**Mushroom Classification**

**DETAILED PROJECT REPORT (DPR)**

CREATOR : LALATENDU DALAI

TABLE OF CONTENTS

● Problem Statement

● Objectives

● Steps Followed

● Visualisations

Problem Statement

The Audubon Society Field Guide to North American Mushrooms contains descriptions of hypothetical

samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family

Mushroom (1981). Each species is labelled as either definitely edible, definitely poisonous, or maybe

edible but not recommended. This last category was merged with the toxic category.

The Guide asserts unequivocally that there is no simple rule for judging a mushroom's edibility,

such as "leaflets three, leave it be" for Poisonous Oak and Ivy.

The main goal is to predict which mushroom is poisonous & which is edible.

We have to build a solution that should able to predict which mushroom is poisonous & which is edible.

Objective

**Objective :**

* The dataset contains the features and characteristics of mushrooms

and the target variable whether they are poisonous or edilble.

* The aim of this project is to use the given data and create a machine learning

model which will predict whether the mushroom is edible or poisonous.

Dataset Information

1. Cap-shape: bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s  
2. Cap-surface: fibrous=f,grooves=g,scaly=y,smooth=s  
3. Cap-color: brown=n,buff=b,cinnamon=c,gray=g,green=r, pink=p,purple=u,red=e,white=w,yellow=y  
4. Bruises?: bruises=t,no=f  
5. Odor: almond=a,anise=l,creosote=c,fishy=y,foul=f, musty=m,none=n,pungent=p,spicy=s  
6. Gill-attachment: attached=a,descending=d,free=f,notched=n  
7. Gill-spacing: close=c,crowded=w,distant=d  
8. Gill-size: broad=b,narrow=n  
9. Gill-color: black=k,brown=n,buff=b,chocolate=h,gray=g, green=r,orange=o,pink=p,purple=u,red=e, . white=w,yellow=y  
10. Stalk-shape: enlarging=e,tapering=t  
11. Stalk-root: bulbous=b,club=c,cup=u,equal=e, rhizomorphs=z,rooted=r,missing=?  
12. Stalk-surface-above-ring: fibrous=f,scaly=y,silky=k,smooth=s  
13. Stalk-surface-below-ring: fibrous=f,scaly=y,silky=k,smooth=s  
14. Stalk-color-above-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y  
15. Stalk-color-below-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y  
16. Veil-type: partial=p,universal=u  
17. Veil-color: brown=n,orange=o,white=w,yellow=y  
18. Ring-number: none=n,one=o,two=t  
19. Ring-type: cobwebby=c,evanescent=e,flaring=f,large=l, none=n,pendant=p,sheathing=s,zone=z  
20. Spore-print-color: black=k,brown=n,buff=b,chocolate=h,green=r, orange=o,purple=u,white=w,yellow=y  
21. Population: abundant=a,clustered=c,numerous=n, scattered=s,several=v,solitary=y  
22. Habitat: grasses=g,leaves=l,meadows=m,paths=p, urban=u,waste=w,woods=d

Steps followed

**Data Extraction**: This step involves extracting the data from different sources relevant to the problem statement or obtaining data from the client.

**Data Preprocessing**: In data preprocessing step, we check if there missing data, duplicate values, and datatypes of each feature. In our dataset, there was not any null and duplicate values

**Exploratory Data Analysis:** This step includes bivariate and univariate analysis of features. Checking outliers using boxplots, and outlier treatment is carried out as well. Distribution of the features are plotted to see to what extent our data is skewed.

**Feature Engineering:** In this part, the datatypes of the features were checked whether it belongs same datatypes or different datatypes. Outliers were checked using boxplot but there is no such major outliers in the dataset.

**Model Implementation:** After train and test splitting, pipeline containing Standard Scaler and Label Encoder was fitted to several models such as RandomForest Classifier,XGB Classifier, KNeighbors Classifier, etc. Their f1 score were obtained and it was determined that KNeighbors Classifier performs better than other models.

