# Toxicity of Mushrooms: A Classification Analysis

David Li, In Son Zeng, Yingdan Zhang

April 16, 2018

**Abstract**

Mushrooms play important roles in ecological balances to maintain assorted cycles in nature, but studying the data from over thousands of species can be a daunting task. Several exploratory and classification analysis methods today may be employed to comprehensively analyze the intricacies of mushroom features. This paper explores the distribution of edible and poisonous mushrooms by plotting 1D, 2D, and 3D graphs about mushroom features. Parametric classification and non-parametric classification models were constructed; performance is measured through misclassification error rates. Ultimately, do sensible procedures of classification exist for mushrooms in numerical and practical aspects?

# 1 Introduction

There exist at least 10,000 mushroom species, with great variation in multiple characteristics including color, shape, toxicity, and edibleness. Because mushrooms are most commonly seen as edible delicacies, an important question arises repeatedly as more mushrooms continue to be discovered: how can one effectively evaluate different traits of mushrooms to predict their edibleness properties? There are many possible approaches, but we will introduce a classification-based perspective to answer this important question. We will analyze statistical patterns, data trends, and apply modern classification methods to bring statistical learning insight to our research. First we will explain how we chose the appropriate dataset in order to provide an adequate foundation for analysis. From that given dataset, we will provide plenty of exploratory data analysis to give a rich background to prior knowledge about the traits and characteristics of known mushrooms. Classification methods will not be focused on deep foundation discussion but rather gently introduced through intuitive explanations, and rigorously applied to the dataset to point out any striking results or interesting discoveries at a technical level. There will be a focus on visual depictions and critical metrics to assess the performance and applicability of each method supplemented with clear explanations of their relevance towards the research question. We will consider strengths and drawbacks of these approaches to highlight the successes and respect the limitations of the analysis. In the end, we may summarize what has been learned and complement the research with future hypothetically questions.

The research question is aptly summarized: Observe relationships and patterns within the mushroom features in addition to reliable classification methods to categorize edibleness?

# 2 Data

The chosen dataset originates from the *National Audubon Society Field Guide to North American Mushrooms book* from 1981, written as a comprehensive guide to many diverse mushroom species. A dataset containing the information on mushroom features was created in 1987, and is now publicly available at the UCI Machine Learning Repository website.

Structural composition is critical in understanding the quantity and quality of the data being analyzed in the methods. The dataset features 8124 observations in total, each detailing a series of 22 mushroom traits and the corresponding edibleness classification. Some of the more prominent mushroom features include cap color, cap shape, odor, stalk shape, growth habitat, and many more listed in Figure 1.

| | | | | |
|---|---|---|---|---|
| Cap Shape | Gill Attachment | Stalk Root | Veil Type | Population |
| Cap Surface | Gill Spacing | Stalk Surface Above Ring | Veil Color | Habitat |
| Cap Color | Gill Size | Stalk Surface Below Ring | Ring Number | |
| Bruises | Gill Color | Stalk Color Above Ring | Ring Type | |
| Odor | Stalk Shape | Stalk Color Below Ring | Spore Print Color | |

Figure 1: Mushroom Features Table, Within the Dataset

Missing data should be deleted to prevent misleading and/or incorrect conclusions from the data exploration and model building steps. 2480 observations showed missing entries for the stalk root feature, thus yielding a new total of 5644 observations going forward. A glance at the data indicates 3488 edible observations, or 61.8% being edible while 38.2% being poisonous. Fortunately, this fairly balanced nature of the dataset bypasses the need to consider complicated structural issues; classifications based upon severely imbalanced datasets can be unreliable due to a natural gravitation towards the correct classification of the dominating class. One prominent technique to alleviate imbalanced datasets is to randomly generate the underrepresented class through bootstrapping.

Figure 2 summarizes a breakdown of categorical distribution for each individual mushroom feature styled such that the green and red bars represent the proportion of edible and poisonous mushrooms, respectively. Note that these observations pertain to the scope of the dataset, and these distributions could change dramatically as new mushroom species are discovered in the future. Some graphs show an equal ratio of edible to poisonous mushrooms within multiple categories in a particular feature, such as with cap surface and cap shape. Conversely, some features such as odor and bruising show uneven distributions of edibility class in certain categories; all foul, musty, pungent, and creosote odor mushrooms are poisonous for instance! Several notable discoveries are worth mentioning: almost all red-capped mushrooms are edible, most scentless mushrooms are edible, non-bruised mushrooms are usually poisonous, all silky-stalked mushrooms are likely to be poisonous, and mushrooms that grow in urban settings are more likely to be poisonous.

