a classification threshold is applied and a confusion matrix is available, model performance is typically evaluated using criteria such as overall accuracy (ACC) or group-specific accuracy, with ACCa for live trees and ACCd for dead ones. However, these metrics are less suitable for imbalanced datasets, as they do not adequately represent model performance. To address this issue, Cohen's Kappa coefficient (K) (Cohen 1960) was proposed in previous studies (Bravo and Bravo-Núñez 2017; Reis et al. 2018) and Matthews correlation coefficient (MCC) (Matthews 1975) has been suggested as more reliable under such conditions (Chicco 2017; Chicco and Jurman

2020). Despite being less explored in forest mortality model evaluation, these standards better account for data imbalances, improving model assessment. This is particularly important in mortality modeling, where "dead" trees, though less frequent, are the most critical group for long-term simulations. Errors in their prediction accumulate over time, potentially compromising forest dynamics simulations (Bircher et al. 2015). Moreover, accurately tracking dead trees is essential for studies on biodiversity, wood decomposition, and carbon sequestration (Friend et al. 2014), among other topics.

While both explanatory variables and the suitability of different algorithms for predicting individual tree mortality have been previously assessed, the criteria for selecting the best-performing model remain underexplored. In this study, we assessed the adequacy of performance metrics for evaluating individual tree mortality models to determine whether model selection outcomes vary depending on the metric used. We used long-term experimental plots of Norway spruce growing under natural conditions, excluding exceptional mortality events such as windthrows, pest outbreaks, and wildfire-induced mortality. To this end, we compared a range of evaluation criteria to understand their behavior and the specific insights each provides for model selection. Model performance was assessed across different combinations of explanatory variables and algorithms, applying various metrics depending on whether a confusion matrix was available. This study addresses the following key questions:

(1) Do different metrics yield varying performance ranges when evaluating the same models?
(2) Does the choice of metric influence which model is selected as the best?
(3) Are AUC and ACC, the standard benchmark for mortality model evaluation, the best choice?

Materials and methods

Data

This study utilized observational data from long-term Norway spruce (*Picea abies* (L.) H. Karst) trials conducted in Bavaria, Germany. These experimental plots comprise even-aged, pure stands established at varying planting densities under distinct thinning treatments. To focus solely on natural mortality processes, only data from unmanaged control plots were included. Thus, the dataset comprises 11,380 tree records from 12 plots, with a natural mortality proportion of 14.6% measured at varying intervals between 1980 and 2018. Tree attributes such as survival status, spatial coordinates, size, and age were recorded during periodic inventories. Ecological variables such as competition indices, growth measurements, and site productivity were derived from the original records, while climate variables were sourced from WorldClim datasets (Fick and Hijmans 2017; Harris et al. 2020). For further details regarding the dataset structure and experimental design, see Vázquez-Veloso et al. (2025).

Analysis

To explore individual tree mortality, a range of covariates was grouped thematically into categories: tree size, stand productivity, competition, growth rates, inventory timing, climatic conditions, and tree social status. These variables were selected and combined based on their ecological relevance and avoiding multicollinearity. Model setups were constructed by systematically combining variables across categories, yielding 180 candidate covariate sets per algorithm. This approach was guided by ecological reasoning, ensuring that the resulting models remained interpretable and biologically meaningful.

Following the covariate selection, several algorithms were used to evaluate its performance. Logistic binary regression (LR) was included as the benchmark algorithm in predicting individual tree mortality (Merkl and Hasenauer 1998; Bravo-Oviedo et al. 2006; Hülsmann et al. 2017; Shifley et al. 2017; Shearman et al. 2019). In addition, various machine learning (ML) algorithms were tested, including decision trees (DT), random forest (RF), naive bayes (NB), K-nearest neighbors (KNN), and support vector machines (SVM). Each combination of covariates and algorithm constituted a candidate model, whose performance was evaluated as described in the following section. All candidate models were trained and tested using the same methodological framework to ensure comparability, with analyses conducted in R (R Core Team 2021) using the caret library (Kuhn 2008). Additional methodological details are available in Vázquez-Veloso et al. (2025).

Evaluation metrics

The metrics used address two different analytical approaches. When no classification threshold was applied, model performance was evaluated by analyzing metric values across all possible thresholds using graphical methods. In these cases, a single classification score was obtained by calculating the area under the curves generated by those methods. For example, the area under the ROC curve (AUC) provides a single value summarizing the true positive rate (TPR) and false positive rate (FPR), representing overall model classification performance. Similarly, the area under the precision-recall curve (AUCPR) condenses the precision and recall values into a single score (Saito and Rehmsmeier 2015). Both AUC and AUCPR were included in this study for model selection without applying a classification threshold. These metrics serve, respectively, as the standard benchmark and the recommended alternative for imbalanced datasets.

When a threshold is applied, a confusion matrix becomes available, allowing the use of different metrics to evaluate model

performance. The confusion matrix allows the visualization of the model predictions by displaying the number of true positives, true negatives, false positives, and false negatives. Each row represents instances of the actual class, while each column represents instances of the predicted class. The main diagonal contains correctly classified cases, whereas off-diagonal elements represent misclassifications. Although threshold-based classification is implemented frequently in LR, all the above algorithms produce class probabilities or scores that can be thresholded to convert continuous results into binary mortality predictions. Each candidate model had its own optimal threshold selected to maximize predictive performance, thereby supporting consistent evaluation across models.

While goodness-of-fit statistics assess how well a statistical model represents the data, performance metrics quantify the quality of its predictions. In this study, goodness-of-fit statistics were not considered, as the main objective was to evaluate model predictability. From the prediction outputs, various performance metrics were derived to assess classification performance across different classes (Table 1). Accuracy (ACC) is a commonly used benchmark to assess how well a binary classification model identifies or excludes a condition, in this case, individual tree mortality. Accuracy is further divided into live tree accuracy (ACCa) and dead tree accuracy (ACCd), both ranging from 0 (worst prediction score) to + 1 (perfect prediction score). These three criteria were included in this study as part of the benchmark and because of their ease of interpretation. Alternatively, Cohen's kappa coefficient (K) (Cohen 1960) measures inter-rated reliability (also intra-rated reliability) for qualitative items, while accounting for the possibility of agreement occurring by chance. It ranges from + 1 (perfect agreement) to 0 (no agreement beyond random chance) and can take negative values when there is no correlation between raters' classifications. The Phi coefficient, also known as the Matthews correlation coefficient (MCC) (Matthews 1975), quantifies the correlation between observed and predicted binary classifications. It returns to a value between − 1 and + 1, where + 1 represents a perfect prediction, 0 indicates a random prediction, and − 1 signifies complete disagreement between predictions and observations. Both K and MCC were included in this study as alternatives to accuracy-based metrics for cases where a threshold-based classification was applied.

