STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–28.
Published online in International Academic Press (www.IAPress.org)

Cost-Sensitive Ordinal Gini: Formulation, Structural Properties, and Empirical Evaluation

Bahtiar [1,*], Armin Lawi [2], Budi Nurwahyu [3], Nirwan Ilyas [4]

[1] *Doctor of Mathematics Study Program, Hasanuddin University, Makassar, Indonesia*
[2] *Information Systems Study Program, Hasanuddin University, Makassar, Indonesia*
[3] *Department of Mathematics, Hasanuddin University, Makassar, Indonesia*
[4] *Department of Statistics, Hasanuddin University, Makassar, Indonesia*

Abstract Ordinal classification arises whenever target categories possess an inherent order, such as in healthcare, risk assessment, and institutional quality assessment. Existing impurity measures for ordinal decision trees, including Ordinal Gini (OGini), assign the same weight to every boundary between adjacent classes, so the splitting criterion cannot express that errors crossing different boundaries may carry different consequences. This study formulates Cost-Sensitive Ordinal Gini (CS-OGini), a generalisation of OGini in which each ordinal boundary is assigned a configurable cost, and shows that OGini is recovered exactly when all boundary weights are equal. Two families of weighting rules are considered: weights specified as a function of the boundary index, and weights derived from the cumulative class proportions of the training data. For the second family we prove that $\beta = 1$ equalises the contribution of all ordinal boundaries to the impurity, and is the unique exponent doing so whenever those contributions are not already equal, while $\beta = 0$ recovers OGini.
We further establish three structural properties that delimit what boundary weighting can achieve. The accumulated boundary cost is monotone in the ordinal distance between classes; the cost-minimising prediction at a leaf is the median of the class distribution irrespective of the boundary weights, so the weights influence the model only through the tree structure; and any impurity defined as the expected cost between two independent labels is invariant to the antisymmetric component of the cost matrix, so asymmetric ordinal costs cannot be represented within this class of criteria.
The empirical study compares nine splitting criteria, including CART, ID3, Ranking Impurity, class-weighted Gini, an ordinal entropy counterpart of OGini, the Frank and Hall decomposition, OGini, and two CS-OGini variants, on fifteen ordinal datasets under an identical training and evaluation protocol. The criteria are statistically indistinguishable: mean accuracy spans only 1.00 percentage point across all nine, five of six omnibus tests are non-significant, no pairwise comparison against OGini survives correction for multiple testing, and the single significant omnibus result is attributable to the two criteria lying outside the closely grouped set rather than to any criterion outperforming the others. Paired confidence intervals bound the difference between CS-OGini and OGini within $[-0.82, +0.30]$ percentage points of accuracy. The contribution of this study is therefore a formal framework for boundary-weighted ordinal impurity together with provable limits on what such weighting can accomplish, supported by evidence that these limits are binding in practice.

Keywords Ordinal Classification, Decision Trees, Ordinal Gini, Splitting Criterion, Boundary Cost, Class Imbalance

AMS 2010 subject classifications 62H30, 68T10, 68T05

DOI: 10.19139/soic-2310-5070-4287

1. Introduction

Many classification problems in the real world involve data with ordered classes, so they cannot be treated as nominal categories that are independent of each other. This type of problem is known as ordinal classification,

*Correspondence to: Bahtiar (Email: mytiar@gmail.com). Doctor of Mathematics Study Program, Hasanuddin University, Jalan Perintis Kemerdekaan Km. 10, Tamalanrea, Makassar, Indonesia (90245)

ISSN 2310-5070 (online) ISSN 2311-004X (print)
Copyright © 202x International Academic Press

which is a classification with labels that have an inherent level or order. These ordinal characteristics are widely found in various fields such as health, transportation, education, industry, assessment systems, and quality control where each class represents a certain level of condition, quality, or risk [25, 19, 27]. In addition, ordinal classification is also widely applied to various fields of pattern recognition in image classification, such as hyperspectral image classification, radar imagery, and image analysis in medical diagnosis systems, as well as credit risk assessment in the banking sector [26, 30].

In contrast to nominal classification, the consequences of prediction errors in ordinal data will differ from one class to another. The magnitude of the error consequences depends on the distance between the actual class and the predicted result class. The farther the distance between the actual class and the predicted class, the greater the severity or cost [6, 19]. In ordinal classification, the relationship between classes and the consequences of prediction errors are important aspects to consider in the model-building process [3, 30].

The Decision Tree method is one of the most widely used methods in machine learning because of its ability to classify and predict and is easy to interpret [6, 19]. These characteristics make Decision Tree widely applied in various fields, such as medical diagnosis, financial crime detection, and credit risk management [2, 21]. However, conventional Decision Tree algorithms generally still treat classes as nominal categories so that the sequence information contained in ordinal data is ignored [3, 9].

To overcome these limitations, various studies have developed Decision Tree methods that consider the ordinality of the data, such as Ordinal Gini (OGini) and Ranking Impurity (RI). These methods already take into account the order of the classes, resulting in more suitable trees for ordinal data [3, 11]. Other studies have also developed ordinal Decision Tree methods by assigning weights to classes based on certain ordinal characteristics [19, 24]. Among these, Ayllón-Gavilán et al. [3] report OGini as the strongest of the ordinal splitting criteria evaluated in their experimental study, particularly with respect to Mean Absolute Error. It is adopted here as the baseline for that reason. However, the formulation still treats each boundary between classes with the same weight, so that every boundary contributes equally to the impurity value. Meanwhile, in many ordinal classification applications, the consequences of prediction errors are determined by the distance or position between classes, so they do not always have the same severity [19].

These differences in severity have led to the development of various cost-sensitive approaches to ordinal classification. A number of studies have utilized class order information and cost-sensitive learning to improve classification performance, but their applications are generally limited to loss functions, learning strategies, active learning, and the model evaluation stage to minimize the cost of prediction errors [2, 31]. Meanwhile, the application of error costs directly into the splitting criterion of ordinal Decision Trees, especially OGini, is still very limited. This condition opens up the opportunity to develop a split evaluation mechanism that not only considers ordinal impurity but also the severity of class prediction errors.

Based on these problems, this study proposes Cost-Sensitive Ordinal Gini (CS-OGini) as a development of OGini by integrating boundary weights based on the misclassification cost between classes into the split evaluation process. Unlike OGini, which assigns the same contribution to every boundary between classes, CS-OGini accommodates differences in the severity of errors across ordinal class transitions, so that split candidates are evaluated by their impurity values together with the misclassification costs between ordinal classes. A second motivation arises from the class distribution itself. Ordinal datasets are frequently imbalanced, and the dominant response to imbalance operates at the data level, through resampling schemes that alter the training distribution before learning begins; comparative evidence for this family is extensive, with Watthaisong et al. [28] evaluating 66 oversampling variants across 50 datasets and three tree-based classifiers. Such schemes, however, are label-order agnostic and leave the splitting criterion untouched. Under the uniform weighting adopted by OGini the contribution of a boundary to the impurity is governed by how evenly that boundary partitions the data. Boundaries adjacent to minority classes therefore contribute very little, and split selection is dominated by the boundaries near the centre of the ordinal scale. This observation motivates a second family of boundary weights, derived from the cumulative class proportions of the training data, which equalises the contribution of all ordinal boundaries and thereby counteracts the bias towards majority classes without modifying the data itself.

Based on this description, this study aims to develop an OGini-based splitting criterion that integrates boundary-based misclassification costs into the split evaluation process of ordinal Decision Trees. The main contributions of this study are as follows:

1. The formulation of Cost-Sensitive Ordinal Gini (CS-OGini) as a generalization of OGini through the integration of boundary costs into the split evaluation mechanism.
2. The formulation of a distribution-based boundary cost scheme, in which the boundary weights are derived from the cumulative class proportions of the training data, together with a proof that the resulting criterion equalises the contribution of all ordinal boundaries and contains OGini as a special case.
3. Three structural properties that delimit what boundary weighting can achieve: the accumulated cost is monotone in the ordinal distance between classes, the cost-minimising label at a leaf is independent of the boundary weights, and any impurity defined as an expected pairwise cost is invariant to the antisymmetric component of the cost matrix, so that asymmetric ordinal costs cannot be represented within this class of criteria.
4. An evaluation of nine splitting criteria on fifteen ordinal datasets under an identical protocol, using metrics that account for both ordinal structure and misclassification cost, which establishes that these limits are binding in practice.

2. Related Work and Fundamental Concepts

2.1. Decision Tree Impurity Measures

Decision Tree is a classification method widely used because of its interpretability and ease of implementation in various decision-making contexts. The decision tree construction process is based on selecting the best attribute to partition the data at each node, which is determined through impurity measures such as Entropy and the Gini Index [3, 18].

Entropy, used in the ID3 and C4.5 algorithms, measures the degree of uncertainty in the class distribution within a node, whereas the Gini Index, used in CART, measures the degree of class impurity. Both impurity measures are designed to maximize class homogeneity in the resulting child nodes [15, 1].

Although effective for general classification problems, these impurity measures treat class labels as nominal categories and therefore ignore the ordinal relationships among classes. Consequently, prediction errors with different levels of severity are treated equally, making them less suitable for ordinal classification problems [3, 9].

2.2. Ordinal Classification and Ordinal Gini (OGini)

In ordinal classification, class labels possess an inherent ordering structure, so prediction errors are not symmetrical. Cohen [6] argued that misclassification weights need not be symmetric when assessing agreement or classification errors. Subsection 3.6.3 shows that an impurity of the present form cannot express such asymmetry, and Subsection 6.3 discusses where in the model it can be introduced instead. To accommodate this characteristic, several approaches have been developed by incorporating ordinal information into the learning process.

One of the most widely used approaches is Ordinal Gini (OGini), which utilizes cumulative probabilities to preserve class ordering during impurity calculation [3, 25]. Unlike the classical Gini Index, OGini computes impurity based on cumulative class distributions, enabling it to capture ordinal structures while maintaining the basic Decision Tree framework.

OGini implicitly reflects ordinal distance through its cumulative formulation. However, the resulting cost structure is uniform and cannot be configured, since every ordinal boundary is assigned an identical weight[3]. Galimberti et al. [10] mathematically demonstrated that the OGini impurity function of Piccarreta [22] is equivalent to the generalized Gini impurity, where the distances between adjacent category scores are fixed.

2.3. Cost-Sensitive Learning in Decision Trees

Cost-sensitive learning is an approach that aims to minimize the total misclassification cost by recognizing that not all prediction errors have the same consequences [8, 20]. In Decision Trees, this approach is generally implemented through error weighting or the incorporation of cost matrices during the learning process [18].

Several decision tree methods have been developed to integrate error costs into the split selection process. However, most of these approaches are designed for nominal or binary classification problems, where the misclassification cost is defined directly for each pair of classes without considering ordinal relationships [18].

Outside the Decision Tree literature, cost sensitivity has also been embedded directly into the objective function of the learner. Hakimi et al. [12] minimise a weighted sum of classification deviations in which the two class costs are derived from the empirical class proportions and normalised to sum to one, so that the penalty attached to the minority class rises as its representation falls. This demonstrates that misclassification costs need not be supplied exogenously by a domain expert, but can be inferred from the data. The formulation is nonetheless binary and label-order agnostic: the cost structure is indexed by class, not by the boundaries separating ordered classes, and it is imposed on a linear decision boundary rather than on a recursive partitioning criterion.

In ordinal classification, existing cost-sensitive approaches are generally applied during the decision-making or evaluation stages rather than within the impurity formulation itself. Consequently, the information contained in misclassification costs is not directly utilized during tree construction [5, 17].

2.4. Boundary-Based Cost in Ordinal Classification

In ordinal classification, relationships among classes can naturally be represented through the boundaries separating adjacent ordinal levels [13, 3]. A boundary-based approach provides a natural representation of misclassification costs, where each prediction error can be interpreted as the accumulation of crossed ordinal boundaries.

Unlike conventional cost-matrix approaches, which define costs directly between class pairs, the boundary-based approach exploits the internal structure of the ordinal scale. Each boundary is assigned a specific weight, while the total misclassification cost is obtained by summing the weights of all crossed boundaries between the actual and predicted classes [16, 10].

This representation offers greater flexibility because boundary weights can be generated by different rules. One possibility is to specify the weights as a function of the boundary index, so that the penalty grows along the ordinal scale according to a chosen pattern [16, 10]. Another possibility, developed in this study, is to derive the weights from the class distribution of the data. The latter is motivated by the observation that ordinal datasets are frequently imbalanced, in which case boundaries adjacent to minority classes contribute disproportionately little to the impurity. Consequently, the boundary weights may be specified a priori from the ordinal scale or inferred from the data. The extent to which such weights can alter the behaviour of the resulting model is examined in Subsection 3.6.

2.5. Research Gap and Positioning

Based on the existing literature, previous studies on ordinal Decision Trees have proposed several impurity functions that consider class ordering, such as OGini. However, the split selection process in these methods is still primarily oriented toward maximizing impurity reduction without explicitly distinguishing the severity of prediction errors among ordinal classes.

