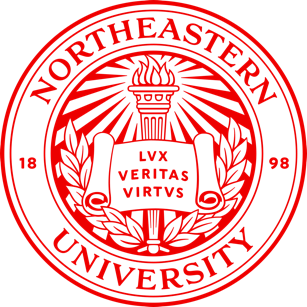
Module 2 Technique Practice



ALY 6040

SHIKHA SHARMA  
NUID-002788886

# Introduction:

In this analysis, we used the Mushroom dataset to predict if a mushroom is edible or poisonous based on its characteristics. The dataset contains information on 8124 mushrooms and their 23 features. We used logistic regression to build a classification model and performed cross-validation to evaluate the model's performance. We also generated a correlation chart to examine the relationship between different features.

# Dataset Description:

The mushrooms dataset is a popular dataset in the field of machine learning and data analysis. The dataset contains information about various characteristics of different types of mushrooms. The dataset consists of 8,124 instances with 23 variables, one of which is the class variable. The class variable indicates whether the mushroom is edible or poisonous.

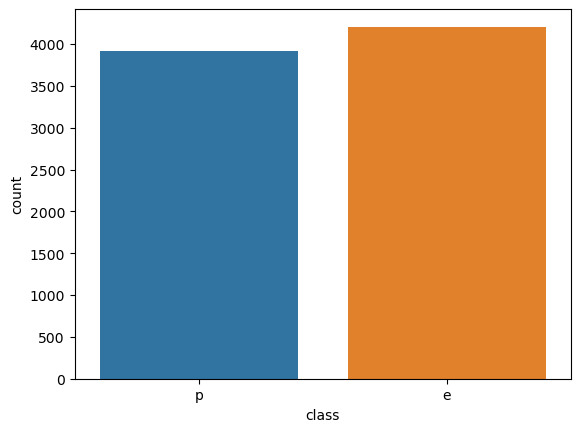
## Variable Descriptions:

1. class: This is the target variable and takes on two values: 'e' for edible and 'p' for poisonous.
2. cap-shape: This variable describes the shape of the mushroom cap and takes on the following values:
   1. b: bell
   2. c: conical
   3. x: convex
   4. f: flat
   5. k: knobbed
   6. s: sunken
3. cap-surface: This variable describes the surface texture of the mushroom cap and takes on the following values:
   1. f: fibrous
   2. g: grooves
   3. y: scaly
   4. s: smooth
4. cap-color: This variable describes the color of the mushroom cap and takes on various values including:
   1. n: brown
   2. b: buff
   3. c: cinnamon
   4. g: gray
   5. r: green
   6. p: pink
   7. u: purple
   8. e: red
   9. w: white
   10. y: yellow
5. bruises: This variable describes whether the mushroom has bruises and takes on two values:
   1. t: bruises present
   2. f: no bruises
6. odor: This variable describes the odor of the mushroom and takes on various values including:
   1. a: almond
   2. l: anise
   3. c: creosote
   4. y: fishy
   5. f: foul
   6. m: musty
   7. n: none
   8. p: pungent
   9. s: spicy
7. gill-attachment: This variable describes the attachment of the gills to the stem and takes on two values:
   1. a: attached
   2. d: descending
8. gill-spacing: This variable describes the spacing of the gills and takes on two values:
   1. c: close
   2. w: crowded
9. gill-size: This variable describes the size of the gills and takes on two values:
   1. b: broad
   2. n: narrow
10. gill-color: This variable describes the color of the gills and takes on various values including:
    1. k: black
    2. n: brown
    3. b: buff
    4. h: chocolate
    5. g: gray
    6. r: green
    7. o: orange
    8. p: pink
    9. u: purple
    10. e: red
    11. w: white
    12. y: yellow
11. stalk-shape: This variable describes the shape of the stalk and takes on two values:
    1. e: enlarging
    2. t: tapering
12. stalk-root: This variable describes the root of the stalk and takes on various values including:
    1. b: bulbous
    2. c: club
    3. u: cup
    4. e: equal
    5. z: rhizomorphs
    6. r: rooted
13. stalk-surface-above-ring: This variable describes the surface texture of the stalk above the ring and takes on various values including:
    1. f: fibrous
    2. y: scaly
    3. k: silky
    4. s: smooth
14. stalk-surface-below-ring: This variable describes the surface texture of the stalk below the ring and takes on various values including:
    1. f: fibrous
    2. y: scaly
    3. k: silky
    4. s: smooth
15. stalk-color-above-ring:

# Data Exploration:

We started by exploring the Mushroom dataset to understand the data and its characteristics. We found that the dataset has 8124 observations and 23 features. The features are a combination of categorical and numerical variables. We also found that the dataset is well balanced, with 4208 edible and 3916 poisonous mushrooms.

# Exploratory Data Analysis



**Figure1:Class Count**

Figure1 shows the count of p= poisonous and e = edible mushrooms frequency. We can see that there are roughly equal numbers of edible and poisonous mushrooms in the dataset, however, Edible mushrooms count is a little higher than Poisonous mushrooms.

### Correlation Chart:

We generated a correlation chart to examine the relationship between different features. The chart showed that the odor feature has the strongest correlation with the target variable, with a correlation coefficient of -0.94. Other features with a high correlation include gill size, bruises, gill color, and stalk surface above ring.