All the metrics were calculated using the R package ROCR (Sing et al. 2005) and irr (Gamer et al. 2019) in the case of Cohen's Kappa coefficient, as well as R software (R Core Team 2021).

Table 1 Summary of the different metrics used to evaluate binary classification analysis

Metric type	Name	Formula
Non-fixed threshold	AUC	Area Under the ROC Curve
	AUCPR	Area Under the Precision-Recall Curve
	ACC	$\frac{(TP+TN)}{(TP+TN+FP+FN)}$
	ACCa	$\frac{TN}{(TN+FN)}$
	ACCd	$\frac{TP}{(TP+FP)}$
Fixed threshold	K	$\frac{2\cdot(TP\cdot TN+FP\cdot FN)}{(TP+FP)\cdot(FP+TN)+(TP+FN)\cdot(FN+TN)}$
	MCC	$\frac{(TP\cdot TN+FP\cdot FN)}{\sqrt{(TP+FP)\cdot(TP+FN)\cdot(TN+FP)\cdot(TN+FN)}}$

The metric type refers to the case study, distinguishing between cases with and without a fixed classification threshold. Mathematical expressions are based on the confusion matrix outputs: *TP* represents True Positives, *TN* represents True Negatives, *FP* represents False Positives, and *FN* represents False Negatives. Note that $TP+TN+FP+FN$ equals the total number of trees in the dataset. An extended version of this table, including additional metrics, is available in Appendix A

Results

The classification performance of the logistic regression (LR) algorithm varied significantly depending on the evaluation metric used, showing different score distributions (Fig. 1). Accuracy (ACC) exhibited a narrow distribution across the full set of candidate models, ranging from 0.84 to 0.91. Accuracy for predicting live trees (ACCa) had the highest values, ranging from 0.86 to 0.95, whereas accuracy for dead trees (ACCd) showed the widest range among the studied metrics, from 0.41 to 0.76. The area under the ROC curve (AUC) indicated relatively high scores from 0.69 to 0.82, while the area under the precision-recall curve (AUCPR) had lower scores with a wider distribution from 0.49 to 0.68. Both Cohen's kappa coefficient (K) and the Matthews correlation coefficient (MCC) had the tightest distributions, ranging from 0.46 to 0.58 and 0.47 to 0.58, respectively.

The choice of the best LR model among candidate models varied significantly depending on the selection criteria. Each metric identified a different model as the best classification performance (Fig. 2). When comparing the scores of these selected models across all criteria, clear trends emerged. The model selected by ACCa registered the poorest score on ACCd, and vice versa. The model selected by ACC also had low predictability for dead trees based on ACCd. The model selected using AUC registered a good score in predicting dead trees but poorer for live ones, ranking among the lowest across AUCPR, K, and MCC. Models selected by AUCPR, K and MCC exhibited more balanced but moderate performance, with prediction scores near the average for both live and dead trees.

Comparing the classification performance of candidate models across algorithms using the study metrics clarified each behavior (Fig. 3). ACC and ACCa ranked models similarly, showing minimal differences between algorithms. ACCd, however, highlighted clearer distinctions in algorithm classification performance. AUC clustered most candidate covariate groups above a score of 0.7 (except SVM), while AUCPR emphasized differences. Both K and

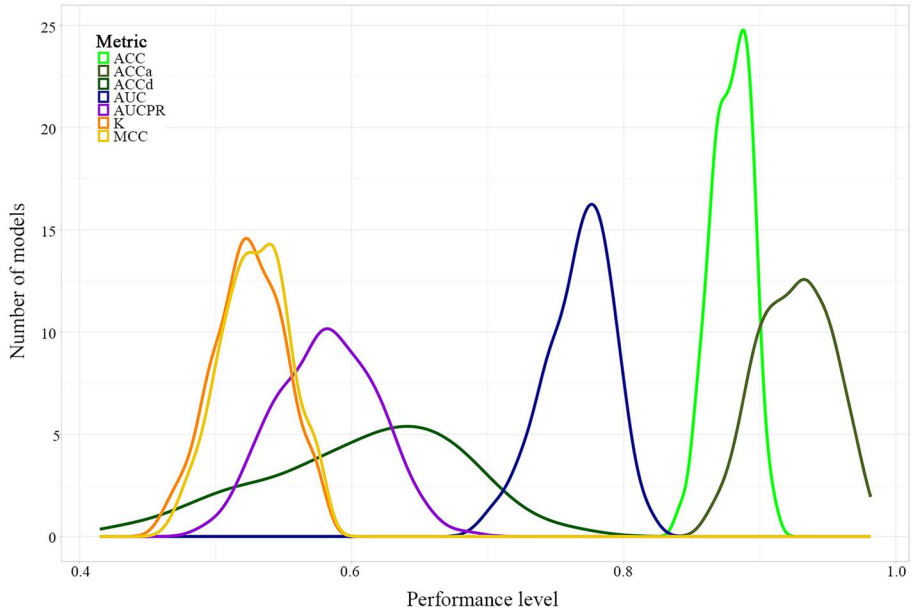


Fig. 1 Classification performance distribution of 180 candidate models fitted using the Logistic Regression algorithm. Metrics assessing model classification performance are represented in different colors

