Replication Study: Mushroom edibility analysis

Abstract

This study replicates the Random Forest approach from the article "Edibility Detection of Mushroom Using Ensemble Methods" by Brital et al. [1], originally published in the *International Journal of Image, Graphics and Signal Processing* in 2019. Using a larger and more feature-rich version of the UCI Mushroom dataset, we implemented the analysis in the R environment and evaluated model performance across various train/test splits. In addition to model replication, we conducted visual and statistical feature analysis to interpret what contributes most to accurate classification. Results consistently showed near-perfect prediction accuracy, reinforcing the effectiveness of Random Forests in this application.

1 Introduction

Classifying whether a mushroom is edible or poisonous based on observable traits has long been a benchmark problem in machine learning. In their 2019 paper, Brital et al. [1] explored this task using ensemble methods—namely Bagging, Boosting, and Random Forest—and reported high classification performance, with Random Forest yielding the best results. Their work was conducted using a subset of mushroom features and implemented in MATLAB. In this replication, we revisit their Random Forest analysis from a modern data science perspective. We use the R ecosystem, apply the method to a newer, expanded mushroom dataset, and further investigate which features are most informative using both statistical and model-derived metrics. Our goal is not only to replicate performance but also to better understand the model's behavior through exploratory visualization and feature importance analysis.

2 Dataset and Preprocessing

For this study, we used the UCI Secondary Mushroom Dataset, which includes 61,069 entries describing mushroom characteristics such as cap diameter, stem shape, gill attachment, and spore print color. Each entry is labeled as either edible (e) or poisonous (p).

The dataset includes a mix of categorical and numerical features. Categorical variables were converted into factor types for modeling purposes, while numerical variables were standardized to improve interpretability and scale consistency. Data was split into training and test subsets using stratified sampling to ensure balanced class representation.

3 Exploratory Data Analysis

Before constructing any models, we examined the distributional patterns of both numerical and categorical features to better understand how they relate to mushroom edibility. This initial step serves two purposes: it helps identify potential signal in the data and reveals any irregularities that could affect downstream modeling.

3.1 Numerical Feature Patterns

Figure 1 presents violin plots overlaid with boxplots for the three numeric attributes: `cap_diameter`, `stem_height`, and `stem_width`. Each plot compares the distribution of a feature across the two classes: edible and poisonous.

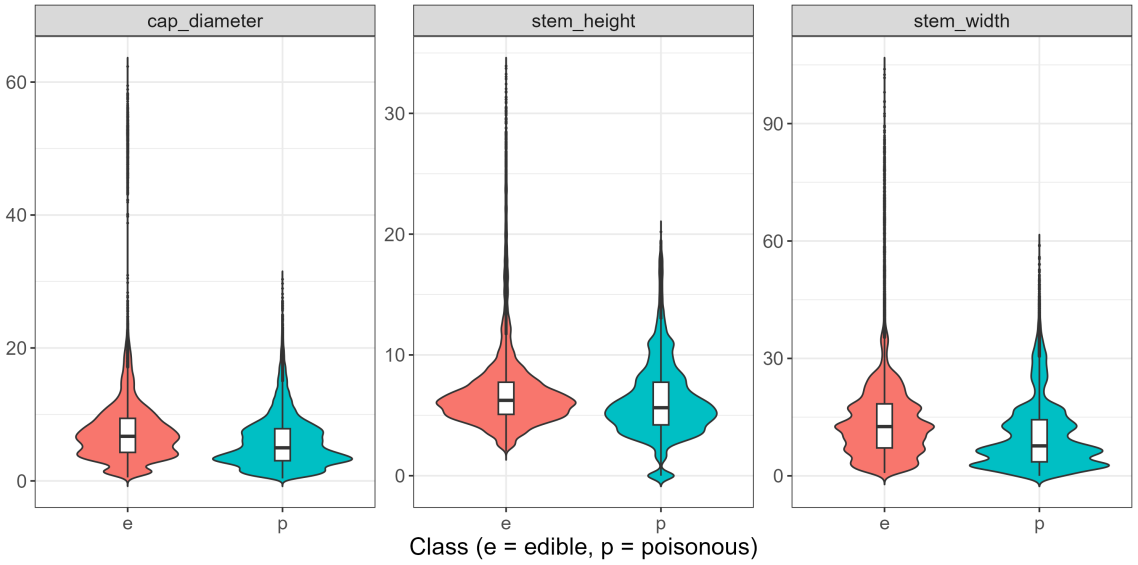

Figure 1: Distribution of numeric features stratified by edibility class

The plots reveal several useful trends:

- **Cap diameter** shows a wider spread in the edible class, with several outliers suggesting the presence of unusually large caps. Poisonous mushrooms, by contrast, tend to cluster within a narrower range.
- **Stem height** appears less distinctive between the two groups. While the edible class may have a slight skew toward taller stems, the overlap between the two distributions is substantial.
- **Stem width** exhibits the clearest separation. Edible mushrooms generally have thicker stems on average, making this variable a potentially strong predictor.