Additionally, it is common for multiple variables to be correlated in effect. Analyzing some of these relationships can be key to determining tendencies of toxicity for a particular species. Note that correlation could be an appropriate measure of "similarity" if the variables involved are quantitative; correlation measurement is not currently appropriate since all variables in the dataset are categorical.

Figure 3 highlights some more interesting pairwise relationships from the list of all mushroom features. These 4 pairs of variables were considered from their relationships being strongly hypothesized to influence toxicity in some degree. Some conclusions from the single feature analysis remained unchanged in the paired analysis, such as most red-capped mushrooms are edible regardless of stalk color. However, some pairwise variable relationships created different conclusive results from the one-dimensional analysis; initially it was considered that urban-grown mushrooms were mostly poisonous, but the pairwise analysis showed that scentless urban mushrooms are edible. Other pairwise patterns were noticed: scentless mushrooms were primarily edible regardless of the growth habitat, while foul odor mushrooms should never be consumed due to their common poisonous nature in all habitats. Mushrooms with yellow caps were generally poisonous except when the stalk color was white. White-capped mushrooms grown in meadows were usually safe to eat, but untrustworthy from an urban environment. Finally, solitary mushrooms were generally evenly distributed to be poisonous and edible but were primarily edible in urban environments.

Simultaneously visualizing 3 mushroom features can be challenging, but some combinations of characteristically dominant features are worthwhile to investigate. Figure 4 & 5 present a 3-dimensional visualization

