Fungi or Foe?: A Statistical Exploration of Classification Problems with Mushroom Edibility

There was once a time when mankind relied upon hunting and gathering as a means of survival, and a large component of this lifestyle was the act of foraging. While foraging today has become less essential, the art is still practiced worldwide. As avid lovers of the outdoors, we have experienced firsthand the triumphs and perils that come with foraging- something as seemingly harmless as a mushroom has the power to kill you. It is with this uncertainty surrounding the edibility of foraged goods, that we set out to create a classification model aiming to determine whether or not a certain type of mushroom is poisonous. This is no easy task, as we must create a model that is as accurate as possible, since when it comes to life or death, there is no room for error.

Research Questions: While establishing goals for our process of creating a classification model, we identified what it is we most wanted to learn from this project. We agreed on prioritizing model accuracy, due to the arguments established above, and thus we wish to learn which types of machine learning models perform the best on our data and what variables are most significant in mushroom classification. We will accomplish this by testing eight different classification models, some of which include logistic regression, K-nearest neighbors, decision trees, and random forest models. We will create both parametric and non-parametric models, in order to see which type performs better on our data overall. Finally, we aim to find a way to utilize our models in real life, and to determine which characteristics are most influential for mushroom classification.

Description of the Data: We are using the "Mushroom Classification" dataset from Kaggle. The data was donated to the UCI Machine Learning Repository in April 1987, and it includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family Mushroom drawn from The Audubon Society Field Guide to North American Mushrooms (1981). This dataset is a CSV file containing 23 different columns of categorical data, all of which are characteristics of such mushrooms. Some notable variables include odor, gill spacing, gill size, stalk root, and habitat. Using the variable "class" as our response, we can classify mushrooms as either edible or poisonous, with unknown edibility being grouped with poisonous.

Figure 1: Important Common Attributes of Mushrooms

Attributes	Description
cap shape	The shape of the upper part of the mushroom (bell, conical, etc.)
cap color	Color of the upper part of the mushroom (brown, green, red, etc.)
cap surface	Texture of the upper part (grooves, smooth, etc.)
odor	How the mushroom smells (musty, pungent, etc.)
gill size	The size of the part under the cap (i.e. the gill)
stalk shape	The shape of the bottom most part (enlarging, tapering)
population	Number of mushrooms grown from one root
habitat	The place where the mushroom is grown (meadows, woods, etc.)

Preliminary Studies:

Data Visualization: In *Figure 2*, we observe that our data is balanced for our dependent variable “class”, with 4,208 edible and 3,916 poisonous mushrooms in our sample. We also see that all 8,124 samples have the same veil-type, partial (p). Thus, we choose to drop veil-type from our variables, as there is no variation. Finally, we observe that for the remaining independent variables, the number of edible and poisonous mushrooms is fairly comparable, with no clear attribute that poisonous mushrooms have over edible mushrooms, and vice versa.