Conversely, studies on cost-sensitive learning have demonstrated that incorporating misclassification costs can improve classification quality. Nevertheless, these approaches generally apply costs during data weighting, learning strategies, or model evaluation, leaving the impurity formulation itself untouched. Consequently, there remains a research gap regarding the direct incorporation of ordinal-distance-based misclassification costs into impurity functions for guiding split selection in ordinal Decision Trees.

To the best of our knowledge, no existing Decision Tree approach simultaneously integrates boundary-based ordinal structures and configurable misclassification costs into a single impurity formulation used directly during split selection.

Table 1. Comparison of impurity measures and cost-sensitive decision tree approaches for ordinal classification

Method	Ordinal Structure	Cost Integration Level	Boundary Cost Scheme	Representative Reference
ID3 / Information Gain	×	–	×	Quinlan (1986)
CART (Gini)	×	–	×	Breiman et al. (1984)
CS-MSD (Mathematical Programming)	×	Objective Function	×	Hakimi et al. (2026)
Weighted Information Gain (WIG)	✓	–	×	Ayllón-Gavilán et al. (2025)
Ranking Impurity (RI)	✓	–	×	Xia et al. (2006)
Cost-Sensitive Ordinal DT	✓	Decision Rule	×	Marudi et al. (2024)
Ordinal Gini (OGini)	✓	Uniform	Uniform	Ayllón-Gavilán et al. (2025)
CS-OGini (This Study)	✓	**Split Criterion**	**Configurable / Distribution-based**	**This Study**

To address this research gap, this study proposes Cost-Sensitive Ordinal Gini (CS-OGini), an extension of OGini that introduces boundary costs as weights assigned to each ordinal boundary. In this formulation, the misclassification distance is no longer defined directly between class pairs but is represented as the accumulated boundary cost crossed between the actual and predicted classes. The impurity is subsequently formulated as the expectation of boundary-cost distances within a node, allowing it to simultaneously represent ordinal class structures and the severity of prediction errors.

A comparative summary of impurity-based and cost-sensitive Decision Tree methods discussed in this section, together with their characteristics, strengths, and limitations, is presented in Table 1.

As summarized in Table 1, existing approaches have not simultaneously integrated boundary-based ordinal structures and cost-sensitive mechanisms within a single impurity formulation. This limitation motivates the development of the proposed Cost-Sensitive Ordinal Gini (CS-OGini) splitting criterion presented in this study.

To clarify the conceptual development of the impurity measures discussed in this section, Figure 1 illustrates the evolution from the classical Gini Index to Ordinal Gini (OGini), and subsequently to the proposed Cost-Sensitive Ordinal Gini (CS-OGini).

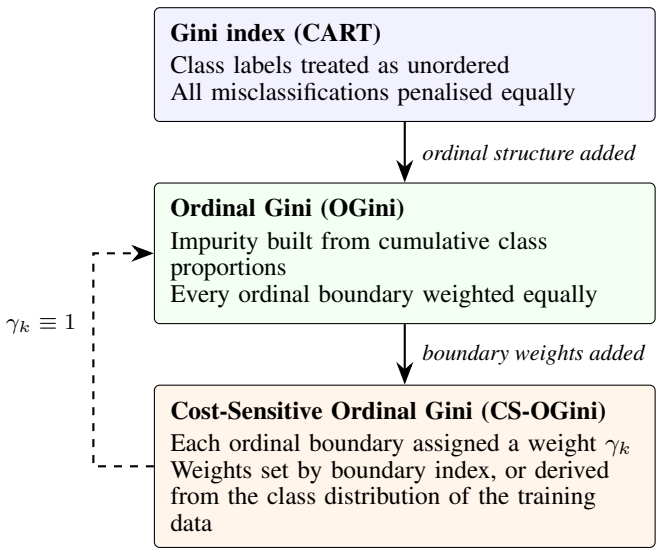

Figure 1. Relationship between the Gini index, Ordinal Gini and the proposed Cost-Sensitive Ordinal Gini. Each criterion adds structure to the one above it; CS-OGini contains OGini as the special case of uniform boundary weights.

3. CS-OGini: Cost-Sensitive Ordinal Gini

This section presents a formal derivation of Cost-Sensitive Ordinal Gini (CS-OGini) splitting criterion. The derivation is carried out by establishing a relationship between the ordinal structure of the class and boundary-based misclassification cost expectations.

3.1. Boundary-Based Ordinal Cost

Let $\mathcal{A} = \{\alpha_1, \alpha_2, \ldots, \alpha_K\}$ be a totally ordered set of ordinal class labels satisfying $\alpha_1 \prec \alpha_2 \prec \cdots \prec \alpha_K$. For the i-th observation, let $(y_i, \hat{y}_i) \in \mathcal{A}^2$ denote its actual and predicted labels, respectively.

Definition 1 (Boundary Cost)
For each pair of consecutive ordinal classes α_k and α_{k+1}, the boundary separating them is assigned a weight

$$\gamma_k \geq 0, \qquad k = 1, \ldots, K-1, \tag{1}$$

which represents the cost of misclassification across that boundary. The accumulated cost between two classes α_i and α_j is obtained by summing the weights of all boundaries lying between them,

$$C(i,j) = \sum_{b=\min(i,j)}^{\max(i,j)-1} \gamma_b, \qquad C(i,i) = 0. \tag{2}$$

In this study, the boundary weight is generated by a weighting rule. Two families of such rules are considered. The first family specifies the weight as a function of the boundary index,

$$\gamma_k = f(k), \qquad k = 1, 2, \ldots, K-1, \tag{3}$$

where the function f determines how the weight changes across boundaries. This function may take various forms with different growth patterns, including the constant form adopted by OGini as well as linear, polynomial, and exponential forms. The functional forms differ in how rapidly the weight grows along the ordinal scale; Subsection 4.3 shows that they yield fewer distinct criteria than their number suggests.

The second family specifies the weight as a function of the class distribution of the training data,

$$\gamma_k = g(P_{m_0,k}), \qquad k = 1, 2, \ldots, K-1, \tag{4}$$

where $P_{m_0,k}$ denotes the cumulative class proportion up to the k-th boundary at the root node. In this case the weight assigned to a boundary is determined by its position relative to the class distribution rather than by its position on the ordinal scale. This family is developed in Subsection 3.4.

Both families produce non-negative weights and satisfy Definition 1. The uniform weighting $\gamma_k = 1$ belongs to both families and recovers the original OGini impurity, so that the two families intersect at the cost-insensitive baseline. Together they provide flexibility in modelling the underlying risk structure, whether that structure is specified a priori through the ordinal scale or inferred from the data.

Definition 2 (Pairwise Boundary Cost)
Let Y and Y' be two independent random variables following the same class probability distribution. The boundary cost between the pair of labels Y and Y' is defined as

$$C(Y, Y') = \sum_{k=1}^{K-1} \gamma_k \, \mathbf{1}_{A_k}, \tag{5}$$

where A_k denotes the event that Y and Y' lie on opposite sides of the k-th boundary, and

$$\mathbf{1}_{A_k} = \begin{cases} 1, & \text{if } A_k \text{ occurs,} \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Thus $C(Y, Y')$ is a random variable representing the total boundary cost separating the random pair of labels Y and Y'. This is consistent with Eq. (2): for a realisation $Y = \alpha_i$ and $Y' = \alpha_j$, the indicator is active precisely for the boundaries lying between α_i and α_j, so that $C(Y, Y') = C(i, j)$.

3.2. Formulation of CS-OGini as an Expected Cost

Let the probability distribution of the i-th class at node (m) in a decision tree be given by

$$p_{m,i} = P(Y = \alpha_i), \qquad i = 1, \ldots, K, \tag{7}$$

and let the cumulative probability up to the k-th boundary be defined as

$$P_{m,k} = \sum_{i=1}^{k} p_{m,i}. \tag{8}$$

Definition 3 (Cost-Sensitive Impurity)
Let Y and Y' be two independent random variables having the same class probability distribution, $\{p_{m,1}, \ldots, p_{m,K}\}$. The cost-sensitive impurity at node m is defined as

$$I_{\mathrm{CS}}(m) = \frac{1}{2}\,\mathbb{E}[C(Y, Y')]\,. \tag{9}$$

The value of $I_{\mathrm{CS}}(m)$ can subsequently be evaluated using Theorem 4.

Theorem 4 (CS-OGini Formulation)
Let γ_k be the boundary weight at the k-th boundary and let $P_{m,k}$ be the cumulative probability up to the k-th boundary at node m. Then the cost-sensitive impurity in Eq. (9) can be expressed as

$$I_{\mathrm{CS}}(m) = \sum_{k=1}^{K-1} \gamma_k\, P_{m,k}\,(1 - P_{m,k})\,. \tag{10}$$

Proof
By Definition 2, the boundary cost between the pair of labels Y and Y' is

$$C(Y, Y') = \sum_{k=1}^{K-1} \gamma_k\, \mathbf{1}_{A_k}, \tag{11}$$

where A_k denotes the event that Y and Y' lie on opposite sides of the k-th boundary.

Applying the linearity of expectation to Eq. (11) gives

$$\mathbb{E}[C(Y, Y')] = \mathbb{E}\left[\sum_{k=1}^{K-1} \gamma_k\, \mathbf{1}_{A_k}\right] \tag{12}$$

$$= \sum_{k=1}^{K-1} \gamma_k\, \mathbb{E}[\mathbf{1}_{A_k}] \tag{13}$$

$$= \sum_{k=1}^{K-1} \gamma_k\, \mathbb{P}(A_k), \tag{14}$$

where Eq. (14) follows from the fact that the expectation of an indicator function equals the probability of the corresponding event.

The event A_k occurs when exactly one of the two labels lies at or below the k-th boundary. These two cases are disjoint, so

$$\mathbb{P}(A_k) = \mathbb{P}(Y \preceq \alpha_k,\ Y' \succ \alpha_k) + \mathbb{P}(Y \succ \alpha_k,\ Y' \preceq \alpha_k)\,. \tag{15}$$

Since Y and Y' are independent and identically distributed, each joint probability factorises, and by Eq. (8) we have $\mathbb{P}(Y \preceq \alpha_k) = P_{m,k}$ and $\mathbb{P}(Y \succ \alpha_k) = 1 - P_{m,k}$. Therefore

$$\mathbb{P}(A_k) = P_{m,k}\,(1 - P_{m,k}) + (1 - P_{m,k})\,P_{m,k} \tag{16}$$

$$= 2P_{m,k}\,(1 - P_{m,k})\,. \tag{17}$$

Substituting Eq. (17) into Eq. (14) yields

$$\mathbb{E}[C(Y, Y')] = 2\sum_{k=1}^{K-1} \gamma_k\, P_{m,k}\,(1 - P_{m,k})\,. \tag{18}$$

Finally, by Definition 3 the cost-sensitive impurity is $I_{\mathrm{CS}}(m) = \frac{1}{2}\,\mathbb{E}[C(Y, Y')]$, so that

$$I_{\mathrm{CS}}(m) = \frac{1}{2} \cdot 2\sum_{k=1}^{K-1} \gamma_k\, P_{m,k}\,(1 - P_{m,k}) = \sum_{k=1}^{K-1} \gamma_k\, P_{m,k}\,(1 - P_{m,k})\,, \tag{19}$$

which completes the proof. □

3.3. Relationship to Ordinal Gini

If all boundary weights are assigned uniformly, that is,

$$\gamma_k = 1, \qquad k = 1, \ldots, K-1, \tag{20}$$

then the proposed CS-OGini impurity reduces to

$$I_{CS}(m) = \sum_{k=1}^{K-1} P_{m,k}\,(1 - P_{m,k})\,, \tag{21}$$

which is identical to the original Ordinal Gini (OGini) impurity.

OGini can be regarded as a special case of the proposed CS-OGini framework, in which all ordinal boundaries contribute equally to the impurity function. In contrast, CS-OGini generalizes OGini by allowing configurable boundary weights that explicitly model the severity of ordinal misclassification.

3.4. Distribution-Based Boundary Cost

In Subsection 3.1, the first family of weighting rules expresses the boundary weight as a function of the boundary index, $\gamma_k = f(k)$, so that the weight is determined solely by the position of the boundary on the ordinal scale. This subsection develops an alternative scheme in which the boundary weight is derived from the class distribution of the training data. Deriving costs from the class distribution rather than from expert judgement has precedent in the cost-sensitive learning literature, where class costs have been set inversely proportional to class frequency and normalised across classes [12]. The scheme developed below adopts the same principle but relocates it from the class to the boundary: the weight is determined by the mass separated by each ordinal cut point, so that the adjustment is indexed by $k = 1, \ldots, K-1$, not by the class label. This distinction is what makes the scheme meaningful for ordinal data, since two datasets with identical imbalance ratios may distribute that imbalance across the ordinal scale in entirely different ways. The motivation is that ordinal classification problems are frequently imbalanced, and that this imbalance is distributed unevenly across the ordinal scale: the ratio ρ_v between the largest and smallest boundary contribution reported in Table 3 ranges from 1.00 to 37.53 across the datasets considered. The analysis below demonstrates that under a uniform weighting the boundaries adjacent to minority classes contribute very little to the impurity value.