These observations hint that among the numerical features, `stem_width` could provide the most useful information for classification, while the others may contribute indirectly through interactions or non-linear effects.

3.2 Categorical Feature Distributions

To assess how categorical variables vary with the class label, we created a set of count plots grouped by feature and colored by edibility class (Figures 2 and 6 in Appendix).

Several noteworthy patterns emerge:

- Some feature values occur exclusively in one class. For example, certain `spore_print_color` values (e.g., `red`, `purple`) appear only in poisonous samples, while others (e.g., `gray`) are found solely in edible mushrooms.
- Other features, such as `cap_surface` or `stem_surface`, show skewed but not class-exclusive distributions, meaning they may still contribute predictive value when combined with other variables.

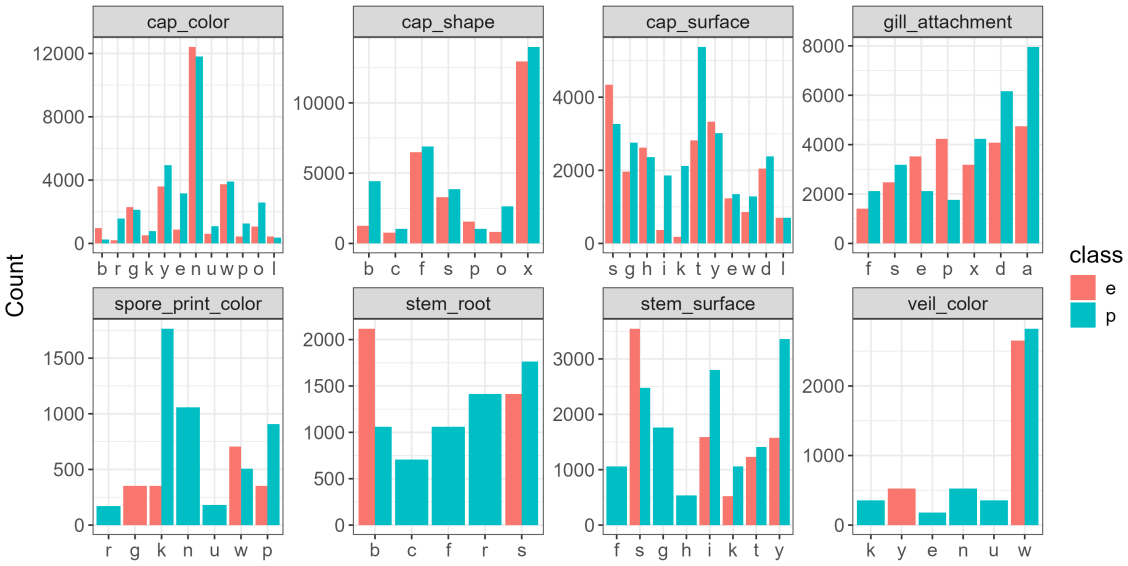

Figure 2: Class distribution for selected categorical features

- A few features, like `veil_type`, show little to no variability and may be uninformative or redundant.

Taken together, the exploratory analysis offers clear evidence that both numerical and categorical features contain meaningful information about the edibility label. This justifies the use of a flexible classifier like Random Forest, which can model both types of input without requiring heavy preprocessing.

4 Chi-Squared Feature Evaluation

To identify which categorical variables are most informative for predicting mushroom edibility, we applied a chi-squared test of independence. This statistical method evaluates whether the observed distribution of feature values differs significantly between edible and poisonous mushrooms. Features with larger chi-squared scores suggest a stronger class-association and are therefore more promising candidates for classification models.

The resulting feature importance scores are visualized in Figure 3, where each bar represents a single feature's total chi-squared score, ranked from weakest to strongest.