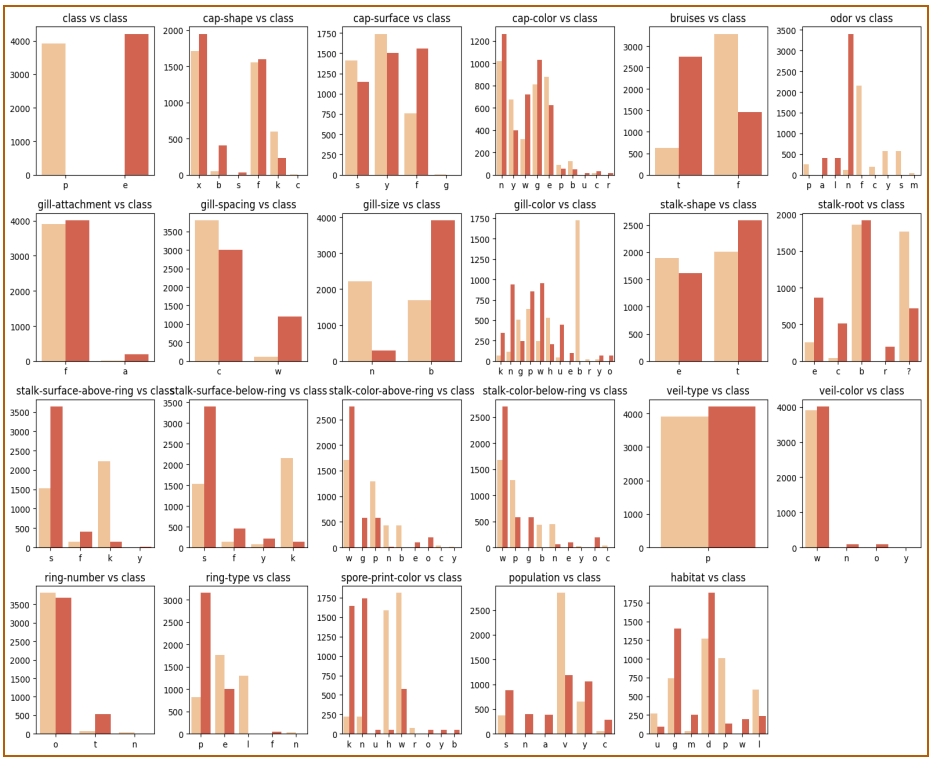

Figure 2: Histograms of our variables

Dimension Reduction and Feature Selection:

To make our dataset usable when constructing the classification models, we converted each of the categorical variables into a numerical format using dummy variables. First, we created single dummy variables for the features with two classes. These include: class, bruises, gill attachment, gill spacing, gill size, and stalk shape. For the remaining variables (all categorical) with more than two classes (say k classes), we generated k-1 dummy variables for each attribute. A one is assigned to the characteristic that the mushroom expresses, while a zero is assigned to the remaining dummy variables for that feature. Transforming the categorical variables into dummy variables establishes a more applicable dataset to use for generating mushroom classification models. After the creation of these dummy variables, we split our dataset into our dependent and independent variables, choosing “class” as our dependent variable, assigning a value of 1 to a poisonous mushroom and a value of 0 to an edible mushroom.

After declaring “class” as our dependent variable, we split our samples into a training and a test set, with 80% of the samples in the training set and the remaining 20% in the test set. The training set is presented to the classification models in order for the model to learn and make

predictions based on the input features and the corresponding output label. We will use our test set to evaluate the performance of our different classification models by obtaining an accuracy score. The splitting of our data into a training set and a test set is necessary to assess how our classification models generalize to new data, thus conveying how the model would operate in real-world scenarios.

Classification Models: We created eight models for our data, described below:

Logistic Regression: Logistic regression is a parametric statistical model used for binary classification, utilized in this context for predicting the probability of a mushroom belonging to a particular class (edible or poisonous). We chose to include logistic regression due to its interpretability, making it suitable for real-life scenarios, as well as the fact that it serves as a baseline against other machine learning models for mushroom classification.

K-Nearest Neighbors (KNN): K-nearest neighbors is a non-parametric machine learning algorithm used for classification and regression tasks, relying on proximity to determine the class of a data point based upon its neighbors. In this project, KNN is selected to assess its effectiveness in classifying mushrooms by considering the similarities to other sample mushrooms in the ecosystem, contributing to the overall goal of creating a model to identify safe mushrooms for consumption.

Decision Tree: A decision tree is a non-parametric tree-like model where each internal node represents a decision based upon a specific feature, leading to leaf nodes representing the final outcome. In our project, we use decision trees for capturing complex relationships within the dataset, aiding in the identification of key features in mushroom classification, and offering interpretability, making them user-friendly for practical applications in real-life scenarios.

Random Forest: Random forest is a non-parametric learning method that constructs multiple decision trees during training and outputs the mode of the classes for classification. We chose to use random forest as it handles complex relationships within the data, reduces overfitting, and provides robust predictions, contributing to our search for a highly accurate model.

Boosting: Boosting is a non-parametric learning algorithm based upon gradient boosting, specifically designed for speed and performance. We valued including boosting in our analysis due to its ability to provide high predictive accuracy, handle complex datasets, and effectively manage feature interactions, as well as its efficiency and scalability.

