Poisonous vs Edible Mushroom Classification With Gradient Boosting

Michael Evans
Department of Computer Science
Old Dominion University
Norfolk, VA 23529
mevan028@odu.edu

Grant Fitch
Department of Computer Science
Old Dominion University
Norfolk, VA 23529
gfitc002@odu.edu

Abstract

Mushroom poisoning is an increasingly common form of toxin-induced-disease and accounts for up to 100 deaths per year in Western Europe alone. Dadpour et al. [2017] Furthermore, mushrooms previously classified as safe to eat are being re-evaluated due to the recent discovery of new syndromes linked to their consumption. White et al. [2019] This increase in mushroom-related deaths could be attributed to the skill required for manual analysis, and demands an automated and robust method for classification. In this paper, we propose a gradient boosting classifier method for binary classification with k-fold cross validation. The dataset used in this study is available in the UC Irvine Machine Learning Repository and contains 173 species of mushroom. Wagner and Hattab [2021] Our classification model achieves an accuracy, precision, recall, and F1 score of 0.99, which shows immense potential for decreasing the annual death rates attributed to incorrect classification.

1 Introduction

Manually classifying mushrooms as poisonous or edible is an ambiguous process that is further complicated by the varying properties among same species samples, leading to differing case reports and uncertainty. Li et al. [2021] Current methods of mushroom classification rely largely on visual identification and biochemical analysis, something not readily available to many foragers. Tutuncu et al. [2022] Previous machine learning methods for this task have been proposed in Zahan et al. [2021], and use the deep convolutional neural network VGG16 for image classification. However, this image based form of classification underperforms our method of numerical features. In the following section, we begin by cleaning the raw dataset and performing encoding on the categorical features. Then, we describe training the gradient boosting classifier (GBC) and hyperparameter tuning. Lastly, we outline our experiments and quantify the performance and features scores of our trained model.

2 Dataset

The raw dataset used for this project was sourced from the UC Irvine Machine Learning Repository. The original dataset was created in 1987 and contained 8,124 samples of 23 species of mushrooms. In 2023, this dataset was expanded to include 61,069 samples from 173 different mushroom species, each characterized by 20 features and a target classification of poisonous or edible. The features focus on specific aspects of mushroom anatomy, such as the measurements, colors, and textures of the mushroom cap, gills, and stem, as well as the veil type and root structure. Additional environmental details, including the habitat and season, were also recorded. Upon initial analysis, the dataset was found to be fairly well-balanced, as shown in Figure 1, with 55.5% of the samples classified as poisonous and 44.5% classified as edible.

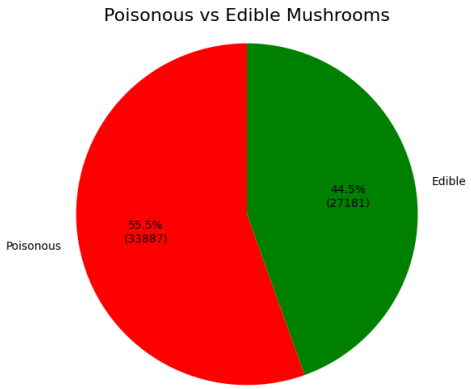

Figure 1: Target Class Distribution of Raw Dataset

2.1 Data Preprocessing

Missing Data Initial exploration revealed that 9 features had missing data, these features can be seen in Figure 2 and observed that several features are missing in over 80% of the samples. We set a threshold of 20% missing data and removed any features exceeding this threshold. This resulted in the removal of several features, including the root structure, veli type, veil color, stem surface, and gill spacing. After dropping these features, the dataset contained 61,069 samples across 13 features and the target class.

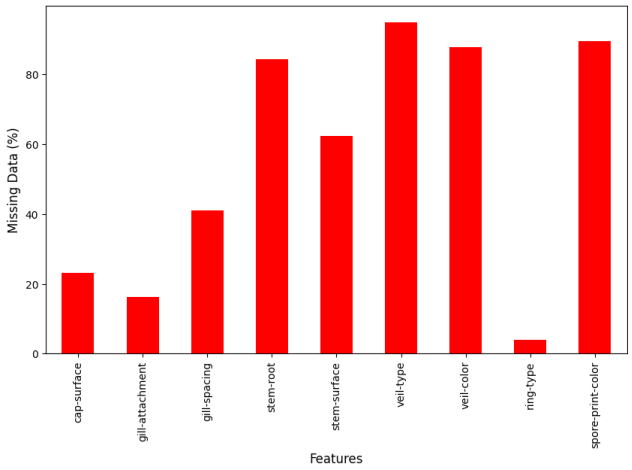

Figure 2: Features With Missing Data

Missing Samples Further examination revealed that 12,002 samples had missing values in one or more features. Several of these samples lacked data for multiple features. Given the large initial dataset, it was decided that dropping these samples would not significantly reduce the robustness of the dataset. As a result, these samples were removed, leaving 49,067 samples with complete data across the 13 remaining features.