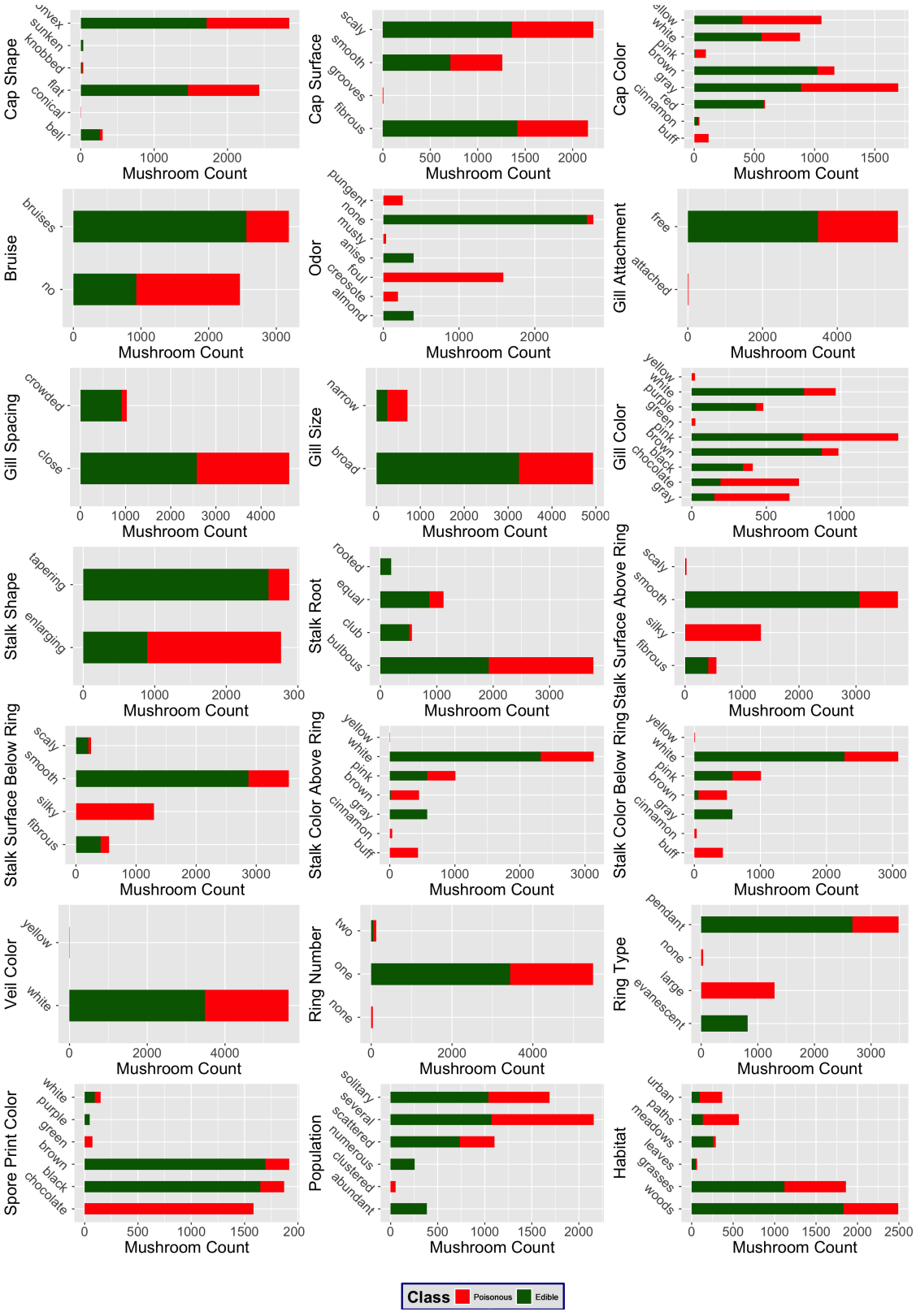


Figure 2: Edibleness of Mushroom based on Classification

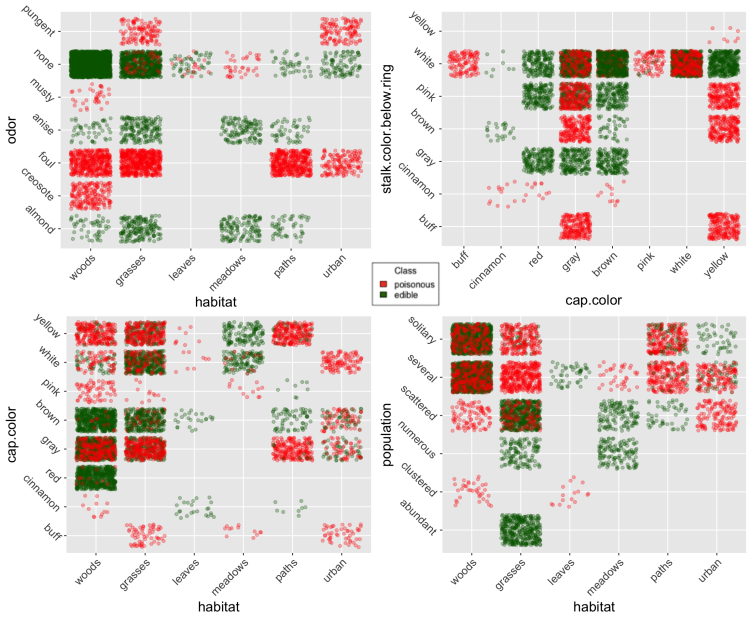


Figure 3: Edibleness of Mushroom based on Pairing Features

of cap shape, surface, and color and a 3-dimensional visualization of cap color, stalk color, and gill color. Example questions to be analyzed can vary: how can distinguishable aspects of the mushroom cap influence toxicity, or how can multiple colors from different parts of a mushroom classify as poisonous or edible.

Judging by these 3-dimensional plots, we can further deduce interesting trends in the data. Figure 4 shows that mushrooms with a sunken cap shape (regardless of cap surface and cap color) tend to be edible. Additionally, mushrooms with a buff cap color are certainly poisonous (regardless of cap surface and cap shape). Mushrooms with red, fibrous caps are likely to be edible (regardless of cap shape). Figure 5 adds that pink-capped mushrooms with white gills are mostly poisonous (regardless of stalk color), and mushrooms with buff or cinnamon colored stalks are certainly poisonous (regardless of either cap or gill color). As seen before, red-capped mushrooms are generally edible except those with cinnamon colored stalks. Lastly, mushrooms with gray-colored stalks are likely to be edible regardless of observed colors from any other part of the mushroom.

