MUSHROOM CLASSIFICATION

DETAILED PROJECT REPORT

AAKASH BHUTE

ABSTRACT

This study focuses on the classification of mushrooms into two categories, namely Poisonous and Edible, using a machine learning model. It aims to determine the significant features that play a crucial role in predicting the edibility or toxicity of mushrooms. Mushrooms have been consumed since ancient times and are highly regarded for their nutritional value. They are low in calories, carbohydrates, fats, and sodium, while being cholesterol-free. Mushrooms offer essential nutrients such as selenium, potassium, riboflavin, niacin, Vitamin D, proteins, and fiber. They have a rich history as a food source and are also recognized for their healing properties in traditional medicine. Various health benefits and potential disease treatments have been associated with mushrooms, including their anticancer and antitumor properties. Moreover, mushrooms exhibit antibacterial effects, enhance the immune system, and assist in lowering cholesterol levels. Furthermore, mushrooms are a valuable source of bioactive compounds. Throughout this machine learning analysis, we will identify the key features that determine whether a mushroom is poisonous or edible.

INTRODUCTION

Purpose of Detailed Project Report (DPR)

- A detailed project report is a very extensive and elaborative outline of a project, which includes essential information such as the resources and tasks to be carried out in order to make the project turn into a success. It can also be said that it is the final blueprint of a project after which the implementation and operational process can occur.
- In this comprehensive project report, we will discuss about the end to end implementation of Mushroom Classification with necessary details like Architecture, Data Visualization, Data Preprocessing, Model Building, Model Performance and Deployment of this project with sample test cases.

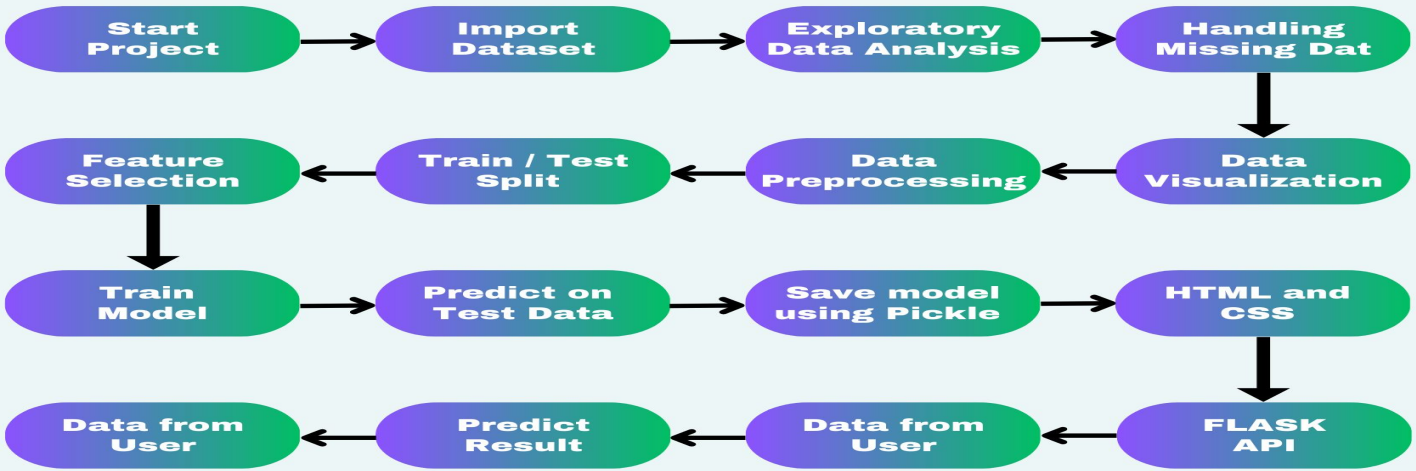
Problem Statement

- 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.

Tools Used


jupyter
pandas
NumPy
python
seaborn
matplotlib
HTML
CSS
Flask
HEROKU
scikit
learn

Architecture Design



Architecture Description

Data Collection

- The data for this project is collected from the Kaggle Dataset, the URL for the dataset is given below: https://www.kaggle.com/datasets/uciml/mushroo m-classification.

Data Description

- This dataset includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family Mushroom drawn from The Audubon Society Field Guide to North American Mushrooms (1981). Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. This latter class was combined with the poisonous one.

Architecture Description

Exploratory Data Analysis

- There are 8124 rows and 23 columns in this data. All the columns are of categorical type. There are two classes present in our target column which are 'p' - poisonous and 'e' - edible.

Handling Missing Data

- At first, we observed that there no missing/null values in the dataset. However, if you go through the data description (check the link) you will find that the missing values in one column is replaced with "?". There are 2480 missing values in 'stalk-root' column. First, we will replace these values with np.nan so that we can handle missing data. we will impute the missing values in 'stalk-root' column using sklearn SimpleImputer with strategy='most_frequent'.

