Team Project Initial Report

Applied Regression Analysis II - MAS 646

Project Topic: Mushroom Edible Prediction

Group Members: Rushil Banerjee, Melanie Cafatti, Austin Abraham

# Contents:

1. Introduction
2. Dataset Description & Objective
3. EDA (Exploratory Data Analysis)
   1. Structure
   2. Checking for Null Values
   3. Fixing Null Values
   4. Statistical Summary
   5. Frequency Distributions
   6. Check for Normality
   7. Pairwise Relationship
4. Model Building
   1. Baseline Logistic Regression Model
   2. Interpretation of Coefficients
   3. Stepwise Selection Model
   4. Interpretation of Coefficients Stepwise Model
5. Model Evaluation
   1. Confusion Matrix Logistic Regression
   2. Confusion Matrix Stepwise Selection
   3. ROC & AUC Logistic Regression
   4. ROC & AUC Stepwise Selection
6. Moving Forward

# 1 Introduction:

# The ability to distinguish between edible and poisonous mushrooms is a critical issue for foragers, chefs, and public health officials. Misidentification can lead to serious health risks, yet visual inspection remains one of the most accessible methods for classification. This project uses a detailed mushroom dataset from Kaggle to develop a logistic regression model that predicts whether a mushroom is edible or poisonous based on its physical characteristics and ecological traits. With over 61,000 observations, the dataset offers a rich source for statistical modeling. By identifying which features are most predictive of edibility, we aim to build an interpretable and actionable model that could support safer mushroom identification and broader understanding of fungal biology.

# 2 Objective:

# To develop a binary classification model that predicts whether a mushroom is edible (e) or poisonous (p) based on a variety of attributes related to morphology (cap, stem, gills), environment (habitat, season), and other physical traits. We will:

# 1. Apply logistic regression to model the probability of a mushroom being poisonous.

# 2. Determine which characteristics have the strongest influence on classification.

# 3. Evaluate the model’s predictive accuracy using appropriate classification metrics.

# DataSet Description:

The dataset contains 61,069 observations, each representing a unique mushroom sample, with a mixture of categorical and numerical attributes that describe its physical features and natural growing conditions. The target variable is class, which indicates whether the mushroom is edible (e) or poisonous (p).

**Target Variable:**

* **class:** Binary indicator of mushroom edibility
  + p = poisonous
  + e = edible

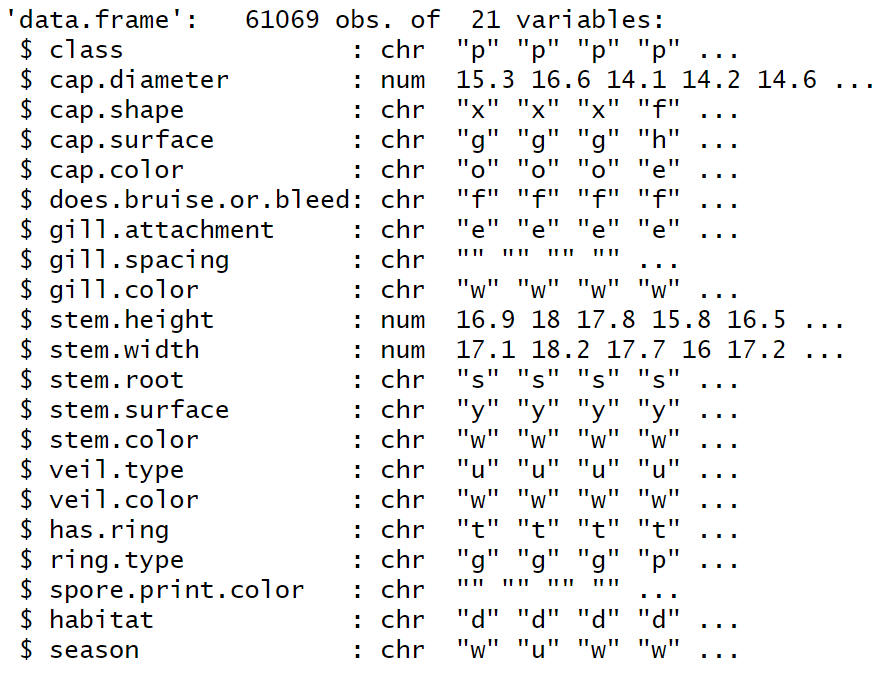
**Key Predictor Variables:**

* **cap-diameter** *(numeric)*: Diameter of the mushroom cap in centimeters
* **cap-shape**: Shape of the cap (e.g., bell, convex, flat, sunken)
* **cap-surface**: Texture of the cap (e.g., fibrous, smooth, scaly, sticky)
* **cap-color**: Color of the cap (e.g., brown, red, white, yellow)
* **does-bruise-or-bleed**: Indicates whether the mushroom bruises or bleeds when damaged
* **gill-attachment**: Describes how the gills are attached to the stem
* **gill-spacing**: Spacing between the gills (e.g., close, distant)
* **gill-color**: Color of the gills
* **stem-height** *(numeric)*: Height of the mushroom stem in centimeters
* **stem-width** *(numeric)*: Width of the stem in millimeters
* **stem-root**: Structure of the stem’s base (e.g., bulbous, cup-shaped, rhizomorphs)
* **stem-surface**: Texture of the stem surface (e.g., smooth, silky, grooved)
* **stem-color**: Color of the stem
* **veil-type**: Type of veil covering the gills (partial or universal)
* **veil-color**: Color of the veil
* **has-ring**: Whether the mushroom has a ring around the stem
* **ring-type**: Type of ring present (e.g., pendant, flaring, grooved, scaly)
* **spore-print-color**: Color left by the mushroom's spore print
* **habitat**: Type of environment the mushroom grows in (e.g., woods, meadows, urban)
* **season**: The season when the mushroom is observed (spring, summer, autumn, winter)

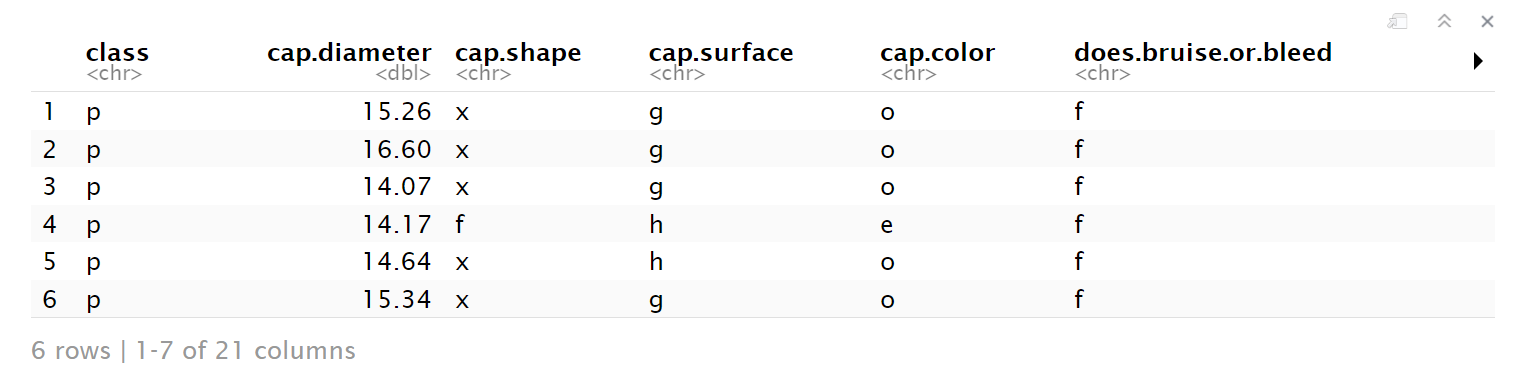
All features are encoded as categorical variables, except for cap-diameter, stem-height, and stem-width, which are numerical. This mix of attribute types makes the dataset highly suitable for a logistic regression classification model.

# 3 EDA:

## 3.1 Structure of Dataset:



First Few Rows of the dataset



## 3.2 Checking for Null Values:



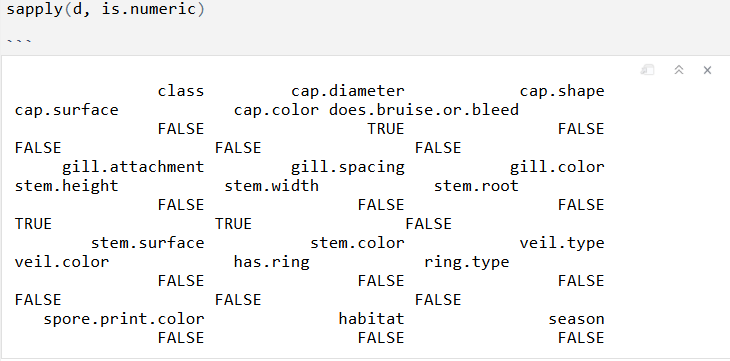
* Issue, where there are no “NA’s”, there are still a lot of blank cells that do not show up, but contain ("")!

