UNIVERSITY OF MILAN

MSc. Data Science and Economics

Course: Machine Learning

Classification Tree

An application on Mushroom Dataset

Professor: Nicolò CESA-BIANCHI

Student: Ly NGUYEN - 988858

Table of Content

1. Introduction

Decision trees are widely used algorithms in supervised machine learning, particularly effective for classification tasks due to their interpretability and ability to handle both categorical and numerical data. In binary classification, decision trees aim to split the dataset into two subsets that are increasingly pure with respect to the target classes. Each split is determined based on an impurity criterion, having applied in this assignment includes gini impurity, entropy, and misclassification rate—which guides the algorithm in selecting the most informative feature with its best value for splitting at each internal node

This assignment focuses on building from scratch binary classification decision trees and applying to the Mushroom dataset[1]. The dataset consists of 61.069 entries, each described by 17 categorical features and 3 numerical features. The target variable, namely class, indicates whether a mushroom is edible (e) or poisonous (p).

Objectives

- The primary objective of this study is to evaluate the effectiveness of decision tree classifiers in predicting mushroom edibility based on observable features.
- Investigate how different impurity criteria—Gini impurity, entropy, and misclassification rate—affect the performance of decision trees.
- Examine the impact of stopping criteria, maximum tree depth and minimum number of samples, per split.
- Identify the combination of splitting and stopping criteria that results in the most accurate and generalizable model.

By comparing different configurations, including hyperparameter tuning, this study seeks to shed light on how decision trees make classifications in a real-world dataset and which features are most predictive of mushroom toxicity.

[1] https://archive.ics.uci.edu/dataset/848/secondary+mushroom+dataset

2. Data Preprocessing

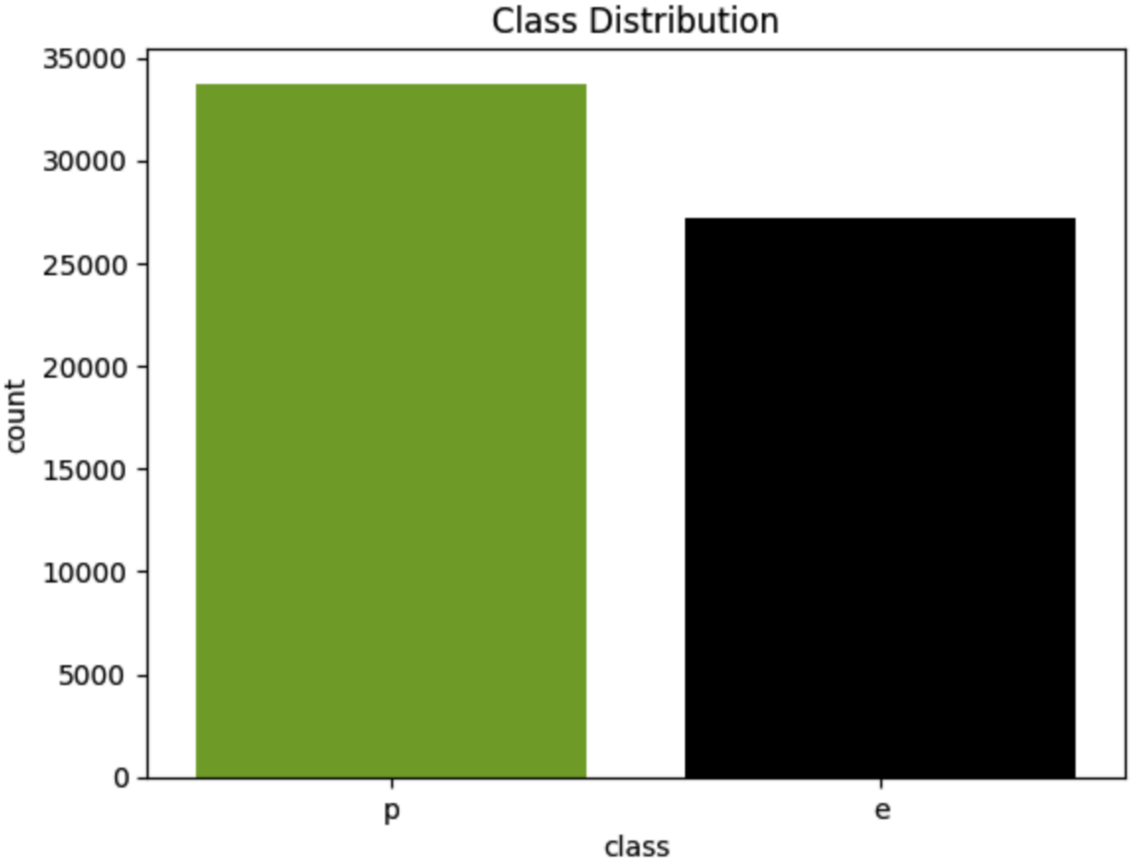
A thorough preliminary examination of the dataset involved assessing its dimensions, data types, and summary statistics, which confirmed the integrity and consistency of the data.

To facilitate interpretability, a dictionary mapping the encoded categorical values to their corresponding meanings was developed.

Missing data, identified as nan values, were replaced with the label “unidentified” to preserve the completeness of the dataset without introducing bias.

Cross-tabulations were employed to uncover underlying patterns and relationships among key features, particularly those known to be biologically relevant, such as odor and gill characteristics.

The distribution of classes was found to be relatively balanced, with approximately 55% of mushrooms labeled poisonous and 45% edible, which mitigates concerns regarding class imbalance in modeling.