3.4.1. Contribution of Each Boundary to the Impurity Let m_0 denote the root node, that is, the node containing the entire training set. Following Eq. (7) and Eq. (8), let $p_{m_0,i}$ be the proportion of class α_i in the training data and let

$$P_{m_0,k} = \sum_{i=1}^{k} p_{m_0,i}, \qquad k = 1, \ldots, K-1, \tag{22}$$

be the cumulative proportion up to the k-th boundary at the root node.

From Theorem 4, the impurity at the root node is

$$I_{\mathrm{CS}}(m_0) = \sum_{k=1}^{K-1} \gamma_k \, P_{m_0,k} \, (1 - P_{m_0,k}) \, . \tag{23}$$

Each term of Eq. (23) represents the contribution of a single ordinal boundary to the impurity. To simplify the notation, let

$$v_k = P_{m_0,k} \, (1 - P_{m_0,k}) \, , \qquad k = 1, \ldots, K-1, \tag{24}$$

so that Eq. (23) becomes $I_{\mathrm{CS}}(m_0) = \sum_{k=1}^{K-1} \gamma_k v_k$.

The quantity v_k attains its maximum value of $1/4$ when $P_{m_0,k} = 1/2$, and approaches zero when $P_{m_0,k}$ approaches either 0 or 1. A boundary that separates a small minority class from the remaining classes produces a highly unbalanced partition of the training data, so that v_k becomes very small and the contribution of that boundary to the impurity becomes negligible.

Table 2 illustrates this effect for the ERA dataset, one of the benchmark problems used in the experiments (Table 3), with K = 9 and N = 750 training observations. The boundary at the upper end of the ordinal scale contributes only 1.63% of the total impurity, whereas the central boundaries contribute 21.64% and 21.62%. The largest contribution exceeds the smallest by a factor of 13.25, computed from the unrounded contributions.

Table 2. Contribution of each ordinal boundary to the OGini impurity for the ERA dataset, whose training class distribution is $(69, 106, 136, 129, 118, 89, 66, 23, 14)$ with $K = 9$ and $N = 750$.

Boundary k	$P_{m_0,k}$	v_k	Contribution (%)
1	0.0920	0.0835	7.45
2	0.2333	0.1789	15.95
3	0.4147	0.2427	21.64
4	0.5867	0.2425	21.62
5	0.7440	0.1905	16.98
6	0.8627	0.1185	10.56
7	0.9507	0.0469	4.18
8	0.9813	0.0183	1.63

3.4.2. Formulation of the Distribution-Based Boundary Cost The imbalance described above can be corrected by requiring that every ordinal boundary contribute equally to the impurity at the root node. This is expressed as

$$\gamma_k \, v_k = c, \qquad k = 1, \ldots, K-1, \tag{25}$$

where c is a constant that is identical for all boundaries. Solving Eq. (25) for γ_k gives $\gamma_k = c/v_k$. As shown later in this subsection, the value of c does not affect the resulting tree, so that c may be set to unity. Relaxing the exponent then yields the following definition.

Definition 5 (Distribution-Based Boundary Cost)
The distribution-based boundary cost is defined as

$$\gamma_k^{(\beta)} = [\, P_{m_0,k} \, (1 - P_{m_0,k}) \,]^{-\beta} \, , \qquad k = 1, \ldots, K-1, \tag{26}$$

where $\beta \geq 0$ is a parameter controlling the strength of the adjustment, and $P_{m_0,k}$ is computed from the training data only.

In contrast to Eq. (3), in which the boundary weight is a function of the boundary index, Eq. (26) defines the boundary weight as a function of the cumulative class proportion. The weight of a boundary is therefore determined by the position of that boundary relative to the class distribution rather than by its position on the ordinal scale alone.

The two limiting cases of Eq. (26) are established in Theorem 6 and Corollary 8.

Theorem 6 (Equality of Boundary Contributions)
Let $\gamma_k^{(\beta)}$ be given by Eq. (26) and let $v_k = P_{m_0,k}(1 - P_{m_0,k})$ with $0 < P_{m_0,k} < 1$ for all $k = 1, \ldots, K-1$. Then:

(i) For $\beta = 1$ every ordinal boundary contributes equally to the impurity at the root node, and

$$I_{CS}(m_0) = K - 1. \tag{27}$$

(ii) If the contributions v_k are not all equal, that is if $\rho_v = \max_k v_k / \min_k v_k > 1$, then $\beta = 1$ is the *unique* exponent in Eq. (26) for which all boundaries contribute equally.

Proof
(i) Substituting $\beta = 1$ into Eq. (26) and applying Eq. (24) gives

$$\gamma_k^{(1)} = \frac{1}{P_{m_0,k}(1 - P_{m_0,k})} = \frac{1}{v_k}. \tag{28}$$

Since $0 < P_{m_0,k} < 1$, it follows that $v_k > 0$ and the expression is well defined. The contribution of the k-th boundary to the impurity in Eq. (23) is therefore

$$\gamma_k^{(1)} v_k = \frac{v_k}{v_k} = 1, \tag{29}$$

which is independent of k. Summing over all boundaries yields

$$I_{CS}(m_0) = \sum_{k=1}^{K-1} \gamma_k^{(1)} v_k = \sum_{k=1}^{K-1} 1 = K - 1. \tag{30}$$

(ii) For a general exponent $\beta \geq 0$, the contribution of the k-th boundary is

$$\gamma_k^{(\beta)} v_k = v_k^{-\beta} v_k = v_k^{1-\beta}. \tag{31}$$

Suppose all contributions are equal. By hypothesis there exist indices j and l with $v_j \neq v_l$, and both are strictly positive. Equality of their contributions requires ${v_j}^{1-\beta} = {v_l}^{1-\beta}$, hence

$$(1 - \beta)\big(\log v_j - \log v_l\big) = 0. \tag{32}$$

Since $v_j \neq v_l$ implies $\log v_j \neq \log v_l$, the second factor is non-zero, so $1 - \beta = 0$, that is $\beta = 1$. Conversely $\beta = 1$ does equalise the contributions by part (i). Hence $\beta = 1$ is the unique such exponent. □

Remark 7
The hypothesis $\rho_v > 1$ in part (ii) is necessary and not merely technical. If all boundaries already contribute equally at the root, so that $\rho_v = 1$, then Eq. (31) is constant in k for *every* β, and no exponent is distinguished. This case occurs in practice: the `balance-scale` dataset in Table 3 has $\rho_v = 1.00$, and for that dataset the distribution-based family degenerates to a rescaling of OGini for all values of β. Uniqueness therefore holds precisely on the datasets for which the adjustment is not already vacuous.

Corollary 8 (OGini as a Special Case)
Let $\beta = 0$ in Eq. (26). Then $\gamma_k^{(0)} = 1$ for all k, and I_{CS} reduces to the OGini impurity.

Proof
For $\beta = 0$ and $v_k > 0$, Eq. (26) gives $\gamma_k^{(0)} = [v_k]^0 = 1$ for every k. Substituting into Theorem 4 yields

$$I_{\mathrm{CS}}(m) = \sum_{k=1}^{K-1} P_{m,k}\,(1 - P_{m,k})\,, \tag{33}$$

which is the OGini impurity given in Eq. (21). □

Corollary 8 shows that the uniform weighting of OGini and the fully adjusted weighting of Theorem 6 are the two limiting members of a single family. Intermediate values $0 < \beta < 1$ produce a partial adjustment, in which the boundary weights lie between these two extremes. The distribution-based scheme preserves the property established in Subsection 3.3, namely that OGini is a special case of CS-OGini.

3.4.3. Properties and Implementation Eq. (26) has several properties that bear on its implementation.

Invariance to scaling. Multiplying all boundary weights by a positive constant c multiplies the impurity of every node by the same constant, since Eq. (23) is linear in γ. Both terms of Eq. (39) are multiplied by c, and the split maximising the gain in Eq. (40) remains unchanged. Consequently, the weights obtained from Eq. (26) may be normalised so that their mean equals unity without altering the resulting tree. This normalisation is applied in the experiments so that the weight values are comparable across datasets.

Estimation of the class proportions. A minority class may be absent from a particular training sample, which yields $P_{m_0,k} \in \{0, 1\}$ and leaves Eq. (26) undefined. The class proportions are estimated using additive smoothing,

$$\hat{p}_{m_0,i} = \frac{n_i + \varepsilon}{n + K\varepsilon}, \qquad \varepsilon = 0.5, \tag{34}$$

where n_i denotes the number of training observations in class α_i and n the total number of training observations.

Choice of the exponent β. Theorem 6 identifies $\beta = 1$ as the unique exponent equalising the contribution of all ordinal boundaries, and Corollary 8 identifies $\beta = 0$ as the exponent recovering OGini. The experiments report the first of these, so that the distribution-based variant evaluated in Section 5 is the fully adjusted member of the family rather than one selected from it. The exponent is fixed rather than tuned. The value $\beta = 1$ is the one the theory distinguishes, so reporting it tests the scheme the derivation actually motivates. Fixing it also leaves every criterion in the comparison with exactly one tuned hyperparameter, the tree depth, so the proposed criterion is not granted a degree of freedom that the baselines lack.

In finite samples $P_{m_0,k}$ is estimated rather than known, and the estimate is least accurate at the extreme boundaries, which are precisely the boundaries receiving the largest weights under $\beta = 1$. The comparison in Subsection 5.3 should be read as an evaluation of the exact adjustment under estimated proportions, which is the form in which a practitioner would apply it.

Boundary weights are computed once. The proportions $P_{m_0,k}$ are computed once from the entire training set and held fixed throughout tree construction. Recomputing the weights at every node would make the boundary cost depend on the position of the node within the tree, which is inconsistent with the interpretation of γ_k as a misclassification cost established in Definition 1. In addition, the property $\mathrm{Gain}(\theta) \geq 0$ is no longer guaranteed when the weights differ between a parent node and its child nodes.

3.5. CS-OGini Split Criterion

Let D_m denote the dataset associated with node m, and let θ denote a candidate split at that node. Applying split θ partitions D_m into two disjoint subsets,

$$D_{m,L}(\theta) \quad \text{and} \quad D_{m,R}(\theta), \tag{35}$$

such that

$$D_{m,L}(\theta) \cup D_{m,R}(\theta) = D_m, \qquad D_{m,L}(\theta) \cap D_{m,R}(\theta) = \varnothing. \tag{36}$$

The impurity after applying split θ is defined as

$$I_{\text{split}}(\theta) = w_L(\theta) I_{CS}(D_{m,L}(\theta)) + w_R(\theta) I_{CS}(D_{m,R}(\theta)), \tag{37}$$

where

$$w_L(\theta) = \frac{|D_{m,L}(\theta)|}{|D_m|}, \qquad w_R(\theta) = \frac{|D_{m,R}(\theta)|}{|D_m|}. \tag{38}$$

The gain of split θ is defined as

$$\text{Gain}(\theta) = I_{CS}(D_m) - I_{\text{split}}(\theta). \tag{39}$$

The optimal split is selected by maximizing the gain value,

$$\theta^* = \arg\max_{\theta} \text{Gain}(\theta). \tag{40}$$

The recursive tree construction procedure is identical to that of the standard Decision Tree algorithm. The splitting process continues until a stopping criterion is satisfied: the maximum tree depth, the minimum number of observations at a node, or the absence of a candidate split with strictly positive gain.

3.6. Structural Properties of the Criterion

This subsection proves three properties that constrain how the boundary weights γ influence the fitted model. The accumulated boundary cost is monotone in the ordinal distance between classes, so that γ induces a genuine ordinal cost structure. The cost-minimising prediction at a leaf is the median of the class distribution and is therefore independent of γ. And any impurity defined as the expected cost between two independent labels is invariant to the antisymmetric component of the cost matrix, so that asymmetric ordinal costs cannot be represented within this class of criteria. The three properties bound the channel through which boundary weighting acts, and Subsection 5.8.1 uses them to interpret the empirical results.

3.6.1. Monotonicity of the accumulated boundary cost

Proposition 9 (Boundary Cost Monotonicity)
Let C be the accumulated boundary cost of Definition 1, with $\gamma_k \geq 0$ for $k = 1, \ldots, K-1$. Then for all $i, j, l \in \{1, \ldots, K\}$:

(i) $C(i,j) = C(j,i) \geq 0$ and $C(i,i) = 0$;
(ii) if $i \leq j \leq l$, then $C(i,l) = C(i,j) + C(j,l)$;
(iii) if $i \leq j \leq l$, then $C(i,j) \leq C(i,l)$ and $C(j,l) \leq C(i,l)$;
(iv) $C(i,l) \leq C(i,j) + C(j,l)$ for arbitrary j, so that C is a pseudometric on $\mathcal{A}$, and a metric if and only if $\gamma_k > 0$ for every k;
(v) if $\gamma_k > 0$ for every k, the inequalities in (iii) are strict whenever $j \neq l$ and $i \neq j$ respectively;
(vi) if $\gamma_k = 1$ for every k, then $C(i,j) = |i - j|$.