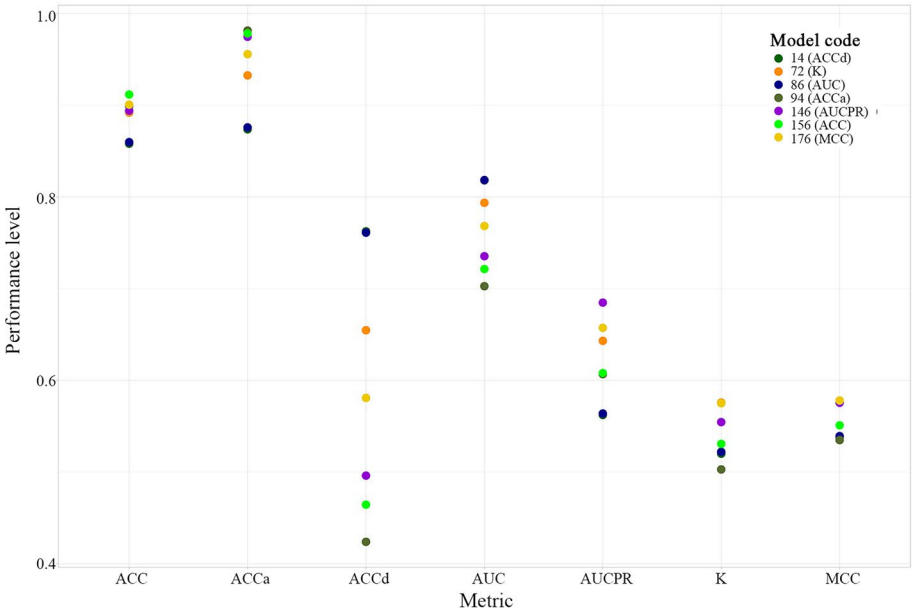


Fig. 2 Comparison of the best Logistic Regression model across different metrics. The best model for each metric was identified (distinguished by color according to the legend) and evaluated using all studied metrics

MCC produced similar results across covariate groups and algorithms, showing differences between them.

While Fig. 1 focused on LR, where each metric selected a different covariate group as the best, two clear patterns emerge when analyzing the best covariate group across all algorithms (Fig. 4). First, the best covariate group selected by ACC, ACCa, and ACCd never coincided across algorithms. Second, for all the algorithms except LR, the same covariate group was consistently chosen by AUCPR, K and MCC.

Discussion

Overview

The implications of fitting individual tree mortality models using different dataset structures were previously reported by Vázquez-Veloso et al. (2025), which also addressed the comparison of various algorithms in terms of predictive performance. The present study complements those findings by focusing on the role of evaluation metrics, an essential aspect for identifying the best-performing model among available alternatives. This section discusses how the perception of model performance varies depending on the metric used and how the metric choice

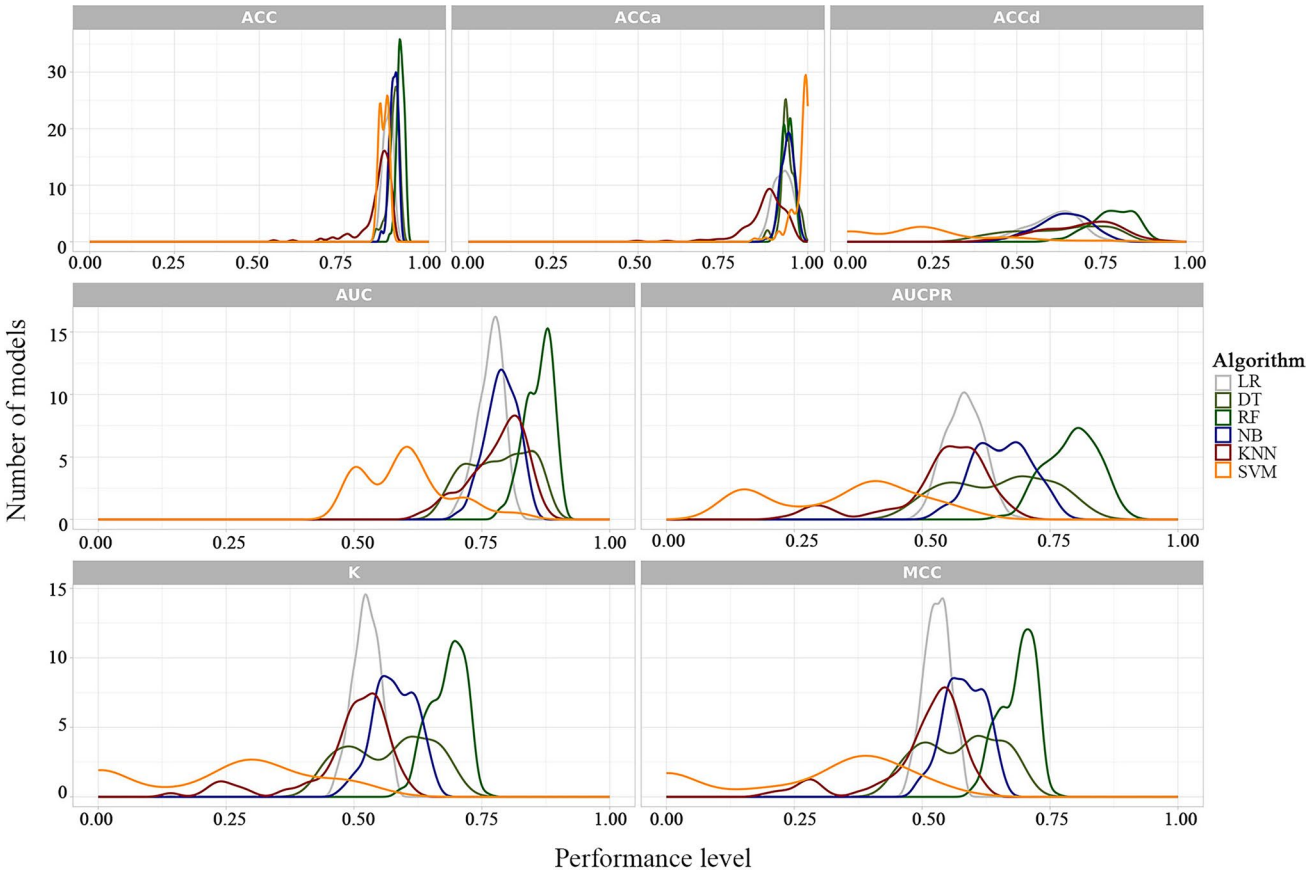
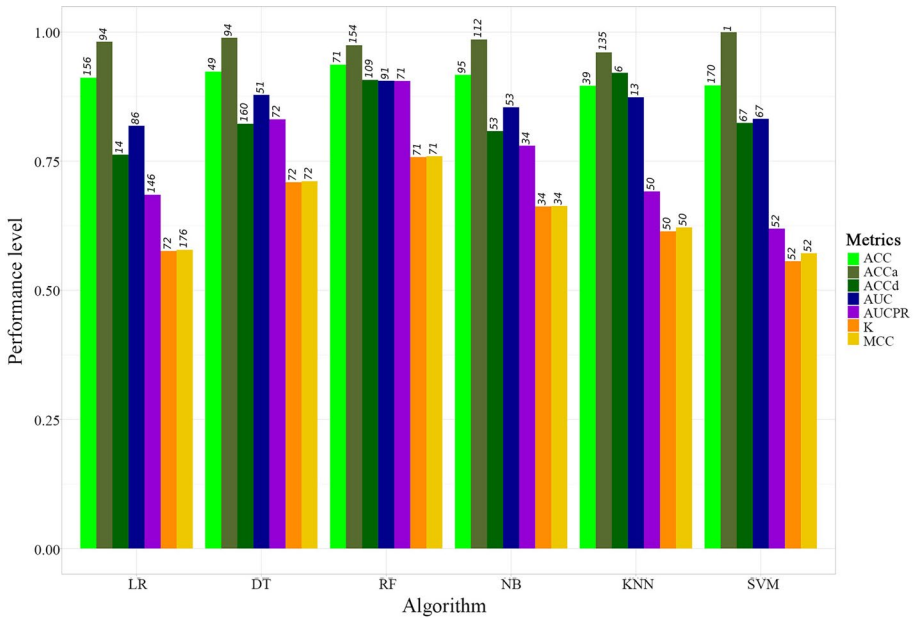