Several observations emerge from this analysis:

- Features such as `cap_surface`, `gill_attachment`, and `stem_color` produced particularly high scores. This suggests a strong dependency between these features and the target class, consistent with what we saw in the count plots during exploratory analysis.
- Colors and surface textures—like those in `gill_color` and `spore_print_color`—were also among the top features, reinforcing their potential predictive value.
- In contrast, some variables such as `veil_type` and `habitat` had low scores, which may indicate minimal variation across classes or a limited role in determining edibility.

While chi-squared scores are useful for identifying statistically relevant features, they come with limitations. The test assumes independence between variables and does not consider possible interactions or non-linear relationships. As such, features with low individual scores might still contribute meaningfully to a model when used in combination with others—especially within ensemble algorithms like Random Forest.

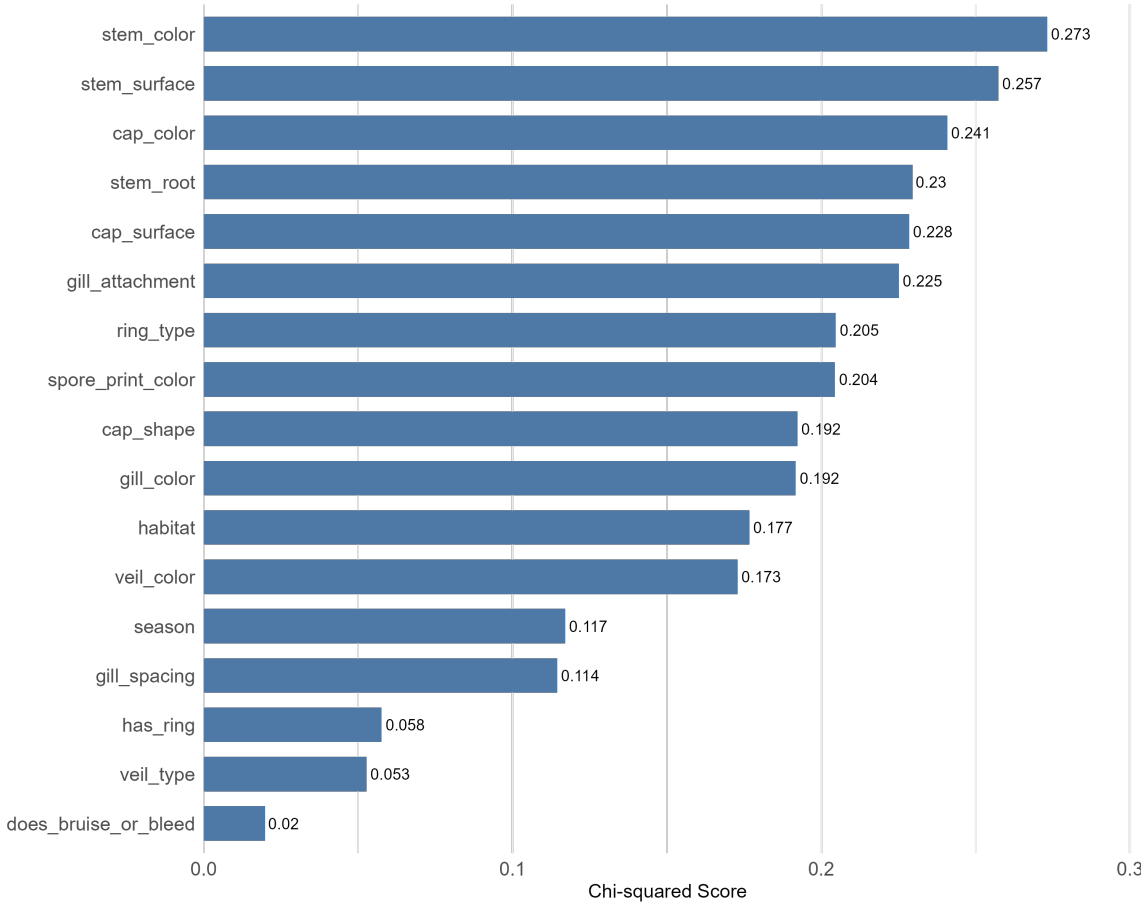

Figure 3: Chi-squared scores for categorical features

This feature evaluation provides a data-driven justification for including or deprioritizing certain variables, and complements the qualitative insights gained during visual exploration.