Architecture Description

Data Preprocessing

- In this step, first we have dropped the column 'veil-type' as it has only one value throughout the data. So, it won't give us much information regarding the class of the mushroom. Next, we mapped our target column to 0 (poisonous) & 1 (edible) values. We used Label Encoder to convert categorical values to numerical then we scaled our data to bring them to same class.

Feature Selection

- After splitting the data into train and test set, we used SelectKBest method with score_func=chi2 to find out which features are most relevant to target column and we found that there are 12 columns out of 21 which we needed for training our model.

Architecture Description

Model Training & Evaluation

- We used XGBClassifier as a model for model training it was very fast compared to the other models and it produced 100% accuracy on train data as well as on test data which is a very good for our project.

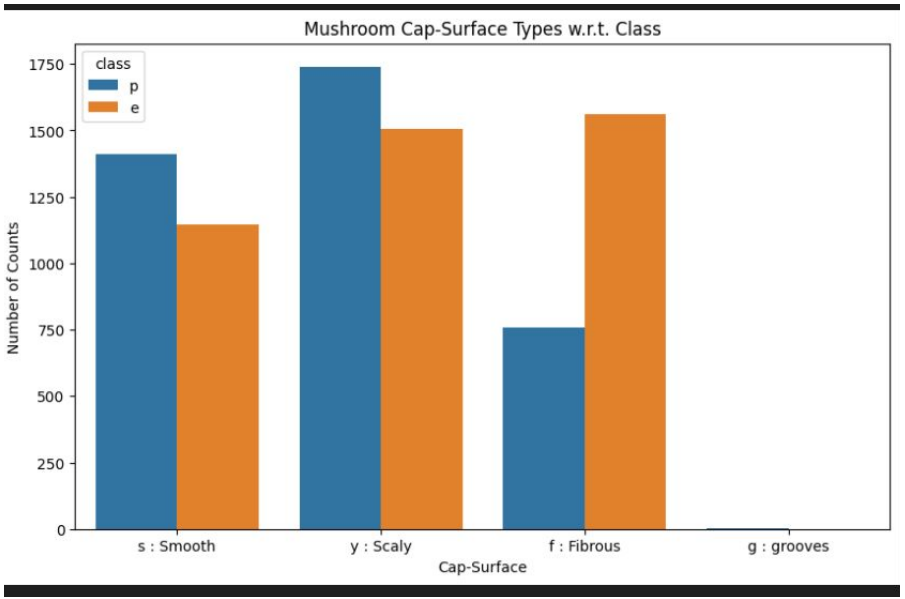
Model Deployment

- We created a webpage using HTML and CSS. We created a Flask web app and first tested in on our local machine. Then we deployed our model using Heroku. We used different combination of input and predicted the output and the results were accurate. The app was working fine and there were no issues found.