Proof
Write $B(i,j) = \{\min(i,j), \ldots, \max(i,j) - 1\}$ for the index set of the boundaries lying between α_i and α_j, so that Eq. (2) reads $C(i,j) = \sum_{b \in B(i,j)} \gamma_b$, with the convention that an empty sum is zero.

(i) The set $B(i,j)$ is symmetric in its arguments and $B(i,i)=\emptyset$, and each $\gamma_b \geq 0$, which gives the three assertions.

(ii) For $i \leq j \leq l$ the index sets $B(i,j)=\{i,\dots,j-1\}$ and $B(j,l)=\{j,\dots,l-1\}$ are disjoint and their union is $B(i,l)=\{i,\dots,l-1\}$. Summing γ_b over this partition gives $C(i,l)=C(i,j)+C(j,l)$.

(iii) Immediate from (ii), since both summands are non-negative.

(iv) Assume without loss of generality $i \leq l$, so $B(i,l)=\{i,\dots,l-1\}$. Three cases arise. If $i \leq j \leq l$, then $B(i,j)\cup B(j,l)=B(i,l)$ by (ii). If $j<i$, then $B(j,l)=\{j,\dots,l-1\} \supseteq B(i,l)$. If $j>l$, then $B(i,j)=\{i,\dots,j-1\} \supseteq B(i,l)$. In every case $B(i,j)\cup B(j,l) \supseteq B(i,l)$, and since $\gamma_b \geq 0$,

$$C(i,j)+C(j,l) \;=\; \sum_{b\in B(i,j)} \gamma_b + \sum_{b\in B(j,l)} \gamma_b \;\geq\; \sum_{b\in B(i,l)} \gamma_b \;=\; C(i,l). \tag{41}$$

Together with (i) this makes C a pseudometric. If $\gamma_k=0$ for some k, then $C(k,k+1)=0$ although $\alpha_k \neq \alpha_{k+1}$, so C is not a metric; conversely if every $\gamma_k>0$ then $B(i,j)\neq\emptyset$ for $i\neq j$ gives $C(i,j)>0$.

(v) Under $\gamma_k>0$, the summands removed in passing from $C(i,l)$ to $C(i,j)$ in (ii) are strictly positive whenever $j\neq l$, and symmetrically for the second inequality.

(vi) With $\gamma_k=1$ the sum counts the elements of $B(i,j)$, of which there are $|i-j|$. □

Proposition 9 justifies the interpretation of γ as a cost structure rather than as an arbitrary reweighting. Part (iii) states that moving the predicted class further from the actual class along the ordinal scale never decreases the cost incurred, which is the defining requirement of an ordinal misclassification cost and is what distinguishes the present formulation from a nominal cost matrix. Part (vi) identifies the uniform weighting as the case in which the accumulated cost reduces to the absolute ordinal distance, so that the Mean Absolute Error is the average accumulated cost under OGini weighting. The monotonicity holds for every admissible γ, and is a property of the framework, not of a particular weighting rule.

3.6.2. Leaf prediction is invariant to the boundary weights The tree returns a single label at each leaf. The natural choice under a cost structure is the label minimising the expected cost incurred at that leaf.

Definition 10 (Cost-Minimising Leaf Prediction)
Let m be a leaf with class distribution $\{p_{m,1},\dots,p_{m,K}\}$ and let C be as in Definition 1. The expected cost of predicting the label α_a at m is

$$R_m(a) \;=\; \sum_{j=1}^{K} p_{m,j}\, C(j,a), \qquad a=1,\dots,K, \tag{42}$$

and the cost-minimising prediction at m is any $a^*(m)\in \arg\min_a R_m(a)$.

Proposition 11 (Invariance of the Leaf Prediction)
Let m be a leaf with cumulative probabilities $P_{m,k}=\sum_{i\leq k} p_{m,i}$ and let $\gamma_k \geq 0$. Then

$$R_m(a+1)-R_m(a) \;=\; \gamma_a\big(2P_{m,a}-1\big), \qquad a=1,\dots,K-1. \tag{43}$$

Consequently:

(i) R_m is quasi-convex in a, and the label

$$a^*_{\mathrm{med}}(m) \;=\; \min\big\{k : P_{m,k} \geq \tfrac{1}{2}\big\} \tag{44}$$

is a cost-minimising prediction for every admissible γ;

(ii) if $\gamma_k>0$ for every k, the set of cost-minimising predictions coincides exactly with the set of medians of the class distribution at m, and $a^*_{\mathrm{med}}(m)$ is the unique minimiser whenever $P_{m,k}\neq\frac{1}{2}$ for all k.

In particular the cost-minimising prediction does not depend on the boundary weights.

Proof
Fix $a \in \{1, \dots, K-1\}$. By Proposition 9(ii), for $j \le a$ the boundary a is added when the prediction moves from α_a to α_{a+1}, so $C(j, a+1) - C(j, a) = \gamma_a$; for $j \ge a+1$ the same boundary is removed, so $C(j, a+1) - C(j, a) = -\gamma_a$. Substituting into Eq. (42),

$$R_m(a+1) - R_m(a) = \sum_{j \le a} p_{m,j}\, \gamma_a \; - \sum_{j \ge a+1} p_{m,j}\, \gamma_a \tag{45}$$

$$= \gamma_a \Big[P_{m,a} - \big(1 - P_{m,a}\big) \Big] \; = \; \gamma_a \big(2P_{m,a} - 1\big), \tag{46}$$

which is Eq. (43).

(i) The sequence $P_{m,1} \le P_{m,2} \le \cdots \le P_{m,K} = 1$ is non-decreasing, so $2P_{m,a} - 1$ changes sign at most once, from negative to non-negative, as a increases. Since $\gamma_a \ge 0$, the increments in Eq. (43) are therefore non-positive for $a < a^*_{\mathrm{med}}(m)$ and non-negative for $a \ge a^*_{\mathrm{med}}(m)$. Hence R_m is non-increasing up to $a^*_{\mathrm{med}}(m)$ and non-decreasing thereafter, which is quasi-convexity, and $a^*_{\mathrm{med}}(m)$ attains the minimum. The index is well defined because $P_{m,K} = 1 \ge \frac{1}{2}$.

(ii) If $\gamma_a > 0$, the sign of the increment in Eq. (43) equals the sign of $2P_{m,a} - 1$, which is a quantity depending only on the class distribution. The increment vanishes if and only if $P_{m,a} = \frac{1}{2}$, which is precisely the condition under which the median of the distribution is not unique. The minimisers of R_m therefore coincide with the medians, and the minimiser is unique when no $P_{m,k}$ equals $\frac{1}{2}$.

In both cases the minimiser is determined by $\{P_{m,k}\}$ alone, so it does not depend on γ. □

Corollary 12 (Weights Act Only Through the Tree Structure)
Let two admissible weight vectors γ and γ', both strictly positive, induce the same sequence of splits and hence the same partition of the feature space. Then the two fitted trees produce identical predictions on every input. Consequently the boundary weights can influence the fitted model only by altering which splits are selected.

Proof
Identical partitions give identical leaves and identical leaf class distributions. By Proposition 11(ii) the label assigned at each leaf is a median of that distribution and does not depend on the weight vector, so the two trees agree on every leaf and therefore on every input. □

By Corollary 12 the boundary weights act on the fitted model only through split selection. The prediction rule at the leaves is closed to cost sensitivity; the pruning objective is examined alongside the experimental results.

3.6.3. Invariance to the antisymmetric component of the cost Definition 3 defines the impurity as one half of the expected cost between two independent labels drawn from the class distribution at the node. The following result shows that this construction alone prevents asymmetric costs from being represented, independently of the particular form of C.

Proposition 13 (Antisymmetric Cost Invariance)
Let $W = (w_{ij}) \in \mathbb{R}^{K \times K}$ be an arbitrary cost matrix, let Y and Y' be independent and identically distributed with class probabilities $\{p_{m,1}, \dots, p_{m,K}\}$, and define

$$I_W(m) \; = \; \tfrac{1}{2}\, \mathbb{E}\big[w_{Y,Y'}\big] \; = \; \tfrac{1}{2} \sum_{i=1}^{K} \sum_{j=1}^{K} p_{m,i}\, p_{m,j}\, w_{ij}. \tag{47}$$

Write $W = S + A$ with $S = \frac{1}{2}(W + W^\top)$ and $A = \frac{1}{2}(W - W^\top)$. Then:

(i) $I_W(m) = I_S(m)$ for every class distribution, so I_W is invariant to A;

(ii) if W is generated by directional boundary costs, namely $u_b \geq 0$ for crossing boundary b upwards and $d_b \geq 0$ for crossing it downwards, so that

$$w_{ij} = \sum_{b=i}^{j-1} u_b \ \ (i < j), \qquad w_{ij} = \sum_{b=j}^{i-1} d_b \ \ (i > j), \qquad w_{ii} = 0, \tag{48}$$

then I_W coincides exactly with the CS-OGini impurity of Theorem 4 under the symmetric boundary weights

$$\gamma_b \ = \ \tfrac{1}{2}\,(u_b + d_b), \qquad b = 1, \ldots, K-1. \tag{49}$$

Proof
(i) By construction $a_{ij} = -a_{ji}$ for all i, j, and in particular $a_{ii} = 0$. The product $p_{m,i}p_{m,j}$ is symmetric under interchange of i and j, so pairing the terms (i, j) and (j, i) in the double sum gives

$$\sum_{i=1}^{K}\sum_{j=1}^{K} p_{m,i}\, p_{m,j}\, a_{ij} = \sum_{i<j} p_{m,i}\, p_{m,j}\, \big(a_{ij} + a_{ji}\big) = 0. \tag{50}$$

Substituting $W = S + A$ into Eq. (47) and applying linearity therefore yields $I_W(m) = I_S(m)$.
(ii) Pairing the terms (i, j) and (j, i) in Eq. (47) and using Eq. (48),

$$I_W(m) = \tfrac{1}{2}\sum_{i<j} p_{m,i}\, p_{m,j}\, \big(w_{ij} + w_{ji}\big) \tag{51}$$

$$= \tfrac{1}{2}\sum_{i<j} p_{m,i}\, p_{m,j} \sum_{b=i}^{j-1}\big(u_b + d_b\big) \ = \ \sum_{i<j} p_{m,i}\, p_{m,j} \sum_{b=i}^{j-1} \gamma_b, \tag{52}$$

with γ_b as in Eq. (49). The inner sum is $C(i, j)$ for these weights, so the right-hand side equals $\frac{1}{2}\sum_{i,j} p_{m,i}p_{m,j}\, C(i, j) = \frac{1}{2}\mathbb{E}[C(Y, Y')]$, which by Definition 3 and Theorem 4 is $\sum_{k=1}^{K-1} \gamma_k P_{m,k}(1 - P_{m,k})$. □

Corollary 14 (Asymmetric Ordinal Costs Are Not Expressible)
Two directional cost structures (u_b, d_b) and (u'_b, d'_b) satisfying $u_b + d_b = u'_b + d'_b$ for every b induce the same impurity at every node, hence the same gain for every candidate split and the same fitted tree. The ratio u_b/d_b, which is the only quantity expressing asymmetry, is therefore invisible to any splitting criterion of the form (47).

The invariance is a consequence of the independence and identical distribution of Y and Y' in Definition 3, not of the boundary-additive structure of C. It applies to any impurity written as an expected pairwise cost, including the classical Gini index and Ranking Impurity. Asymmetry is not thereby ruled out of ordinal decision trees altogether: by Proposition 11 the leaf prediction responds to the cost structure whenever the increment in Eq. (43) is replaced by a directional one, and this is the point at which asymmetric costs can act. That route is developed in Subsection 6.3.

4. Experiments

4.1. Datasets

This study uses fifteen ordinal datasets. Twelve were selected from the public ordinal classification repository of [3], available at https://www.uco.es/grupos/ayrna/ordinal-trees, by the stratified rule described below. Three further datasets are included as fixed inclusions: the SNP-Accreditation dataset, the Contraceptive Method Choice (CMC) dataset, and the Nursery dataset.

The SNP-Accreditation dataset is drawn from *Capaian 8 SNP berdasarkan Konversi Rapor Pendidikan* (2026), whose eight standard attainment scores are the predictor features; the target variable is the school accreditation status, which consists of three ordinal categories. CMC and Nursery are widely used benchmark problems in the ordinal classification literature and are included so that the results can be compared with published work.

Table 3. Datasets used in the experimental study. The first twelve were selected from the ordinal classification repository of [3] by the stratified rule described below; the last three are fixed inclusions. Q is the number of ordinal classes, IR the imbalance ratio, and $\rho_v = \max_k v_k / \min_k v_k$ the ratio between the largest and smallest boundary contribution to the OGini impurity, where $v_k = P_{m_0,k}\,(1 - P_{m_0,k})$. All quantities are computed from the training partition.