# 3 Methods

## 3.1 Data Processing

Part of the statistical learning modeling process ideally requires future data to test the accuracy and applicability of the classification models. When other data is not available, an alternative procedure is to partition the original dataset into a training dataset for model building and a test dataset for model testing, so as to assess the model performance. We chose to partition the mushroom dataset in a 80 - 20 ratio for

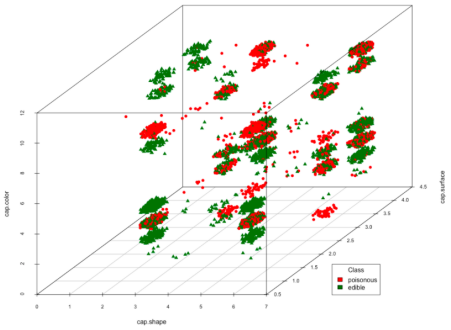


Figure 4: 3D Plot of cap shape, cap surface, and cap color

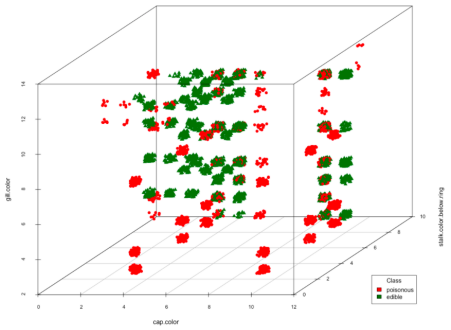


Figure 5: 3D Plot of cap color, stalk color below ring, and gill color

training and test data from the 5644 observations. The statistical models for the later four classification methods were built from the training data, with misclassification rates reported upon the training and test data. Furthermore, "veil-type" mushroom feature contained the same value for all 5644 observations so it will not be considered in many model building procedures, giving a final total of 21 mushroom features going forward.

An alternative approach to work with the mushroom data is to produce a dimension-reduced dataset through Principal Component Analysis (PCA). Since working simultaneously on all 21 mushroom features can be difficult and cumbersome, using PCA-applied data has a clear advantage. Therefore we may "compress" the data into lower dimensions through PCA for easier visualization through the composite dimensions, where other new unseen interpretations may arise. This compression is not easily understood intuitively due to some loss of interpretability, but preserves most of information in the data. Performance assessments on the classification methods are evaluated upon both the original and PCA-applied data.

## 3.2 Multidimensional Scaling

Multidimensional Scaling (MDS) is a modern computational method purposed for measuring similarity between individual cases in a dataset, usually through visualization. In this case, a dissimilarity function is appropriate to represent the "distance" between observations; the current dissimilarity matrix counts the

number of differences between observations through a binary "different" or "not different". For observation a and b with features i = [1,2,.....,21], the dissimilarity function is

$$d(a,b) = \sum_{i=1}^{21} \mathbb{1}(a_i \neq b_i) \tag{1}$$

Principal Component Analysis (PCA) is simply a specific technique derived from MDS that accomplishes the desired dimension reduction while preserving the "distance" between observations with minimal information loss. Once the MDS plot is constructed for low-dimensional visualization, we can generalize conclusions about the original data with high dimensions.

### 3.3 LDA and Logistic Regression

Linear Discriminant Analysis (LDA) and Logistic Regression are linear parametric classification algorithms used to assign observations into a discrete classes. The parameters of interest were fit through respectively maximizing the conditional likelihood and joint likelihood. Particularly, LDA uses the first two discriminant directions to help project the data into lower dimensions based on between-class and within-class variance, while logistic regression estimates the probability (transformed from log-odds) for each observation and categorizes them based on probabilistic decision boundaries. Note that LDA has less flexible parametric assumptions, requiring Gaussian-distributed data and homoscedasticity. Since there are 21 mushroom features in our dataset, we specifically use Multiple Logistic Regression model to linearly separate the observations. The discussion section further below delves more into the accuracy and error rate of the models and comments on some conclusive results from this application.