Data Visualization

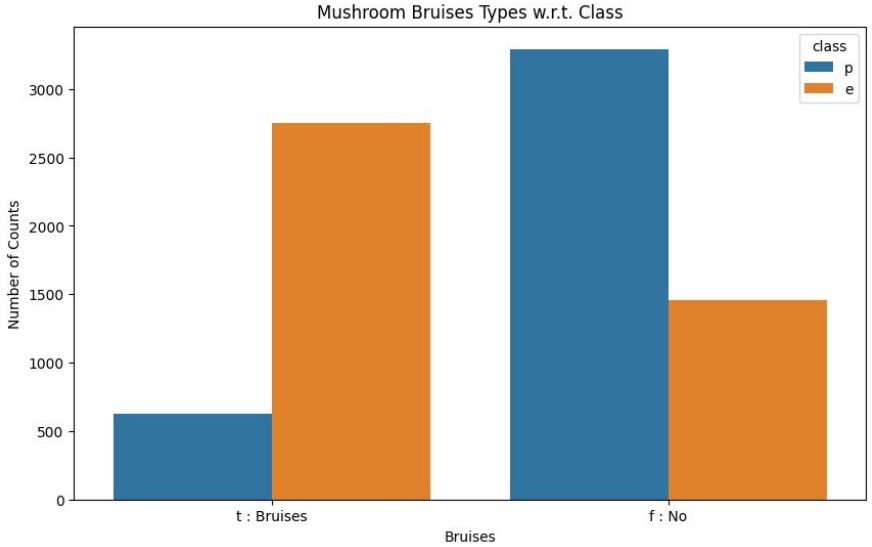
Mushroom Class Type

Mushroom Cap Surface

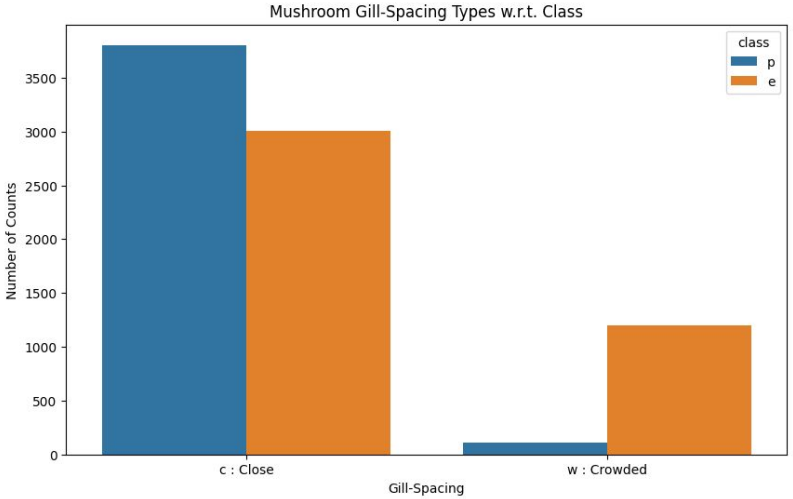


Data Visualization

Mushroom Bruises Types

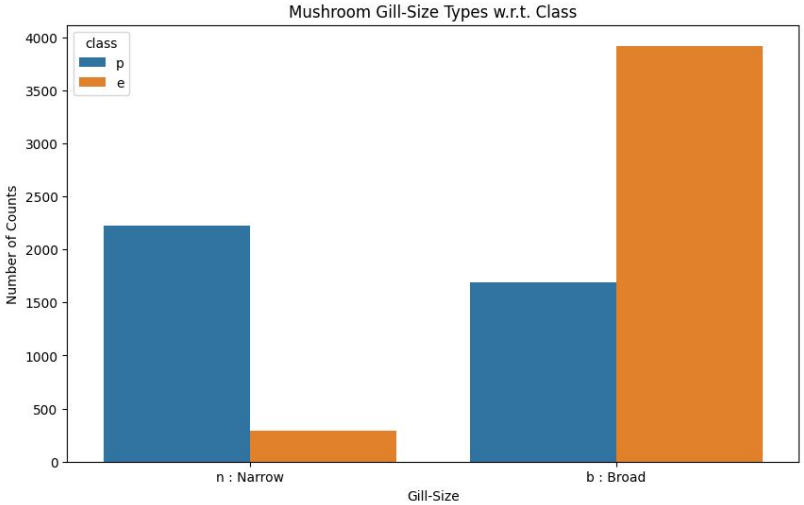


Mushroom Gill Spacing

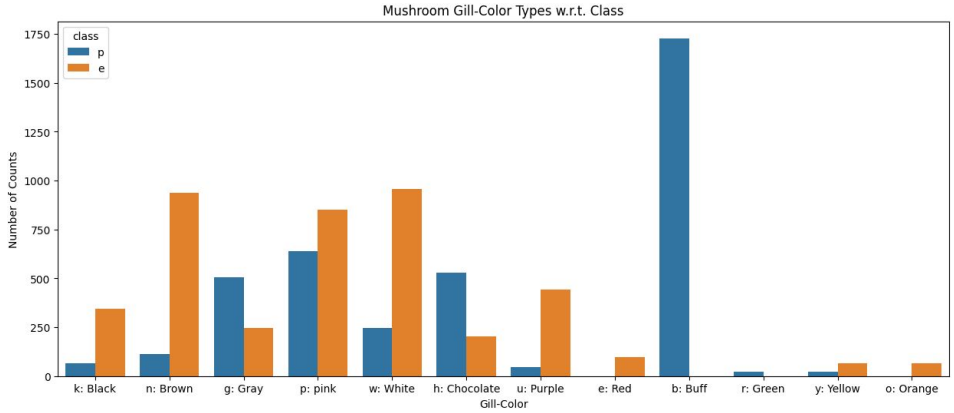


Data Visualization

Mushroom Gill Size

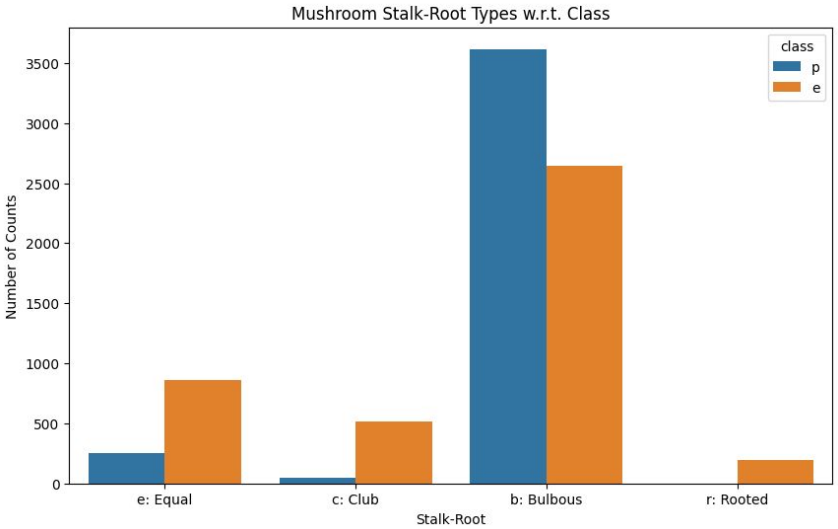


Mushroom Gill Color

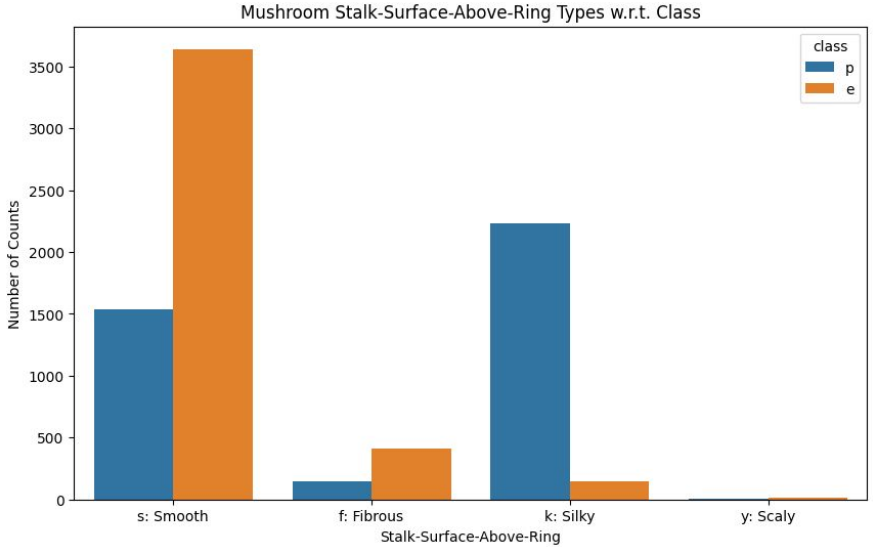


Data Visualization

Mushroom Stalk Root

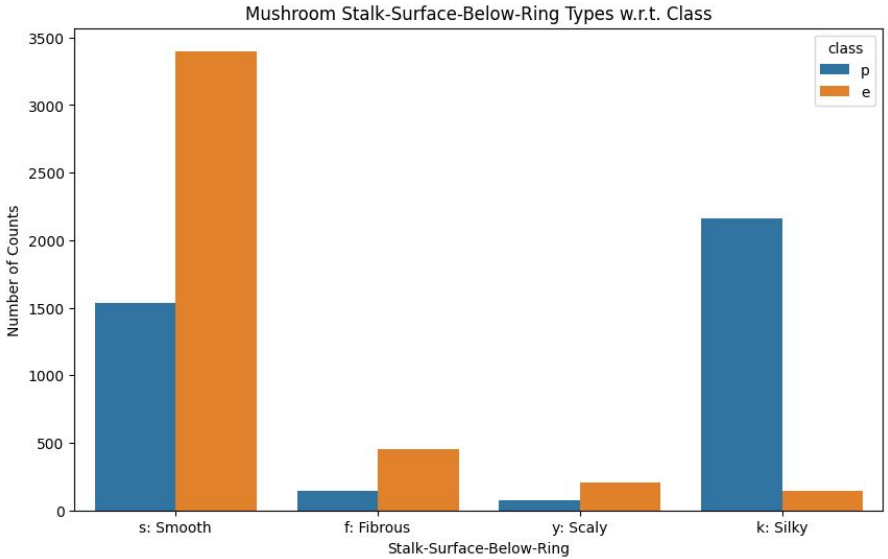