## 

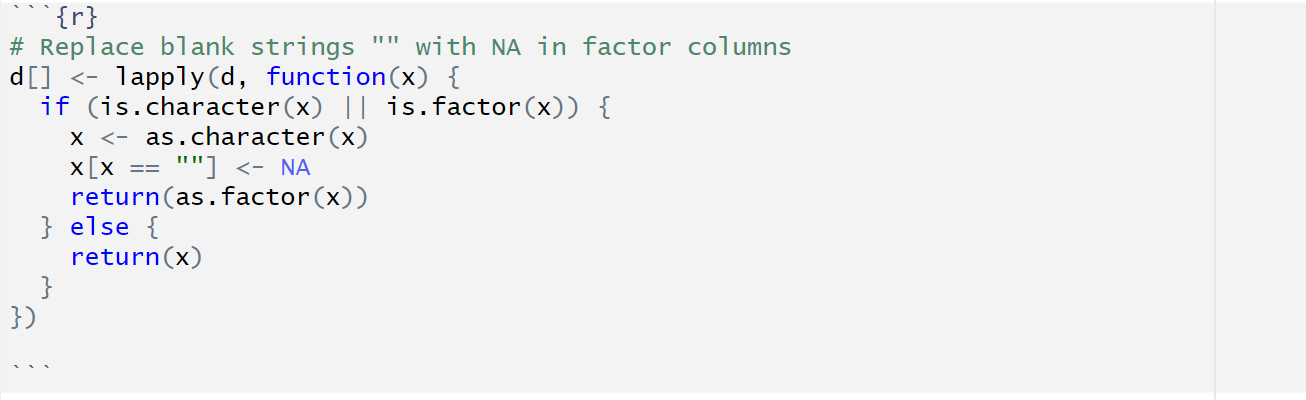
## A summary of the dataset’s numeric features shows that all three variables (cap.diameter, stem.height, and stem.width) have considerable range and potential outliers, especially in cap.diameter and stem.width, suggesting the importance of scaling and checking for skewness before modeling.

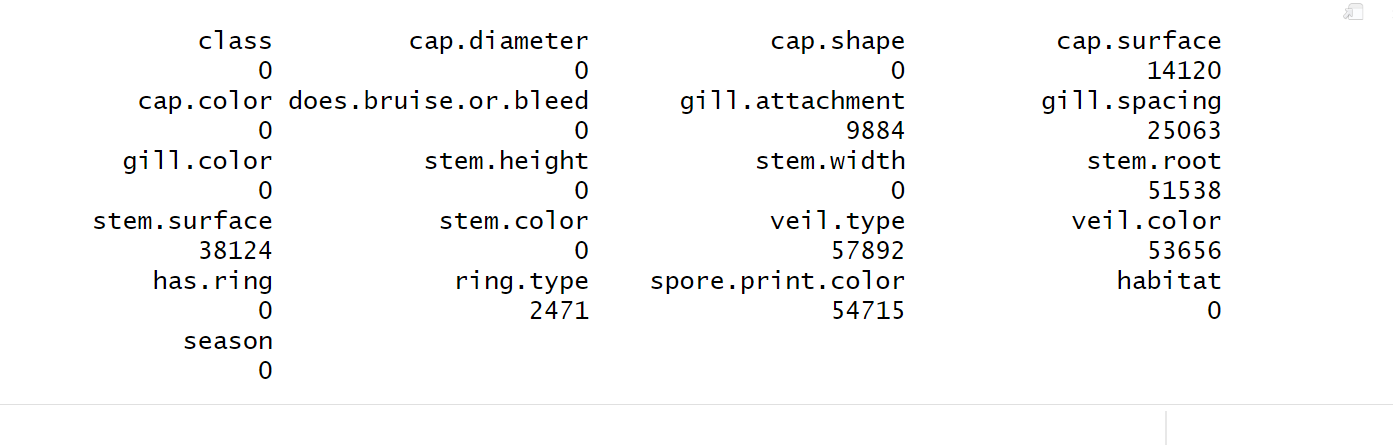
Checking for numeric columns:-

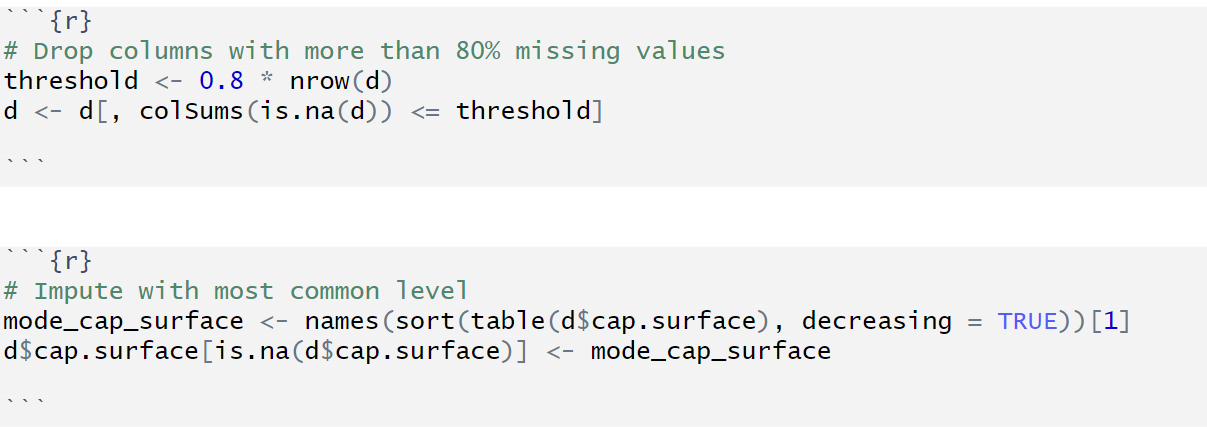
Only three variables—cap.diameter, stem.height, and stem.width—are numeric; the rest are categorical.



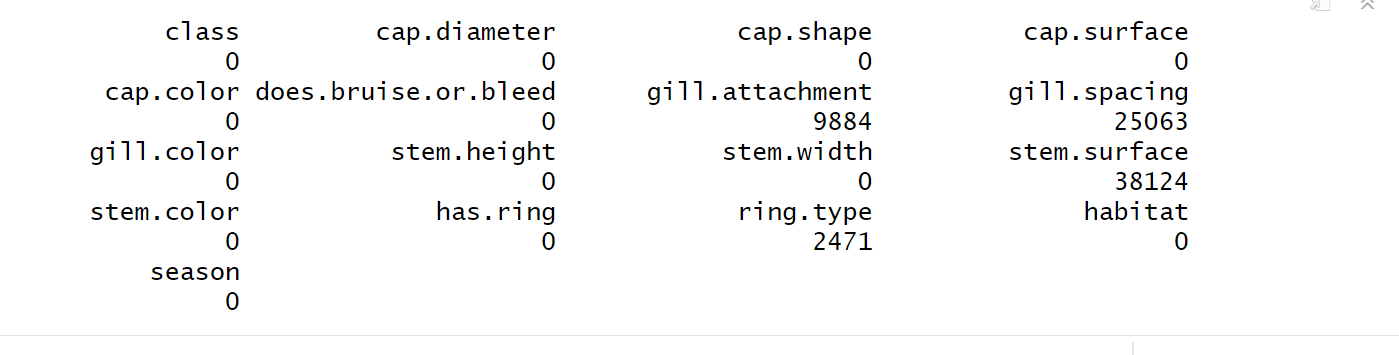
## 3.3 Fixing Null Values:

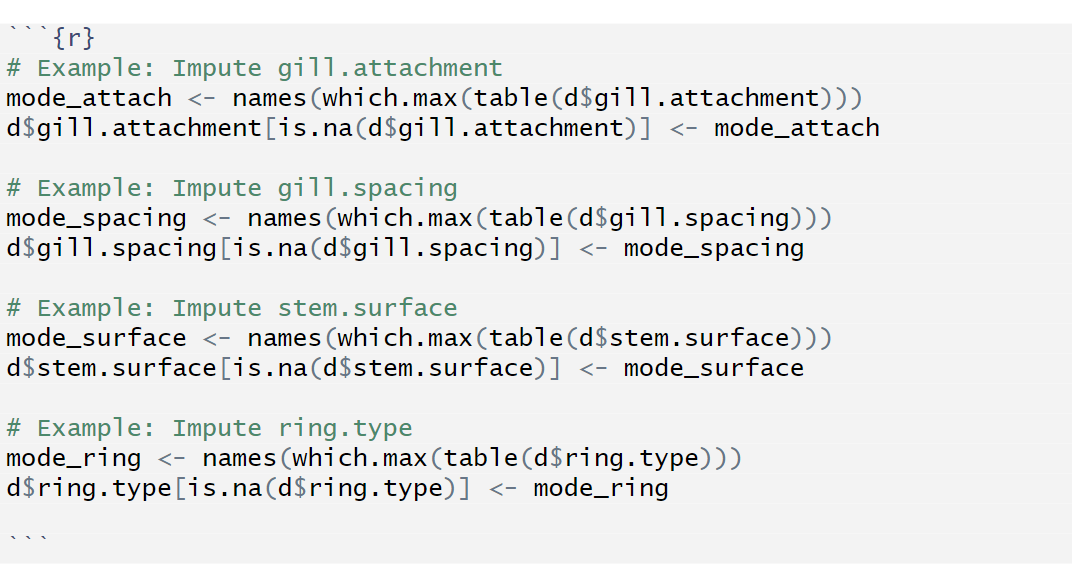


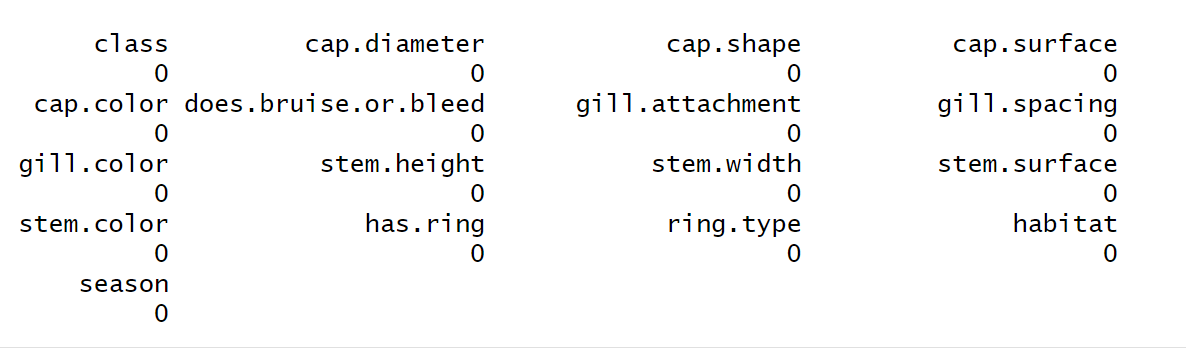




Columns with more than 80% missing values were dropped. For remaining columns, missing values (like in cap.surface) were imputed using the most frequent category (mode).

