Mushroom Classification Guide for Vegetarians

Introduction

According to Forage Hyperfoods (Staicu, 2023), mushroom foraging is popular among vegetarians in Canada, offering a sustainable source of nutritious food. However, identifying whether a mushroom is poisonous, or edible can be a life-or-death decision. To support vegetarian foragers, we aim to create a reliable guide that uses data-driven insights to classify mushrooms based on key features.

This analysis utilizes the mushroom_cleaned dataset from Kaggle to identify the critical variables that determine mushroom edibility. By utilizing predictive models such as decision trees, logistic regression, and neural networks built in Python, we will pinpoint the significant features that distinguish poisonous mushrooms from edible ones.

The results of this study will equip vegetarians with a scientifically grounded tool to make safer foraging decisions.

Data Selection and Target Variable

Dataset: The dataset was sourced from Kaggle and contains 9 variables that explain mushroom features such as cap diameter, cap shape, class, gill attachment, gill color, season, stem height, stem width, and stem color.

Target Variable: We identified our target variable as the mushroom class which was made up of a binary set of data that denotes whether a mushroom is poisonous (0) or edible (1).

Rejected Variables: All the variables in our dataset were important and therefore we did not need to reject any variable at the beginning of the analysis.

Modelling

This guide provides vegetarians with data-driven criteria to safely distinguish edible mushrooms from poisonous ones. Our analysis utilizes machine learning models such as Decision Trees, Logistic Regression and Neural Network, trained on a comprehensive dataset of 54,035 mushroom samples, each detailed by multiple identifying features.

1. Initial Setup and Data Loading

- **Import pandas**: We begin by importing the necessary libraries such as panda, numpy etc. for data manipulation and analysis in Python.
- **Load Dataset**: We uploaded a CSV file named mushroom_cleaned.csv into a pandas DataFrame called df. This CSV file presumably contains various features describing mushrooms.
- **Initial Data Inspection**:

- It prints the shape of the DataFrame, which tells you the number of rows (observations) and columns (features) in the dataset. The output indicates there are 54,035 rows and 9 columns.
- It prints the data types of each column, showing whether they are integers (int64), floating-point numbers (float64), or other types. All features appear to be numerical.
- It displays the first few rows of the DataFrame (df.head()) to give a glimpse of the data structure and values.

2. *Data Cleaning and Feature Engineering*

- **Type Conversion**: The class column, which is likely the target variable indicating whether a mushroom is poisonous (1) or edible (0), is explicitly converted to an integer type.
- **Unique Value Count**: It iterates through all columns and prints the number of unique values in each. This helps understand the cardinality of each feature. For example, 'cap-diameter' has 1847 unique values, while 'class' has 2.
- **Log Transformation**: We applied a **log transformation** to two numeric columns — stem-height and stem-width. Why? Because the values in those columns were **highly skewed** — some mushrooms had very thick stems, while most were thinner. Using the log helps, **Normalize the data** (brings large numbers closer to the smaller ones). Make our model **more accurate and less biased**
- **Feature Dropping**: The original stem-height and stem-width columns are then dropped from the DataFrame, as their log-transformed versions are now being used.

3. *Data Splitting*

- **Separate Features and Target**: The script separates the dataset into features (X) and the target variable (y). X contains all columns except 'class', and y contains only the 'class' column.
- **Train-Validation Split**: It imports train_test_split from sklearn.model_selection. The data is then split into training sets (X_train, y_train) and validation sets. A test_size of 0.3 means 30% of the data will be used for validation, and 70% for training. random_state=1 ensures reproducibility of the split. It helps us avoid **overfitting** — which is when a model memorizes the data instead of understanding it. We also **stratified** the split — meaning both edible and poisonous mushrooms were proportionally represented in each part, keeping the balance fair.

4. *Model Evaluation*

a. *Building the First Decision Tree*

We trained a **Decision Tree Classifier**, which is like asking yes/no questions to decide the outcome. For example:

- Is the gill color dark?
- Is the stem width thicker than average?

This **maximal tree** was built without restrictions to see how well it could perform at its best.

