**Mushroom Classification**

By: Sonali Shintre

I chose to go with the Classification problem:

1. **Problem Statement:**

Using mushroom data set to classify

* Mushroom are poisonous or edible
  + What types of machine learning models perform best on this dataset?
  + Which features are most indicative of a poisonous mushroom?

1. **Data Set:**

Data set is all about the feature of mushrooms like type, color, shape of the data which will use to classify whether the mushroom is edible or not, shape od the data is rows 8124, and 23 columns

**2.1 Data Features**:

cap-shape, cap-surface, cap-color, bruises, odor, gill-attachment, gill-spacing, gill-size, gill-color, stalk-shape, stalk-root, stalk-surface-above-ring, stalk-surface-below-ring, stalk-color-above-ring, stalk-color-below-ring, veil-type, veil-color, ring-number, ring-type, spore-print-color, population, habitat

**2.2 Target Column:**

Class label poisonous (p) or edible (e)

1. **Exploratory Data Analysis:**

There were no null values in the dataset, Data was clean, there

* Target column class had 52% of the data samples says that the mushroom are poisonous and 48% data samples says that the mushroom are edible. No issues like unbalance data

Did EDA to the top feature columns like

* Cap-shape of Mushroom- 45% of mushroom have conical shape and also most edible mushroom,
* Cap-color of Mushroom- 28% of mushroom are in brown in color and also most of them are edible mushroom
* Cap-surface of Mushroom- 40% of mushroom are had scaly surface and also most of them are not edible mushroom

1. **Data visualization**:

* Used pie and bar chart to see the data distribution and the different class feature for feature extraction and analysis

1. **Preprocessing:**

* Using sklearn Label Encoder normalize data. used to transform non-numerical labels
* sklearn Standardization to scale all of the values in the dataset such that the mean value is 0 and the standard deviation is 1
* Data was split on the train set as 70% and 30% for test data set

1. **Model Evaluation**:

* Used Logistic regression, support vector classifier and Random Forest for model evaluation in which Logistic regression gave the accuracy score 95% on the train and test data.
* support vector classifier and Random Forest the accuracy score was 100% on train and test data which has the problem of overfitting

1. **Cross Valuation:**

* Spit the data into the 5-fold and calculated the root mean square error, mean absolute error
* Logistic regression has performed well compared well compared with other ML algorithms