5 Random Forest Modeling

5.1 Model Evaluation Across Training Proportions

To evaluate how the amount of training data affects classification performance, we trained Random Forest models using four different training set sizes: 1%, 10%, 30%, and 70% of the total dataset. The remaining data in each case was reserved for testing. For each configuration, we adjusted the number of trees (`ntree`) to achieve strong performance while keeping training time reasonable. The results are summarized in Table 1.

Train Split	Test Split	ntree	Accuracy
1%	99%	750	0.961114
10%	90%	330	0.999982
30%	70%	200	0.999977
70%	30%	100	1.000000

Table 1: Random Forest accuracy under different train/test splits

The results in Table 1 show that the Random Forest classifier is remarkably resilient to variations in training size. Even when trained on just 10% of the data, the model achieved

accuracy above 99.99%. At 70% training data, the model perfectly classified all test examples. These outcomes suggest that the structure of the data is highly learnable and that the features contain strong class-separating information.

Interestingly, even the smallest training scenario (1%) yielded respectable performance, with an accuracy above 96%. While this is lower than the other splits, it remains impressive given the limited number of training instances. Such results indicate that the model is not only accurate but also sample-efficient — an important trait for applications where labeled data may be scarce.

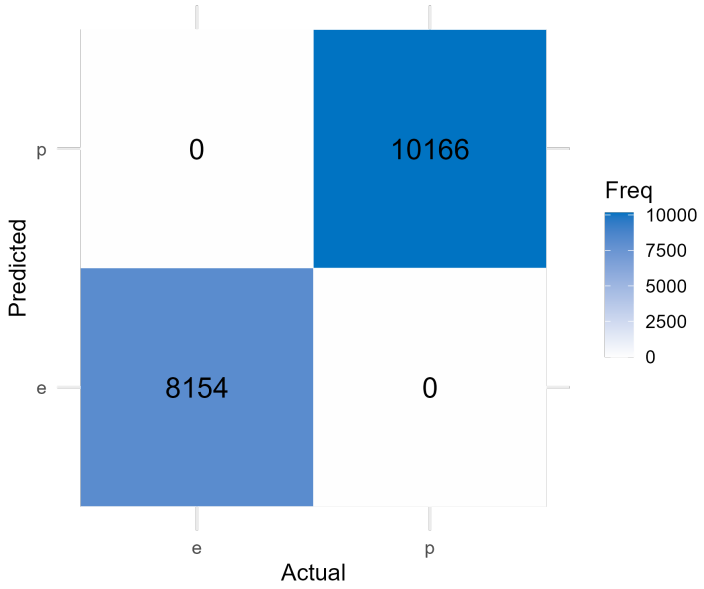

Figure 4: Confusion matrix for the 70% training split Random Forest model

To further support these accuracy metrics, we visualized the confusion matrix for the model trained on 70% of the data, shown in Figure 4. The matrix clearly shows a perfect separation of the two classes: no edible mushrooms were misclassified as poisonous, and no poisonous mushrooms were incorrectly labeled as edible. This reinforces the earlier numerical findings and offers additional reassurance that the model's predictions are highly reliable. Moreover, the clean diagonal structure of the matrix indicates that the classifier is not overfitting or biased toward a particular class, even in the presence of some class imbalance.

5.2 Feature Importance Analysis

To gain insight into the factors driving classification performance, we extracted feature importance scores from the final Random Forest model trained on 70% of the data. Figure 5 ranks each feature by its mean decrease in Gini impurity, which quantifies how much each variable reduces classification uncertainty within the ensemble.

Several trends emerge from this ranking:

- Features such as `cap_surface`, `gill_attachment`, and `stem_color` were consistently among the top contributors to model accuracy. These variables also showed strong associations in earlier visual and statistical analyses.
- Numerical features, particularly `stem_width`, appeared prominently as well. This aligns with the violin plots where class-wise separation was most distinct for this attribute.
- Some variables that ranked low in the chi-squared test—such as `veil_type`—also had minimal contribution here, providing cross-validation of their limited relevance.