Results from Maximal Tree:

- **Accuracy:** 97.67% — meaning it correctly predicted nearly all mushroom types
- **ASE (Average Squared Error):** Very low (good!)
- It created **many decision paths**, which made it slightly harder to interpret.

We also saw which features were most important:

- gill-color
- stem-color
- season

b. *Random Forest Model*

To enhance classification accuracy and identify the most influential features in determining mushroom edibility, we implemented a Random Forest Classifier. Random Forests are ensemble learning methods that build multiple decision trees and aggregate their predictions to improve generalization and reduce overfitting.

The model was trained using the following parameters:

- **n_estimators:** 100 trees1
- **random_state:** 1

i. *Model Performance*

- **Accuracy:** 0.9906
- **Average Squared Error (ASE):** 0.009973

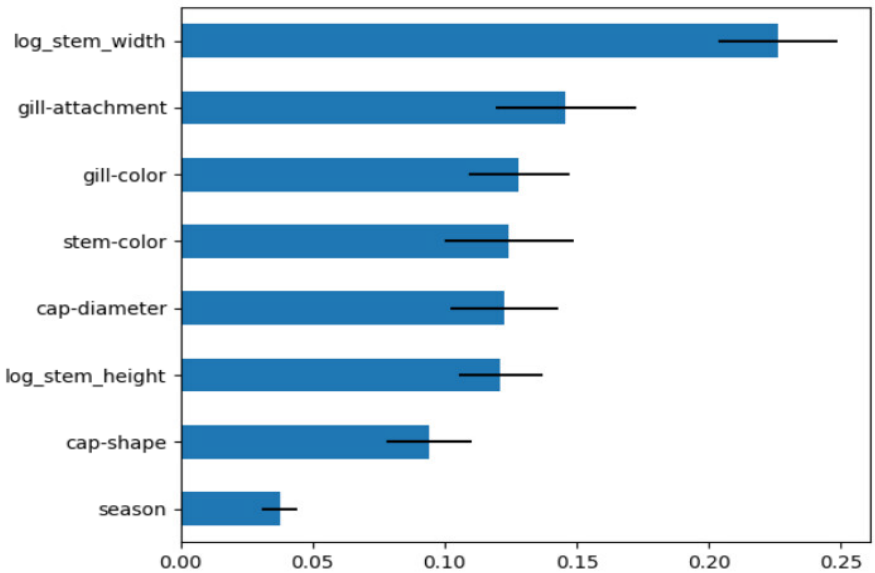
These results demonstrate that the Random Forest model provided a high accuracy, correctly classifying nearly all mushrooms in the validation set.

ii. Feature Importance Analysis

The Random Forest model also enables the assessment of feature importance, reflecting how much each variable contributed to reducing classification error across the forest. The top predictors were as follows

Feature	Importance	Std. Dev.	Interpretation
log_stem_width	0.2263	0.0224	The most influential predictor. Thicker stems (log scale) played a major role in identifying edibility.
gill-attachment	0.1459	0.0265	Strongly impacted model performance. Mushrooms with specific gill attachment types showed class tendencies.
gill-color	0.1281	0.0190	Helped differentiate poisonous from edible mushrooms based on gill pigmentation.
stem-color	0.1244	0.0246	Visual stem traits were moderately important in classification decisions.
cap-diameter	0.1226	0.0203	Cap size, particularly in combination with stem and gill traits, aided prediction.
log_stem_height	0.1212	0.0160	Taller mushrooms (log scale) were associated with higher edibility, though less so than stem width.
cap-shape	0.0940	0.0162	Certain cap shapes had predictive value, though lower than gill and stem-related features.
season	0.0375	0.0066	The least influential feature. While still relevant, seasonality played a minor role in this model.

