Mushrooms - Edible or Poisonous?

Project Report

Anna Bicelli, Alicja Dorobis, and Maximilian Lemberg
Universitat Autònoma de Barcelona (UAB)
Barcelona, Spain

December 16, 2024

Instructors: Dimosthenis Karatzas & Oguz Mulayim
Course Code: MO71635
Semester: Fall 2024

Contents

1 Introduction

This project aimed to classify mushrooms as either edible or poisonous using a simulated dataset. We conducted extensive data pre-processing, including handling missing values and scaling features, followed by detailed data analysis involving correlation checks, feature importance assessments, and dimensionality reduction.

The dataset contains characteristics such as cap shape, stem width, and gill colour, making this a binary classification task.

We applied various machine learning techniques, including logistic regression, decision trees, and support vector machines, comparing their performance to identify the most accurate and interpretable model. However, in this report, we focused mainly on logistic regression and random forest classifiers.

This report discusses the pre-processing steps, data analysis, model selection, and results.

2 Data Pre-processing

The dataset was loaded using `read_csv`, specifying the correct delimiter `;`. The dataset contains 61,069 entries and 21 features, ensuring statistical reliability. Initial inspections using `.head()`, `.shape`, and `.info()` revealed some inconsistencies, so we performed several cleaning steps.

We renamed columns by replacing hyphens with underscores for easier use. Missing values, which were not present in numerical features, were filled with the mode for the majority of categorical columns. Five variables with over 50% missing data were removed.

Duplicates were addressed for both, the entire data set and within individual columns. To handle categorical features, we applied one-hot encoding using `OneHotEncoder` from `sklearn`, converting them into binary columns to allow the model to treat them as numeric inputs.

Furthermore, the target variable `class` was label-encoded, assigning 1 to poisonous mushrooms and 0 to edible ones, making it suitable for machine learning algorithms.

The dataset was split into training, validation, and test sets to ensure robust evaluation and prevent data leakage. Since the dataset is not perfectly balanced (45/55), we stratified the target variable y when splitting into test and train sets. The training set was used for model training and hyperparameter tuning, the validation set for model selection, and the test set for assessing final performance on unseen data.

We analyzed the numerical features (`cap_diameter`, `stem_height`, and `stem_width`) and identified 2,262 outliers using Z-Score statistics. Instead of using the Standard Scaler, which centres the data to a mean of 0 and a variance of 1 but is sensitive to outliers, we opted for the Robust Scaler.

The Robust Scaler scales the data using the median and interquartile range, making it resistant to the influence of outliers [1]. This approach allowed us to retain the potential value that outliers might bring to our model while significantly reducing their impact. By compressing the scale of extreme values, the Robust Scaler prevents outliers from dominating the feature

distributions and influencing the model disproportionately.

We also created new features to capture complex relationships:

- **Mushroom Size:** Defined as `cap_diameter` × `stem_height`.
- **Stem Size:** Combined stem dimensions into a single feature.

3 Data Analysis

We proceeded with the data analysis part of the project, where we first checked the class distribution to see if the dataset was balanced. Poisonous mushrooms made up 55.5% of the dataset, while edible mushrooms constituted 44.5%, indicating a fairly balanced dataset, reducing the risk of bias in model training.

We identified features with low variance as they provide minimal predictive information. Columns with variances below a threshold, in our case 0.01, such as `ring_type_m` and `stem_color_b`, were flagged for potential removal.

Then, we computed a correlation matrix to explore relationships between features and visualized it using a heatmap. No single feature strongly correlated with the target variable `class`, but some features were highly related. For example, `has_ring_f` and `has_ring_t` were perfectly negatively correlated, indicating redundancy, so we dropped one of them.

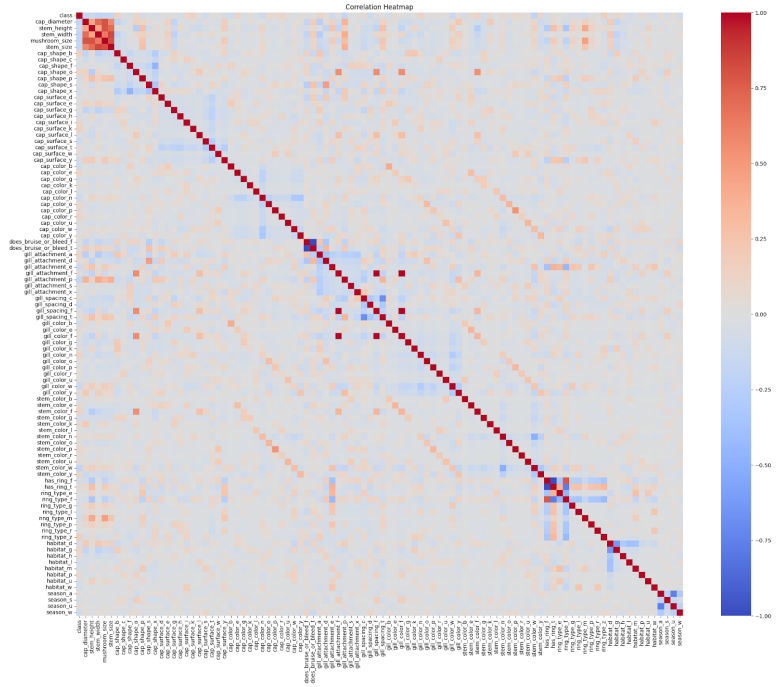

Figure 1: Coefficient matrix highlighting relationships between features.