Fig. 3 Classification performance distribution of 180 candidate models across different algorithms. Each box represents the classification performance distribution for a specific metric, with colors indicating the algorithm used for model training

Fig. 4 Comparison of the best covariate group selected per algorithm across different metrics. The best model per algorithm and metric was identified (distinguished by color according to the legend), and the corresponding covariate group can be found in the graph (number). Each selected covariate group was then evaluated using all studied metrics across algorithms



influences the selection of the most suitable model. It concludes with practical recommendations for users aiming to fit new individual tree mortality models.

Distribution of classification performance among metrics

The results demonstrate clear differences in classification performance across metrics, revealing consistent trends. Accuracy metrics were used to assess model behavior in classifying live and dead trees, each providing a different interpretation of their score. Among the 180 candidate models fitted using LR, ACCa had the highest scores in absolute values, followed by ACC, while ACCd had the lowest and most variable for the same models. Differences in predictability between live (ACCa) and dead (ACCd) trees were expected due to class imbalance, as models generally predicted live trees more accurately. This trend is attributed to the higher frequency of live tree records in the dataset, while dead tree predictions were more unstable, showing a wider classification performance range. Although ACC provides a general indication of prediction reliability, it can sometimes be misleading. In imbalanced datasets, ACC scores are biased toward the majority class, which may result in apparently high performance even when the minority class is poorly predicted (Naidu et al. 2023; da Rocha et al. 2018; Reis et al. 2018). ACC, which accounts for both classes, yielded lower scores than ACCa, reflecting the penalty associated with poor dead tree predictability.

When evaluating candidate covariate groups across different algorithms, accuracy metrics followed similar trends. Although classification performance ranges varied by algorithm, all exhibited consistent patterns, with dead trees remaining the most challenging group to predict.

When comparing metric scores across candidate models, consistent patterns emerged among algorithms. AUC exhibited a broader classification performance range and consistently higher than AUCPR when evaluating the same models across algorithms. However, this outcome was expected due to differences in their mathematical formulation. AUC is based on TPR and TNR, optimizing the correct classification of both positive and negative values. Still, it was reported that AUC can mask poor performance in imbalanced datasets, since its values are not sufficiently attenuated under such conditions (Jeni et al. 2013). In contrast, AUCPR relies on Precision and Recall, focusing solely on the accurate classification of positive cases (in our study, dead trees). Since natural processes often involve sparse datasets with few positive instances (e.g., dead trees), excluding negative values in the prediction score is often preferable. In such cases, AUCPR provides a more reliable model evaluation score (Chicco 2017; Chicco and Jurman 2020).

The remaining metrics, K and MCC, were also evaluated. Despite their conceptual differences, both yielded similar scores across covariate groups and remained consistent across algorithms. Compared to previous metrics, their scores were more conservative, emphasizing the importance of dead tree predictions in the overall evaluation of the model. This makes them preferable for assessing model performance, as noted by de la Cruz Huayanay et al. (2025).

Therefore, while K and MCC produced comparable scores, classification performance ranges varied significantly across the criteria studied when evaluating all setups tested for each algorithm.

Best candidate model selection among metrics

Among the 180 candidate models fitted using the LR algorithm, each metric selected a different best-performing model. When these models were evaluated across all metrics, distinct patterns emerged. For example, the model that maximized ACCa typically showed the lowest ACCd score, and vice versa, illustrating the trade-off between predicting one class well at the expense of the other, a pattern reported by Vujovic (2021). Similarly, the model selected by overall accuracy (ACC) underperformed on ACCd, reflecting its bias toward the majority class (live trees). This trade-off arises not from fitting separate models for survival and mortality, as each model directly estimated mortality probabilities with survival defined as the complement, but rather from the effects of thresholding under class imbalance. This underscores the importance of using balanced metrics such as AUCPR, K, and MCC that reflect performance across both classes and avoids misleading conclusions based on single-class optimization (Chicco 2017; Chicco and Jurman 2020; de la Cruz Huayanay et al. 2025).

Focusing again on LR results, the model selected based on AUC achieved high accuracy in predicting dead trees but at the cost of reduced live tree predictability. Consequently, it showed a lower score when evaluating with AUCPR, K, and MCC. A similar pattern was observed across all algorithms, with two cases (NB and SVM) where the models selected by ACCd and AUC were the same. Although prioritizing dead tree prediction is essential, excessively misclassifying live trees as dead leads to overestimating stand-level mortality. In individual tree mortality simulations, the objective is to predict dead trees accurately while maintaining a realistic mortality rate at the stand level. When models are used iteratively, such as in long-term forest management simulations (Bravo et al. 2025), misclassifications have different long-term consequences. If a tree is misclassified as alive, the mortality model re-evaluates it in the next simulation step, potentially correcting the error. However, if a tree is misclassified as dead, it is permanently removed from the simulation, eliminating its influence on stand dynamics and propagating the error across all subsequent projections. This highlights the unequal costs of misclassifying the positive versus negative class, as exemplified by Naidu et al. (2023) in the context of distinguishing sick from healthy individuals. To address

this issue in individual tree mortality, an alternative approach is to develop models that return a probability value rather than a binary classification (see, for example, Bravo-Oviedo et al. 2006; Hülsmann et al. 2016). The predicted mortality probability can then be multiplied by the tree expansion factor, which may improve the representation of mortality at the stand level. This approach aligns with common simulation practices in which trees represent cohorts within sample plots and mortality is applied proportionally rather than by deterministic removal. In such settings, the exact identity of dead trees is less important than the aggregated mortality response. Thus, while our study uses threshold-based classification to evaluate tree-level discrimination, we acknowledge that, in operational forest simulators, applying predicted probabilities directly may yield more stable stand-level dynamics.