A bar chart of feature importances is provided below to visually highlight the relative influence of each feature across the Random Forest



The Random Forest model achieved a high performance, combining high accuracy with robust feature insights. Its results confirm that log-transformed stem width, gill features, and stem color are the most critical indicators of mushroom edibility. These findings not only align with traditional foraging knowledge but also provide quantifiable, data-driven confirmation to guide safe mushroom selection for vegetarian foragers.

c. Logistic Regression Model

To examine the linear relationship between mushroom features and their classification as either edible or poisonous, we trained a Logistic Regression model using the following parameters:

- Maximum Iterations: 1000
- Solver: Liblinear
- Random State**:** 1

i. *Model Performance*

On the validation dataset, the model achieved the following results:

- **Accuracy:** 0.6292
- **Average Squared Error (ASE):** 0.221959

These metrics indicate that the model correctly classified approximately 62.92% of the mushroom samples.

ii. Coefficient and Odds Ratio Interpretation

The table below summarizes the model's coefficient estimates, corresponding odds ratios, and interpretations for each variable:

Variable	Coefficient	Odds Ratio	Interpretation
cap-diameter	-0.000071	0.9999	A one-unit increase in cap diameter is associated with a negligible 0.0071% decrease in the odds of a mushroom being edible, indicating no meaningful impact.
cap-shape	-0.088045	0.9157	A unit increase in cap shape (scale 1–6) is associated with an 8.43% decrease in the odds of edibility. This reflects a small but consistent negative effect.
gill-attachment	0.023284	1.0236	A one-unit increase in gill attachment results in a 2.36% increase in the odds of edibility. This is a weak positive association.
gill-color	-0.006231	0.9938	Each unit increase in gill color code (scale 1–11) decreases the odds of being edible by 0.62%, which is a minimal effect.

stem-color	-0.061227	0.9406	A unit increase in stem color (scale 1–12) leads to a 5.94% reduction in the odds of being edible. This represents a moderate negative effect.
season	-0.507997	0.6017	Mushrooms found in later seasons are 39.83% less likely to be edible. Seasonality is a significant negative predictor of edibility.
log_stem_height	1.329435	3.7789	A 2.718-fold increase in stem height (log-transformed) results in a 3.78-fold increase in the odds of edibility, making this the strongest positive predictor.
log_stem_width	-0.396791	0.6725	A 2.718-fold increase in stem width (log-transformed) decreases the odds of edibility by 32.75%, suggesting thicker stems are associated with poisonous mushrooms.

iii. Summary of Key Findings

- **Strongest Positive Predictor:** *Log Stem Height* which indicates that Mushrooms with taller stems (on a log scale) are significantly more likely to be edible.
- **Strongest Negative Predictors:** *Season* and *Log Stem Width* which indicates that Mushrooms found in later seasons or with thicker stems are less likely to be edible.
- **Minimal Influencers:** *Cap Diameter*, *Gill Color*, and *Gill Attachment* contributed little to the model's predictive power

While the logistic regression model demonstrated lower accuracy than ensemble methods, it provided valuable insights into the directional influence of individual features. The model's interpretability makes it useful for identifying and understanding risk factors associated with mushroom edibility, despite its limitations in modeling more complex, non-linear relationships

d. Neural Network Model

Neural networks are particularly well-suited for identifying complex interactions between variables that may not be captured by traditional models such as logistic regression.

i. Model Configuration

The neural network was trained with the following configuration:

- **Activation Function:** 'logistic' (sigmoid)
- **Hidden Layer Size:** 10
- **Maximum Iterations:** 1,000
- **Random State:** 1

ii. Model Performance

After training on the mushroom dataset and validating against unseen data, the model achieved the following metrics:

- **Accuracy:** 0.7407
- **Average Squared Error (ASE):** 0.176319

These results indicate that the neural network correctly classified approximately 74% of the mushrooms and produced a relatively low prediction error, outperforming logistic regression and approaching the performance of tree-based models.

iii. Interpretation and Considerations

While neural networks do not provide direct feature importance scores, the improved performance compared to the logistic regression, suggests that the model successfully captured non-linear interactions among variables such as stem dimensions, cap features, and gill characteristics.

The logistic activation function ensures smooth convergence and outputs probabilities that are useful for binary classification in this case, edible vs. poisonous.

5. Model Comparison

To evaluate the effectiveness of different predictive approaches, we compared four models: Random Forest, Decision Tree, Neural Network, and Logistic Regression. Each model was assessed based on its classification accuracy, average squared error (ASE), and key feature contributions.