### 3.4 KNN

K-Nearest Neighbors (KNN) is a non-parametric supervised learning method which takes no assumption on the underlying distribution of the data. Due to its non-parametric nature, KNN is extremely powerful and works well with most models except those with an exceedingly high proportion of outliers. It classifies the objects by first taking the distance matrix on feature similarities and a number of nearest neighbors parameter, then assigning a label for each observation via a "majority rule" from their neighbors' class. The optimal selection of the nearest neighbors parameter (through trial and error) seeks to minimize cross-validation errors and determines the best final KNN model of interest.

### 3.5 Decision Trees

Decision trees are another non-parametric supervised learning method, built upon a more straightforward classification approach. If the data is notably splittable by certain features, a recursive splitting procedure can be outlined by certain cutoffs to determine the likely classification; this procedure ends when no further

sensible splits are needed. Graphically, each node represents the likely class label while branches represent the notable classification splits and decision rules involved. Overall, these nodes and branches compose a decision tree and show a mapping of all predictors' values applied through decision rules to predict the final class.

# 4 Results and Discussions

Separate sections address each classification method, each displaying initial appropriate visualization with useful remarks. A numerical summary of the obtained various misclassification error rates is appended to provide supplemental evidence for each classification method.

## 4.1 Multidimensional Scaling

Figure 6 displays the MDS graph to visualize and interpret the 21 features of mushrooms in 2 dimensions, with two colors representing the true class labels. Although some difficulty arises when commenting on the influence of individual mushroom features, it is clear that most edible and poisonous mushrooms are grouped closely together; this indicates that predictions will likely be accurate and the risk of classification confusion should be low. However, there is still a minimal area of overlapping classes where it could be difficult to predict an observation's edibleness class; prediction confidence should be otherwise fairly high. Finally, there is a notable isolated cluster of poisonous observations and a similarly isolated cluster for edible observations which indicates that certain patterns of mushroom features strongly point to a particular class.

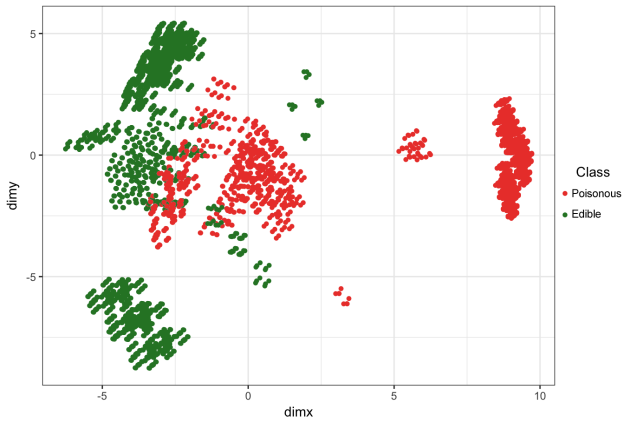


Figure 6: Classical MDS Based on Dissimilarity

## 4.2 LDA and Logistic Regression

In Figure 7, plots in the first, second and third columns respectively matched the training, test, and overall data, whereas plots in the first and second rows respectively matched the original and PCA-preprocessed

data. For the original data, we chose to plot odor and spore print color of the mushrooms as the two discriminant directions due to their hypothesized strong influence on class identification. For the PCA-preprocessed data, we plotted the first and second principal components as the two discriminant directions since they explain the most amount of variance in the complete data.