2.2 Data Encoding and Transformation

Boolean Features The target class of poisonous or edible and the Boolean features, including if the mushroom has a veil ring or if it bruises were encoded using binary values: 1 for poisonous or true, and 0 for edible or false.

Categorical Features Categorical features, such as cap shape, gill attachment, habitat, season, and colors of the cap, stem, and gills, were target encoded. Target encoding replaces categorical values with the mean of the target variable. For this dataset, that is the probability of the mushroom being poisonous. Figure 3 shows the final encoding for the season feature, each season category was replaced with the likelihood of the mushroom being poisonous for that season. This encoding method provides additional insight into the dataset, such as the observation that mushrooms found in the summer are more likely to be poisonous with just under 60% probability of being poisonous, especially when compared to those found in the winter having roughly a 35% probability of being poisonous.

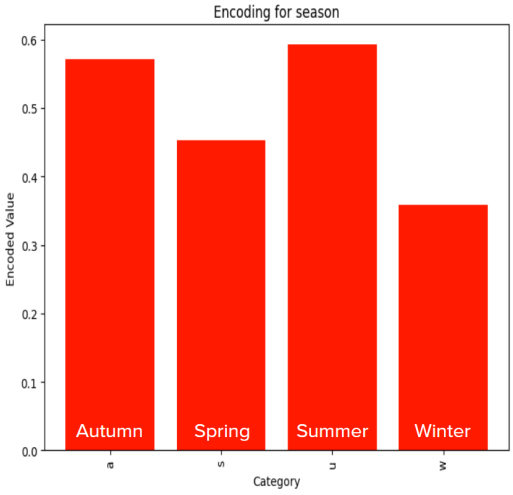

Figure 3: Target Encoding for Season

Numerical Features Numerical features like stem height and width were regularized for consistency and better model performance. To handle outliers, we calculated z-scores for these features and identified outliers with values exceeding a threshold of 3. This led to the removal of 2,016 outliers and can be visualized in Figure 4. This figure shows the extreme outliers for cap diameter and stem width, once these outliers are removed, the data still covers a wide range, but the extreme outliers seen in the raw data have been removed, resulting in a much cleaner and tighter final data set for these values.

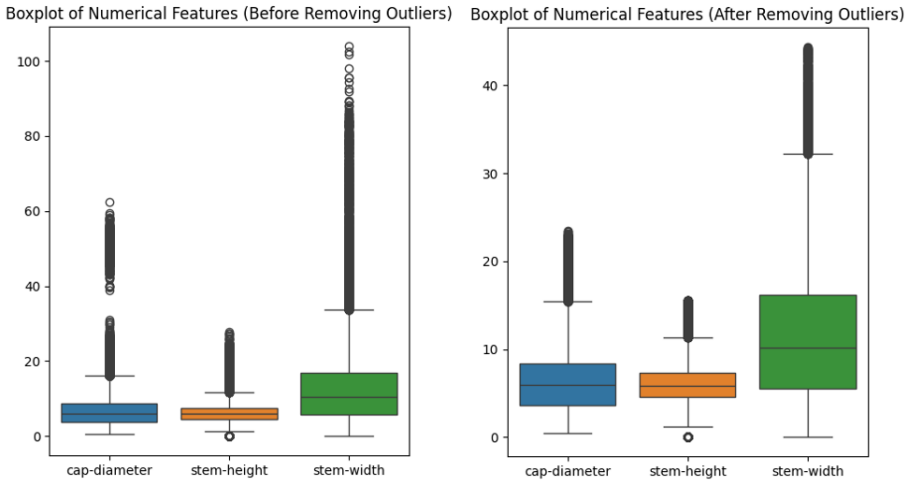

Figure 4: Numerical Features Before and After Preprocessing

After handling outliers, the numerical features were standardized using StandardScaler, ensuring that each feature had a mean of 0 and a standard deviation of 1. The scaling parameters of the sample mean and standard deviation were then saved in an encoding dictionary for potential future use.