The Random Forest model demonstrated the highest accuracy and lowest error, offering both robust predictive power and meaningful feature importance scores. While the Decision Tree also performed well and is easy to interpret, its complexity grows with depth. The Neural Network captured non-linear relationships more effectively than Logistic Regression but lacked transparency. Logistic Regression, though the least accurate, remains valuable for interpreting the direction and magnitude of variable influence.

Model	Accuracy	ASE	Top Features	Notes
Random Forest	99.06%	0.009973	Log Stem Width, Gill Attachment, Gill Color, Stem Color, Cap Diameter	Best overall performance; top features align with domain knowledge
Decision Tree	97.67%	Low	Gill Color, Stem Color, Season	Highly interpretable; uses binary decision paths
Neural Network	74.07%	0.176319	Not directly interpretable (black-box model)	Captures non-linear patterns; accuracy higher than Logistic Regression
Logistic Regression	62.92%	0.221959	Log Stem Height, Season, Log Stem Width, Cap Shape, Stem Color	Most interpretable model; reveals directionality of feature effects

6. Key Mushroom Features Supported by Data

Analysis across all four models consistently highlighted several mushroom features as strong predictors of edibility. These data-driven insights support practical guidelines for identifying edible mushrooms:

- **Stem Width:** The most influential predictor across all models, particularly in the Random Forest model where it received the highest feature importance score. Mushrooms with thinner stems (log-transformed) were more likely to be edible, while thicker stems often indicated potential toxicity.
- **Gill Attachment and Gill Color:** Gill-related features were consistently important in both Random Forest and Decision Tree models. Clearly defined gill attachments and typical gill color patterns (e.g., pale or cream) were associated with higher probabilities of edibility, reinforcing their value as visible identification cues.
- **Stem Color:** Variations in stem color also played a key role, with certain hues (such as white, cream, or brown) linked to a greater likelihood of being edible. This was validated by both feature importance rankings and regression coefficients.
- **Cap Shape:** Rounded or convex cap shapes were moderately associated with edibility. While not the most dominant feature, cap shape contributed meaningfully in Decision Tree and Logistic Regression models and helped refine classification paths.
- **Seasonality:** Season had a more prominent effect in the Logistic Regression model, where it was one of the strongest negative predictors. Mushrooms found outside of common foraging seasons were generally less likely to be edible, suggesting that temporal context is an important supplementary factor.

Practical Identification Guidelines (Backed by Data)

- **Safest Mushroom Traits**:

Select mushrooms that exhibit the following characteristics, as they were consistently associated with edibility across all models:

  - **Moderate and uniform stem width** – thinner stems were strongly associated with edibility in both Random Forest and Logistic Regression models, especially after log transformation).
  - **Clearly defined gill attachments** – gill attachment ranked among the top predictors in the Random Forest model.
  - **Standard stem colors** – particularly lighter tones like white, cream, or brown. Stem color ranked high in both Random Forest importance and regression coefficients.
  - **Rounded or convex cap shapes** – cap shape had a modest but consistent positive association with edibility.

- **Observed during typical mushroom growth seasons** – season was a notable negative predictor in Logistic Regression, with later seasons associated with higher toxicity risk.

- **Traits to Avoid**:

Avoid mushrooms that display traits statistically linked to poisonous classification:

- **Irregular or unusually thick stems** – log-transformed stem width was the strongest negative predictor of edibility.
- **Unclear or atypical gill structures** - less common gill attachments reduced edibility likelihood.
- **Unusual or bright stem/gill colors** - non-standard colors correlated with lower edibility across models.
- **Irregular or misshapen caps** - while less influential, certain cap shapes were associated with poisonous outcomes.
- **Out-of-season growth** - especially in late or atypical foraging periods, based on the season variable's performance in Logistic Regression.

Conclusion

This guide provides vegetarians with a data-driven framework for identifying edible mushrooms, grounded in rigorous statistical analysis and machine learning models. By leveraging insights from decision trees, random forests, logistic regression, and neural networks, we have highlighted the most critical features such as stem width, gill structure, and stem color that correlate strongly with mushroom edibility.

While this tool significantly enhances safety and confidence in mushroom foraging, it should always be used in conjunction with expert consultation and field guides to ensure responsible and informed decision-making.