Given the dataset's well-structured nature and comprehensive coverage, minimal feature engineering was necessary, allowing for a straightforward application of decision tree classification methods.

3. Methodology

3.1. Model Architecture

The classification model was implemented as a binary decision tree, constructed from the ground up to provide full control over its design and interpretability. Each tree node is represented by a Node class, which encapsulates key attributes such as the splitting feature, threshold value, information gain, and prediction label for terminal leaves. The main decision tree class, DecisionTree, manages the recursive construction, prediction, and evaluation of the model.

3.2. Splitting and Stopping Criteria

The decision tree supports three impurity measures as splitting criteria to evaluate the quality of candidate splits:

- Gini Impurity: Measures the likelihood of incorrect classification if a sample is randomly labeled according to the distribution of labels in the node.
- Entropy: Represents the Shannon entropy, quantifying the disorder or uncertainty in the node.
- Misclassification Error: Reflects the proportion of misclassified instances in the node, focusing on the most frequent class

Two stopping criteria are enforced during tree construction to prevent overfitting and ensure computational efficiency:

- Maximum Depth: The recursion terminates when the maximum allowed depth is reached.
- Minimum Samples per Node: Splitting ceases if a node contains fewer samples than the specified threshold.

3.3. Training Procedure

The training algorithm recursively partitions the dataset by selecting, at each node, the feature and corresponding threshold that maximize the information gain based on the chosen criterion. If no beneficial split is found or stopping criteria are met, the node becomes a leaf with a prediction assigned as either it is pure or the majority class of samples within that node.

The dataset was divided into three stratified subsets: training, validation, and test sets, maintaining the class distribution to ensure balanced representation.

3.4. Hyperparameter Tuning

Hyperparameter optimization was conducted specifically for the tree trained with the Gini criterion. A grid search was performed over a range of maximum depths and minimum samples per node using the validation set. The objective was to identify the parameter combination that maximized validation accuracy, thereby improving generalization performance.

3.5. Model Evaluation

The training and test errors are calculated using the 0-1 loss function, a fundamental and widely used metric in classification tasks. The 0-1 loss measures the proportion of misclassified samples by assigning a penalty of 1 for each incorrect prediction and 0 for correct predictions.

This measure intuitively reflects the model's misclassification rate, representing the fraction of incorrectly classified instances. It provides a straightforward and interpretable evaluation of model performance, helping to foresee underfitting / overfitting, complementing other metrics such as accuracy, precision, recall, and F1-score used in this study to comprehensively assess the binary classification trees.

4. Result and Conclusion

4.1. Result from general trees with max depth of five and min sample of forty.

	Gini	Entropy	Misclassification
Training Error	0,2490	0,2561	0,224
Testing Error	0,2492	0,2546	0,225
Accuracy	0,75	0,74	0,78
Precision	0,79	0,69	0,73
Recall	0,60	0,78	0,79
F1	0,68	0,73	0,76

In terms of training and testing error, the tree using the Misclassification criterion demonstrates the lowest error rates (22.4% for training and 22.5% for testing), indicating better generalization compared to Gini (24.9% / 24.9%) and Entropy (25.6% / 25.5%). This suggests that the misclassification-based tree was more effective at finding splits that minimize incorrect predictions on unseen data. Looking at accuracy, the Misclassification tree again leads with 78%, followed by Gini (75%) and Entropy (74%).

When examining precision and recall, we observe an interesting trade-off. The Entropy-based tree achieves the highest recall (78%), indicating better sensitivity to the positive class (‘edible’), but it does so at the expense of precision (69%), implying more false positives. Conversely, the Gini tree has the highest precision (79%) but the lowest recall (60%), suggesting it is more conservative, favoring fewer false positives at the cost of missing true positives. The Misclassification tree strikes a balanced performance, achieving both strong precision (73%) and high recall (79%), resulting in the highest F1-score (76%), which harmonizes both metrics.

Overall, the tree built using the Misclassification criterion offers the best balance between bias and variance, and demonstrates superior performance across most evaluation metrics. This highlights the potential of using misclassification as a splitting criterion when working with well-structured datasets like the mushroom dataset.

4.2 Result from tuning the gini tree

Max depth	Min sample	Accuracy
3	5	70,4%
3	60	70,4%
3	360	70,4%
7	5	81,8%
7	60	81,8%
7	360	81,8%
11	5	94,9%
11	60	94,9%
11	360	93,9%

The best-performing Gini-based decision tree, obtained through hyperparameter tuning, demonstrates strong predictive capability. The optimal configuration was identified as max_depth=11 and min_samples=5, which allowed the tree sufficient depth and flexibility to capture complex patterns in the training data without overfitting.

The tree achieves a very low training error of 5.5% and a similarly low test error of 5.3%, calculated using the 0–1 loss function. The close alignment between training and test errors indicates excellent generalization, suggesting that the model learned patterns that are robust and not overly tailored to the training data.

Classification Metrics

- Accuracy: At 94.67%, the accuracy is notably high, confirming the model's overall effectiveness in correctly classifying the majority of test instances.
- Precision: The precision of 98.19% for the positive class ('edible') reflects the model's high reliability in positive predictions, i.e., when it predicts a mushroom is edible, it is rarely wrong.

- Recall: The recall of 89.72% shows that the model successfully identifies a large proportion of edible mushrooms, though it still misses some.
- F1 Score: The F1 score of 93.76% reflects a strong balance between precision and recall, making the model both effective and safe—a particularly important consideration when the cost of false positives or false negatives is non-trivial (e.g., misclassifying poisonous mushrooms as edible).