Support Vector Machine Classification (SVM): SVM classification is a machine learning tool that seeks to find the optimal hyperplane to separate different classes in a high-dimensional space. It is a

parametric model, relying on the assumption that the data is linearly separable or can be transformed into a higher-dimensional space where a linear separation exists. SVM is included for its ability to handle non-linear relationships in the data and its effectiveness in binary classification tasks.

Gaussian Naive-Bayes: Naive-Bayes classification is a parametric probabilistic model based on Bayes' theorem, assuming that features are conditionally independent given the class label and following a Gaussian distribution. In the context of our project, Naive-Bayes was chosen for its simplicity, efficiency with high-dimensional data, and its ability to provide probabilistic predictions.

Linear Discriminant Analysis (LDA): LDA is a supervised classification process that computes a discriminant function for each class, based on linear combinations of features, aiming to maximize the separation between classes. It is a parametric model, making assumptions about the normality and equality of covariance matrices of the features within each class. We choose to use LDA for its effectiveness in feature extraction and dimensionality reduction.

Model Comparison:

Results/Discussion: For each model, we used accuracy for the training and test sets as a means of model comparison. Half of our models have an accuracy of one, indicating a perfect classification rate on both the training and test sets. This is likely due to the fact that our samples covered only 23 species of mushrooms, and these samples were based off of qualities found in The Audubon Society Field Guide to North American Mushrooms (1981). Thus, once our model accurately identifies what species a mushroom is, it can easily classify that species in the future as edible or poisonous.

Figure 3: Accuracies for our models

Model	Training Accuracy	Test Accuracy	Training Mean	Test Mean
Naive Bayes	0.942453	0.948923	0.941068	0.911993
Decision Tree	0.995999	0.996307	0.995999	0.995690
LDA	0.999692	0.998769	0.999692	0.998152
LogisticRegression	0.999846	0.998154	0.999231	0.996921
RandomForest	1.000000	1.000000	1.000000	1.000000
XGBOOST	1.000000	1.000000	1.000000	1.000000
KNN	1.000000	1.000000	1.000000	1.000000
SVM	1.000000	1.000000	1.000000	1.000000

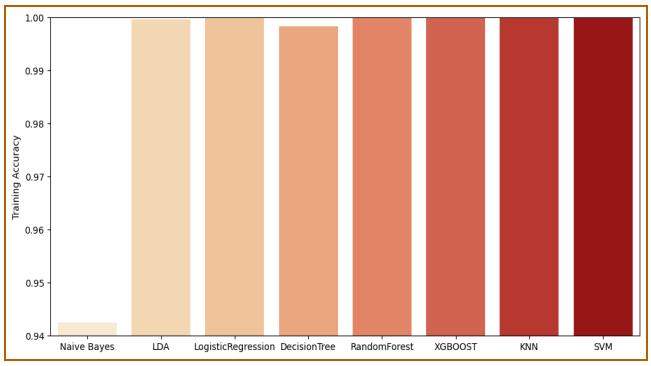

Figure 4: Comparing Training Accuracies

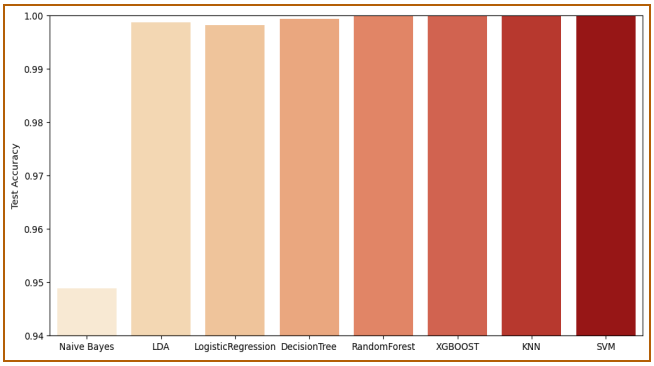

Figure 5: Comparing Test Accuracies

Our Model Choices: We strategically selected two highly accurate finalist models for our data, one parametric and one not, shown below. Models with an accuracy of one were not considered for our final models, as they have the risk of overfitting, as well as losing interpretability and ease of use.