Stalk Surface Above Ring

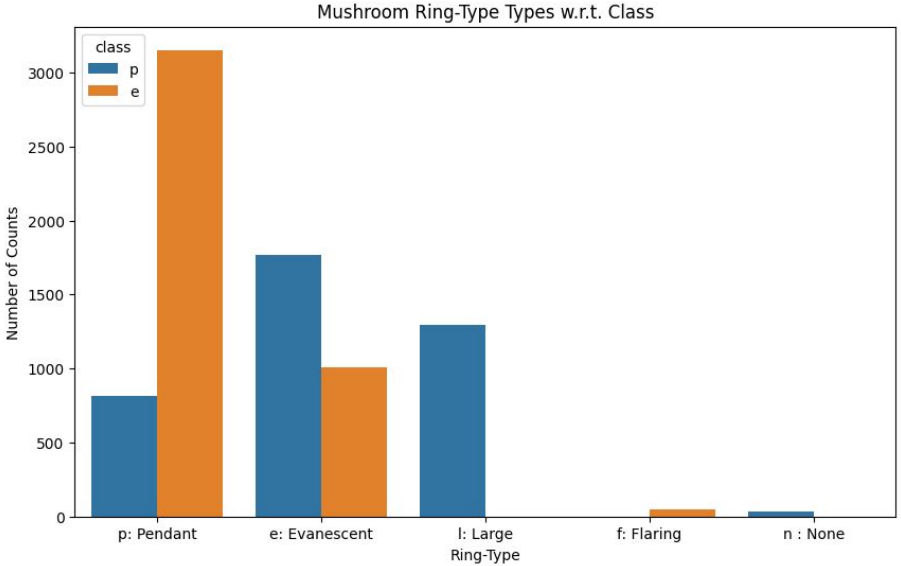


Data Visualization

Stalk Surface Below Ring

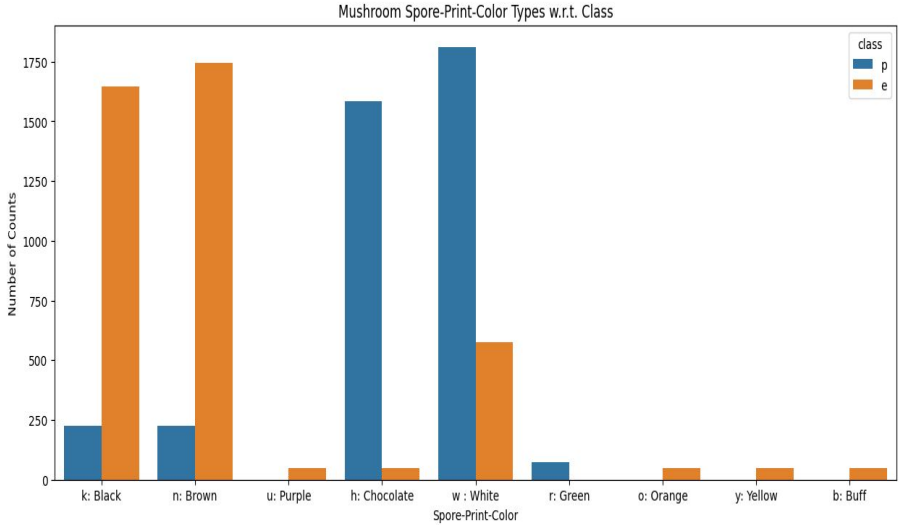


Mushroom Ring Type

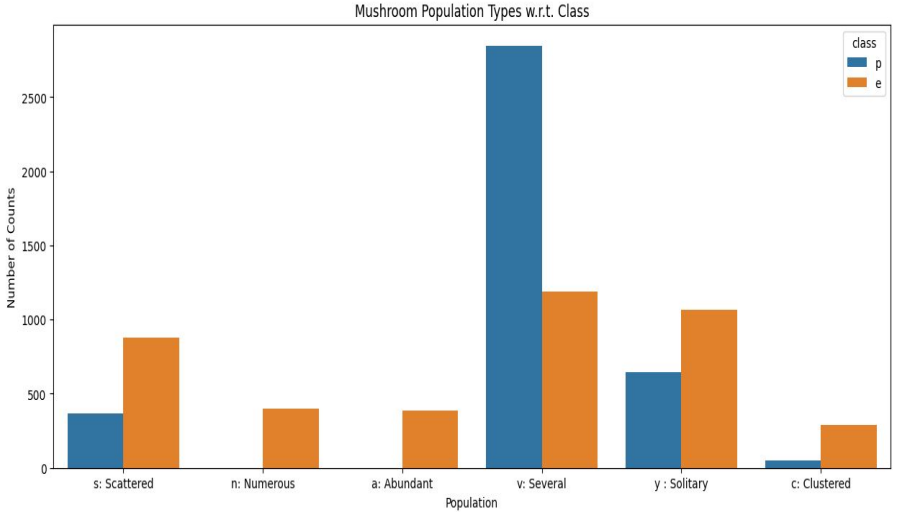


Data Visualization

Spore Print Color

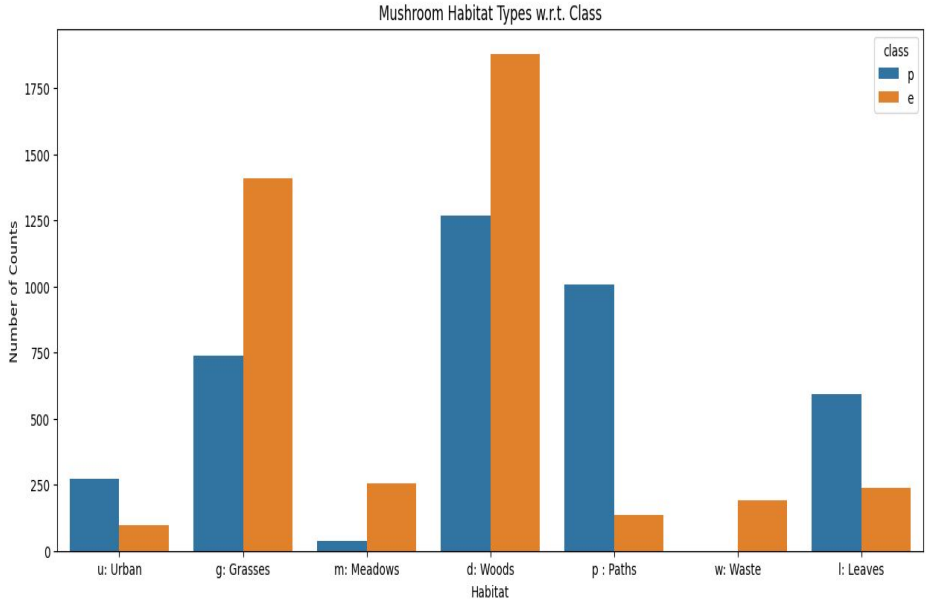


Population

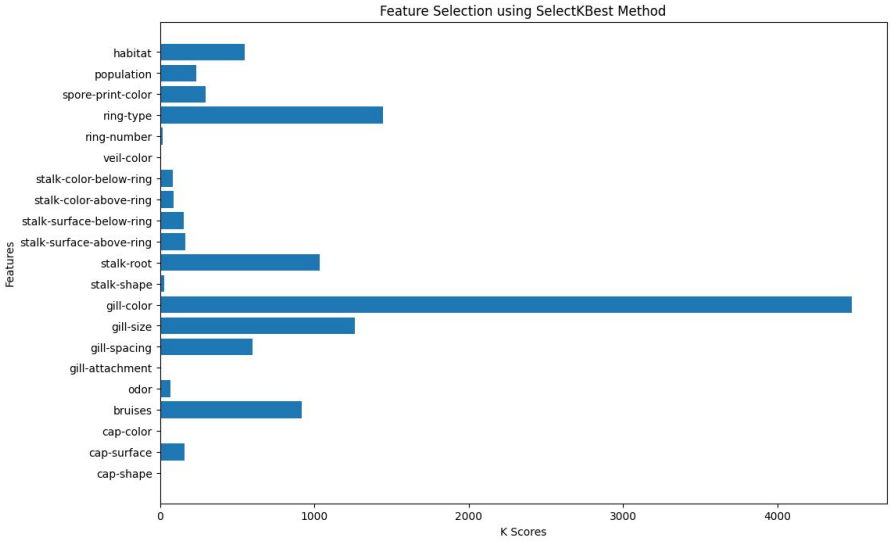


Data Visualization

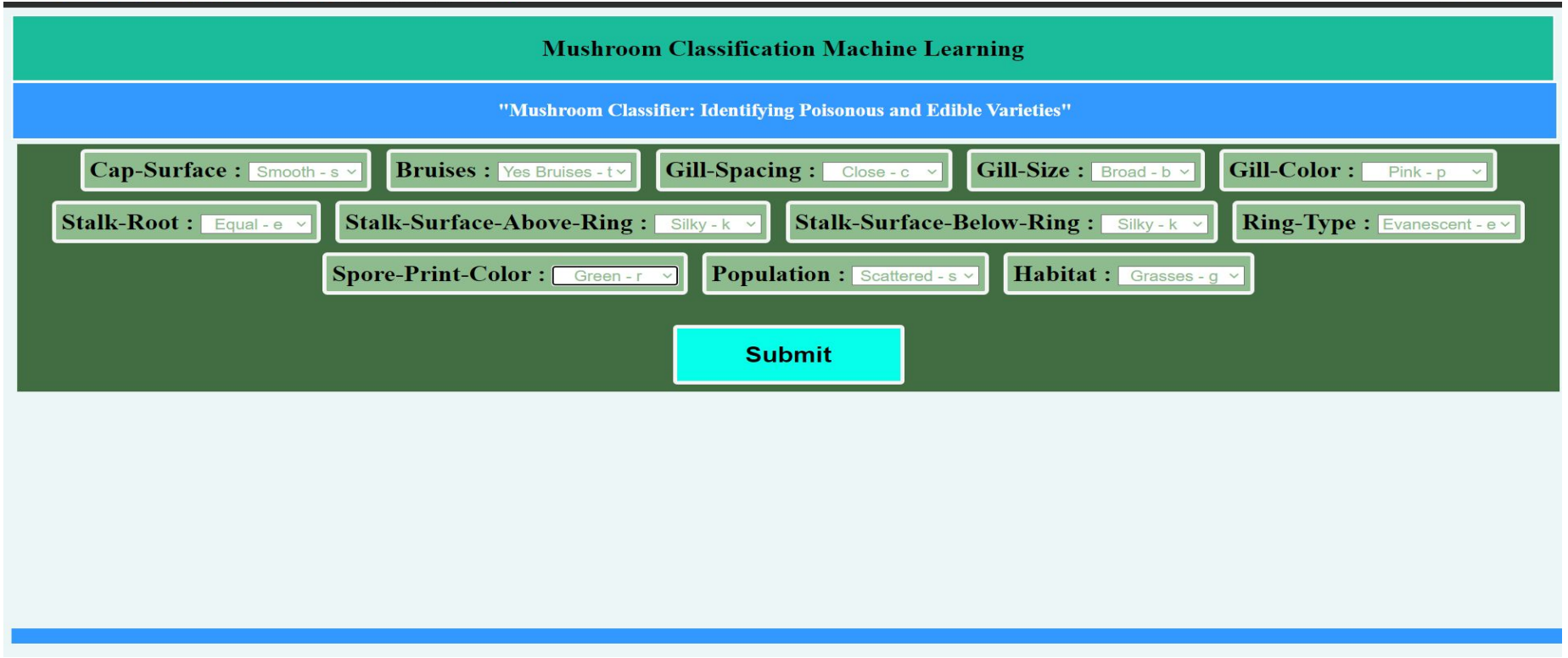
Habitat



Feature Selection

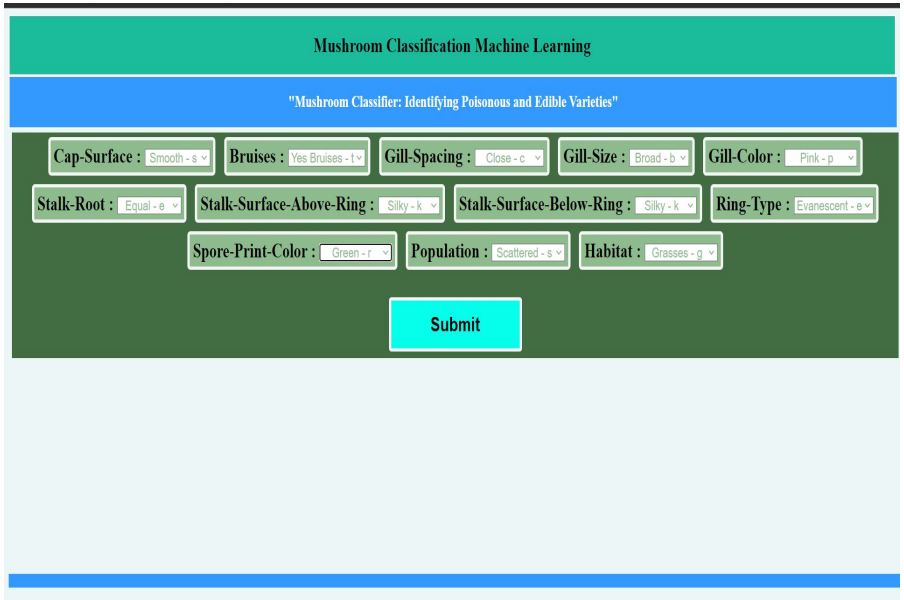


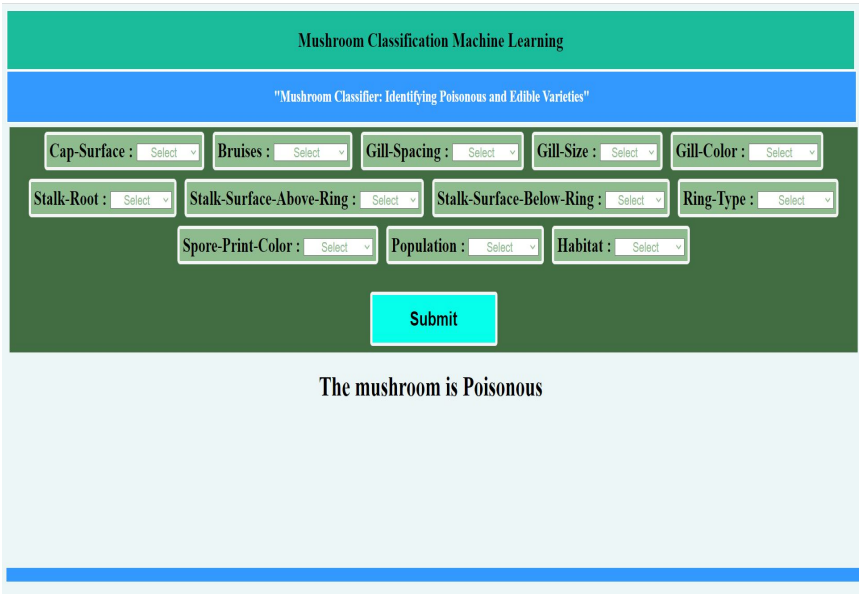