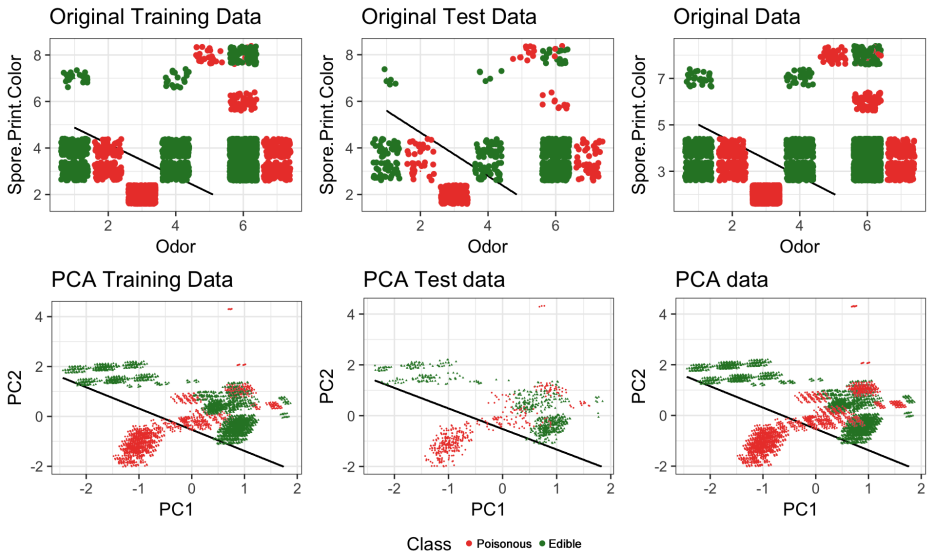


Figure 7: LDA Projections onto the first two discriminant directions

It is visually evident that basing the current classification on linear separation is very ineffective; even from an intuitive standpoint, it is impossible to draw a linearly separating boundary to perfectly classify the observations into the correct class labels. Moreover, recall that LDA requires a Gaussian assumption and homoscedasticity, which is unlikely to be fulfilled by the 21-dimensional categorical mushroom data. Therefore, low LDA misclassification error rates such as 0.05 from Figure 10 may seem like good news, but they fail to recognize the inappropriate model construction from the severe assumption violations. Generally it is obvious that LDA should not be considered as an appropriate model.

The logistic regression performance is also seen in Figure 10, and shows equally poor conclusive results due to similar structural issues described in LDA. Both of the training and test errors of logistic regression are 0, which are also misleading. Again, a model based on linearly separation is inappropriate in the current results.

## 4.3 KNN

The Figure 8 shows the training, cross-validation and test misclassification errors when the number of nearest neighbors parameter ranges from 1 to 30. Recall that the greater value the parameter is, the greater the number of nearest neighbors is considered for class labeling. For the original data, all three classification errors increased gradually when k increased; the error rates started from 0 and rose to slightly more than 0.01. For the PCA-applied data, the training error similarly grew as k increased with the CV and Test

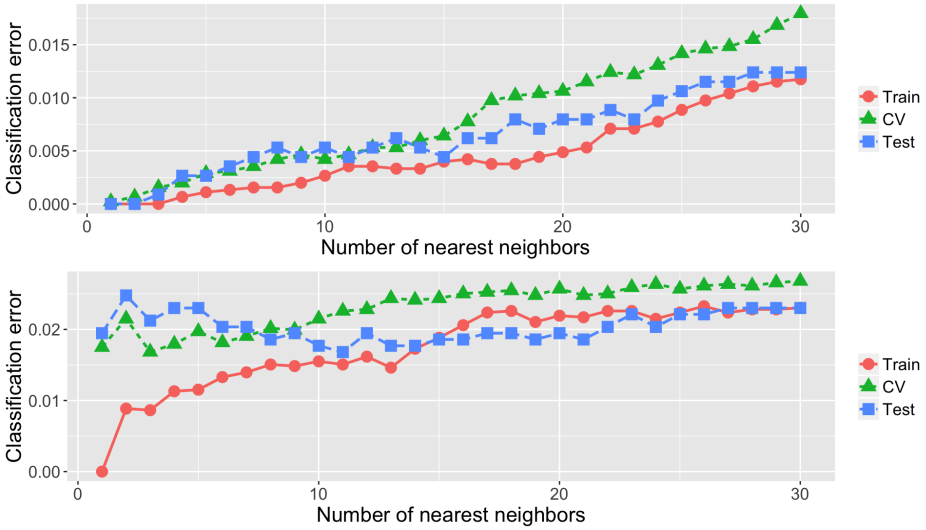


Figure 8: KNN Error Rate for Original and PCA Pre-Processed Data

errors fluctuating between 0.015 and 0.025. The optimal number of nearest neighbors in each setting was determined through trial and error and comparing the cross-validation error rates, giving K=1 for the original data and K=6 for the PCA-applied data for the lowest cross-validation error.

Within context of the mushroom data, KNN performed extremely well. The training and test errors were all 0. Unlike LDA and Logistic Regression, the assumptions are not as stringent due to the non-parametric nature of KNN. KNN only requires careful consideration when the data contains many significant outliers, which is not a problem as evidenced by the MDS plot. Therefore it suffices to conclude that KNN is an excellent classification method for predicting edibleness from the 21 specified mushroom features, due to the excellent performance shown through low prediction error rates of the KNN model.