Stratum	Dataset	Q	IR	ρ_v	#Train	#Test	#Feat.
$Q = 3$, $\rho_v < 2$	balance-scale	3	6.00	1.00	468	157	4
	mammoexp	3	3.20	1.68	276	136	5
$Q = 4$–5, $\rho_v < 2$	abalone	5	1.00	1.50	1000	3177	10
	stock	5	1.00	1.50	600	350	9
	support	5	39.20	1.05	489	241	19
$Q = 4$–5, ρ_v 2–10	LEV	5	15.10	9.01	750	250	4
	car	4	18.51	5.77	1296	432	21
	nhanes	5	10.67	6.64	3499	1724	30
$Q = 6$–9, $\rho_v > 10$	ERA	9	9.71	13.25	750	250	4
	winequality-red	6	63.75	37.53	1199	400	11
$Q \geq 10$, ρ_v 2–10	abalone10	10	1.00	2.78	1000	3177	10
	stock10	10	1.00	2.78	600	350	9
Fixed inclusions							
	SNP-Accreditation	3	4.59	1.16	5739	1434	8
	CMC	3	1.89	1.08	1180	293	9
	Nursery	5	3456	1.07	10369	2591	8

The twelve repository datasets were selected according to a rule fixed before any experiment was run, so that the selection cannot depend on the results. Two hard filters were applied: at least three ordinal classes, and at least 150 training observations, the latter to exclude datasets too small for stable cross-validation. The remaining datasets were then stratified along the two axes that are relevant to the proposed criterion. The first is the number of ordinal classes Q, which determines the number of configurable boundaries. The second is the ratio between the largest and the smallest boundary contribution to the OGini impurity,

$$\rho_v \;=\; \frac{\max_k v_k}{\min_k v_k}, \qquad v_k = P_{m_0,k}\big(1 - P_{m_0,k}\big), \tag{53}$$

which measures directly how unevenly the ordinal boundaries influence split selection. The imbalance ratio was not used for stratification, since a dataset may be strongly imbalanced while all of its boundaries contribute almost equally; across the datasets in the repository the rank correlation between the two quantities is only 0.47. The `support` dataset illustrates the point, with an imbalance ratio of 39.20 but $\rho_v = 1.05$. Both quantities are computed from the training data alone. One dataset was drawn from each non-empty stratum, taking the largest training sample as a deterministic tie-breaker, and further rounds then drew the next-largest dataset from each stratum until the target of twelve was reached, with ties across strata again broken by training size. Strata for which the repository contains no dataset passing the hard filters are omitted. This procedure yielded the twelve datasets listed in Table 3, two or three per stratum. All selected datasets are reported in the results, and none was discarded after the experiments were conducted.

The resulting collection has two features that bear on the analysis. `abalone` and `abalone10`, and likewise `stock` and `stock10`, are discretisations of the same underlying regression problem at different numbers of classes, so the fifteen datasets do not constitute fifteen independent blocks for the rank-based tests in Section 5. A sensitivity analysis retaining only one member of each pair is reported in Subsection 5.7. The Nursery dataset contains a class represented by one training observation, which is below the number of folds used for cross-validation; Subsection 4.5 describes the treatment of this case.

4.2. Splitting Criteria Compared

Nine criteria are compared. Six serve as impurity-function baselines within a single decision tree implementation; one is a decomposition method included as a reference point; and two are the proposed CS-OGini variants,

which are also impurity functions. Throughout, m denotes a node, $p_{m,i}$ the proportion of class α_i at m, and $P_{m,k} = \sum_{i \le k} p_{m,i}$ the cumulative proportion up to the k-th boundary, as in Eq. (7) and Eq. (8).

Nominal baselines. The Gini index of CART [4] and the entropy of ID3 [23] treat the class labels as unordered,

$$I_{\mathrm{Gini}}(m) = 1 - \sum_{i=1}^{K} p_{m,i}^2, \qquad I_{\mathrm{Ent}}(m) = -\sum_{i=1}^{K} p_{m,i} \log_2 p_{m,i}, \tag{54}$$

with the convention $0 \log_2 0 = 0$. They are included to quantify how much, if anything, is gained by using the ordinal structure at all.

Ordinal impurity baselines. Ordinal Gini [3] replaces the class proportions by the cumulative proportions,

$$I_{\mathrm{OGini}}(m) = \sum_{k=1}^{K-1} P_{m,k}\big(1 - P_{m,k}\big), \tag{55}$$

and is the special case $\gamma_k \equiv 1$ of Eq. (10). Its entropy counterpart applies the binary entropy to each cumulative proportion,

$$I_{\mathrm{OEnt}}(m) = -\sum_{k=1}^{K-1} \Big[P_{m,k} \log_2 P_{m,k} + \big(1 - P_{m,k}\big) \log_2\big(1 - P_{m,k}\big) \Big], \tag{56}$$

with the same convention $0 \log_2 0 = 0$, so that a boundary separating no observations contributes nothing. This criterion stands in the same relation to Eq. (55) as entropy stands to the Gini index. Ranking Impurity [29] accumulates the ordinal distance over all pairs of observations at the node,

$$I_{\mathrm{RI}}(m) = \sum_{i=1}^{K} \sum_{j=i+1}^{K} (j - i)\, n_{m,i}\, n_{m,j}, \tag{57}$$

where $n_{m,i}$ is the number of observations of class α_i at m. It differs from the other criteria compared here in two respects. Its evaluation is quadratic in K, in contrast to the linear cost of Eq. (55), and it is defined on counts, not proportions, so it is not invariant to the size of the node. The impurity after a split is obtained by summing Eq. (57) over the two children, which is the aggregation used by its originators, instead of the size-weighted average of Eq. (37) used for the seven proportion-based criteria. Applying Eq. (37) to Eq. (57) would define a different criterion. This difference is intrinsic to the measure and not a choice of the present study, but it means that comparisons involving Ranking Impurity are comparisons of criteria that aggregate over a split in different ways.

Imbalance baseline. Class-weighted Gini addresses class imbalance through the priors rather than through the ordinal structure. Each class is assigned a weight w_i inversely proportional to its frequency in the training set,

$$w_i = \frac{n}{K\, n_i}, \qquad \tilde{p}_{m,i} = \frac{w_i\, p_{m,i}}{\sum_{j=1}^{K} w_j\, p_{m,j}}, \qquad I_{\mathrm{CWG}}(m) = 1 - \sum_{i=1}^{K} \tilde{p}_{m,i}^2, \tag{58}$$

and the impurity is the Gini index of the reweighted distribution. It is included because the distribution-based scheme of Definition 5 is also a response to imbalance, but applied at the boundaries rather than at the classes. Since it does not use the class order, it is grouped with the nominal impurity functions in Tables 5 and 6.

Decomposition baseline. The approach of Frank and Hall [9] is not an impurity function but a decomposition: the K-class ordinal problem is replaced by $K-1$ binary problems, the k-th predicting whether the label exceeds α_k, and the class probabilities are recovered from the $K-1$ binary posteriors by differencing. Each binary problem is solved by a decision tree using the nominal Gini index of Eq. (54). The decomposition changes the model class, not the splitting criterion, and it predicts from the class probabilities reconstructed at the leaves instead of from a single leaf label. Differences involving Frank and Hall therefore cannot be attributed to the impurity formulation alone. It is reported throughout as a reference point, appearing in every table and in the grouping of Subsection 5.2, but no claim about the effect of an impurity function rests on a comparison involving it.

Proposed criteria. The two CS-OGini variants use Eq. (10),

$$I_{CS}(m) = \sum_{k=1}^{K-1} \gamma_k \, P_{m,k}\big(1 - P_{m,k}\big), \tag{59}$$

with the linear index-based weights $\gamma_k = k$ and with the distribution-based weights $\gamma_k^{(\beta)}$ of Definition 5. The choice of these two among the schemes of Subsection 4.3 is explained there.

Table 4 summarises the nine criteria.

Table 4. The nine criteria compared. K is the number of ordinal classes.

Criterion	Type	Uses class order	Defining equation
CART (Gini)	Impurity	No	Eq. (54), Gini term
ID3/C4.5 (Entropy)	Impurity	No	Eq. (54), entropy term
Class-Weighted Gini	Impurity	No	Eq. (58)
OGini	Impurity	Yes	Eq. (55)
Ordinal Entropy	Impurity	Yes	Eq. (56)
Ranking Impurity	Impurity	Yes	Eq. (57)
CS-OGini (Linear)	Impurity	Yes	Eq. (10), $\gamma_k = k$
CS-OGini (Distribution)	Impurity	Yes	Eq. (10), Def. 5 with $\beta = 1$
Frank & Hall	Decomposition	Yes	$K-1$ binary trees

4.3. Boundary Cost Schemes

The index-based family of Eq. (3) was instantiated with three functional forms, alongside the uniform baseline and the distribution-based scheme:

$$\underbrace{\gamma_k = 1}_{\text{uniform (OGini)}}, \quad \underbrace{\gamma_k = k}_{\text{linear}}, \quad \underbrace{\gamma_k = k^2}_{\text{polynomial}}, \quad \underbrace{\gamma_k = 2^k}_{\text{exponential}}, \quad \underbrace{\gamma_k^{(\beta)} = \big[P_{m_0,k}(1 - P_{m_0,k})\big]^{-\beta}}_{\text{distribution-based}}. \tag{60}$$

The first four assign weights according to the boundary index and are identical for every dataset with the same number of classes. The distribution-based scheme assigns weights according to the cumulative class proportions of the training data, so the resulting weights differ across datasets.

The index-based family provides fewer distinct criteria than the number of functional forms suggests. Since the impurity of Eq. (10) is linear in γ, it is unchanged by a positive rescaling of the weights, so for $K = 3$ the criterion depends only on the ratio γ_2/γ_1. The linear scheme then gives $\gamma = (1, 2)$ and the exponential scheme gives $\gamma = (2, 4)$; both yield a ratio of two and are therefore the same criterion. Four of the fifteen datasets have $K = 3$.

Two schemes are carried forward into the comparison of Section 5. The index-based schemes of Eq. (60) differ only in the rate at which the weight grows along the ordinal scale, and the argument above shows that this family is degenerate at small K: on the four datasets with $K = 3$ the linear and exponential schemes are literally the same criterion. The linear scheme is reported as the representative of that family, and the distribution-based scheme as the representative of the second. Reporting one member of each family ties the multiplicity of the comparison,

and hence the severity of the correction applied in Subsection 5.3, to the number of distinct mechanisms the study contains, and not to the number of functional forms available.

4.4. Evaluation Metrics

The evaluation uses accuracy, Macro-F1, balanced accuracy, mean absolute error, quadratic weighted kappa, and expected misclassification cost. Let n denote the number of test observations, y_i and $\hat{y}_i$ the actual and predicted class indices of the i-th observation, and O_{ij} the confusion matrix entry counting observations of actual class α_i predicted as α_j.

Accuracy is the proportion of correct predictions. *Macro-F1* is the unweighted mean of the per-class F_1 scores, and *Balanced Accuracy* the unweighted mean of the per-class recalls; both give equal weight to every class irrespective of its frequency, and are reported because several of the datasets are strongly imbalanced.

Mean Absolute Error measures the average ordinal distance of the errors,

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|. \tag{61}$$

Quadratic Weighted Kappa measures agreement corrected for chance, with a penalty growing quadratically in the ordinal distance,

$$\mathrm{QWK} = 1 - \frac{\sum_{i,j}\omega_{ij}O_{ij}}{\sum_{i,j}\omega_{ij}E_{ij}}, \qquad \omega_{ij} = \frac{(i-j)^2}{(K-1)^2}, \tag{62}$$

where E_{ij} is the expected confusion matrix under independence of the actual and predicted marginals. Values are reported multiplied by 100.

Expected Misclassification Cost is the average cost incurred on the test set under a fixed cost matrix,

$$C^{\mathrm{eval}}(i,j) = \sum_{b=\min(i,j)}^{\max(i,j)-1} b \;=\; \frac{|i-j|\,(i+j-1)}{2}, \qquad C^{\mathrm{eval}}(i,i) = 0, \tag{63}$$

that is, the boundary-additive cost of Definition 1 evaluated at the linear boundary weights $\gamma_b = b$. The evaluation cost matrix C^{eval} is fixed in advance and is *identical for every criterion*, and in particular is independent of the boundary weights γ used during training. This is essential for the fairness of the comparison: were C^{eval} derived from the γ of the criterion under evaluation, each criterion would be scored against its own objective. Note that C^{eval} is not the absolute ordinal distance, since ECOST would then coincide with Eq. (61); because the weights grow with the boundary index, an error of a given ordinal distance is penalised more heavily at the upper end of the scale than at the lower end. The two metrics are reported separately.

The linear weighting $\gamma_b = b$ used to define C^{eval} is also the weighting of the CS-OGini (Linear) variant, so that this one criterion is trained under a cost structure identical to the one against which ECOST scores it, while the remaining eight are not. The comparison on ECOST therefore favours that variant. It nonetheless attains a higher mean rank on ECOST than OGini (5.00 against 4.30), with a paired difference of -0.0178 and a 95% confidence interval of $[-0.0592, +0.0237]$. The conclusion of Subsection 5.3 thus holds also for the criterion whose training weights coincide with the evaluation cost.