For the LR algorithm, the models selected by AUCPR, K, and MCC differed from each other but showed similar performances across all study metrics. When evaluated with AUCPR, K, and MCC, these models outperformed those selected by ACC, ACCa, ACCd, and AUC, while still achieving acceptable accuracy when predicting live and dead trees separately. For the remaining algorithms, these metrics consistently selected the same covariate groups, reinforcing their similar behavior. Although their class-specific accuracy was not the highest, the models chosen by AUCPR, K, and MCC maintained a balanced trade-off, optimizing overall model predictability.

Therefore, AUCPR, K, and MCC consistently selected the same or similar model that optimized overall predictability. In contrast, AUC favored models that maximized dead tree predictability but at the cost of reduced accuracy for live trees, negatively affecting overall model classification performance independently of the algorithm employed.

Best metric choice for evaluating individual tree mortality models

Because binary classification models may be evaluated with or without a confusion matrix, the discussion on optimal evaluation metrics must address each scenario separately.

When no confusion matrix is available, models predict a mortality probability between 0 and 1, assigning trees a degree of mortality risk rather than a binary classification. In such cases, AUC is the most used measure according to the literature (Bravo-Oviedo et al. 2006; Adame et al. 2010; Hülsmann et al. 2018; Salas-Eljatib and Weiskittel 2020; McNellis et al. 2021). AUC tends to favor models that prioritize dead tree prediction at the expense of correctly identifying live trees, due to its known limitations in imbalanced datasets (Jeni et al. 2013; Chicco 2017; Chicco and Jurman 2020). A more appropriate alternative is the AUCPR, a standard more commonly used in other scientific fields. In our study, models selected by AUCPR achieved more balanced predictions for both classes. Moreover, when evaluated with K and MCC (both of which often selected the same best model), AUCPR-based models maintained comparable performance, reinforcing their suitability for mortality modeling. In simulation settings, such probability-based models are applied by multiplying the predicted mortality probability by the tree's expansion factor or frequency, enabling continuous estimation of stand-level mortality (Bravo et al. 2001; Bravo-Oviedo et al. 2006). This avoids the need for discrete classification, preserves structural realism, and reduces the bias introduced by thresholds. In that simulation context, we suggest that AUCPR be considered a preferred metric for evaluating mortality models, given its suitability for imbalanced data and the tendency of AUC to mask poor performance (Jeni et al. 2013).

When a confusion matrix is available based on a threshold value, trees are strictly classified as "dead" or "alive", eliminating intermediate probability values. In these cases, ACC provides an overall assessment of model classification performance, while ACCa and ACCd evaluate predictability for each group separately. However, ACC tends to be misleading due to the high proportion of live trees (often exceeding 90%), as it produces scores biased toward the majority class (Naidu et al. 2023). To address these limitations, K and MCC serve as more robust alternatives (de la Cruz Huayanay et al. 2025). While their underlying principles differ, our study found only minor differences in their performance scores, and both consistently identified the same best-performing model for each algorithm (excluding LR, but both models showed similar behaviors). Although previous studies recommend MCC for imbalanced datasets (Chicco 2017; Chicco and Jurman 2020), our findings do not strongly support favoring MCC over K. However, we confirm that both are superior to accuracy-based metrics for evaluating mortality models, in line with de la Cruz Huayanay et al. (2025). In a simulation context, models selected using the aforementioned metrics aimed to achieve accurate dead tree prediction while maintaining realistic stand-level mortality rates. We recommend selecting models with the most balanced predictability between classes, reflected in measures such as K and MCC, to improve the simulation of stand dynamics and reduce error propagation over time.

Conclusion

Two different approaches were analyzed to determine the most suitable standards for evaluating individual tree mortality models. Our analysis of long-term experimental plots demonstrates that metric selection significantly impacts model assessment and best model identification. When no confusion matrix is available, AUCPR proves to be a more reliable alternative to AUC, as it provides a

more balanced evaluation in imbalanced datasets, where dead trees represent a minority class but the most critical one. In contrast, AUC tends to favor models that maximize dead tree predictability at the expense of overall classification accuracy. When a confusion matrix is available, accuracy-based metrics (ACC, ACCa, and ACCd) can be misleading due to the high prevalence of living trees. Instead, K and MCC are better alternatives, as they account for both classes more effectively and provide a fairer evaluation of model classification performance. These findings, which were consistent across the six algorithms studied, highlight the importance of carefully selecting evaluation metrics to ensure reliable predictions of individual tree mortality. Given that these models are often used in simulation frameworks, selecting metrics that promote balanced predictability between classes can help improve the accuracy and stability of long-term forest dynamics projections. Future studies should further explore the impact of metric selection across different data structures, forest types, and management scenarios to refine mortality model evaluation methods. In addition, the adoption of a standardized and consistent evaluation framework represents an important topic for future research.

Acknowledgements The authors thank the Bayerischen Staatsforsten (BaySF) for supporting the establishment of the plots and the Bavarian State Ministry for Nutrition, Agriculture, and Forestry for permanent support of the project W 07 "Long-term experimental plots for forest growth and yield research" (# 7831-22209-2013).

Author contributions Aitor Vázquez-Veloso: conceptualization, data curation, formal analysis, methodology, resources, writing – original draft, writing – review & editing; Andrés Bravo Núñez: conceptualization, supervision, writing – original draft, writing – review & editing; Astor Toraño-Caicoya: resources, methodology, data curation, supervision; Hans Pretzsch: resources, conceptualization, funding acquisition, methodology, supervision. Felipe Bravo: conceptualization, funding acquisition, methodology, supervision.

Funding Open access funding provided by FEDER European Funds and the Junta de Castilla y León under the Research and Innovation Strategy for Smart Specialization (RIS3) of Castilla y León 2021-2027.

Data availability Original data, results, code, figures, and bibliography used in this study are available in the following links: Zenodo: https://doi.org/10.5281/zenodo.17387888. GitHub (linked to Zenodo DOI): https://github.com/aitorvv/metrics_for_individual_tree_mortality_models

Declarations

Conflict of interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A

An extended version of Table 2 (following the main text labels) is provided in this section, including the mathematical expressions of selected metrics mentioned in the text, based on values extracted from the confusion matrix.