## 4.4 Decision Trees

Figure 9 shows the binary outcomes of splitting mushrooms based on decision rules of certain mushroom features; for each node, we split the observations by "condition fulfilled" branching to the left and "condition unfulfilled" branching to the right. For instance, the top internal node corresponds to splitting mushrooms based a mushroom's spore print colors. Mushrooms with chocolate colored spore prints branch to the left and are directly classified as poisonous, while the other colors such as black, brown, green, purple, white are distributed to the next node for splitting. The second internal node split the mushrooms based on the gill size; mushrooms with narrow gills satisfy the splitting condition and branch to the left, while mushrooms with broad gills brach to the right to apply the next splitting rule. Overall, the splitting rules render a nicely interpretable final result: the decision trees are likely to classify mushrooms as poisonous when most of the tree's splitting conditions were met and edible otherwise. The classification performance for this decision tree is impressive, yielding misclassification error rates below 1% for both the original training and test data

as well as below 4% for both the PCA-applied training and test data. No assumptions about decision trees were known to be violated, so these results are reasonable to gauge decision trees as a viable edibleness classification method from the 21 mushroom features.

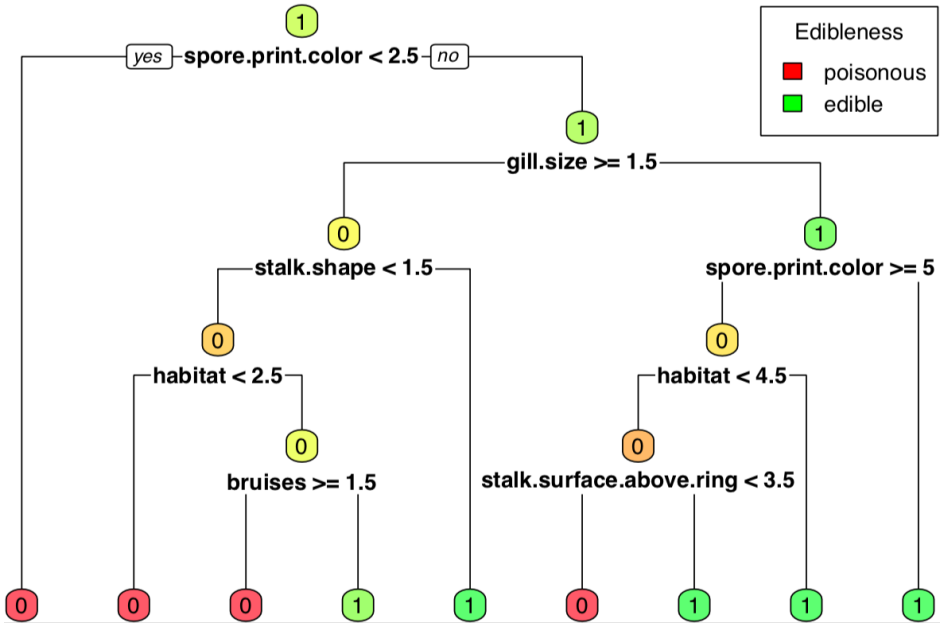


Figure 9: Classification Tree for Edibleness of Mushrooms

## 4.5 Classification errors (Training and Test error)

Figure 10 summarizes the misclassification errors of the training and test data for the Logistic regression, LDA, KNN, and decision tree methods. Generally, the PCA-preprocessed data performed worse with higher misclassification error rates in all 4 classification methods. This is verifiable since using the first 3 principal components will not capture all the variance in the data. Logistic regression and KNN "perfectly" classify the mushrooms for original data, while LDA gave the error rate approximately as high as 0.05. It was discussed earlier that linear-based models with violated assumptions will have unreliable conclusions thus deeming their error rates nonsensical. For PCA-preprocessed data, the classification errors are highly stratified for parametric and non-parametric models. That is, the error rates for Logistic regression and LDA are higher at approximately 0.18, whereas the error rates for KNN and decision trees are only between 0.0133 and 0.031. It tells that non-parametric classification methods not only are more appropriate for the mushroom dataset, but also perform more accurately in correctly identifying the edibleness.