4.5. Training and Evaluation Protocol

Partitions. The twelve repository datasets are used exactly as distributed, including their predefined training and test partitions and their predefined encoding of the predictor variables. The resulting partition sizes are reported in Table 3 and range from 24% to 75% of each dataset allocated to training; they are not re-partitioned, so that the results remain comparable with published work using the same repository. For the three fixed inclusions, which are not part of the repository, an approximately 80%–20% split is drawn by stratified sampling to preserve the class distribution; the exact partition sizes, which differ slightly from the nominal ratio because the allocation is rounded within each class, are reported in Table 3.

Each dataset is evaluated over five independent partitions. For the twelve repository datasets these are the first five of the partitions distributed with the repository. For the three fixed inclusions the first partition is the original hold-out split and the remaining four are drawn by repeated stratified sampling over the combined training and test data at the same test ratio, each from an independently seeded generator derived from a fixed base seed, so that the partitions are reproducible. Within each dataset all nine criteria use identical partitions and identical cross-validation folds, so the difference in provenance between the two groups does not affect the comparison.

For the three fixed inclusions, all predictor variables are treated as discrete ordinal variables and encoded as ordered integers according to the documented level ordering of each attribute. The twelve repository datasets retain the encoding distributed with the repository, which for some of them expands a categorical attribute into several indicator columns; the number of features reported in Table 3 is the number of columns actually presented to the learner in each case.

Common protocol. Within each dataset, every criterion uses the same training and test partitions, the same cross-validation folds for parameter selection, the same leaf assignment rule, and the same implementation of every evaluation metric. The leaf label is the majority class of the training observations reaching the leaf, applied identically to all nine criteria. By Proposition 11 the cost-minimising label of Definition 10 is likewise independent of γ, so neither choice of leaf rule can create a difference between the criteria compared here.

Parameter selection. The maximum tree depth is selected for every criterion over the grid $\{2, 3, 4, 5, 6, 8, 10, 16\}$ by stratified five-fold cross-validation on the training data, using the same folds for every criterion and the same selection rule, namely the minimisation of the cross-validated mean absolute error. The choice of a single, ordinal selection criterion for all nine methods is deliberate: selecting each criterion under its own objective would confound the comparison of the impurity functions with the comparison of their tuning objectives. A node is split only if it contains more than five training observations; nodes at or below that size become leaves. A candidate split is admissible only if both children receive at least one observation, and a split is accepted only if its gain is strictly positive, so that splits leaving the impurity unchanged are rejected. All three constraints, together with the depth limit, are identical for the nine criteria. For the distribution-based scheme the exponent is fixed at $\beta = 1$, the value distinguished by Theorem 6, and the boundary weights are computed once from the training class distribution of each dataset by Eq. (26) with the additive smoothing of Eq. (34), then normalised to unit mean. No search over β is performed, so the depth is the only quantity selected by cross-validation for any of the nine criteria. The selected depth is driven mainly by the dataset, not by the criterion: a variance decomposition of the selected depths attributes 81.3% of their variance to the dataset and partition and 18.7% to the criterion, and the median selected depth is five or six for every criterion. Individual selections nonetheless span the full grid. The tuning grids, the fold construction and the selection rule are implemented in the repository, together with the complete cross-validation log from which the selections were made.

The final model is retrained on the entire training set at the selected parameters and evaluated once on the test set.

Degenerate strata. The Nursery dataset contains a class with fewer training observations than the number of folds, so a stratified 5-fold partition cannot place a representative of that class in every fold. The fold construction assigns that observation to the validation set of one fold and to the training set of the remaining four, so all five folds are well defined and none is discarded. Because the folds are built once per dataset and shared by all nine criteria, the effect is identical for every criterion and does not bias the comparison. Since the class is likewise represented by a single observation in the test partition, and no observation is assigned to that class by any criterion in any partition, its per-class F1 and its recall are both zero throughout. Macro-F1 and balanced accuracy on this dataset are bounded above by 80% for every criterion. The bound is identical for all nine criteria and affects the level of these two metrics on this dataset, not the comparison between criteria.

5. Results and Discussion

Corollary 12 establishes that boundary weighting acts on the fitted tree through one channel, the selection of splits. The experiments reported below measure how much that channel carries. Subsections 5.2 and 5.3 bound the difference that boundary weighting produces against the unweighted baseline and against six further criteria, and Subsection 5.4 tests whether a practitioner can steer the errors of the fitted tree by choosing γ. The comparison criteria calibrate the scale on which such a difference would be visible rather than forming the object of study.

5.1. Experimental Protocol

All results reported in this section were obtained under a single protocol. The protocol is that of Subsection 4.5; the statistical treatment is as follows.

Metrics are first averaged over partitions within a dataset, and the resulting values are then compared across datasets. Because the metrics are not on a common scale across datasets, comparisons between criteria use the mean rank computed within each dataset, following the protocol of [7]. Mean values are reported alongside the ranks, since ranks alone convey the consistency of a difference but not its magnitude.

Statistical comparison proceeds in two stages. The Friedman test assesses whether the criteria differ overall on a given metric. Pairwise Wilcoxon signed-rank tests at the dataset level then compare each criterion against the OGini baseline; these are computed for every metric irrespective of the omnibus outcome, and the Holm–Bonferroni correction is applied jointly across the resulting forty-eight comparisons, which are reported in Subsection 5.3. Testing the full set rather than only the metrics on which the omnibus test is significant is the more conservative choice, since it enlarges the family over which the correction is made. Because a non-significant test does not by itself establish equivalence, we also report paired confidence intervals, which bound the magnitude of any difference that remains compatible with the data.

The boundary-control test of Subsection 5.4 is assessed separately, by a two-sided Wilcoxon signed-rank test with Holm correction across the forty-five dataset-by-factor comparisons. A one-sided test in the direction the framework predicts could not detect an effect in the opposite direction, and whether such an effect occurs is part of what the test is intended to determine.

5.2. Comparison of Splitting Criteria

Table 5 reports the mean value and mean rank of each of the nine splitting criteria across the fifteen datasets. Seven of the nine are closely grouped: CART, ID3, OGini and ordinal entropy, both CS-OGini variants, and the Frank and Hall decomposition, whose status as a change of model class is noted in Subsection 4.2. Their mean accuracy ranges from 62.25 to 63.00 percent and their mean Macro-F1 from 50.50 to 50.79 percent, and their mean ranks on accuracy span 4.17 to 6.50. Within this group the lowest mean rank belongs to ID3 on four of the six metrics, to the Frank and Hall decomposition on QWK, and to CART on ECOST. None of ID3's four paired differences against OGini excludes zero, however, and its mean accuracy exceeds that of OGini by 0.02 percentage points, so the lowest mean rank does not establish a leading position. Two criteria lie outside the group. Ranking Impurity and class-weighted Gini have mean ranks above the other seven on five of the six metrics each, the exception in both cases being accuracy. On accuracy the two lie inside the range spanned by the grouped set, because the Frank and Hall decomposition holds the highest mean rank there at 6.50; the exception therefore reflects the position of the decomposition rather than any strength of these two criteria.

Across all nine criteria the range of mean accuracy is 1.00 percentage point, and the range of mean MAE is 0.0172. For comparison, the standard deviation of accuracy across the five partitions within a single dataset has a median of 1.61 percentage points and exceeds the full range of 1.00 for 72.6% of the dataset-by-criterion combinations, so the differences between criteria are smaller than the variability induced by the choice of partition.

Rank and value do not order the criteria identically. Ordinal entropy attains the highest mean accuracy of all nine criteria at 63.00 percent while ranking fourth on that metric, and Frank and Hall attains the lowest mean ECOST while ranking seventh. A mean rank records how often one criterion exceeds another and a mean value records by how much, so a disagreement between them indicates that the margins of the wins are smaller than the margins of the losses. Among the five criteria with the lowest mean rank on accuracy the mean values span 0.22 percentage

Table 5. Mean value and mean rank of each splitting criterion across the fifteen datasets. Ranks are computed within each dataset and averaged; rank 1 is best. Lower is better for MAE and ECOST. Criteria are listed in the order of Table 4 and are not sorted by any performance column: as Subsection 5.3 shows, no difference against OGini survives correction for multiple testing on any metric.

	Acc		Macro-F1		BAcc		QWK		MAE		ECOST	
Method	Val	Rank	Val	Rank	Val	Rank	Val	Rank	Val	Rank	Val	Rank
Nominal impurity												
CART (Gini)	62.95	4.37	50.62	4.60	51.55	4.67	63.21	5.00	0.5454	4.10	1.7068	4.20
ID3/C4.5 (Entropy)	62.96	4.17	50.76	3.90	51.73	3.90	63.57	4.37	0.5446	3.83	1.7082	4.23
Class-Weighted Gini	62.00	5.93	47.39	6.47	49.28	5.80	61.40	6.87	0.5596	6.67	1.7723	6.93
Ordinal impurity												
OGini	62.94	4.47	50.70	4.50	51.53	4.83	63.17	4.63	0.5424	4.50	1.6907	4.30
Ordinal Entropy	63.00	4.43	50.79	4.17	51.86	3.97	63.38	4.23	0.5448	4.63	1.6981	4.70
Ranking Impurity	62.24	5.63	47.98	7.07	49.01	6.80	61.78	5.93	0.5559	6.40	1.7292	5.87
Proposed												
CS-OGini (Linear)	62.78	4.30	50.64	4.73	51.55	4.60	63.42	5.13	0.5466	5.03	1.7084	5.00
CS-OGini (Distribution)	62.68	5.20	50.50	5.37	51.58	5.23	62.87	4.83	0.5461	4.80	1.7055	4.63
Decomposition												
Frank & Hall	62.25	6.50	50.71	4.20	50.95	5.20	63.44	4.00	0.5437	5.03	1.6851	5.13
Friedman p	0.1685		0.0135		0.1060		0.1077		0.0636		0.1129	

points, against a median within-dataset standard deviation of 1.61 across partitions, so neither ordering carries information at this resolution.

Five of the six Friedman tests are not significant at the 5% level. The single significant result, on Macro-F1 ($p = 0.0135$), does not indicate that any criterion outperforms the others: the two criteria furthest from the mean rank on this metric are Ranking Impurity (7.07) and class-weighted Gini (6.47), the same two that lie outside the closely grouped set. Recomputing the test on the remaining seven criteria alone raises the p-value to 0.6146, and their mean ranks then span only 3.50 to 4.90. The Friedman test detects that the set of criteria is not homogeneous, which in this case reflects the presence of these two rather than the superiority of any particular one.

5.3. Comparison with the OGini Baseline

Table 6 reports the paired difference between each criterion and OGini over the fifteen datasets, together with a 95% confidence interval. The sign convention is such that a positive value indicates that the criterion performs better than OGini on that metric. For the proposed criterion the comparison bounds the difference rather than resolving it in either direction. Any advantage of either CS-OGini variant over OGini exceeding 0.30 percentage points of accuracy is excluded at the 95% level, as is any deficit exceeding 0.82 points. This is consistent with the structural results of Subsection 3.6, which confine boundary weighting to a single channel, and is the tightest statement the design supports.

Across the forty-eight pairwise comparisons, six intervals exclude zero, against the 2.4 expected under the global null hypothesis. The excess is attributable to the two criteria lying outside the closely grouped set in Subsection 5.2: five of the six involve Ranking Impurity or class-weighted Gini. After Holm correction across all forty-eight tests no comparison survives, and none of the six involves the proposed criterion.

The intervals differ in width as well as in location. For Ranking Impurity and class-weighted Gini the point estimates are large and the intervals wide, indicating a genuine deficit whose magnitude is imprecisely estimated. For the remaining six criteria the intervals are narrow and centred close to zero, the one exception being the accuracy of the Frank and Hall decomposition, whose interval excludes zero by a margin of 0.05 points and does not survive the correction.

For the proposed criterion the intervals are narrow throughout. Any advantage of the distribution-based scheme over OGini exceeding 0.30 percentage points of accuracy, 0.30 points of Macro-F1, 0.53 points of balanced accuracy, or 0.60 points of QWK can be excluded at the 95% level; the corresponding bounds are 0.0068 for MAE and 0.0165 for ECOST. The linear scheme yields the same picture, with a bound of 0.29 points of accuracy.

Table 6. Paired difference from OGini with 95% confidence intervals over fifteen datasets. Positive values indicate that the criterion performs better than OGini; for MAE and ECOST, on which lower values are better, the sign is reversed accordingly. OGini itself is omitted, being the reference. Rows follow the order of Table 4, grouped by criterion type. Intervals excluding zero are marked with an asterisk. No comparison survives Holm correction across the forty-eight tests.