Web Interface



Test Cases

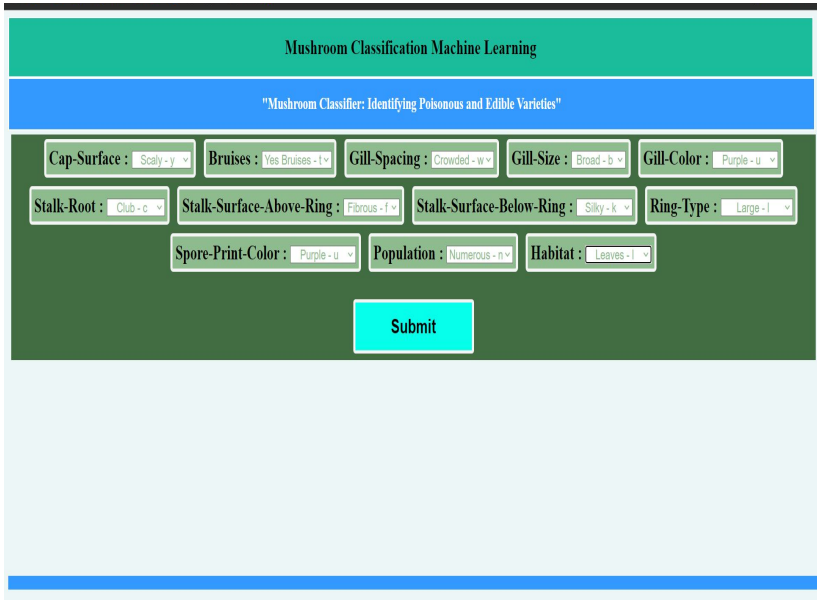
Poisonous Mushroom Example

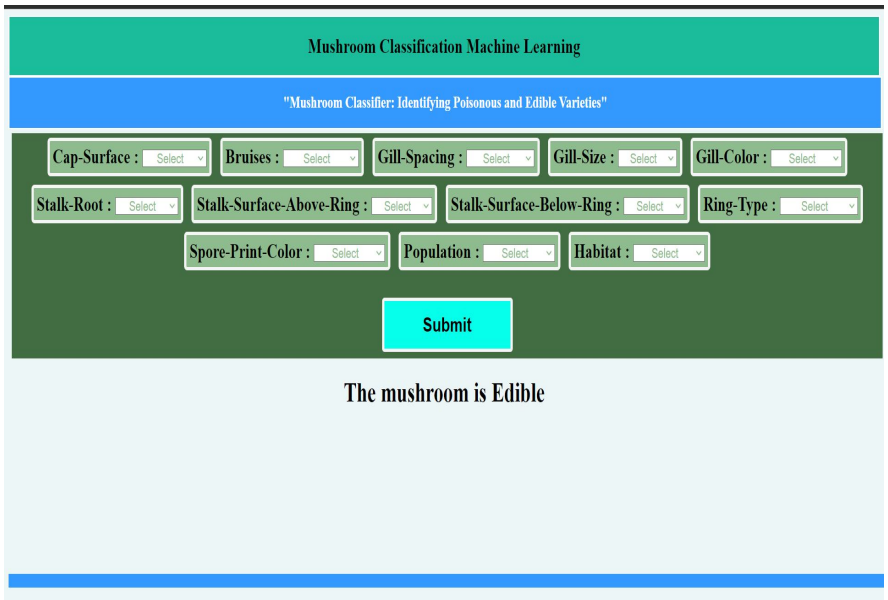




Test Cases

Edible Mushroom Example





Summary

- The target column has 2 class type one is 'poisonous' which has 3916 counts and second is 'edible' which has 4208 counts so we have nearly equal counts for poisonous and edible classes in our data. Hence we can say that our data is balanced.