**Model Evaluation:** Test dataset is used to evaluate the model. 20% of dataset was separated for testing. Predicted results of the model are compared with the actual data to check the amount of error. As there was no considerable change after , it helped us to overcome overfitting and perform better on new data.

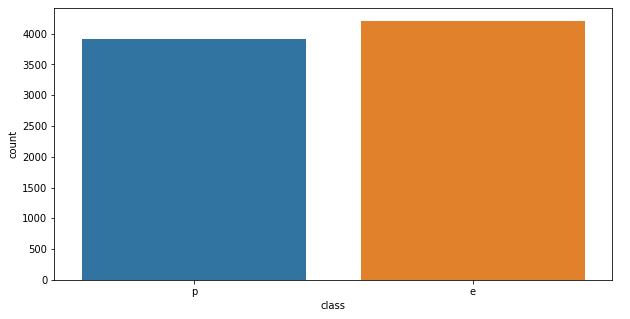
**Designing UI:** For this project, a user interface is built using Apache Air Flow.It is a open-source platform for developing, scheduling, and monitoring batch-oriented workflows.

**Designing a server:** A server should be created to run the UI application continuously. Amazon Web Service EC2 instance is used to create a virtual server for the application.Elastic Compute Cloud (EC2) is a virtual server in AWS for running applications on the AWS infrastructure.

**Deployment:** The codes for this machine learning model should be deployed to the cloud, so that when data is entered into the application, our code runs, and a user gets the result online. In this stage, we containerized the code using Docker and deploy the model to AWS.