Model 1- Logistic Regression (Parametric): We choose logistic regression for its interpretability, as it provides a clear linear decision boundary and its coefficients can be directly interpreted in terms of feature importance. It is also easy to implement (providing direct probabilities for each sample), computationally efficient, and works well for binary classification problems.

Most Influential Predictors: By viewing the coefficients for each predictor, the top 5 variables that had the largest effect on class included: odor, spore print color, gill size, stalk root, and stalk surface above ring

---

Model 2- Decision Tree (Non-Parametric): We choose decision trees for their ease of use in real life scenarios, as they can be visually represented and therefore easily interpreted. Decision trees also provide clear decision boundaries, thus classification is easier for those without a statistical background. These factors make our decision tree model the perfect balance between interpretability and capturing complex decision boundaries.

Most Influential Predictors: By using the feature importance function for decision trees, the top 5 variables that had the largest effect on class included: odor, population, stalk surface below ring, habitat, and spore print color.

Observations: We observe that the top five most important/influential variables between our two models differ slightly, however we see that both include odor and spore print color. So when looking to classify a mushroom when foraging in reality, we suggest looking at these qualities first. We also notice that our Decision Tree model has a slightly higher test accuracy and slightly lower training accuracy than logistic regression, however in the end this difference in accuracy is not significant.

Applicability to Real World:

Sample: In order to test our models in a real-life scenario, we obtain a mushroom of the species Agaricus Bisporus, commonly known as the Portobello mushroom, and input its qualities into each of our models. This mushroom is applicable in this experiment as it is of the Agaricus family and is native to the grasslands in North America. Knowing that this mushroom is edible, we utilize our models to confirm this. All of our models return a value of 0 (representing the mushroom is edible), except for our Naive-Bayes model, which has the lowest accuracy of our eight models. Therefore, confirming that our models are both accurate and applicable to the real-world.

Figure 6: Predictions for Agaricus Bisporus

Model	Prediction
Logistic Regression	0
KNN	0
Decision Tree	0
Random Forest	0
Boosting	0
SVM	0
Naive-Bayes	1
LDA	0

Conclusion: While we accurately created a plethora of models that can aid in mushroom classification, a key takeaway from this project is that there is no shortcut in determining the edibility of mushrooms; it is a complex science that requires extensive knowledge on the species' features. Each model we created prioritized slightly different mushroom characteristics, and it is for this reason that a machine learning model for mushroom classification is necessary. With the implementation of our models, the general population can forage for mushrooms without needing to memorize the characteristics of each species they might encounter. It is with this that given a wider variety of samples and funding, we could make an informative program that can save lives. Overall, we observe that our nonparametric models perform better on our dataset, with higher accuracies on average compared to our parametric models. When choosing a final model, we agreed that logistic regression and decision trees were comparable, but ultimately the model we would foresee utilizing is our decision tree. Decision trees require no computational background to use, and thus can provide a simple way of classifying mushrooms for the average user, who could utilize a physical copy of our decision tree while in nature. Also, our decision tree model contained no "falsely edible" samples- a mushroom predicted to be edible but in reality is poisonous- while logistic regression had four. These cases are potentially dangerous, and thus we wish to avoid them at all costs in our model. In conclusion, our efforts toward creating a machine learning model for mushroom edibility lay the foundation for a safer, more accessible, and precise foraging experience for the average person.

Works Cited

Lincoff, Gary H. The Audubon Society Field Guide to North American Mushrooms. Knopf, 1981.

"Mushroom Classification." Kaggle, UCI Machine Learning, 1 Dec. 2016, www.kaggle.com/datasets/uciml/mushroom-classification/data.

Petruzzello, Melissa. "Portobello Mushroom." Encyclopædia Britannica, Encyclopædia Britannica, inc., 31 May 2023, www.britannica.com/topic/portobello-mushroom.

Ria, Nushrat, et al. ICCCNT, 2021, pp. 1–5, State of Art Research in Edible and Poisonous Mushroom Recognition.