Panel A Method	Accuracy	Macro-F1	Balanced Acc.
Nominal impurity			
CART (Gini)	$+0.01\ [-0.64, +0.66]$	$-0.08\ [-0.72, +0.56]$	$+0.02\ [-0.66, +0.69]$
ID3/C4.5 (Entropy)	$+0.02\ [-0.71, +0.75]$	$+0.06\ [-0.69, +0.81]$	$+0.20\ [-0.68, +1.08]$
Class-Weighted Gini	$-0.94\ [-1.86, -0.03]^*$	$-3.31\ [-6.43, -0.19]^*$	$-2.25\ [-5.65, +1.15]$
Ordinal impurity			
Ordinal Entropy	$+0.06\ [-0.33, +0.45]$	$+0.09\ [-0.73, +0.91]$	$+0.33\ [-0.31, +0.98]$
Ranking Impurity	$-0.70\ [-1.49, +0.09]$	$-2.72\ [-4.24, -1.20]^*$	$-2.52\ [-4.14, -0.89]^*$
Proposed			
CS-OGini (Linear)	$-0.17\ [-0.62, +0.29]$	$-0.06\ [-0.50, +0.39]$	$+0.02\ [-0.49, +0.52]$
CS-OGini (Distribution)	$-0.26\ [-0.82, +0.30]$	$-0.20\ [-0.69, +0.30]$	$+0.05\ [-0.44, +0.53]$
Decomposition			
Frank & Hall	$-0.69\ [-1.34, -0.05]^*$	$+0.01\ [-0.89, +0.92]$	$-0.57\ [-1.52, +0.37]$

Panel B Method	QWK	MAE	ECOST
Nominal impurity			
CART (Gini)	$+0.04\ [-0.94, +1.02]$	$-0.0030\ [-0.0192, +0.0132]$	$-0.0161\ [-0.0602, +0.0279]$
ID3/C4.5 (Entropy)	$+0.40\ [-0.93, +1.72]$	$-0.0023\ [-0.0173, +0.0128]$	$-0.0175\ [-0.0716, +0.0365]$
Class-Weighted Gini	$-1.77\ [-4.75, +1.21]$	$-0.0172\ [-0.0411, +0.0066]$	$-0.0816\ [-0.1853, +0.0220]$
Ordinal impurity			
Ordinal Entropy	$+0.22\ [-0.86, +1.29]$	$-0.0024\ [-0.0119, +0.0071]$	$-0.0074\ [-0.0389, +0.0241]$
Ranking Impurity	$-1.39\ [-2.89, +0.12]$	$-0.0136\ [-0.0246, -0.0025]^*$	$-0.0385\ [-0.0802, +0.0032]$
Proposed			
CS-OGini (Linear)	$+0.25\ [-0.65, +1.16]$	$-0.0042\ [-0.0131, +0.0047]$	$-0.0178\ [-0.0592, +0.0237]$
CS-OGini (Distribution)	$-0.30\ [-1.20, +0.60]$	$-0.0037\ [-0.0143, +0.0068]$	$-0.0149\ [-0.0462, +0.0165]$
Decomposition			
Frank & Hall	$+0.27\ [-0.78, +1.33]$	$-0.0013\ [-0.0127, +0.0101]$	$+0.0055\ [-0.0403, +0.0514]$

A non-significant test alone would not establish equivalence between the two criteria. The intervals additionally constrain any difference between them to be small.

5.4. Behaviour of the Boundary Weights

Three measurements bound what split selection alone carries.

Weights act only through the tree structure. Proposition 11 and Corollary 12 confine the weights to split selection. The channel that remains is narrow. When a single boundary weight is raised by a factor of eight while the others are held at unity, 81.1% of test predictions remain identical to those of the uniform weighting; at a factor of sixteen the figure is still 78.7%.

Raising a boundary weight does not reliably reduce errors at that boundary. The framework is intended to allow the practitioner to express that some ordinal boundaries matter more than others. This was tested directly. For each of the sixty-six ordinal boundaries in the collection, the weight γ_k was raised from unity to each of the values 4, 8 and 16 while the remaining weights were held at unity, so that the comparison is against the OGini baseline and the effect of a single boundary is isolated. Combining the sixty-six boundaries with five partitions, three tree depths (4, 6 and 8) and three weight magnitudes gives 2970 controlled comparisons. The crossing-error rate at the targeted boundary decreased in 33.7% of them, and the mean change was $+0.00047$, that is, a slight

increase. The mean change at the remaining boundaries was +0.00223 per boundary, measured on the same scale, so the intervention disturbs the boundaries it does not target by about five times as much as it moves the one it does.

The direction of the effect varies across the collection. On seven of the fifteen datasets the targeted crossing rate falls on average and on eight it rises. Of the forty-five dataset-by-factor comparisons, eleven are significant after Holm correction under a two-sided test: six in which the targeted rate falls, confined to `LEV`, `car` and `stock10`, and five in which it rises, confined to `nhanes` and `abalone10`. The two groups are comparable in magnitude: averaged over all boundaries, depths, factors and partitions of a dataset, the largest mean decrease is -0.0080 on `stock10` and the largest mean increase $+0.0061$ on `nhanes`. Raising the weight of a boundary therefore does produce a detectable change in the errors crossing that boundary, but whether the change is favourable is a property of the dataset, not of the intervention. The eleven significant comparisons describe this collection and do not support an inferential claim about ordinal problems in general. Within each dataset-by-factor cell the units entering the test are boundaries, partitions and depths, which are not mutually independent, so the nominal significance overstates the evidence. The quantity carrying the conclusion is the sign of the mean change, which is negative on seven datasets and positive on eight however the units are weighted. Raising the factor from four to sixteen moves the overall proportion of decreases only from 33.2% to 34.7%, so the outcome is not an artefact of an insufficiently large weight.

Index-based weights collapse for small K. The degeneracy established in Subsection 4.3 is a further limit on what the index-based family can express: at $K = 3$ the linear and exponential schemes are the same criterion, so four of the fifteen datasets contribute nothing to a comparison between them.

5.5. Adequacy of the Experimental Design

The principal finding is the absence of a difference, so the sensitivity of the design requires examination.

The observed differences are small relative to the noise in their own measurement. On Macro-F1 the paired difference between the distribution-based scheme and OGini is -0.196 points with a standard error of 0.232, so the effect is smaller than one standard error. The between-dataset standard deviation of that difference is 0.899 points, while the standard deviation of Macro-F1 across the five partitions within a single dataset has a median of 2.6 points; the variability observed between datasets is no larger than what repeated partitioning of the same dataset already produces, and the estimated genuine heterogeneity between datasets is negligible.

The instability of the rankings at this number of repetitions was confirmed directly. A bootstrap over partitions shifts the mean rank on accuracy by up to 1.87 positions depending on which partitions are drawn, with a median maximum shift of 0.77; the linear scheme attains a better mean rank than OGini in 44% of the resamples. The ordering within the closely grouped set is therefore not stable at five partitions, which is itself evidence that the differences it encodes are small.

Additional repetitions would sharpen these estimates without altering them. Extending the study to further datasets would not change the conclusion either: with fifteen datasets and the observed effect size the paired test has a power of approximately 0.12 at the 5% level, and roughly doubling the number of datasets would raise this only to about 0.20. Moreover, the additional datasets available in the repository are largely discretisations of the same underlying regression problems, so they do not supply independent blocks for the rank-based tests; the sensitivity analysis in Subsection 5.7 quantifies the effect of the two such pairs already present in the collection.

5.6. Computational Complexity

5.6.1. Analytical complexity Consider the evaluation of all candidate splits at a node containing n observations, with d predictor features and K ordinal classes. For each feature the observations are sorted once, at cost $O(n \log n)$, after which the class counts on either side of every candidate threshold are obtained incrementally. Evaluating a single threshold requires the cumulative class proportions and their weighted sum, both of which are $O(K)$. The total cost at one node is therefore

$$O(d\,(n \log n + nK)) \tag{64}$$

for OGini and for CS-OGini alike. The two criteria differ only in that CS-OGini multiplies each of the $K-1$ boundary terms by its weight γ_k, which adds $K-1$ multiplications per threshold and does not change the asymptotic order. The boundary weights themselves are computed once, at the root, at cost $O(K)$ for both families of weighting rules.

CS-OGini introduces no asymptotic computational overhead relative to OGini. For comparison, Ranking Impurity evaluates a quadratic form over the class counts and therefore requires $O(d\,nK^2)$ at one node, while the Frank and Hall decomposition trains $K-1$ separate binary trees and thus incurs a cost that grows linearly with the number of classes.

5.6.2. Empirical running time Table 7 reports the time required to construct a single tree of depth eight, with the OGini and CS-OGini times taken as the minimum of seven repetitions. The ratio between CS-OGini and OGini ranges from 0.97 to 1.06 across all configurations. Since CS-OGini performs strictly more arithmetic than OGini, ratios below unity are measurement noise rather than a genuine speed-up, and their presence indicates that the difference between the two criteria is smaller than the resolution of the measurement. This confirms the analysis above: incorporating boundary weights into the impurity does not increase the cost of tree construction in practice.

Table 7. Tree construction time in seconds (depth = 8) on synthetic data with controlled n, d and K. The OGini and CS-OGini times are the minimum of seven repetitions; the remaining columns are a single run and are reported for scale only. n is the number of observations, d the number of features, and K the number of ordinal classes; RI denotes Ranking Impurity. All criteria were measured on the same machine with the same implementation, so absolute times are machine-dependent and only the ratios between criteria are meaningful.

n	d	K	OGini	CS-OGini	Ratio	CART	RI	Frank–Hall
500	10	3	0.202	0.198	0.98	0.308	0.454	0.504
500	10	9	0.243	0.235	0.97	0.346	0.440	1.865
2000	10	3	0.471	0.480	1.02	0.769	0.981	1.386
2000	10	9	0.549	0.560	1.02	0.784	1.053	5.259
2000	30	9	1.605	1.659	1.03	4.008	5.197	19.312
8000	10	9	1.480	1.562	1.06	2.028	1.735	13.456

The effect of the number of classes is small enough to be obscured by measurement noise at this problem size. With $n = 2000$, $d = 10$ and depth six, the construction time of OGini rises from 0.195 s at $K = 3$ to 0.275 s at $K = 12$, while CS-OGini and Ranking Impurity vary non-monotonically over the same range. This is consistent with the term nK in Eq. (64) being dominated by the sorting term $n \log n$ for the class cardinalities encountered in practice: the predicted growth in K is smaller than the run-to-run variation of the measurement.

5.7. Sensitivity to the Non-Independent Dataset Pairs

Four of the fifteen datasets form two pairs of discretisations of the same underlying problem: `abalone` and `abalone10` share the same predictors and response, as do `stock` and `stock10`. The fifteen blocks entering the rank-based tests are not fully independent, and the effective number of distinct problems is thirteen. The analysis was repeated twice, once removing the ten-class member of each pair and once removing the original, so that neither choice of representative drives the result.

The conclusions are unchanged. Removing the ten-class members leaves five of the six omnibus tests non-significant, with Macro-F1 again the exception ($p = 0.0084$); removing the originals leaves four, with MAE joining Macro-F1 (MAE $p = 0.0274$, Macro-F1 $p = 0.0031$). Under both choices Ranking Impurity and class-weighted Gini occupy the two highest mean ranks on five of the six metrics, the exception being accuracy, exactly as on the full collection. The omnibus results therefore continue to reflect the presence of these two rather than any criterion outperforming the others.

The position of both CS-OGini variants within the closely grouped set is unchanged under either choice, as is that of OGini and the remaining criteria: on accuracy the linear scheme moves from a mean rank of 4.30 to 4.35 and 4.42, the distribution-based scheme from 5.20 to 5.38 and 5.23, and OGini from 4.47 to 4.62 and 4.77. The

exercise also illustrates the instability discussed in Subsection 5.5: removing two of fifteen datasets is enough to exchange which criterion holds the lowest mean rank on three of the six metrics under one choice of representative and on four of the six under the other.

5.8. Discussion

5.8.1. Why boundary weighting has little effect The structural properties established in Subsection 3.6 account for the experimental findings. Cost sensitivity can enter a decision tree at three points: the splitting criterion, the prediction rule at the leaves, and the pruning objective. The present framework places it at the first. Proposition 11 shows that the second point is closed to it, since the cost-minimising label is the median of the leaf distribution irrespective of γ. Cost-complexity pruning was evaluated separately, using the Bayes leaf rule that the pruning criterion requires and an evaluation cost matrix fixed independently of the boundary weights. Several members of the distribution-based family were compared against uniform weighting under that protocol, and the two remain indistinguishable: the omnibus test across the weighting schemes gives $p = 0.915$ on accuracy and $p = 0.555$ on Macro-F1 over the fifteen datasets, and the smallest paired p-value against the uniform baseline is 0.131. Adding the pruning objective therefore does not open a channel that split selection alone had closed. By Proposition 11 the Bayes leaf label is the median irrespective of γ, so the change of leaf rule cannot itself create or remove a difference between the weightings. The configuration and its output are included in the repository. The only remaining channel is the choice of splits. Split selection is a ranking problem, so a boundary weight affects the tree only when it reverses the ordering of two candidate splits, and in doing so it moves the tree towards a split that the unweighted criterion judged inferior.

This also explains the direction of the effect. Weighting a boundary heavily makes the impurity approximate a binary problem at that boundary, discarding information carried by the remaining $K - 2$ boundaries. On the training data the targeted boundary is separated more sharply; on unseen data the tree generalises less well. The variance introduced by this specialisation is of the same order as the bias it removes, which is what the near-zero mean differences in Table 6 reflect.

5.8.2. Class imbalance and boundary contributions The distribution-based scheme was motivated by an imbalance in how boundaries contribute to the impurity, and that imbalance is real. On the ERA dataset the largest boundary contribution exceeds the smallest by a factor of 13.25 (Table 2), and across the datasets considered the ratio ρ_v ranges from 1.00 to 37.53. Setting $\beta = 1$ removes this imbalance exactly, by Theorem 6.