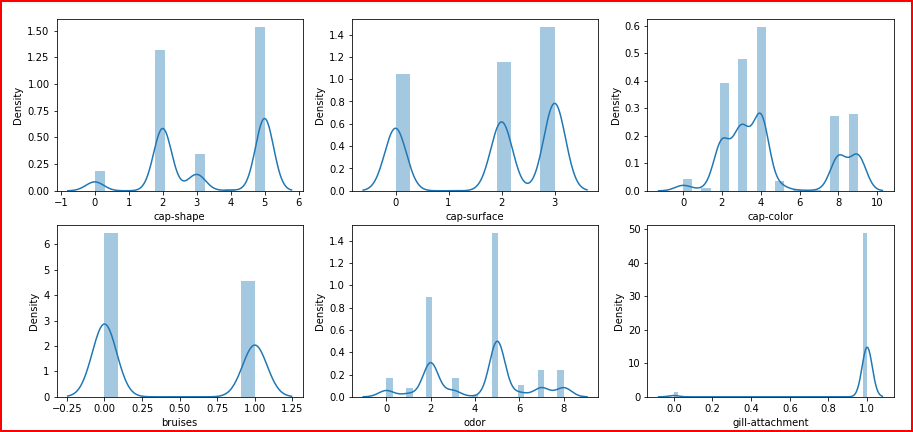
Visualisation

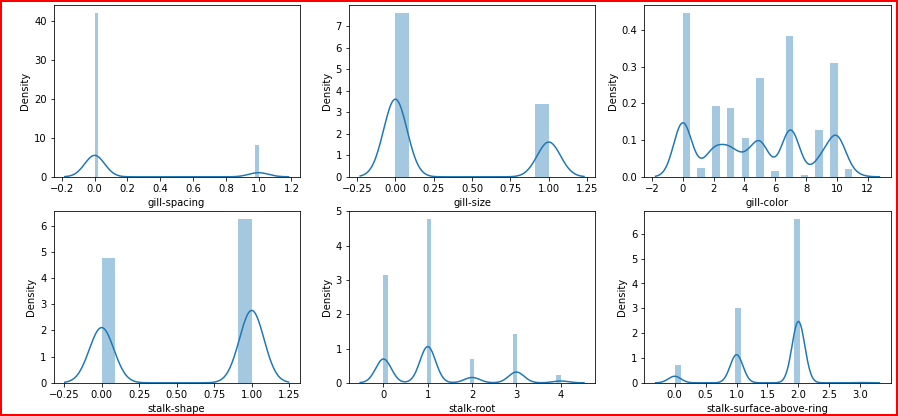
Mushroom class



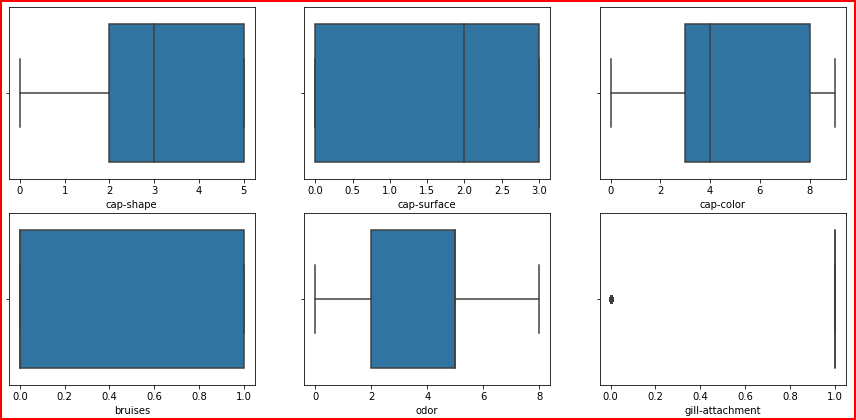
# 

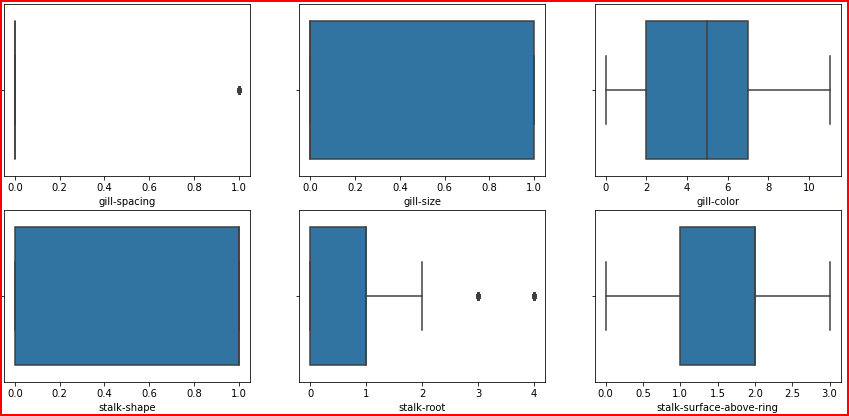
Distributions



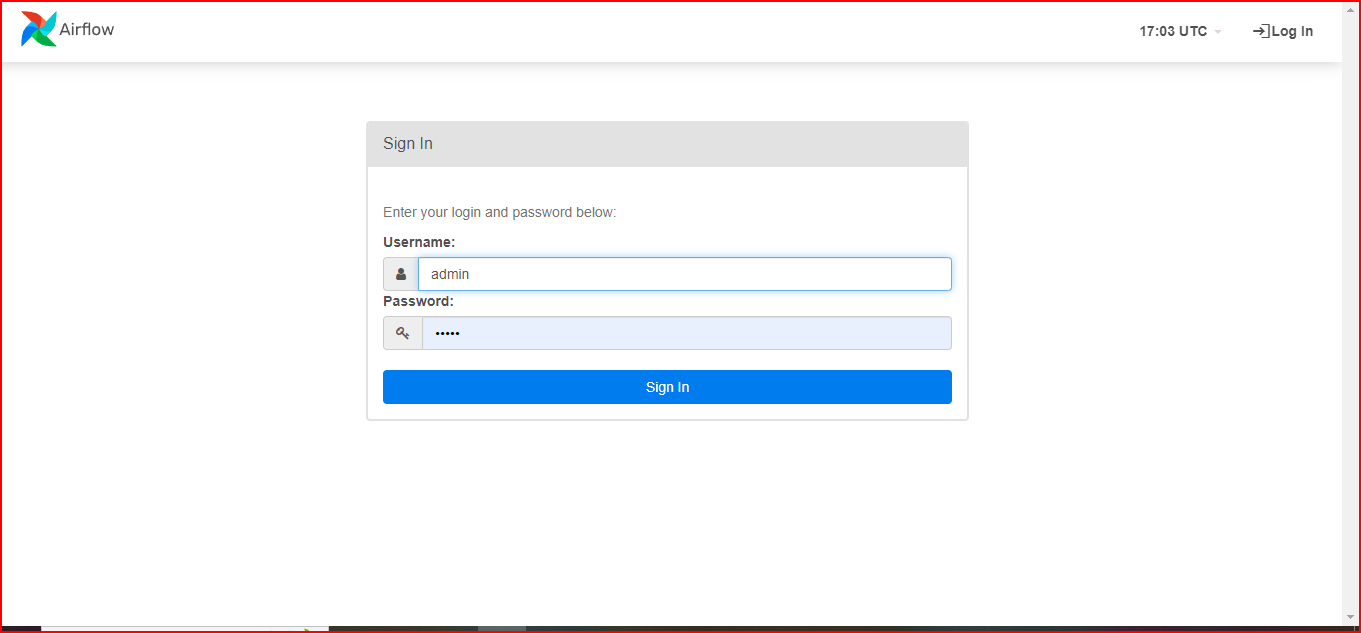


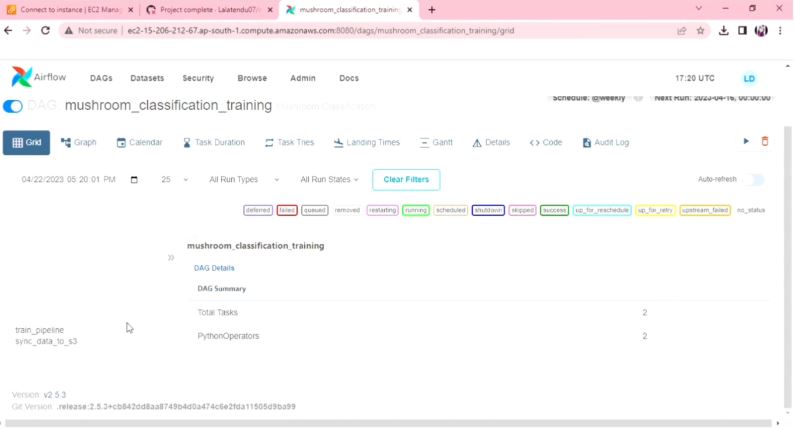
Checking Outliers





Web interface





**App link**: http://ec2-15-206-212-67.ap-south-1.compute.amazonaws.com:8080/

Summary

* The target column has 2 class type one is ‘poisonous’ which has 3916 counts and second is edible which 4208 counts so we have nearly equal counts for poisonous and edible classes in our data.Hence we can say our data is balanced.
* There are 4 types of cap-surface in a mushroom and also it suggests that edible mushroom do not have cap-surface.
* The mushroom can have gill spacing as close or crowded but still it could be poisonous or edible.
* The mushroom can have gill size as narrow or broad but still it could be poisonous or edible.
* The dataset does not contain any missing values and all the columns have same data types.
* The dataset does not contain any major outlier.
* The KNeighbors Classifier model has the best accuracy on both train and test data.

Q & A

* What is the type of the data ?

The type of the data is categorical.

* Which cloud platform is used for the deployment?

Amazon AWS is used for the deployment.

* What is the source of the data?

The dataset is taken from UCI repository.

* How logs are managed?

We are using different logs as per the steps that we follow in training and prediction like model training logs and prediction logs ,etc.

* Is your model 100% sure about whether the mushroom is poisonous or edible?

Looking at the results, yes we are sure. However it is recommended that you also take help from someone who is expert as some characteristics are same for edible ande poisonous mushroom.

**Thank You**