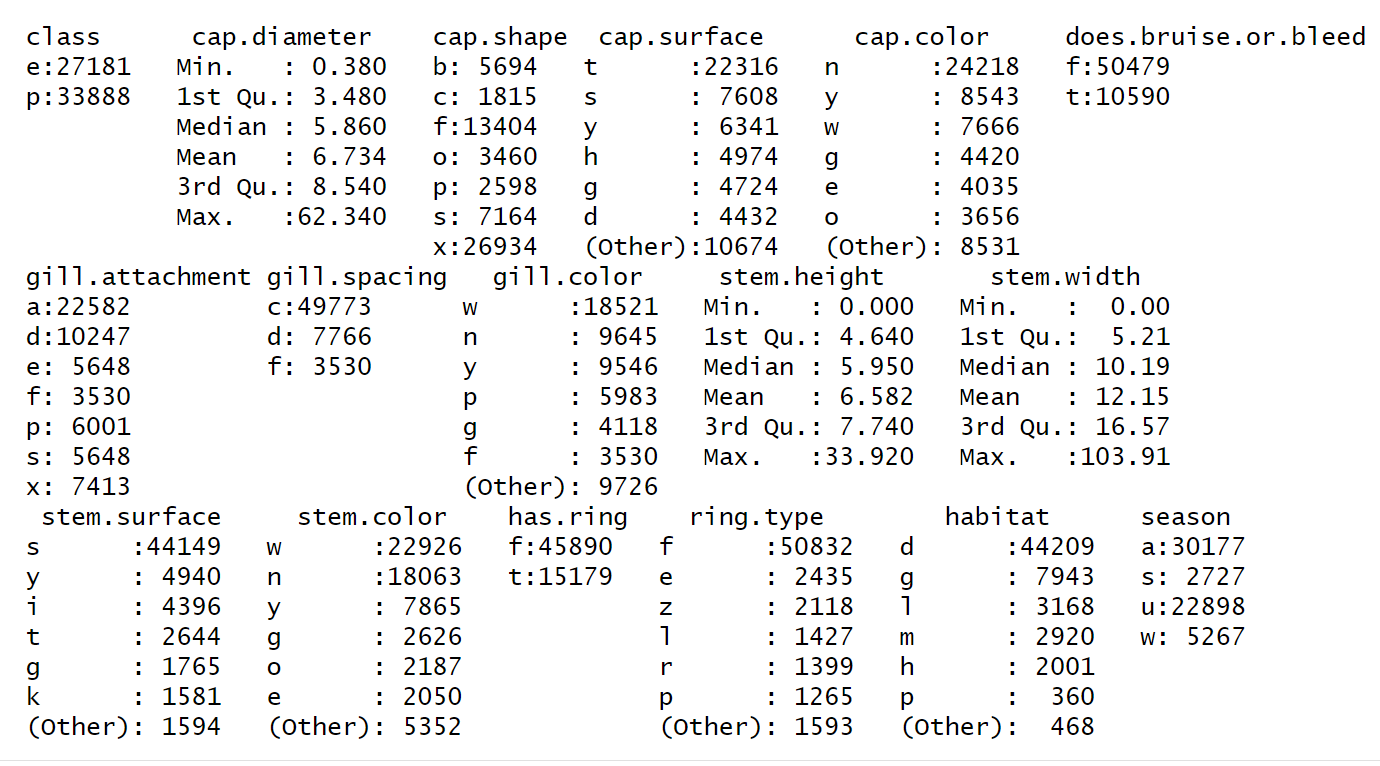






After one last round of imputing, the variables gill.attachment, gill.spacing, stem.surface, and ring.type are imputed using this method. The top portion of the image also summarizes the number of missing values in each column, highlighting which features required imputation.

## 3.4 Statistical Summary:

Continuous features like cap.diameter, stem.height, and stem.width are summarized using descriptive statistics (mean, median, quartiles), while categorical variables such as cap.shape, gill.attachment, and ring.type are broken down by frequency. This dual-format summary helps identify class imbalances, detect unusual values (e.g., a maximum cap.diameter of 62.340), and spot potential data quality issues—crucial steps before applying any machine learning models.

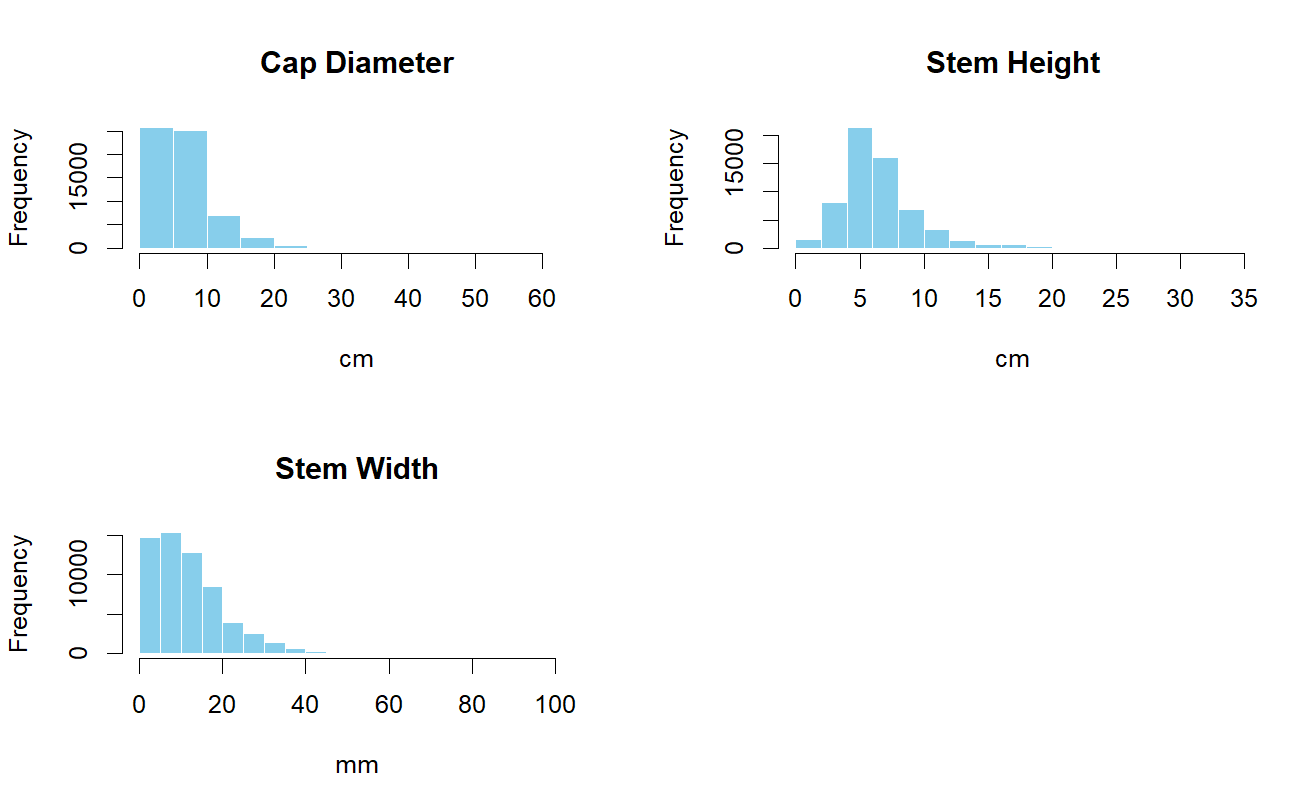
## 

## 

## 

## 3.5 Frequency Distribution:

The histograms visualize the distributions of key continuous features: Cap Diameter, Stem Height, and Stem Width. All three variables exhibit right-skewed distributions, indicating that while most mushroom specimens fall within a lower range of measurements, there are a few outliers with significantly larger dimensions. These visualizations help identify potential anomalies (e.g., unusually large stem widths approaching 100 mm) and reinforce the need for normalization or transformation before applying algorithms sensitive to scale and skewness.