- There are 4 types of cap-surface in a mushroom and also it suggests that 'edible' mushrooms do not have 'capsurface' : 'g : grooves' according to our data.
- The mushroom may or may not have bruises but still it could be poisonous or edible according to our data.
- The mushroom can have Gill Spacing as Close or Crowded but still it could be poisonous or edible according to our data.
- The mushroom can have Gill Size as Narrow or Broad but still it could be poisonous or edible according to our data.
- The 'edible' mushroom do not have Gill Color : Buff, Green and 'poisonous' mushroom do not have Gill Color : Red, Orange according to our data.

Summary

- The 'poisonous' mushroom do not have Stalk Root as Rooted type according to our data.
- The mushroom can have Stalk-Surface-Above-Ring as Smooth, Fibrous, Silky or Scaly but still it could be poisonous or edible according to our data.
- The mushroom can have Stalk-Surface-Below-Ring as Smooth, Fibrous, Silky or Scaly but still it could be poisonous or edible according to our data.
- The 'edible' mushroom do not have Ring-Type as Large and None and 'poisonous' mushroom do not have RingType as Flaring according to our data.
- The 'edible' mushrooms do not have Spore-Print-Color as Green and 'poisonous' mushrooms do not have SporePrint-Color as Purple, Orange, Yellow, Buff according to our data.
- The 'poisonous' mushrooms do not have Population Type as Numerous and Abundant according to our data.
- The 'poisonous' mushrooms do not have Habitat Type as Waste according to our data.
- The XGBoost Classifier model has 100% accuracy on both training data and test data.