2.3 Final Dataset

After preprocessing and encoding, the final dataset consists of 47,051 samples, each with complete data across the 13 features and target class. The final balance can be seen in Figure 5 with 55.8% of the samples classified as poisonous and 44.2% being classified as edible. This final dataset maintains the balance of the original raw dataset and is ready to be used in the training and testing of our model. The encoding dictionary that was created during this process can later be used to encode new data for classification purposes, ensuring the accuracy of the model.

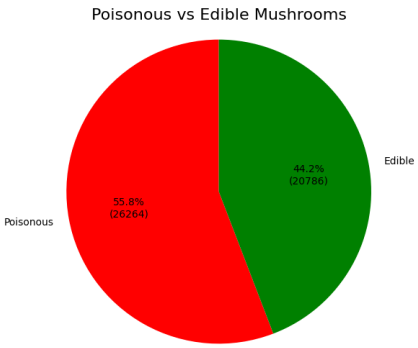

Figure 5: Target Class Distribution of Cleaned Dataset

3 Methods and Experiments

3.1 Importing and Splitting Data

We implement our proposed solution with the scikit-learn GBC Python module, Jupyter Notebook, pandas, numpy, and matplotlib for data visualization. Following the data preprocessing described in the previous section, the cleaned mushroom CSV data is loaded into a dataframe. From this dataframe, we create a matrix, X, containing the data columns, and a column vector, y, to store the labels. From here, we create an 80/20 train/test split using sklearn.model_selection.train_test_split.

3.2 Hyperparameter Tuning and Cross-Validation

The GBC hyperparameter, n_estimators, adjusts the number of boosting stages to perform during training. We test n_estimators values ranging from 10 to 1500 during our cross-validation training loop. As gradient boosting is resistant to overfitting, we find that larger values of n_estimators result in better testing performance. To provide a more robust and reliable estimate of the model's performance, we test each n_estimators value with 5-fold cross validation. The mean accuracy for each hyperparameter value is recorded, shown below in Figure 6, and the highest is selected for use.

3.3 Model Evaluation

A primary motivator for choosing the GBC model for this classification problem is its resilience against over-fitting and ability to handle non-linear relationships. With the best n_estimator value selected from the previous step, we fit the model with this parameter and evaluate its performance. We apply the following metrics from the sklearn.metrics module to evaluate the model's performance: precision_score, recall_score, f1_score, and accuracy_score. We are also interested in the feature importances of the model, which will be discussed in the following section.

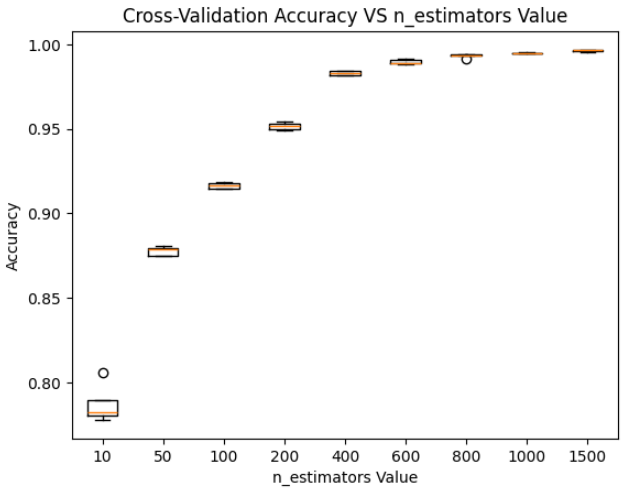

Figure 6: Hyperparameter Tuning: Cross-Validation Accuracy vs n_estimators Value

4 Results

The evaluation metrics provided by sklearn indicate that our GBC achieved optimal performance, with precision, recall, F1-score and accuracy all scoring over 0.99 across the testing dataset as seen in Table 1. These results highlight the effectiveness of the model for classifying edible and poisonous mushrooms.

Class	Precision	Recall	F1-Score	Accuracy	Support
Edible	0.9964	0.9959	0.9961	-	4128
Poisonous	0.9968	0.9972	0.9966	-	5283
Overall	0.9966	0.9966	0.9966	0.9966	9411

Table 1: Evaluation Metrics.

To further analyze the model's predictions, a confusion matrix was generated and can be seen in Figure 7. This matrix shows a significantly low error rate with only 32 samples being mislabeled overall.