Appendix B

Complementary graphs supporting the figures presented in the text. The graph included in the text is repeated here to facilitate direct comparison. Figure 5 complements Fig. 4 of the manuscript, while Figs. 6 and 7 complements Fig. 6.

Table 2 Summary of different support metrics used to evaluate binary classification analysis, and its mathematical formula based on confusion matrix

Name	Formula
True Positive Rate (TPR) or Sensitivity	$TP/(TP+FN)$
True Negative Rate (TNR) or Specificity	$TN/(TN+FP)$
False Positive Rate (FPR)	$FP/(FP+TN)$
False Negative Rate (FNR)	$FN/(FN+TP)$
Precision	$TP/(TP+FP)$
Recall	$TP/(TP+FN)$

According to the confusion matrix, TP represents True Positives, TN represents True Negatives, FP represents False Positives, and FN represents False Negatives

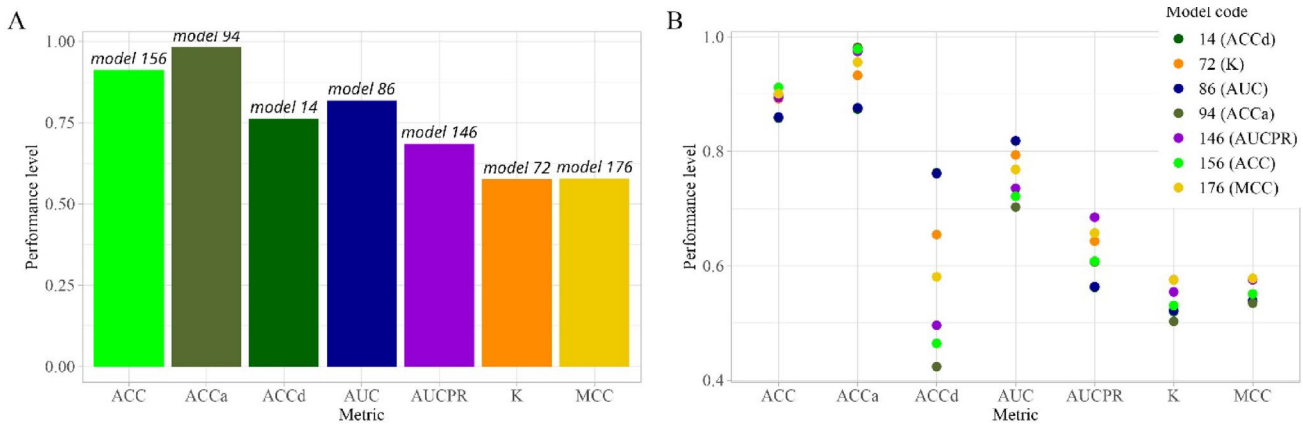


Fig. 5 Comparison of the best Logistic Regression model across different metrics. **A** Selection of the best model based on the highest score for each metric, **B** Evaluation of each metric's selected best model using all studied metrics. Metrics assessing model classification performance are represented in different colors

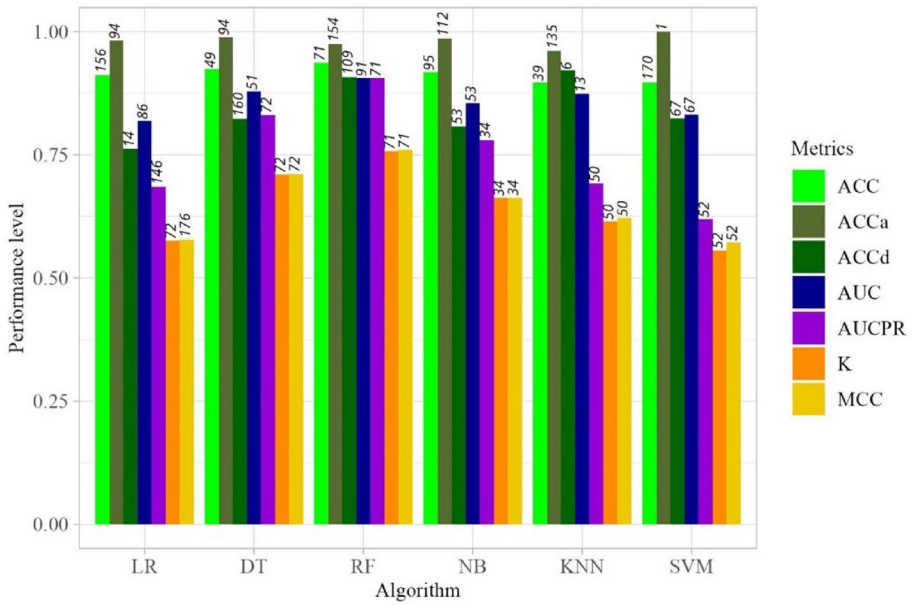


Fig. 6 Comparison of the best covariate group selected per algorithm across different metrics: selection of the best covariate group based on the highest score for each metric across algorithms