We extracted and examined highly correlated feature pairs, removing redundant features like `gill_color_f`, `gill_spacing_f`, and `gill_attachment_f`, which were perfectly correlated, to avoid multicollinearity.

We performed PCA on the training set. The first principal component explained 37.4% of the variance and was strongly influenced by size-related features such as `mushroom_size`, `stem_size`, `stem_height`, `cap_diameter`, and `stem_width`.

Furthermore, the features `stem_height`, `cap_diameter`, `stem_width`, `stem_size`, and `mushroom_size` are some of the most important according to the Random Forest model. This shows they each capture valuable and unique information for the classification task. As `stem_size` and `mushroom_size` are created from `stem_height`, `cap_diameter`, and `stem_width`, their similarity shown in the PCA comes from their design, not redundancy. Therefore, combining these features into a single measure could reduce the unique information they provide. We chose not to combine these features to keep the model easy to interpret and maintain its performance, even though they have similar contributions in the first principal component.

We also visualized the feature distributions. Edible mushrooms tended to have larger cap diameters and taller stems, while poisonous mushrooms often had thicker stems and smaller overall sizes. Therefore, we decided to go back and add the features `mushroom_size` and `stem_size` to capture the non-linear relationship.

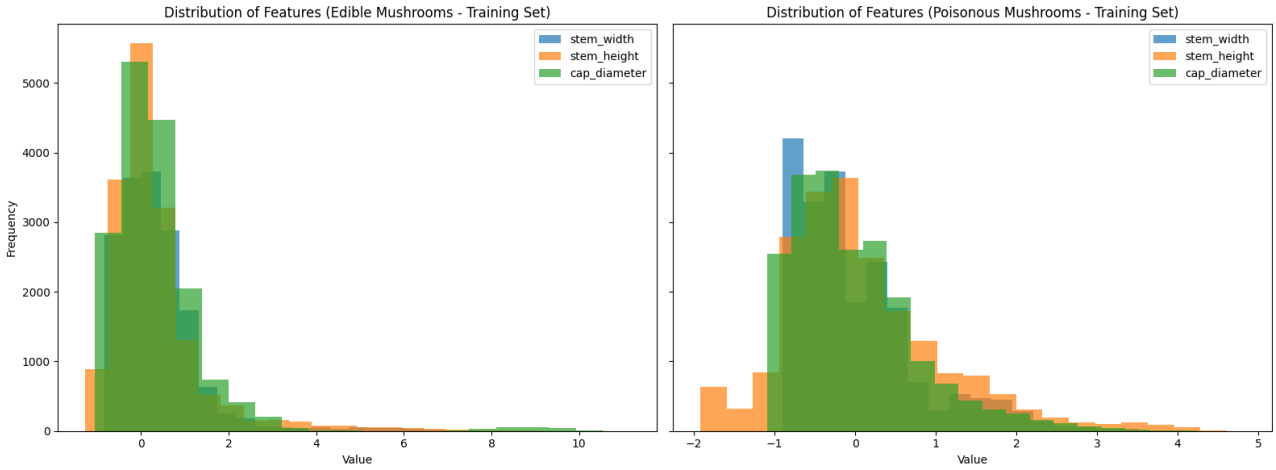

Figure 2: Real-world mushroom appearance visualized

4 Models

4.1 Logistic Regression

We optimized the logistic regression model using `GridSearchCV` to tune parameters such as C, penalty (L1), and solver. The model was evaluated using accuracy, precision, recall, F1 score, and ROC-AUC, providing a comprehensive performance assessment.

The best model applied L1 regularization, shrinking less important coefficients to zero. Notably, engineered features like `mushroom_size` and `stem_size` were retained, indicating their distinct contribution even in the presence of correlated features (e.g., `cap_diameter`, `stem_height`).