## 

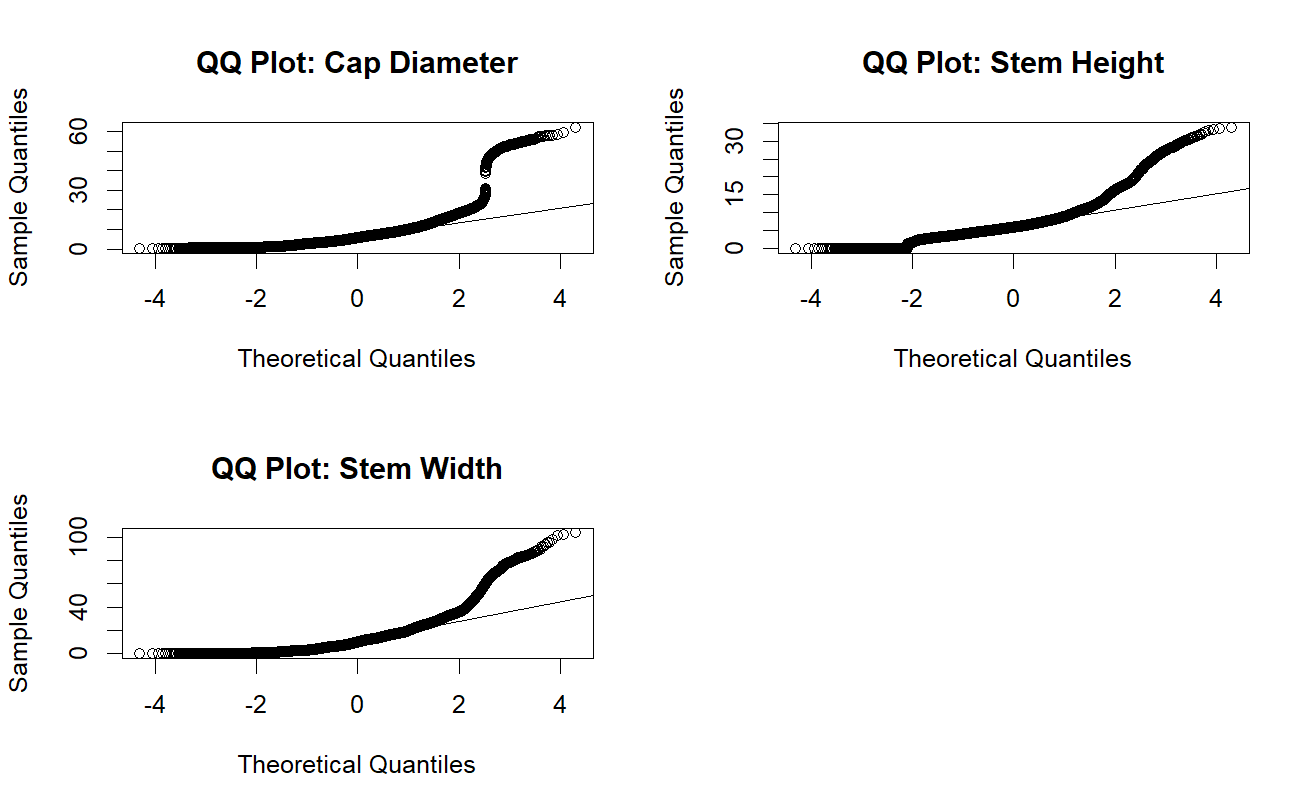
## 

## 3.6 Check for Normality:

To assess whether the variables follow a normal distribution, we used Quantile-Quantile (QQ) plots for *Cap Diameter*, *Stem Height*, and *Stem Width*. A QQ plot compares the distribution of the data to a theoretical normal distribution. If the data points lie roughly along the reference line, the distribution is approximately normal.

From the plots above:

* All three variables show noticeable deviations from the straight line, particularly at the tails.
* This suggests that *Cap Diameter*, *Stem Height*, and *Stem Width* do not follow a normal distribution.
* As a result, non-parametric methods or data transformations may be needed for further analysis.

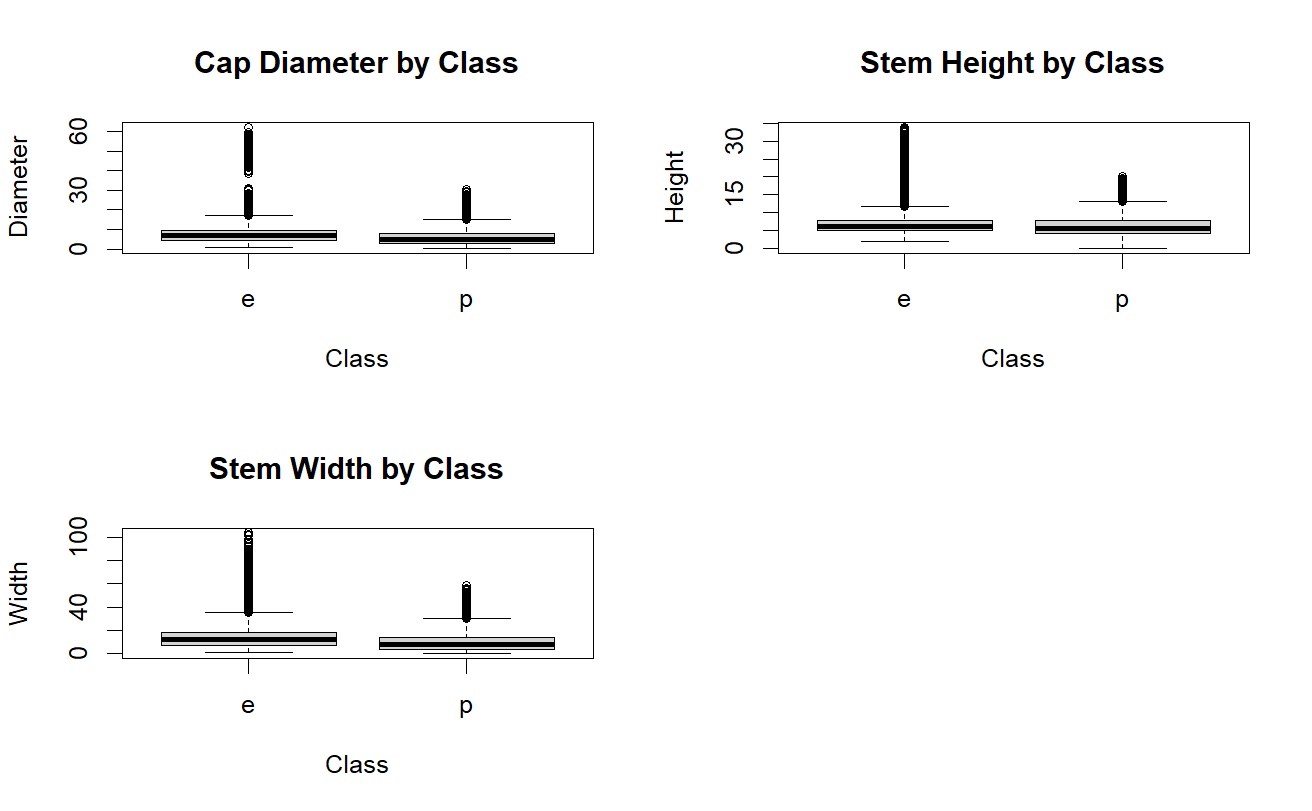


## 

## 3.7 Pairwise Relationship:

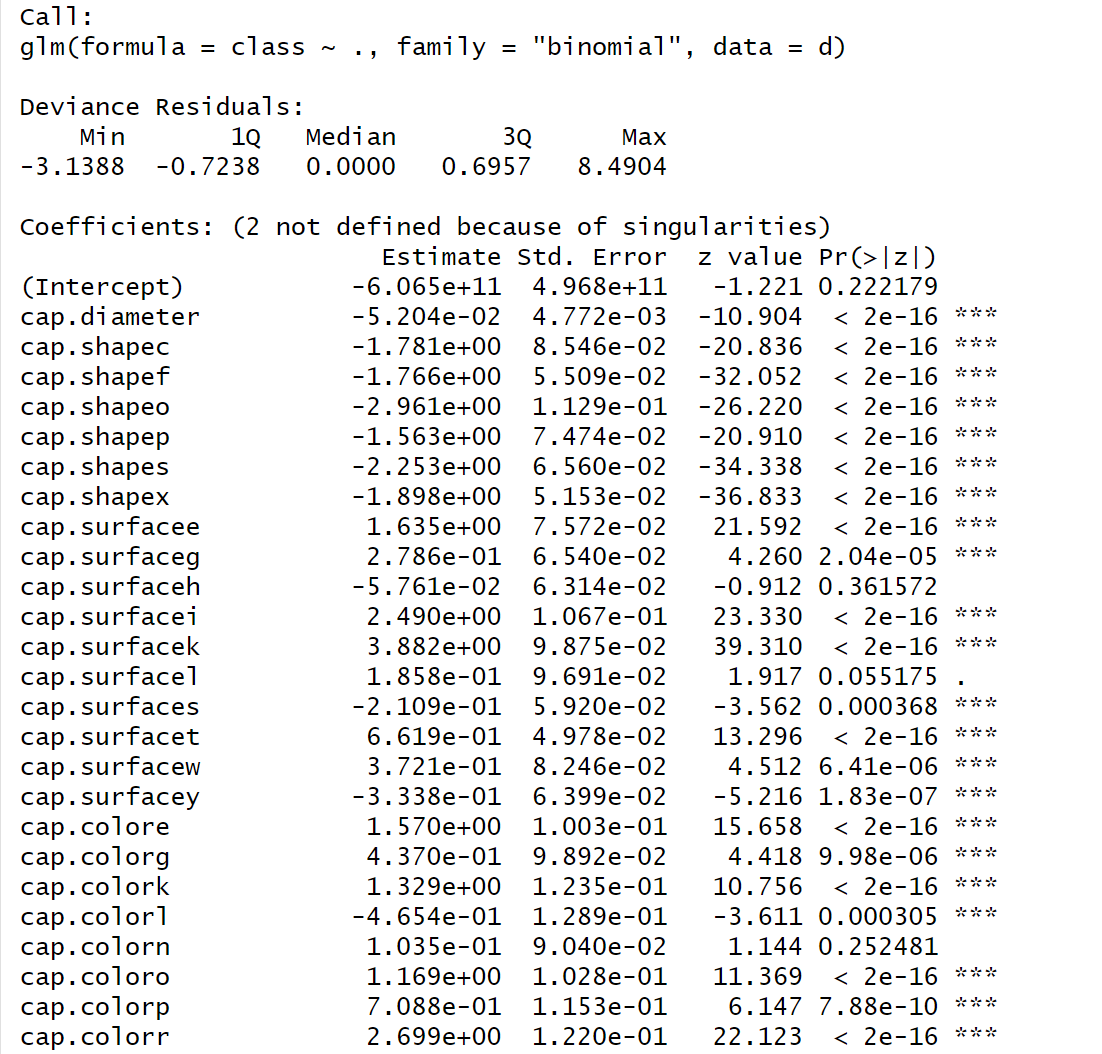
To explore how the features vary by class, we used boxplots to visualize the distributions of *Cap Diameter*, *Stem Height*, and *Stem Width* across the two classes: edible (e) and poisonous (p).