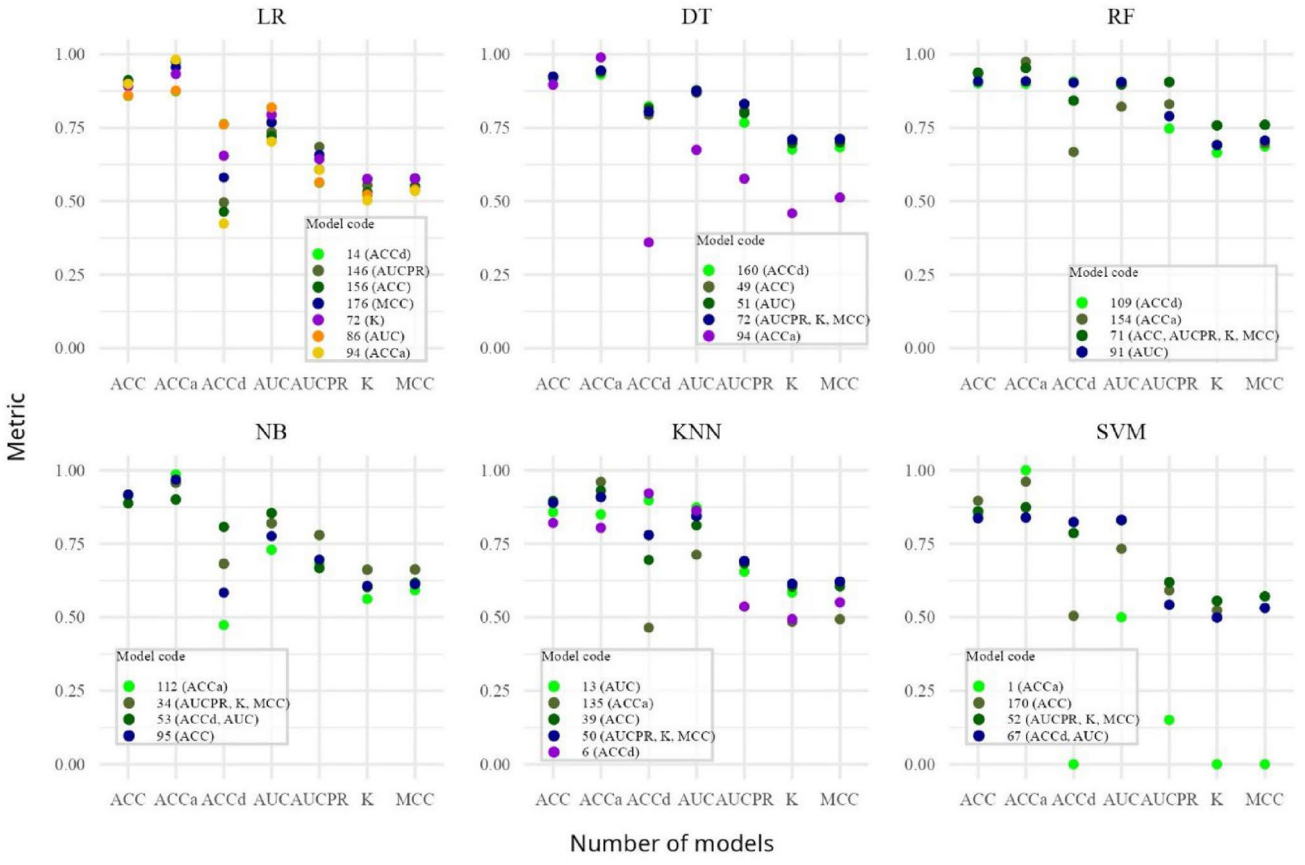


Fig. 7 Comparison of the best covariate group selected per algorithm across different metrics: evaluation of each metric's selected best covariate group using all studied metrics across algorithms

References

Adame P, del Río M, Cañellas I (2010) Modeling individual-tree mortality in Pyrenean oak (*Quercus pyrenaica* Willd.) stands. Ann for Sci 67(8):810. https://doi.org/10.1051/forest/2010046

Bircher N, Cailleret M, Bugmann H (2015) The agony of choice: different empirical mortality models lead to sharply different future forest dynamics. Ecol Appl 25(5):1303–1318. https://doi.org/10.1890/14-1462.1

Bravo F, Hann DW, Maguire DA (2001) Impact of competitor species composition on predicting diameter growth and survival rates of Douglas-fir trees in southwestern Oregon. Can J for Res 31(12):2237–2247. https://doi.org/10.1139/x01-164

Bravo F, Ordóñez C, Vázquez-Veloso A, Michalakopoulos S (2025) SIMANFOR cloud decision support system: structure, content, and applications. Ecol Model 499:110912. https://doi.org/10.1016/j.ecolmodel.2024.110912

Bravo F, Bravo-Núñez A (2017) Clasificación de la calidad de estación forestal mediante técnicas de aprendizaje automático (machine learning). In: Actas 7º Congreso Forestal Español. Sociedad Española de Ciencias Forestales, Cáceres. (in Spanish). https://www.congresoforestal.es/fichero.php?t=41725&i=5686&m=2185

Bravo-Oviedo A, Sterba H, del Río M, Bravo F (2006) Competition-induced mortality for Mediterranean *Pinus pinaster* Ait. and *P. sylvestris* L. For Ecol Manage 222(1–3):88–98. https://doi.org/10.1016/j.foreco.2005.10.016

Bugmann H, Seidl R, Hartig F, Bohn F, Brůna J, Cailleret M, François L, Heinke J, Henrot AJ, Hickler T, Hülsmann L, Huth A, Jacquemin I, Kollas C, Lasch-Born P, Lexer MJ, Merganič J, Merganičová K, Mette T, Miranda BR, Nadal-Sala D, Rammer W, Rammig A, Reineking B, Roedig E, Sabaté S, Steinkamp J, Suckow F, Vacchiano G, Wild J, Xu CG, Reyer CPO (2019) Tree mortality submodels drive simulated long-term forest dynamics: assessing 15 models from the stand to global scale. Ecosphere 10(2):e02616. https://doi.org/10.1002/ecs2.2616

Chicco D (2017) Ten quick tips for machine learning in computational biology. BioData Min 10(1):35. https://doi.org/10.1186/s13040-017-0155-3

Chicco D, Jurman G (2020) The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 21(1):6. https://doi.org/10.1186/s12864-019-6413-7

Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20(1):37–46. https://doi.org/10.1177/001316446002000104

da Rocha SJSS, Torres CMME, Jacovine LAG, Leite HG, Gelcer EM, Neves KM, Schettini BLS, Villanova PH, da Silva LF, Reis LP, Zanuncio JC (2018) Artificial neural networks: modeling tree survival and mortality in the Atlantic Forest biome in Brazil. Sci Total Environ 645:655–661. https://doi.org/10.1016/j.scitotenv.2018.07.123

de la Cruz Huayanay A, Bazán JL, Russo CM (2025) Performance of evaluation metrics for classification in imbalanced data. Comput Stat 40(3):1447–1473. https://doi.org/10.1007/s00180-024-01539-5

Fick SE, Hijmans RJ (2017) WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int J Climatol 37(12):4302–4315. https://doi.org/10.1002/joc.5086

Fravolini G, Tognetti R, Lombardi F, Egli M, Ascher-Jenull J, Arfaioli P, Bardelli T, Cherubini P, Marchetti M (2018) Quantifying decay progression of deadwood in Mediterranean mountain forests. For Ecol Manage 408:228–237. https://doi.org/10.1016/j.foreco.2017.10.031

Friend AD, Lucht W, Rademacher TT, Keribin R, Betts R, Cadule P, Ciais P, Clark DB, Dankers R, Falloon PD, Ito A, Kahana R, Kleidon A, Lomas MR, Nishina K, Ostberg S, Pavlick R, Peylin P, Schaphoff S, Vuichard N, Warszawski L, Wiltshire A, Woodward FI (2014) Carbon residence time dominates uncertainty in terrestrial vegetation responses to future climate and atmospheric CO_2. Proc Natl Acad Sci U S A 111(9):3280–3285. https://doi.org/10.1073/pnas.1222477110