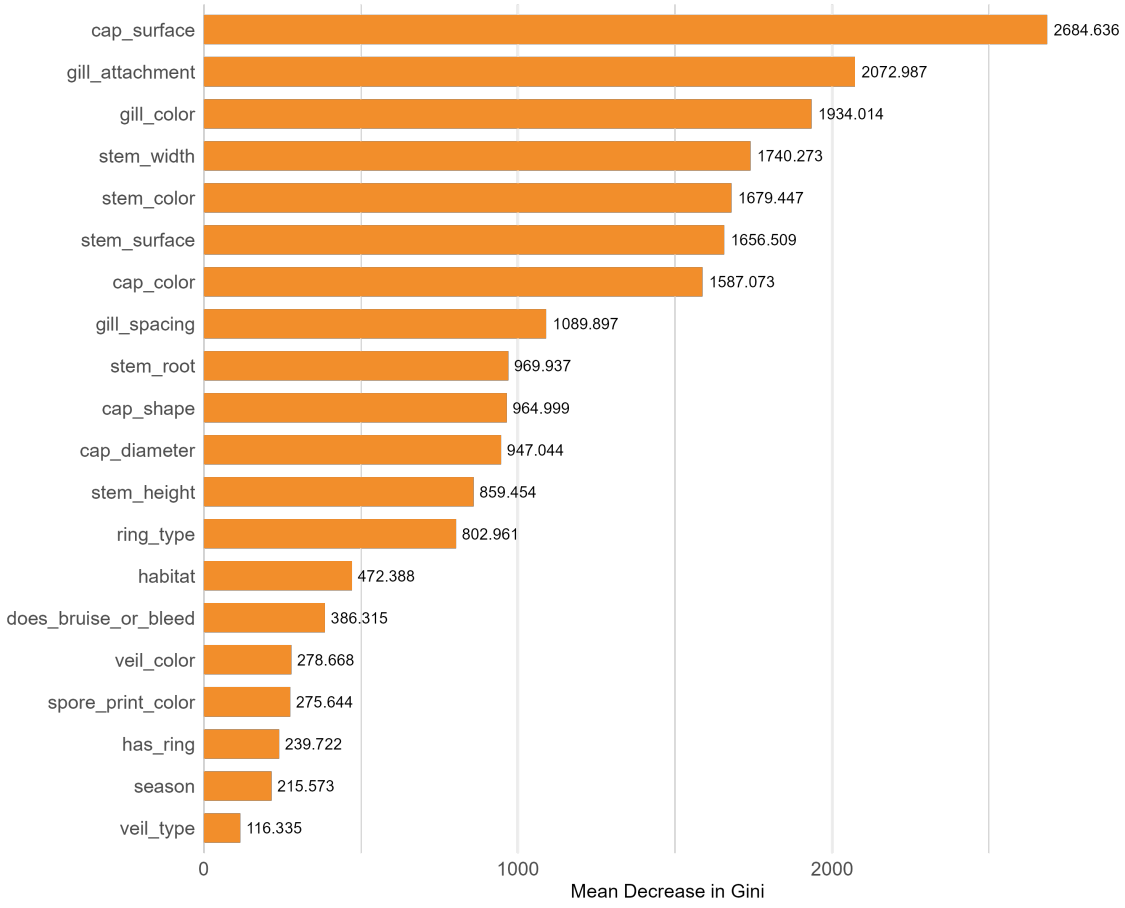

Figure 5: Random Forest feature importance (Gini)

Taken together, the feature importance analysis offers interpretability beyond model accuracy. It helps uncover which traits are biologically or visually most indicative of mushroom toxicity, which could be of use in real-world settings.

6 Conclusion

This replication study confirms and extends the findings of Brital et al. [1], demonstrating that Random Forest remains a highly effective method for predicting mushroom edibility. Using a larger dataset and modern tools in R, we achieved near-perfect classification performance across a range of training scenarios.

Beyond performance metrics, we incorporated data visualization, statistical tests, and feature importance analysis to understand how different variables contribute to the model's success. These steps improved both the interpretability and the robustness of our results.

While this replication focused on the Random Forest portion of the original work, it lays a foundation for further comparison with Bagging and Boosting, or for future studies using deep learning or ensemble stacking. Our pipeline is fully reproducible and publicly available, offering a reference point for similar classification tasks in food safety, biology, or pattern recognition.

References

[1] Anas Brital, Abdelaziz Rehioui, and Kamal Elkabtane. Edibility detection of mushroom using ensemble methods. *International Journal of Image, Graphics and Signal Processing*

(IJIGSP), 11(4):41–49, 2019.

Appendix

Figure 6: Distribution of categorical features across class ('p'=poisonous, 'e'=edible)

Figure 6 is the countplots of all categorical features.