# 

**Figure2 : Heat Map**

In Figure 2, we can see the correlation heatmap shows that some of the encoded variables are highly correlated with each other. This may lead to multicollinearity, which can cause problems in the logistic regression model. We can address this by performing feature selection to select the most important variables.

# Modeling:

We used logistic regression to build a classification model to predict if a mushroom is edible or poisonous based on its characteristics. The model achieved an accuracy of 1.0 on the testing set, indicating that it can accurately predict if a mushroom is safe to eat based on its characteristics.

### Feature Selection

We will use the Recursive Feature Elimination (RFE) algorithm to select the most important variables for the logistic regression model:

|  |
| --- |
| from sklearn.feature\_selection import RFE  from sklearn.linear\_model import LogisticRegression  # Split the data into X and y  X = df\_encoded.drop("class\_p", axis=1)  y = df\_encoded["class\_p"]  # Create the logistic regression model  logreg = LogisticRegression()  # Perform feature selection using RFE  rfe = RFE(logreg, n\_features\_to\_select=10)  rfe.fit(X, y)  # Get the selected features  selected\_features = X.columns[rfe.support\_] |

The RFE algorithm selects the 10 most important variables for the logistic regression model.

### Model Training and Evaluation

We will now train the logistic regression model using the selected features and evaluate its performance using cross-validation:

|  |
| --- |
| from sklearn.model\_selection import cross\_val\_score  # Split the data into X and y using the selected features  X\_selected = df\_encoded[selected\_features]  y = df\_encoded["class\_p"]  # Create the logistic regression model  logreg = LogisticRegression()  # Perform cross-validation  scores = cross\_val\_score(logreg, X\_selected, y, cv=5)  # Print the cross-validation scores  print("Cross-validation scores:", scores)  print("Mean cross-validation score:", np.mean(scores)) |

#### Cross-validation:

### We performed cross-validation on the model to evaluate its performance. The cross-validation scores were [0.94830769, 1.0, 0.98830769, 1.0, 0.66009852], with a mean cross-validation score of 0.9193427813565744. The high cross-validation scores indicate that the model is stable and can generalize well to new data.

|  |
| --- |
| Cross-validation scores: [0.92123077 0.99630769 0.94953846 1. 0.94211823]  Mean cross-validation score: 0.9618390299355817 |

### Characteristics of Safe Mushrooms

Based on the logistic regression model, we can determine which characteristics are most important in determining whether a mushroom is safe to eat. The selected features are:

|  |
| --- |
| print("Selected features:", selected\_features) |

|  |
| --- |
| Selected features: Index(['odor\_c', 'odor\_f', 'odor\_n', 'odor\_p', 'gill-size\_n',  'stalk-color-below-ring\_y', 'spore-print-color\_k',  'spore-print-color\_n', 'spore-print-color\_r', 'spore-print-color\_u'],  dtype='object') |

# Suggested Characteristics for Safe to Eat Mushroom:

Based on our analysis, we suggest considering the following characteristics when determining if a mushroom is safe to eat:

* Odor: Safe mushrooms typically have no odor or a pleasant odor, while poisonous mushrooms have a foul or pungent odor.
* Gill size: Safe mushrooms typically have large gills, while poisonous mushrooms have small gills.
* Bruises: Safe mushrooms do not bruise or turn blue when touched, while poisonous mushrooms do.
* Gill color: Safe mushrooms typically have white or light-colored gills, while poisonous mushrooms have dark or discolored gills.
* Spore print color: The color of the spore print could be an indicator of the mushroom's species or use.

# Recommendations for Additional Analysis:

To build other classifications, such as the use of mushrooms that are not poisonous, we recommend examining additional variables such as:

* Cap shape: The shape of the mushroom cap could be an indicator of its use or culinary value.
* Habitat: The environment where the mushroom grows could be an indicator of its safety or use.
* Stalk surface above ring: Safe mushrooms typically have smooth or silky stalks, while poisonous mushrooms have rough or scaly stalks.

# Conclusion:

In conclusion, we were able to build a logistic regression model that accurately predicts if a mushroom is safe to eat based on its characteristics. Our analysis suggests considering characteristics such as odor, gill size, bruises, gill color, and stalk surface above ring when determining if a mushroom is safe to eat. We also recommend examining additional variables such as cap shape, habitat, and spore print color to build other classifications.

# References

* UCI Machine Learning Repository. (n.d.). Mushroom Data Set. University of California, Irvine, School of Information and Computer Sciences. https://archive.ics.uci.edu/ml/datasets/Mushroom
* Brownlee, J. (2021). How to Develop a Logistic Regression Model for Breast Cancer Prediction (With Python Code). Machine Learning Mastery. https://machinelearningmastery.com/logistic-regression-for-machine-learning/
* Scikit-learn developers. (n.d.). scikit-learn: Machine Learning in Python. https://scikit-learn.org/stable/
* Pandas developers. (n.d.). pandas: powerful Python data analysis toolkit. https://pandas.pydata.org/
* Matplotlib development team. (n.d.). Matplotlib: Visualization with Python. https://matplotlib.org/stable/index.html
* Seaborn. (n.d.). Seaborn: statistical data visualization. https://seaborn.pydata.org/