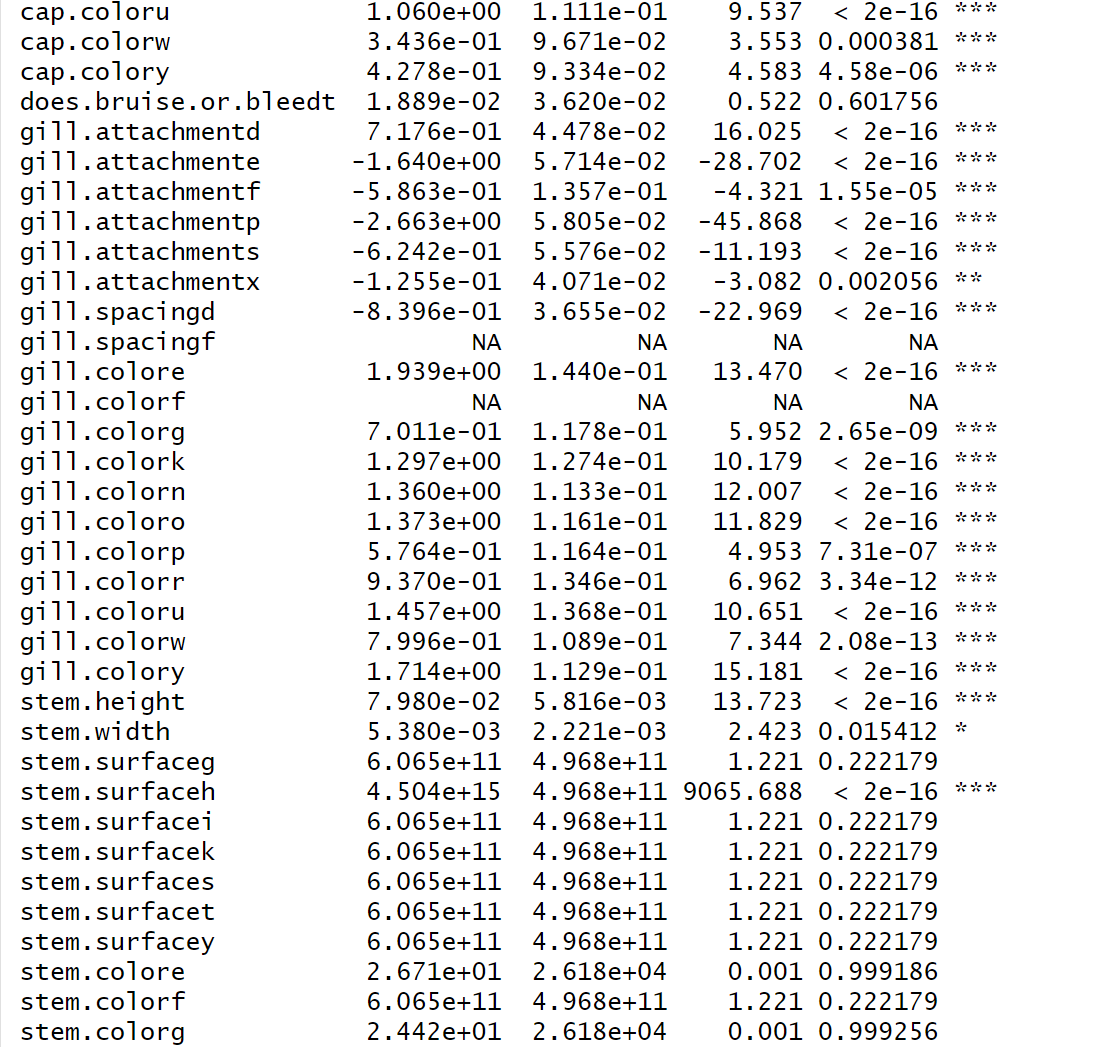
* The plots show slight differences in the distribution of all three variables between the two classes.
* While the medians are relatively close, the spread and outliers differ, particularly for *Cap Diameter* and *Stem Width*.
* These visualizations help identify potential patterns or distinguishing characteristics between edible and poisonous mushrooms.

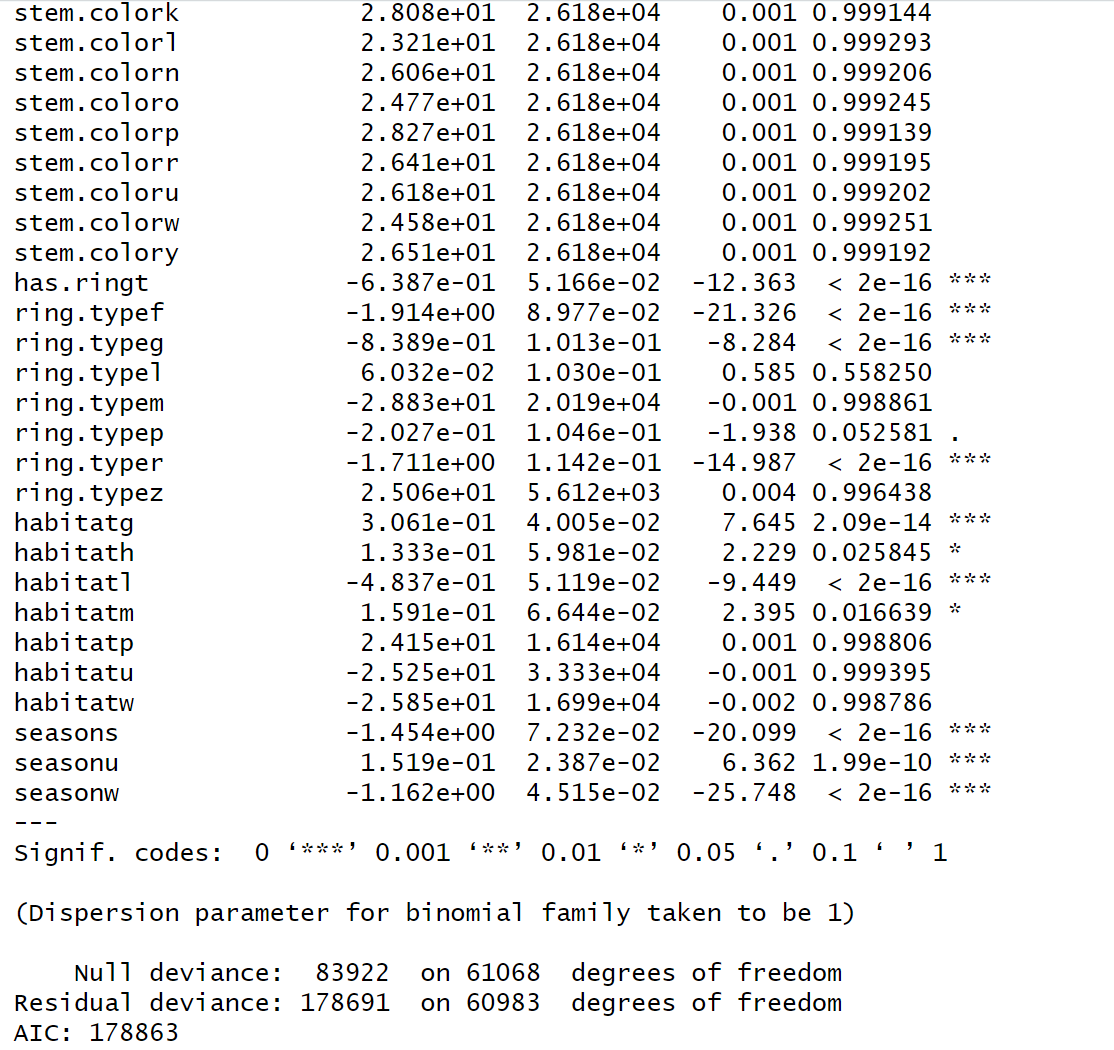


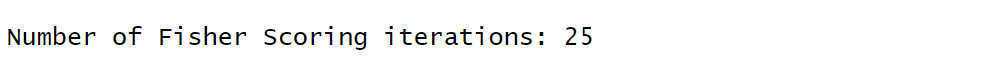
# 4 Model Building:

## 4.1 Baseline Logistic Regression Model:









## 

## 

## 4.2 Interpretation of Coefficients:

## **Statistical Significance and Effect Direction**

1. **Highly significant predictors**: Many variables display \*\*\* significance markers, indicating p-values below 0.001. This suggests these features have very strong statistical relationships with the classification outcome. Nearly all gill attachment types, most cap surface features, and several color variables fall into this category.
2. **Effect direction interpretation**:
   * Negative coefficients decrease the log odds of the positive outcome category
   * Positive coefficients increase the log odds of the positive outcome category
   * For example, cap.shapeo (-2.961) substantially decreases the probability of a positive classification, while cap.surfacek (3.882) dramatically increases it