Removing the imbalance does not, however, improve prediction. The same holds for class-weighted Gini, which addresses class imbalance through altered priors instead of boundary weights, and whose mean rank lies above that of the unweighted Gini index on all six metrics. Two different corrections for imbalance, applied at different points in the impurity, both leave the comparison against the criterion they correct unchanged or slightly unfavourable.

5.8.3. Implications for practice The practical implication is that the choice among the splitting criteria considered here is of secondary importance within the closely grouped set. Their mean accuracies differ by at most 0.75 percentage points, and no pairwise comparison against OGini survives correction for multiple testing on any metric. Practitioners working with ordinal decision trees may select a criterion on grounds of interpretability or computational cost, since expected accuracy does not distinguish them, and direct their effort towards other components of the modelling pipeline. For the two criteria outside this group the separation is larger but also less precisely estimated, and it was obtained under a protocol chosen for the comparison as a whole rather than tuned to any individual criterion; Ranking Impurity additionally carries a computational cost quadratic in the number of classes, which may matter independently of accuracy when K is large.

The results locate the proposed criterion. CS-OGini generalises OGini and recovers it exactly at $\gamma = 1$; it introduces no asymptotic computational overhead, with measured running times within 0.97 to 1.06 times those of OGini (Subsection 5.6); and its performance is statistically indistinguishable from OGini and from the established baselines. It does not provide the practitioner with usable control over which ordinal errors occur. The intervention has a measurable effect on the targeted boundary, but its sign cannot be specified in advance.

5.8.4. Threats to validity Three limitations qualify these findings.

First, the comparison covers a single decision tree learner. Tree-based ensembles are routinely reported to improve on a single tree in applied settings [14], and whether the criteria compared here behave equivalently within an ensemble is untested. There is reason to expect otherwise: a criterion that produces slightly weaker but more diverse individual trees may perform better in a random forest than in a single tree.

Second, the study evaluates symmetric boundary costs. By Corollary 14 any two directional cost structures with the same boundary sums $u_b + d_b$ induce identical trees, so asymmetric costs cannot be expressed within this class of impurity functions at all and the conclusion applies to the symmetric case only.

Third, the datasets are drawn from a single public repository, and a substantial proportion of the problems with many classes are obtained by discretising regression targets. Ordinal problems arising natively with ten or more classes are underrepresented, and the behaviour of the criteria on such problems remains open.

6. Conclusions

6.1. Summary

This study formulated Cost-Sensitive Ordinal Gini (CS-OGini), a splitting criterion for ordinal decision trees in which each ordinal boundary is assigned a configurable cost, and established its properties both theoretically and empirically.

The formulation expresses the impurity at a node as half the expected boundary-additive cost between two independently drawn labels, and yields the closed form $I_{CS}(m) = \sum_k \gamma_k P_{m,k}(1 - P_{m,k})$. Ordinal Gini is recovered exactly when all boundary weights are equal, so CS-OGini is a strict generalisation. Two families of weighting rules were considered: weights specified as a function of the boundary index, and weights derived from the cumulative class proportions of the training data. For the second family the exponent $\beta = 1$ equalises the contribution of all ordinal boundaries to the impurity at the root, and is the unique exponent doing so whenever those contributions are not already equal, while $\beta = 0$ recovers OGini, so the family interpolates between the two.

Three structural properties delimit what boundary weighting can achieve within this framework. By Proposition 9 the accumulated boundary cost is monotone in the ordinal distance between classes, so the weights induce a genuine ordinal cost structure. By Proposition 11 the cost-minimising prediction at a leaf is the median of the class distribution irrespective of the boundary weights, so the weights influence the model only through the tree structure. By Proposition 13, any impurity defined as the expected cost between two independent labels is invariant to the antisymmetric component of the cost matrix, and for directional boundary costs the impurity collapses exactly onto the symmetric weights $\gamma_b = \frac{1}{2}(u_b + d_b)$, so asymmetric ordinal costs cannot be represented within this class of criteria at all.

Empirically, nine splitting criteria were compared on fifteen ordinal datasets under an identical protocol. No criterion differs significantly from OGini on any metric once multiplicity is accounted for: mean accuracy spans 1.00 percentage point across all nine, five of six omnibus tests are non-significant, and no pairwise comparison survives Holm correction across the forty-eight tests performed. Paired confidence intervals bound the difference between the distribution-based CS-OGini variant and OGini within $[-0.82, +0.30]$ percentage points of accuracy, and the linear variant within $[-0.62, +0.29]$. A direct test of the framework's intended capability found that raising the weight of a specific boundary reduced the errors crossing that boundary in only 33.7% of 2970 controlled comparisons.

6.2. Contributions

The contribution of this study is a formal characterisation of boundary-weighted ordinal impurity. It comprises: a boundary-weighted impurity framework for ordinal decision trees that contains OGini as a special case; a derivation of boundary weights from the class distribution, together with a proof that $\beta = 1$ is the unique exponent equalising the contribution of all boundaries; three structural properties that delimit what such weighting can accomplish; and an evaluation across nine criteria and fifteen datasets showing that these limits are binding in practice. The

implementation, including an automated test suite exercising the three structural properties and the complete cross-validation log of the parameter selection, is available at https://github.com/tiar77/cs-ogini.

6.3. *Future Work*

The structural results point beyond the present study, at the decision rule and at the level of the ensemble.

The first is asymmetric ordinal costs. Corollary 14 establishes that such costs cannot enter the impurity, but they can enter the decision rule. Replacing the symmetric increment of Eq. (43) by its directional counterpart under costs u_k for over-prediction and d_k for under-prediction gives

$$R_m(a+1) - R_m(a) \;=\; u_a P_{m,a} - d_a\big(1 - P_{m,a}\big), \tag{65}$$

which is non-negative precisely when $P_{m,a} \geq \tau_a$ with $\tau_a = d_a/(u_a + d_a)$. When the ratio u_k/d_k is constant in k, the threshold $\tau_a \equiv \tau$ does not vary, the argument of Proposition 11(i) applies unchanged, and the cost-minimising label is the τ-quantile $\min\{k : P_{m,k} \geq \tau\}$ rather than the median. When the ratio varies across boundaries, Eq. (65) may change sign more than once, R_m need not be quasi-convex, and the minimiser must be obtained by direct evaluation over the K candidate labels, which costs $O(K)$ per leaf. Since the prediction then does respond to the cost structure, this is the natural place for asymmetric ordinal costs to act, and applications such as accreditation or severity assessment, where over- and under-prediction carry different consequences, provide the motivating cases.

The second is ensemble learning. Corollary 12 shows that the boundary weights act on the fitted model only by perturbing split selection, and the empirical results show that this perturbation does not improve single-tree accuracy. Taken together, these suggest using the weights as a source of diversity instead. Randomising the boundary weights across the trees of a random forest would decorrelate them at no additional computational cost, and the accuracy of a forest depends on both the strength of its individual trees and the correlation between them. Whether the resulting trade-off is favourable is an empirical question that the present single-tree design cannot answer.

Finally, the diagnostic quantities used here to characterise datasets, in particular the ratio ρ_v between the largest and smallest boundary contribution, are computed from the training data alone and may prove useful beyond the present study for anticipating whether cost-sensitive modifications are likely to affect a given problem.

REFERENCES

1. Abed Alsulami, Reda Khalifa, and Wajdi Alghamdi. Improving performance and accuracy in decision trees: A literature survey on impurity functions. *International Journal of Advanced Computer Science & Applications*, 17(4):985, 2026.
2. Imane Araf, Ali Idri, and Ikram Chairi. Cost-sensitive learning for imbalanced medical data: a review. *Artificial Intelligence Review*, 57(4):1, 2024.
3. Rafael Ayllón-Gavilán, Francisco José Martínez-Estudillo, David Guijo-Rubio, César Hervás-Martínez, and Pedro A Gutiérrez. Splitting criteria for ordinal decision trees: an experimental study. *Pattern Recognition*, 168:112273, 2025.
4. Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. Wadsworth, Belmont, CA, 1984.
5. Hui Chen and Mengru Ji. Experimental comparison of classification methods under class imbalance. *EAI Endorsed Transactions on Scalable Information Systems*, 8(33), 2021.
6. Jacob Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213, 1968.
7. Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.
8. Charles Elkan. The foundations of cost-sensitive learning. In *International Joint Conference on Artificial Intelligence*, pages 973–978. Lawrence Erlbaum Associates Ltd, 2001.
9. Eibe Frank and Mark Hall. A simple approach to ordinal classification. In *European conference on machine learning*, pages 145–156. Springer, 2001.
10. Giuliano Galimberti, Gabriele Soffritti, and Matteo Di Maso. Classification trees for ordinal responses in r: the rpartscore package. *Journal of Statistical Software*, 47:1–25, 2012.
11. Bita Ghasemkhani, Kadriye Filiz Balbal, Kokten Ulas Birant, and Derya Birant. Ordinal random tree with rank-oriented feature selection (ORT-ROFS): a novel approach for the prediction of road traffic accident severity. *Mathematics*, 13(2):310, 2025.
12. Redouane Hakimi, Badreddine Benyacoub, and Mohamed Ouzineb. A Mathematical Programming Model with Cost Sensitivity in the Objective Function for Imbalanced Datasets Challenges. *Statistics, Optimization & Information Computing*, 15(2):1099–1118, 2026.

13. Deniu He. Active learning for ordinal classification based on expected cost minimization. *Scientific Reports*, 12(1):22468, 2022.
14. Teguh Herlambang, Mohamad Yusak Anshori, Bambang Suharto, Zuraini Othman, Mohd Sanusi Azmi, and Mochammad Romli Arief. A Comparative Statistical Analysis of Decision Tree and AdaBoost Ensemble for Employee Performance Classification in the Hospitality Sector. *Statistics, Optimization & Information Computing*, 16(2):1813–1824, 2026.
15. Andrei P Kirilenko. Decision trees. In *Practical Data Mining with AI for Social Scientists*, pages 79–109. Springer, 2025.
16. Sotiris B Kotsiantis and Panagiotis E Pintelas. A cost sensitive technique for ordinal classification problems. In *Hellenic Conference on Artificial Intelligence*, pages 220–229. Springer, 2004.
17. Charles X Ling and Victor S Sheng. Cost-sensitive learning and the class imbalance problem. *Encyclopedia of machine learning*, 2011(2008):231–235, 2008.
18. Susan Lomax and Sunil Vadera. A survey of cost-sensitive decision tree induction algorithms. *ACM Computing Surveys (CSUR)*, 45(2):1–35, 2013.
19. Matan Marudi, Irad Ben-Gal, and Gonen Singer. A decision tree-based method for ordinal classification problems. *IISE Transactions*, 56(9):960–974, 2024.
20. Hany Osman. Cost-sensitive learning using logical analysis of data. *Knowledge and Information Systems*, 66(6):3571–3606, 2024.
21. George Petrides and Wouter Verbeke. Cost-sensitive ensemble learning: a unifying framework. *Data Mining and Knowledge Discovery*, 36(1):1–28, 2022.
22. Raffaella Piccarreta. Classification trees for ordinal variables. *Computational Statistics*, 23(3):407–427, 2008.
23. J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.
24. Lior Rabkin, Ilan Cohen, and Gonen Singer. Resource allocation in ordinal classification problems: A prescriptive framework utilizing machine learning and mathematical programming. *Engineering Applications of Artificial Intelligence*, 132:107914, 2024.
25. Gerhard Tutz. Ordinal trees and random forests: Score-free recursive partitioning and improved ensembles. *Journal of Classification*, 39(2):241–263, 2022.
26. Haomin Wang, Gang Kou, and Yi Peng. Multi-class misclassification cost matrix for credit ratings in peer-to-peer lending. *Journal of the Operational Research Society*, 72(4):923–934, 2021.
27. Jinhong Wang, Jintai Chen, Jian Liu, Dongqi Tang, Danny Z Chen, and Jian Wu. A survey on ordinal regression: Applications, advances and prospects. *arXiv preprint arXiv:2503.00952*, 2025.
28. Tanawan Watthaisong, Khamron Sunat, and Nipotepat Muangkote. Comparative Evaluation of Imbalanced Data Management Techniques for Solving Classification Problems on Imbalanced Datasets. *Statistics, Optimization & Information Computing*, 12(2):547–570, 2024.
29. Fen Xia, Wensheng Zhang, and Jue Wang. An Effective Tree-Based Algorithm for Ordinal Regression. *IEEE Intelligent Informatics Bulletin*, 7(1):22–26, 2006.
30. Ayfer Ezgi Yilmaz and Haydar Demirhan. Weighted kappa measures for ordinal multi-class classification performance. *Applied Soft Computing*, 134:110020, 2023.
31. Xinyi Zhu, Hongbing Zhang, Quan Ren, Lingyuan Zhang, and Yihang Ge. Cost-focused intelligent reservoir logging interpretation: evaluation of cost-sensitive active learning strategies. *Engineering Applications of Artificial Intelligence*, 158:111630, 2025.