Performance Metrics

Simplest Model (Validation Set):

- Accuracy: 78.14%
- F1 Score: 80.04%
- ROC-AUC: 85.96%

Best Model (Test Set):

- Accuracy: 78.25%
- F1 Score: 80.53%
- ROC-AUC: 85.88%

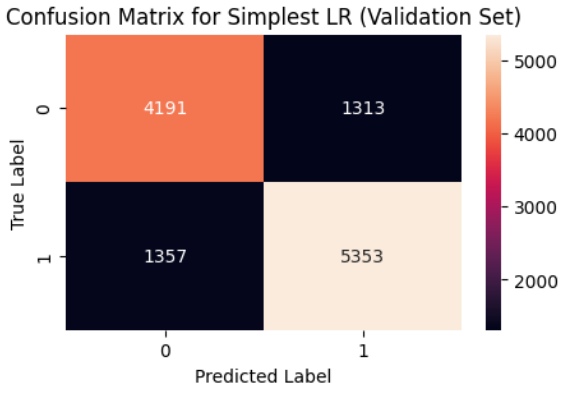

Figure 3: LR Confusion Matrix for Simplest Model (Validation Set)

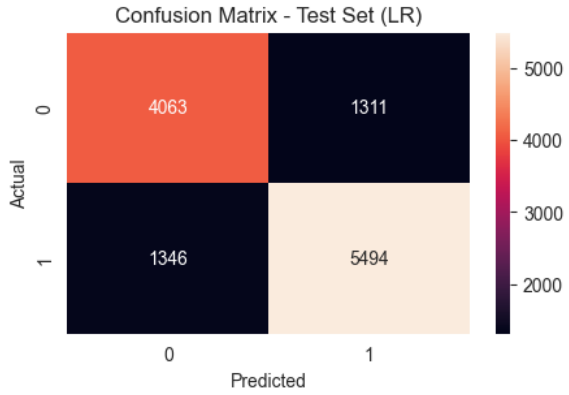

Figure 4: LR Confusion Matrix (Test Set)

Overall, the regularized logistic regression model provided a strong baseline, showing consistent performance on both validation and test sets, with notably high F1 and ROC-AUC scores. However, the overall improvements from further data pre-processing, tuning, and regularization were modest, suggesting that exploring more complex models or accepting dataset limitations may be necessary.

4.2 Random Forest

We next employed a Random Forest (RF) classifier, where `GridSearchCV` and 5-fold cross-validation guided our hyperparameter tuning, including the number of trees, depth, and split criteria. These steps ensured a fair and robust evaluation, minimizing the risk of overfitting and providing a more reliable assessment of the model's true performance.

Our final RF model reached 98% accuracy, with strong F1 and ROC-AUC scores on both validation and test sets. Notably, these metrics remained consistent across multiple splits and folds, indicating the model's ability to generalize well.

Performance Metrics

Validation Set:

- Accuracy: 98.0%
- F1 Score: 98.25%
- ROC-AUC: 99.84%

Test Set:

- Accuracy: 98.0%
- F1 Score: 98.0%
- ROC-AUC: 99.86%

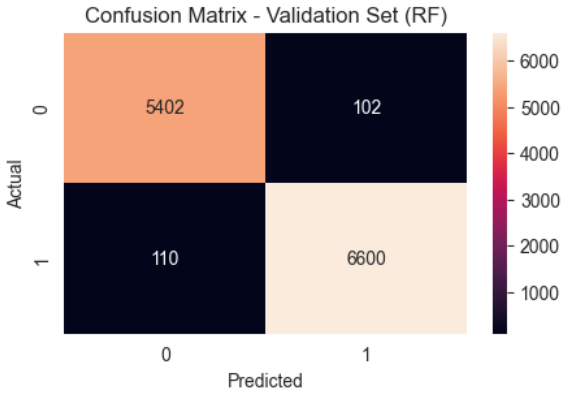

Figure 5: RF Confusion Matrix (Validation Set)

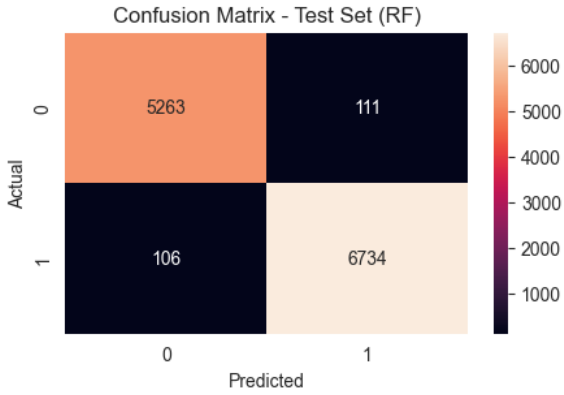

Figure 6: RF Confusion Matrix (Test Set)

These consistent, high-quality results demonstrate that our Random Forest model not only surpasses the logistic regression baseline but also maintains its strong performance without overfitting, confirming the value of ensemble methods for this project.

5 Conclusion

This project aimed to classify mushrooms as edible or poisonous using a combination of data cleaning, preprocessing, and feature engineering. Our initial logistic regression model established a solid baseline, but improvements through tuning and regularization were modest.

By introducing a Random Forest classifier, we captured complex feature interactions and achieved higher accuracy, F1, and ROC-AUC scores. Cross-validation confirmed that the Random Forest did not overfit and generalized well to unseen data.

These results highlight that tree-based ensemble methods were well-suited for this task, providing both, strong predictive performance and interpretable feature importance. Overall, the project successfully identified an effective model and delivered reliable results.

References

[1] Andreas C. Müller and Sarah Guido. *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media, Inc., 2016.