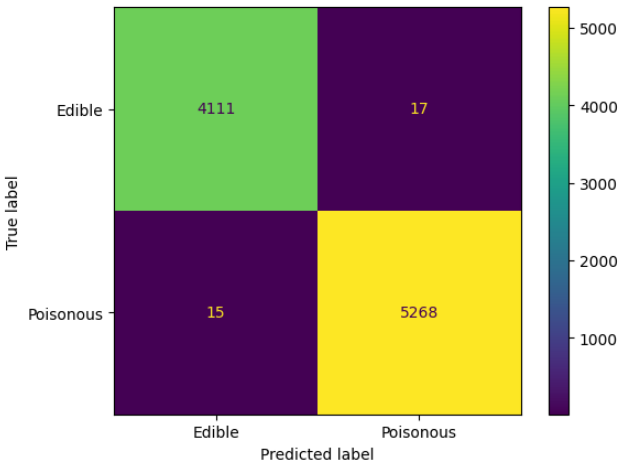

Figure 7: Confusion Matrix for GBC Model

Feature importances were also extracted from the model using sklearn (Figure 8) and revealed that attributes such as stem width and color, as well as gill attachment and color, are the most influential predictors. These features suggest that broader physical traits are key to distinguishing mushroom categories. On the other end, attributes like habitat and season, while less influential, may play a role in refining classifications. Features such as cap color and shape, which are closely ranked, may also exhibit correlations that enhance the model's predictive accuracy.

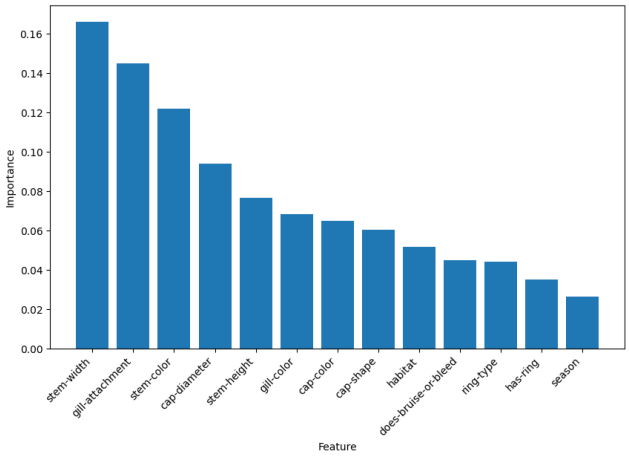

Figure 8: Confusion Matrix for GBC Model

5 Conclusion

To combat the recent increase in toxin-induced-disease attributed to consuming poisonous mushrooms, we propose a robust and generalizable solution to mushroom classification. In this paper, we trained a gradient boosting classifier on 47,051 mushroom samples with 13 features, and achieved a score of 0.99 on the evaluation metrics of accuracy, precision, recall, and F1-score. With contemporary machine learning methods for this task under performing on image data, we propose a numerical method for training and prediction. As an extension to this work, we could compare the performance of GBC against other common machine learning methods such as Naive Bayes, or SVM.

References

Bita Dadpour, Shahrad Tajoddini, Maliheh Rajabi, and Reza Afshari. Mushroom poisoning in the northeast of iran; a retrospective 6-year epidemiologic study. *Emergency*, 5(1), 2017.

Huili Li, Yang Tian, Nelson Menolli Jr, Lei Ye, Samantha C Karunarathna, Jesus Perez-Moreno, Mohammad Mahmudur Rahman, Md Harunur Rashid, Pheng Phengsintham, Leela Rizal, et al. Reviewing the world's edible mushroom species: A new evidence-based classification system. *Comprehensive reviews in food science and food safety*, 20(2):1982–2014, 2021.

Kemal Tutuncu, Ilkay Cinar, Ramazan Kursun, and Murat Koklu. Edible and poisonous mushrooms classification by machine learning algorithms. In *2022 11th Mediterranean Conference on Embedded Computing (MECO)*, pages 1–4. IEEE, 2022.

Heider D. Wagner, Dennis and Georges Hattab. Secondary Mushroom. UCI Machine Learning Repository, 2021. DOI: https://doi.org/10.24432/C5FP5Q.

Julian White, Scott A Weinstein, Luc De Haro, Regis Bédry, Andreas Schaper, Barry H Rumack, and Thomas Zilker. Mushroom poisoning: A proposed new clinical classification. *Toxicon*, 157:53–65, 2019.

Nusrat Zahan, Md Zahid Hasan, Md Abdul Malek, and Sanjida Sultana Reya. A deep learning-based approach for edible, inedible and poisonous mushroom classification. In *2021 international conference on information and communication technology for sustainable development (ICICT4SD)*, pages 440–444. IEEE, 2021.