## 4.6 Classification errors (Within class)

As shown in Figure 11, within class errors focus on the classification errors separately by each distinct class, which may provide alternative interpretations of the classification performance. For instance, we might

| | Original | | PCA-preprocessed | |
|---|---|---|---|---|
| | Training error | Test error | Training error | Test error |
| Logistic Regression | 0.0000 | 0.0000 | 0.1863 | 0.1743 |
| LDA | 0.0410 | 0.0504 | 0.1883 | 0.1761 |
| KNN | 0.0000 | 0.0000 | 0.0133 | 0.0204 |
| Decision Trees | 0.0011 | 0.0027 | 0.0177 | 0.0310 |

Figure 10: Table of Training and Test Errors

question if certain classes carry diverse results of classification errors. Notice that the classification errors in poisonous mushrooms is always higher than those in edible mushrooms. In extreme cases such as with logistic regression and LDA, the within class error for poisonous mushrooms can misclassify as often as 1 in every 3 observations. In addition, the robustness of non-parametric methods is remarkable: even considering the instability in PCA-preprocessed data, we find that the classification errors of KNN and decision trees never exceed 0.04 for all two classes.

Also as before, most PCA-preprocessed data performed worse than the original data in all of the statistical learning models. Using three PCA components still excels at reducing the data to low dimensions, but finds issues for inference and interpretability as some information within the higher composite components is lost.

| | Original | | PCA-preprocessed | |
|---|---|---|---|---|
| | Training error | Test error | Training error | Test error |
| **Logistic Regression** | | | | |
| Poisonous | 0.0000 | 0.0000 | 0.3132 | 0.2986 |
| Edible | 0.0000 | 0.0000 | 0.1079 | 0.0974 |
| **LDA** | | | | |
| Poisonous | 0.0435 | 0.0556 | 0.3306 | 0.3079 |
| Edible | 0.0394 | 0.0473 | 0.1004 | 0.0946 |
| **KNN** | | | | |
| Poisonous | 0.0000 | 0.0000 | 0.0145 | 0.0301 |
| Edible | 0.0000 | 0.0000 | 0.0122 | 0.0143 |
| **Decision Trees** | | | | |
| Poisonous | 0.0029 | 0.0069 | 0.0180 | 0.0394 |
| Edible | 0.0000 | 0.0000 | 0.0176 | 0.0258 |

Figure 11: Table of Within Class Errors

# 5 Conclusions

In summary, the most accurate and appropriate classification approaches were the KNN and decision tree models. These non-parametric models are much more flexible, which make sense to apply to the mushroom dataset since the mushroom features are unlikely to follow any parametric assumptions or distributions required in parametric models. For any sort of desired prediction from the specified 21 mushroom traits, we can reasonably consult the expected classification class from these methods. Consequently, the linear-based classification methods performed the worst by misclassification error rates. All of these models and

algorithms were built from this dataset, but may change with more data or as more new mushroom species are discovered over time.

A data-relevant limitation of our research alludes to the fact that our scope of the data is restricted only to North America. Expanding the data to mushrooms worldwide may lead to different results, due to the diverse environments and climates. Missing data truncated certain categories in some variables such as spore print color. In a practical sense, classification of mushrooms extends beyond a binary definition of edible or poisonous; some species may be edible but with an unappealing taste, and some species may only be slightly discomforting to ingest but not poisonous.

The design of the distance matrix is another considerable limitation, which is currently based on the assumption that differences in mushroom features are equally important; a more realistic weighted distance matrix by variable importance is highly desired for improved results. However, weighted distance matrices are computationally expensive, especially when the data is large; the process may be resemblant of iterating through the distance matrix to update weights per computational step until convergent weights are obtained. In addition, further testing of the weighted distance matrix is required to ensure validity in application such as unbiasedness.

# References

*The Audubon Society Field Guide to North American Mushrooms* (1981). G. H. Lincoff (Pres.), New York: Alfred A. Knopf

Dua, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning data mining, inference, and prediction* (2nd ed., Springer Series in Statistics). New York: Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Springer Texts in Statistics). NY: Springer.