### **Specific Feature Categories**

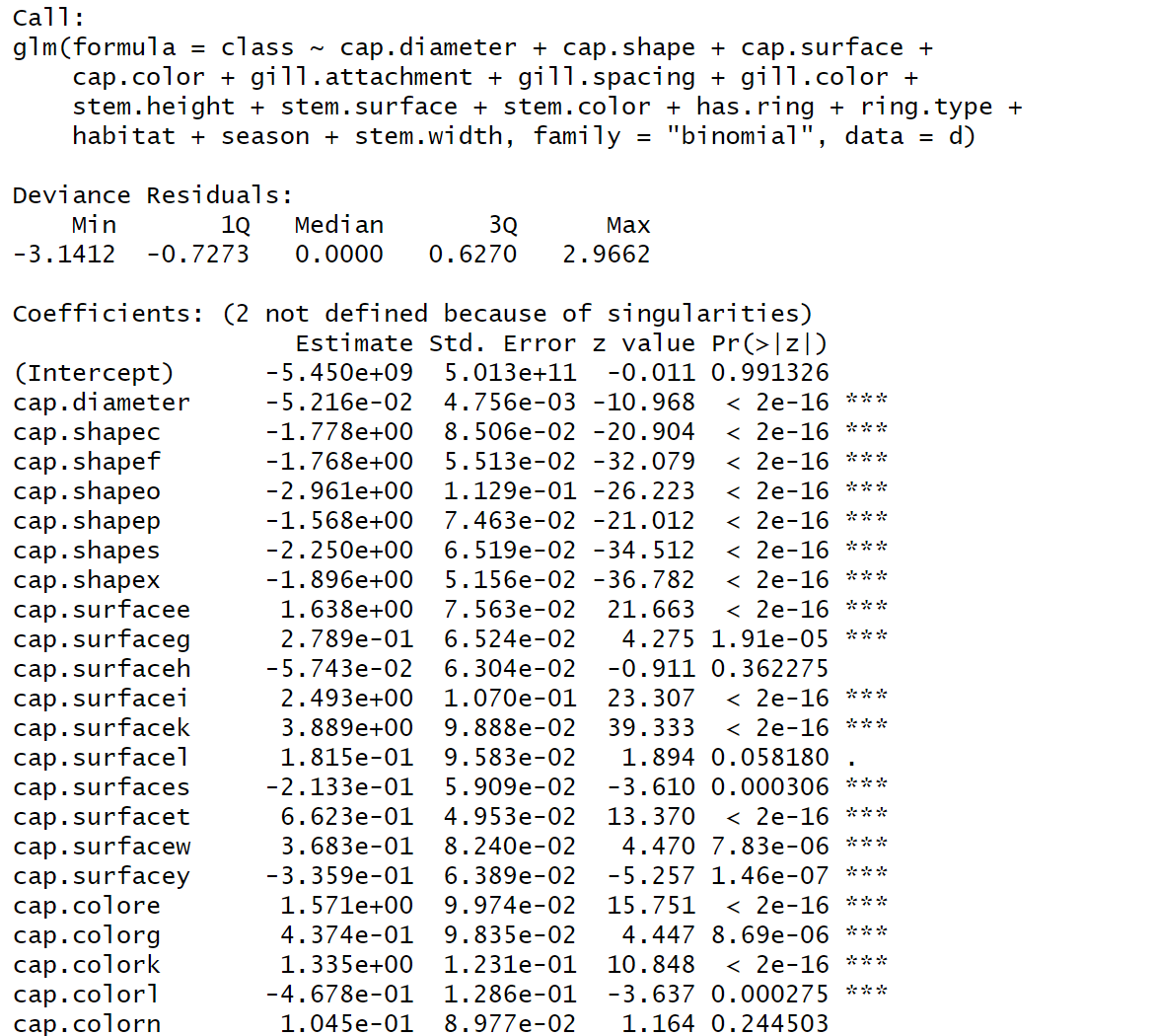
1. **Cap characteristics**:
   * Cap diameter shows a negative effect (-5.204e-02), suggesting smaller caps are associated with the positive outcome
   * Cap shape variables predominantly show negative coefficients, with cap.shapes (-2.253) and cap.shapex (-1.898) having particularly strong negative effects
   * Cap surface characteristics show varied effects, with some strongly positive (cap.surfacek: 3.882) and others strongly negative
2. **Gill characteristics**:
   * Gill attachments show some of the strongest effects in the model, with gill.attachmentp (-2.689) and gill.attachmente (-1.640) being notably negative
   * Gill spacing (gill.spacingd: -8.396e-01) significantly impacts classification
   * Gill color variables show predominantly positive effects, with gill.colory (1.714) and gill.colore (1.939) having particularly strong positive associations
3. **Stem characteristics**:
   * Most stem surface variables show extremely large coefficients with high standard errors, suggesting potential model estimation issues
   * Stem width (5.380e-03) has a statistically significant but relatively small effect compared to other features
4. **Habitat and season**:
   * Habitat variables show moderate effects, with habitatg (3.061e-01) having a significant positive effect
   * Seasonal factors are strongly significant, with seasons (-1.454) showing a strong negative effect and seasonu (1.519e-01) showing a positive effect