Gamer M, Lemon J, Puspendra Singh IF (2019) irr: Various Coefficients of interrater reliability and agreement. CRAN. https://doi.org/10.32614/CRAN.package.irr

Harris I, Osborn TJ, Jones P, Lister D (2020) Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. Sci Data 7:109. https://doi.org/10.1038/s41597-020-0453-3

Hartmann H, Bastos A, Das AJ, Esquivel-Muelbert A, Hammond WM, Martínez-Vilalta J, McDowell NG, Powers JS, Pugh TAM, Ruthrof KX, Allen CD (2022) Climate change risks to global forest health: emergence of unexpected events of elevated tree mortality worldwide. Annu Rev Plant Biol 73:673–702. https://doi.org/10.1146/annurev-arplant-102820-012804

Hülsmann L, Bugmann HKM, Commarmot B, Meyer P, Zimmermann S, Brang P (2016) Does one model fit all? Patterns of beech mortality in natural forests of three European regions. Ecol Appl 26(8):2465–2479. https://doi.org/10.1002/eap.1388

Hülsmann L, Bugmann H, Brang P (2017) How to predict tree death from inventory data—lessons from a systematic assessment of European tree mortality models. Can J for Res 47(7):890–900. https://doi.org/10.1139/cjfr-2016-0224

Hülsmann L, Bugmann H, Cailleret M, Brang P (2018) How to kill a tree: empirical mortality models for 18 species and their performance in a dynamic forest model. Ecol Appl 28(2):522–540. https://doi.org/10.1002/eap.1668

Jeni LA, Cohn JF, De La Torre F (2013) Facing imbalanced data—recommendations for the use of performance metrics. In: 2013 humaine association conference on affective computing and intelligent interaction. Geneva, Switzerland. IEEE. https://doi.org/10.1109/acii.2013.47

Kuhn M (2008) Building predictive models in *R* using the caret Package. J Stat Soft. https://doi.org/10.18637/jss.v028.i05

Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta BBA Protein Struct 405(2):442–451. https://doi.org/10.1016/0005-2795(75)90109-9

McNellis BE, Smith AMS, Hudak AT, Strand EK (2021) Tree mortality in western U.S. forests forecasted using forest inventory and random forest classification. Ecosphere 12(3):e03419. https://doi.org/10.1002/ecs2.3419

Merkl D, Hasenauer H (1998) Using neural networks to predict individual tree mortality. In: Proceedings of the int'l conference on engineering applications of neural networks, pp 10–12

Naidu G, Zuva T, Sibanda EM (2023) A review of evaluation metrics in machine learning algorithms. In: Artificial intelligence application in networks and systems, Springer International Publishing, Cham, pp 15–25. https://doi.org/10.1007/978-3-031-35314-7_2

Paletto A, De Meo I, Cantiani P, Ferretti F (2014) Effects of forest management on the amount of deadwood in Mediterranean oak ecosystems. Ann for Sci 71(7):791–800. https://doi.org/10.1007/s13595-014-0377-1

Pretzsch H, Grams T, Häberle KH, Pritsch K, Bauerle T, Rötzer T (2020) Growth and mortality of Norway spruce and European beech in monospecific and mixed-species stands under natural episodic and experimentally extended drought. Results of the KROOF throughfall exclusion experiment. Trees 34(4):957–970. https://doi.org/10.1007/s00468-020-01973-0

Pretzsch H, del Río M, Arcangeli C, Bielak K, Dudzinska M, Ian Forrester D, Kohnle U, Ledermann T, Matthews R, Nagel R, Ningre F, Nord-Larsen T, Szeligowski H, Biber P (2023) Competition-based mortality and tree losses. An essential component of net

primary productivity. For Ecol Manage 544:121204. https://doi.org/10.1016/j.foreco.2023.121204

R Core Team (2021) R: a language and environment for statistical computing. R foundation for statistical computing. Vienna, Austria. https://www.R-project.org/

Reis LP, de Souza AL, dos Reis PCM, Mazzei L, Soares CPB, Miquelino Eleto Torres CM, da Silva LF, Ruschel AR, Rêgo LJS, Leite HG (2018) Estimation of mortality and survival of individual trees after harvesting wood using artificial neural networks in the Amazon rain forest. Ecol Eng 112:140–147. https://doi.org/10.1016/j.ecoleng.2017.12.014

Saito T, Rehmsmeier M (2015) The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE 10(3):e0118432. https://doi.org/10.1371/journal.pone.0118432

Salas-Eljatib C, Weiskittel AR (2020) On studying the patterns of individual-based tree mortality in natural forests: a modelling analysis. For Ecol Manag 475:118369. https://doi.org/10.1016/j.foreco.2020.118369

Senf C, Buras A, Zang CS, Rammig A, Seidl R (2020) Excess forest mortality is consistently linked to drought across Europe. Nat Commun 11:6200. https://doi.org/10.1038/s41467-020-19924-1

Shearman TM, Varner JM, Hood SM, Cansler CA, Hiers JK (2019) Modelling post-fire tree mortality: can random forest improve discrimination of imbalanced data? Ecol Model 414:108855. https://doi.org/10.1016/j.ecolmodel.2019.108855

Shifley SR, He HS, Lischke H, Wang WJ, Jin WC, Gustafson EJ, Thompson JR, Thompson FR, Dijak WD, Yang J (2017) The past and future of modeling forest dynamics: from growth and yield curves to forest landscape models. Landsc Ecol 32(7):1307–1325. https://doi.org/10.1007/s10980-017-0540-9

Sing T, Sander O, Beerenwinkel N, Lengauer T (2005) ROCR: visualizing classifier performance in R. Bioinformatics 21(20):3940–3941. https://doi.org/10.1093/bioinformatics/bti623

Vázquez-Veloso A, Toraño Caicoya A, Bravo F, Biber P, Uhl E, Pretzsch H (2025) Does machine learning outperform logistic regression in predicting individual tree mortality? Ecol Inform 88:103140. https://doi.org/10.1016/j.ecoinf.2025.103140

Venturas MD, Todd HN, Trugman AT, Anderegg WRL (2021) Understanding and predicting forest mortality in the western United States using long-term forest inventory data and modeled hydraulic damage. New Phytol 230(5):1896–1910. https://doi.org/10.1111/nph.17043

Vujovic ŽĐ (2021) Classification model evaluation metrics. Int J Adv Comput Sci Appl. https://doi.org/10.14569/ijacsa.2021.0120670