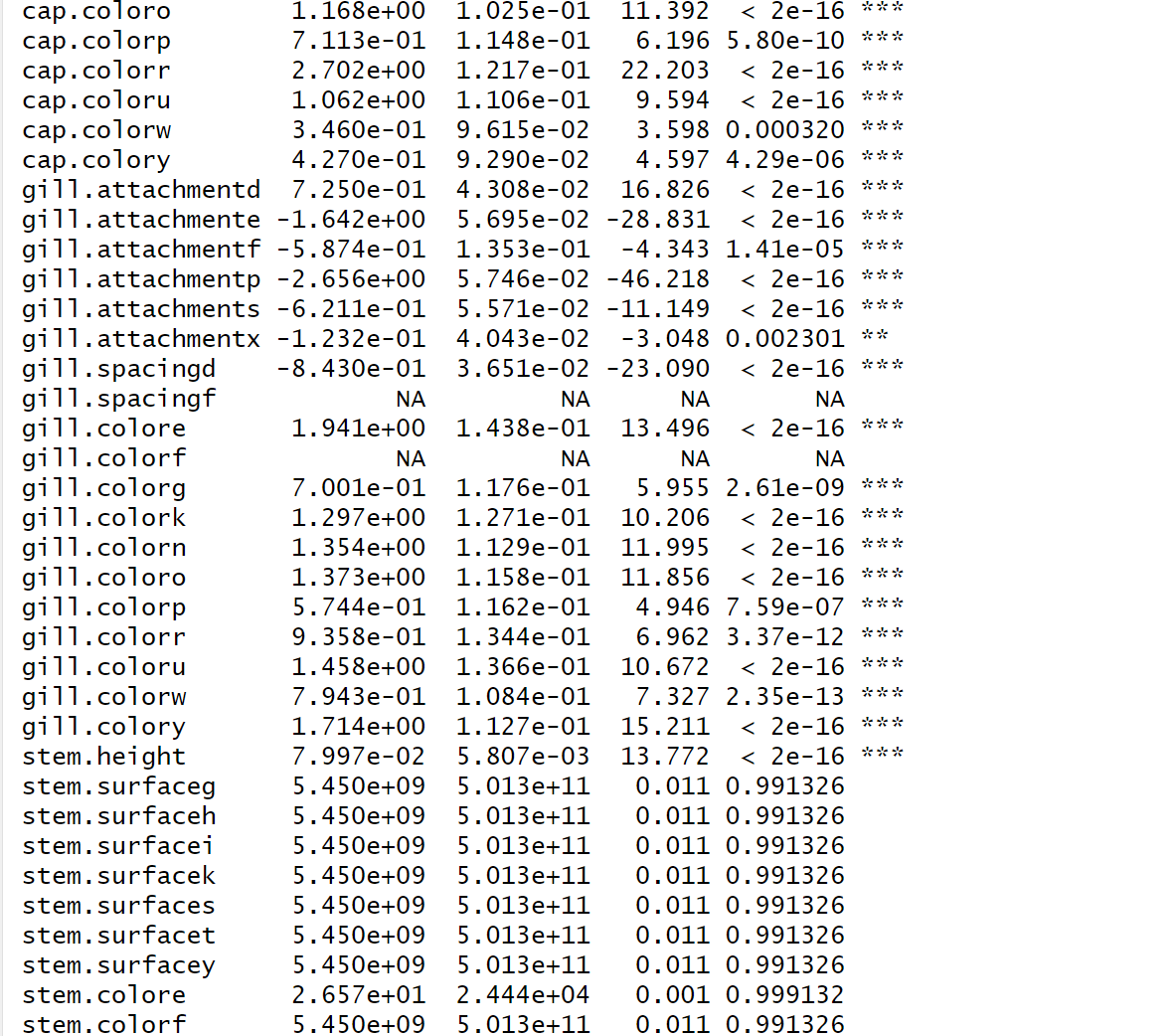
### **Model Diagnostics**

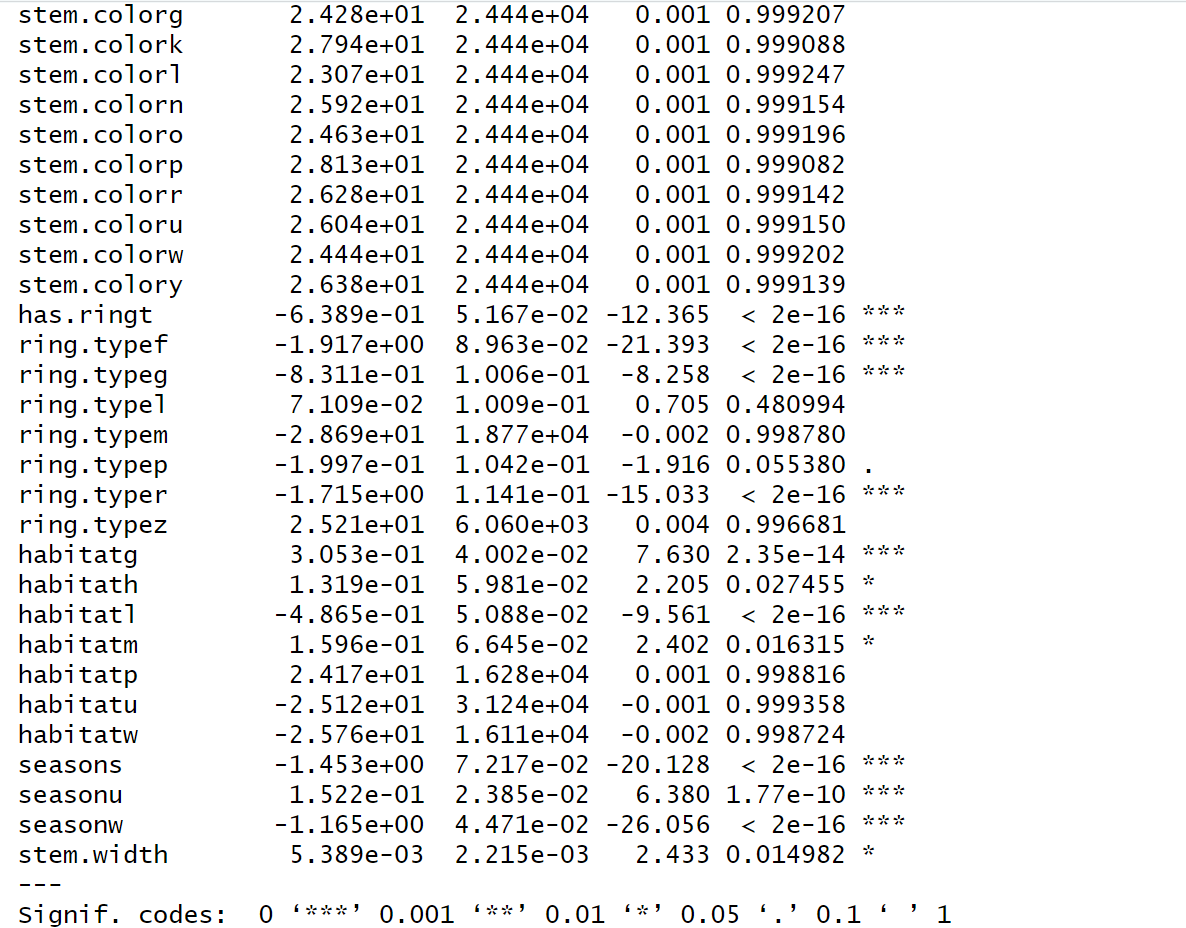
1. **Model fit assessment**:
   * The null deviance (83922) compared to residual deviance (178691) suggests the model explains some variance, but substantial unexplained variation remains
   * AIC of 178863 indicates a complex model with many parameters
   * The note about "2 not defined because of singularities" suggests some collinearity issues in the predictors
2. **Practical implications**:
   * Features with the largest absolute coefficient values have the strongest influence on classification outcomes
   * The model identifies specific characteristics that could serve as key visual indicators for mushroom classification
   * The dispersion parameter fixed at 1 is consistent with the binomial family assumption

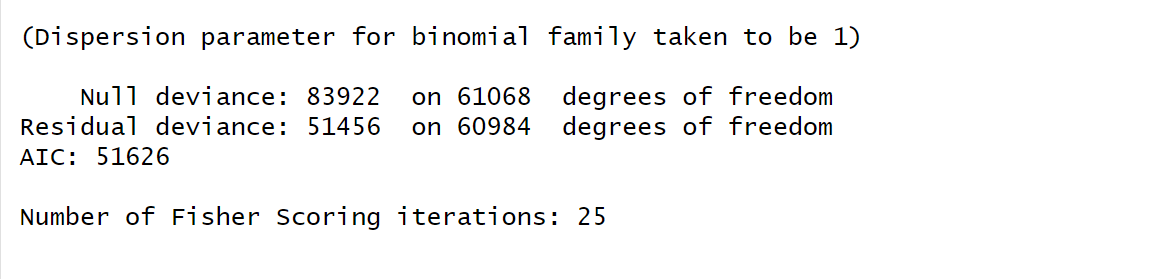
This comprehensive analysis of coefficients reveals the complex interplay of various mushroom morphological features in determining classification. The strongest predictors appear to be related to gill attachment patterns, cap surface textures, certain color variations, and seasonal growth patterns, which collectively form a statistical basis for distinguishing between mushroom categories.

## 4.3 Stepwise Selection Model:









## 

## 

## 4.4 Interpretation of Coefficients Stepwise Model:

1. **Cap characteristics**:
   * Cap diameter has a significant negative effect (-5.216e-02), suggesting smaller caps slightly increase likelihood of a particular classification
   * Cap shape variables show strong negative associations, especially cap.shapes (-2.250) and cap.shapeo (-2.961)
   * Cap surface variables show mixed effects with some strongly positive (cap.surfacek: 3.889) and others negative
   * Several cap colors significantly impact classification, with cap.colorr (2.702) having one of the strongest positive effects
2. **Gill characteristics**:
   * Gill attachment types strongly predict classification, with gill.attachmentp (-2.655) having a large negative effect
   * Gill colors show predominantly positive coefficients, with gill.colory (1.714) having the largest positive effect
   * Gill spacing (gill.spacingd: -8.430e-01) significantly reduces probability of the outcome
3. **Other features**:
   * Ring characteristics are important predictors, especially ring.typer (-1.715) and ring.typef (-1.917)
   * Seasonal patterns strongly influence classification, with seasons (-1.454) and seasonw (-1.165) both negative
   * Habitat variables show mixed effects of moderate strength
4. **Model fit**:
   * Improved from previous model with AIC reduced to 51626 (from 178863)
   * Null deviance (83922) vs. residual deviance (51456) indicates the model explains a substantial portion of the variation

The coefficients represent changes in log odds of classification when a particular feature is present. The most predictive features appear to be cap shape, gill attachment type, cap surface, and specific colors across different mushroom parts.

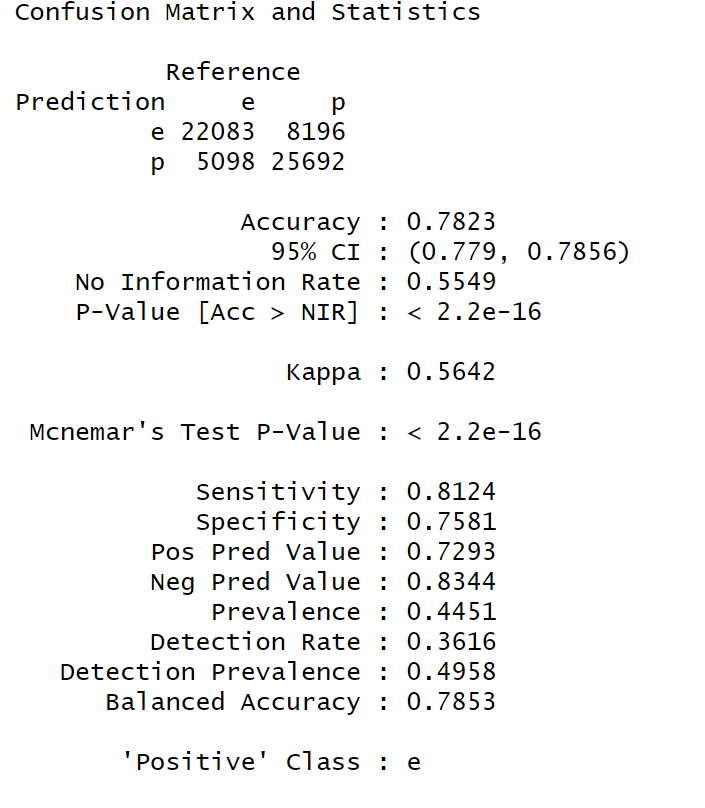
# 5 Model Evaluation:

## 5.1 Confusion Matrix Logistic Regression:

# **Interpretation of Confusion Matrix**

* **Accuracy: 78.23%** with 95% CI (77.9%, 78.56%)
* **Sensitivity: 81.24%** - Strong at identifying edible mushrooms
* **Specificity: 75.81%** - Reasonably good at identifying non-edible mushrooms
* **PPV: 72.93%** - When predicting "edible," correct 73% of the time
* **NPV: 83.44%** - When predicting "non-edible," correct 83% of the time
* **Kappa: 0.5642** - Moderate agreement beyond chance

Matrix shows 22,083 true positives, 25,692 true negatives, with 8,196 false positives and 5,098 false negatives. The model performs significantly better than the baseline (p < 2.2e-16), with balanced performance across both classes.



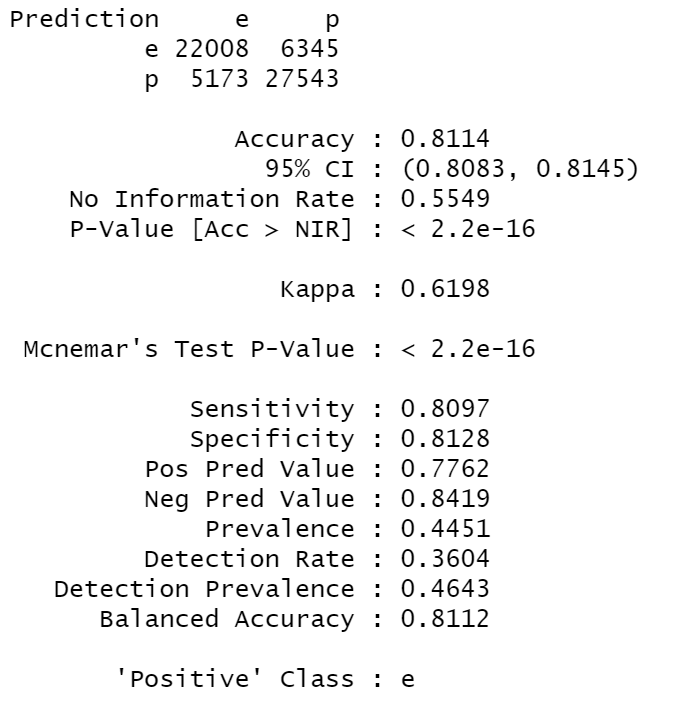
## 5.2 Confusion Matrix Stepwise Selection:

# **Interpretation of Confusion Matrix**

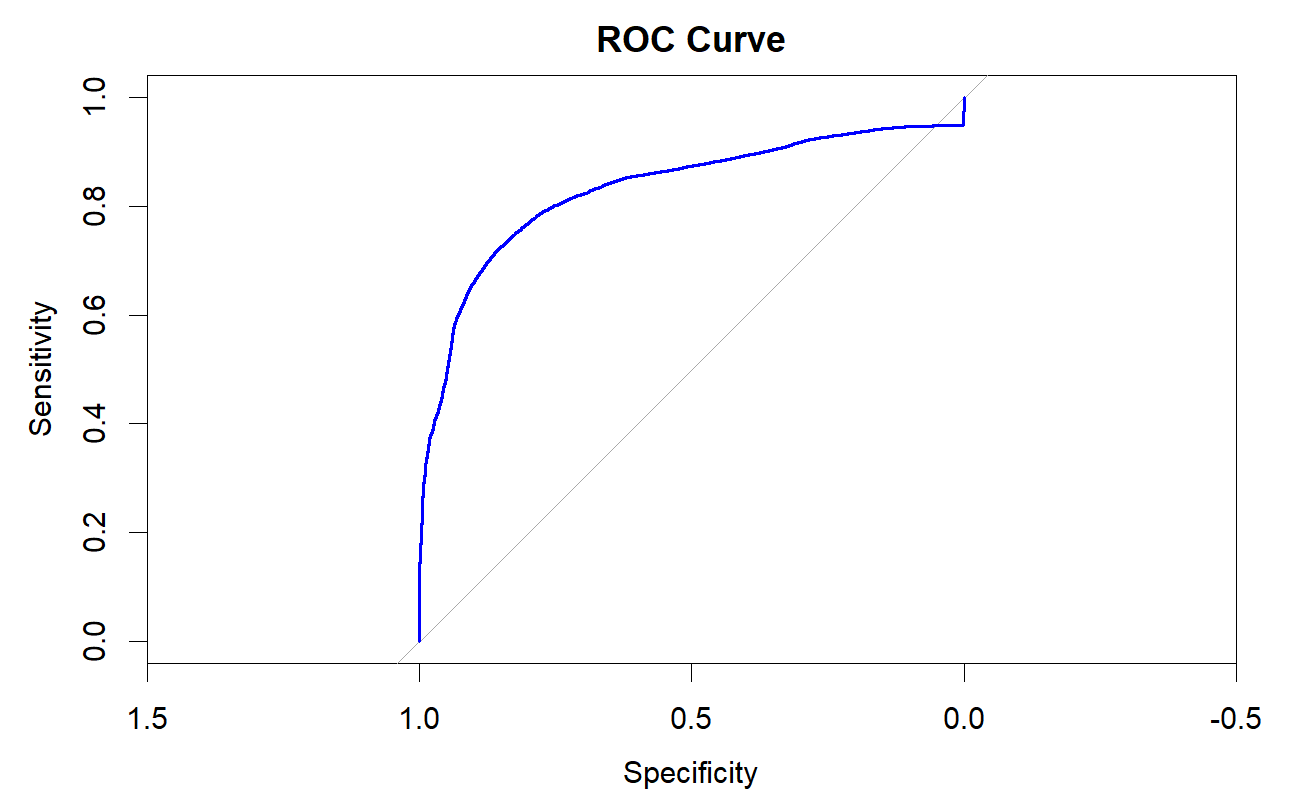
This confusion matrix shows an improved classification model with:

* **Accuracy: 81.14%** (CI: 80.83%-81.45%) - Better than the previous model
* **Sensitivity: 80.97%** - Strong identification of edible mushrooms
* **Specificity: 81.28%** - Improved identification of non-edible mushrooms
* **PPV: 77.62%** - Higher precision when predicting "edible"
* **NPV: 84.19%** - Reliable when predicting "non-edible"
* **Kappa: 0.6198** - Substantial agreement beyond chance

The matrix shows 22,008 true positives and 27,543 true negatives, with fewer misclassifications (6,345 false positives, 5,173 false negatives) than the previous model. This model demonstrates more balanced performance with nearly identical sensitivity and specificity, suggesting reliable predictions for both edible and poisonous mushrooms.



## 5.3 ROC & AUC Logistic Regression:



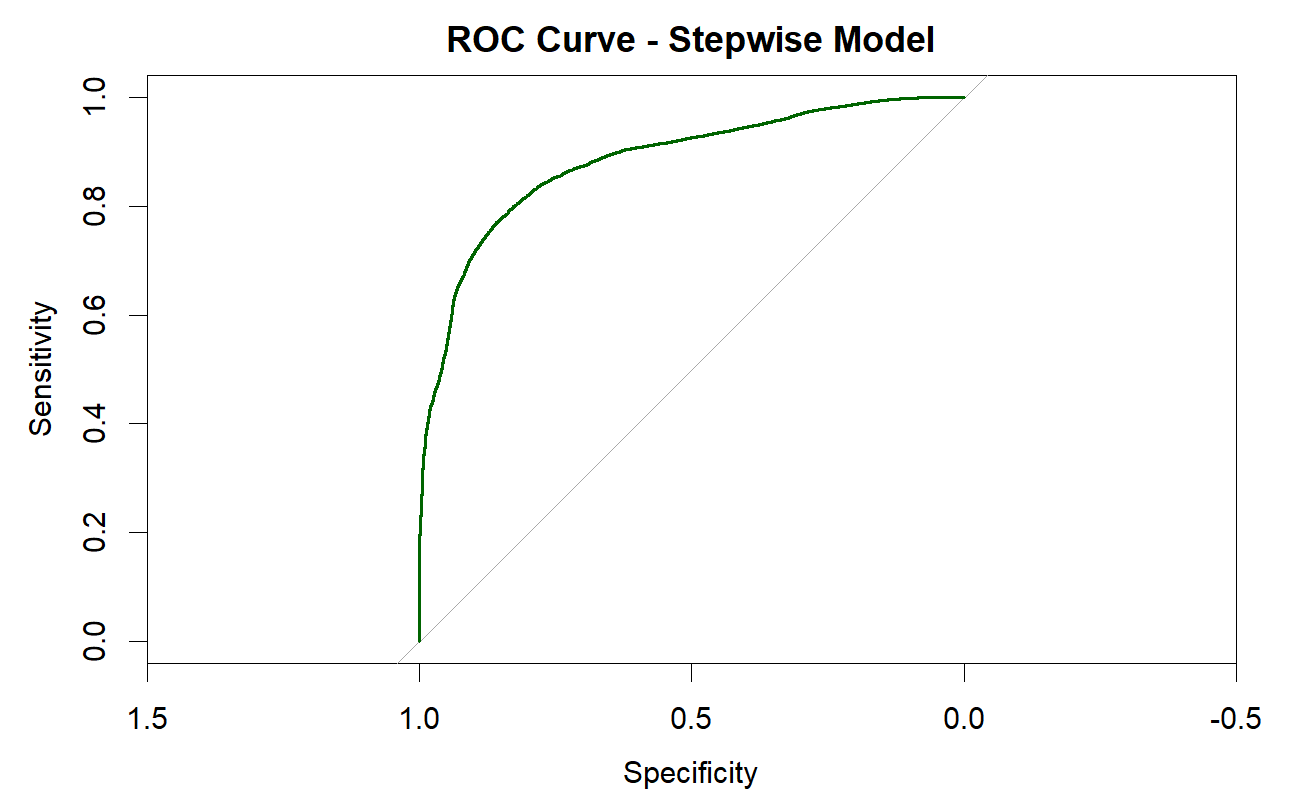
AUC: 0.8315

# **Interpretation of ROC Curves**

## **Initial model**

* **AUC: 0.8315** - Shows good discriminatory power
* The blue curve demonstrates how the model balances sensitivity and specificity across different threshold values
* Strong lift from the diagonal reference line indicates the model performs much better than random guessing
* The curve shows relatively steep improvement in the high-specificity region (left side)

## 5.4 ROC & AUC Stepwise Selection:



AUC Stepwise Model - 0.8835

## **Stepwise Model**

* **AUC: 0.8835** - Shows excellent discriminatory power
* The green curve represents an improved model using stepwise feature selection
* Higher AUC (0.8835 vs 0.8315) indicates this model is more effective at distinguishing between edible and poisonous mushrooms
* The curve reaches closer to the top-left corner (ideal classification point)

The stepwise model offers substantial improvement over the initial model, suggesting that feature selection effectively identified the most predictive mushroom characteristics. With an AUC of 0.8835, the stepwise model demonstrates strong predictive performance that would be highly reliable for mushroom classification.

# 6 Moving Forward:

In advancing our model for predicting customer satisfaction based on service attributes, we plan to enhance its predictive performance and business relevance through a structured refinement process. By leveraging model selection criteria such as the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), we aim to achieve an optimal balance between explanatory power and simplicity. Stepwise selection techniques—forward selection, backward elimination, and bidirectional elimination—will guide the identification of key service factors that drive satisfaction, minimizing noise and avoiding overfitting. To ensure model stability and applicability across different market segments, we will employ cross-validation techniques for robust evaluation and generalizability.